# Visual Generative AI for Extended Reality: A Survey

Guoxi Huang

School of Computer Science and Digital Technologies, Aston University
Birmingham, B4 7ET, United Kingdom.

Corresponding author. E-mail: g.huang2@aston.ac.uk

**Abstract**

Visual generative models create objects, spatial scenes, time-varying 3D content, and action-conditioned observations for extended reality (XR). Using these outputs requires knowing which viewing, editing, and state-changing operations they support, and which system components supply those operations. This critical narrative review catalogues 204 methods and systems and traces generated outputs into specific XR operations. Four comparison axes organize content units, generative mechanisms, representations, and operating stages; common mathematical definitions clarify conditioning and update rules. An evidence map records eight capability dimensions for a purposively selected subset of 31 methods and systems. The synthesis combines these profiles with source-specific quantitative comparisons and deployment studies, identifying the components, configurations, and endpoints supporting capability claims. Building on paradigm-level comparisons, this analysis distinguishes the evidence needed to generate editable structure, preserve an accepted modification, and respond to tracked input. Visual recall and modification retention require different references, while generation throughput and input-to-display latency describe different parts of the response path. Worked task comparisons explain how these distinctions change the choice between prepared geometry with local rendering and continuing observation generation. We derive task-specific selection criteria and propose matched evaluations of binocular viewing, preview replacement, persistent changes, response scheduling, and physical-space alignment. The resulting framework connects model selection to explicit task requirements and identifies the additional tests needed to assess XR use. The survey homepage is available at https://edhuang1.github.io/xr-survey/.

**Keywords:** Visual generative models, Extended reality, 3D asset generation, Scene generation, Interactive world models, XR evaluation

Preprint manuscript. Not peer reviewed.

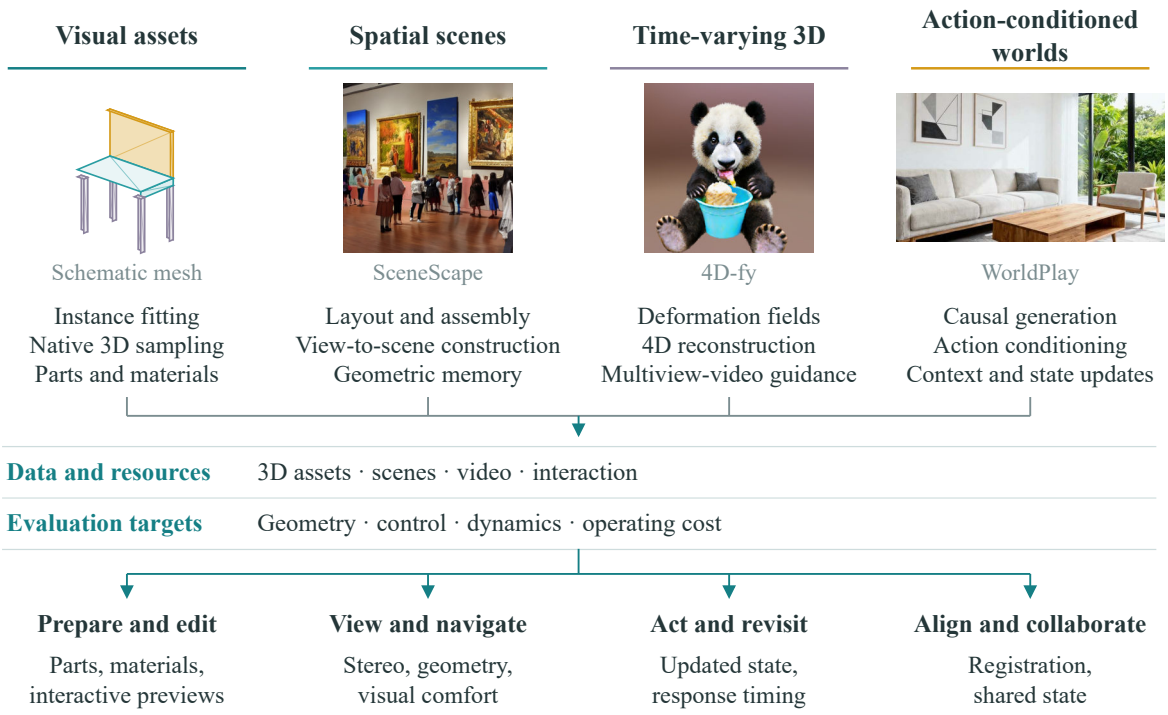



**Fig. 1** Review map linking content families, mechanisms, data and evaluation targets, and XR operations. The asset mesh is schematic. Images from Fridman et al. (2023); Bahmani et al. (2024); Sun et al. (2026a) show scenes, time-varying 3D, and action-conditioned observations.

## 1 Introduction

Visual generative models contribute geometry, appearance, layout, motion, and continuing observations to extended reality (XR). In virtual reality (VR), augmented reality (AR), and mixed reality (MR), users also expect this content to respond to navigation and manipulation. A generated chair may look convincing in one view, yet placing it, walking around it, and moving its parts require different structures and controls. This review asks which generated outputs and supporting components enable the operations an XR application needs.

The answer depends on how content is constructed and what remains available during use. DreamFusion optimizes 3D content through an image-generation prior, whereas TRELLIS samples structured 3D latents (Poole et al. 2023; Xiang et al. 2025). At scene scale, Text2Room extends a textured mesh through generated views and geometric fusion; Holodeck assembles retrieved assets using generated spatial constraints (Höllein et al. 2023; Yang et al. 2024d).

Temporal generation introduces another distinction: 4D-fy and DreamGaussian4D prepare time-varying representations, while WorldPlay generates observations from actions and history (Bahmani et al. 2024; Ren et al. 2023; Sun et al. 2026a). The former retain content that can be rendered at a selected time and viewpoint; WorldPlay synthesizes new observations from its available context. This difference places rendering and generation at different points in the response path.

XR deployments reveal where these operations are introduced. CoGen3D adds headset loading and manipulation to Hunyuan3D 2.0 outputs; Dreamcrafter uses proxies while final generation runs asynchronously (Zhao et al. 2025; Jiang et al. 2026; Vachha et al. 2025). Evaluation must follow content through this handoff. Generation benchmarks test defined content tasks (Duan et al. 2025; Lu et al. 2026b), whereas XR applications may require manipulation, paired eye views, or responses to tracked input. Generation throughput measures output rate; input-to-display latency can also include scheduling, transfer, decoding, rendering, and display.

**Relation to existing surveys.** Existing reviews already connect generative techniques to XR devices, workflows, and evaluation (Ning et al. 2026). The visual GenAI/MR survey cross-codes models, interaction modalities, and evaluation; it also examines hybrid systems and reversible spatial proxies (Wang et al. 2026a). Reviews of immersive human–AI interaction and embodied 3D generation address user intent, iterative processing, simulator compatibility, physical validity, and task performance (Nguyen et al. 2026; Ye et al. 2026). Complementary surveys cover VR applications, multimodal generation, and immersive-video creation, delivery, and display (Rahimi et al. 2025; Taha 2026; Hu et al. 2025b).

World-model surveys organize representations, prediction targets, and downstream tasks (Kong et al. 2025; Li et al. 2025d; Shang et al. 2025); an action–state–observation analysis relates visual generation directly to game-engine state (Li et al. 2026b). Most closely, Xie et al. compare four scene-generation paradigms against seven capability criteria, provide individual-method capability profiles, group quantitative results by protocol, and map application requirements to paradigms. Their analysis includes XR efficiency, persistence, and the distinction between prescribed 4D motion and interaction (Xie et al. 2026a). We build on these comparisons in the configuration-level synthesis below.

Our unit of comparison is the path from a generated output to a specified XR operation. We identify the representation exposed to the application, the component supplying the operation, and the configuration and endpoint actually tested. Evidence for generating separate parts concerns the available editing unit; evidence for manipulating imported objects concerns an implemented operation. Retention requires testing whether the accepted edit survives leaving and returning. These distinctions determine whether a reader can select a generator, needs an additional interaction component, or must still test retention. Sections 5.1–5.6 connect such evidence across model and system studies; Section 6 proposes matched tests where that connection remains unmeasured. This synthesis extends capability profiling and interaction-workflow analysis through explicit evidence chains.

This critical narrative review catalogues 204 methods and systems and 37 resources, with targeted retrieval through 23 October 2026. A purposively selected subset of 31 methods and systems receives detailed capability coding. Appendix A explains selection, source inspection, and coding boundaries. Figure 2 counts catalogue papers by bibliographic year: formal publication year, or initial preprint year for preprint-only records. Each work is counted once, merging preprints with subsequent publications. The counts summarize the review catalogue.

**Contributions.** The review develops three connected analyses:

1. **Mechanisms and operational structure.** Four axes describe content units, generative mechanisms, representations, and operating stages. Common mathematical definitions connect generated units to their conditioning inputs, retained representations, and update operators (Sections 2.1 and 3). This shared description separates camera/time queries, observation generation, and state-changing operations.
2. **Evidence across models and deployments.** Capability evidence is assessed separately from the descriptive axes. A source-linked map records evidence types across eight capability dimensions for 31 methods and systems under inspected configurations. Protocol-specific quantitative comparisons accompany these profiles. Model-to-system traces identify where delivery, interaction, and display components supply the reported operation (Figure 10; Sections 4 and 5.1).
3. **Task-specific selection and evaluation.** Worked comparisons relate the required editable unit, retained state, and response path to qualifying evidence (Section 5.6). Five proposed matched evaluation designs target transfers from model-level results to XR use (Section 6).

**Reading routes.** Readers entering XR can start with Section 2.2, then use the analytical framework in Section 2.1 to orient the method comparisons. Sections 2.3 and 2.4 supply representation and learning background as needed. Readers choosing a pipeline can start with the worked tasks in Section 5.6, then follow the method, evidence-profile, and timing references.

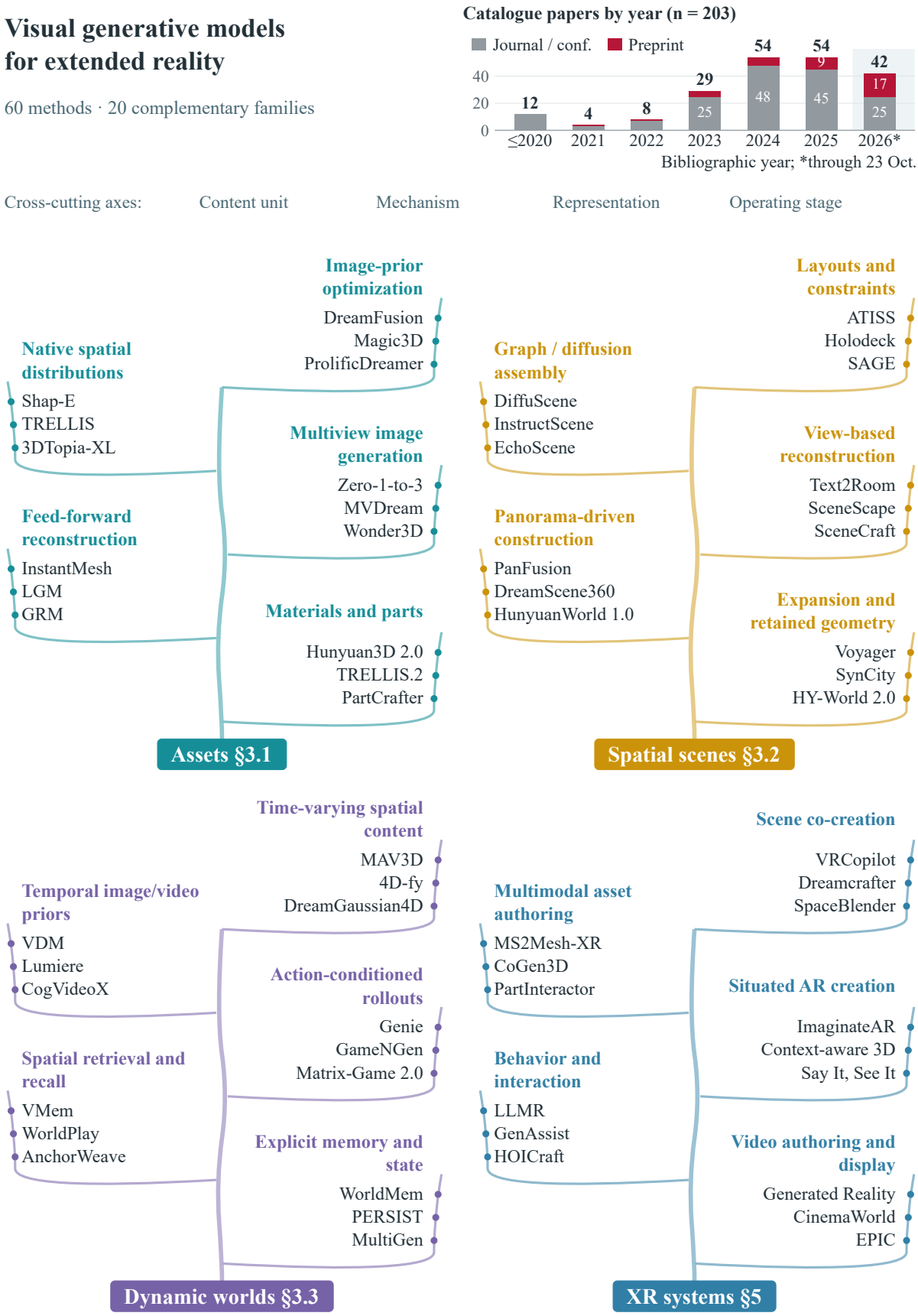



**Fig. 2** Taxonomy of 60 representative methods in 20 complementary families. Names link to references; domain and axis labels link to sections. The inset counts 203 catalogued papers by bibliographic year, excluding the Genie 3 blog; 2026 runs through 23 October.

## 2 Preliminaries and Review Framework

We define the comparison axes, then introduce the XR, spatial-representation, and generative-learning concepts needed to interpret them. Scene-generation and world-model surveys provide complementary introductions to representations and predictive modeling (Xie et al. 2026a; Kong et al. 2025; Li et al. 2025d; Shang et al. 2025).

### 2.1 Analytical Framework for the Review

We compare each method along four axes: content unit, generative mechanism, representation, and operating stage. Content units organize the method families; the other axes support comparisons within and across them. A method configuration may combine several mechanisms, representations, or operating stages. Figure 3 contrasts generated observations with retained content that an application can query or update.

#### 2.1.1 Content units and types of change

**Content units.** An asset supplies an object or localized element; a spatial scene supplies an arrangement or connected environment. Time-varying 3D content supplies a prepared temporal representation; an action-conditioned world produces observations in response to inputs during use.

These units organize Section 3 by the output exposed to an application. A video may supervise a prepared 3D animation or serve as an observation generated during interaction. For each output, we distinguish viewpoint changes, object dynamics, action responses, and information retained for later queries.

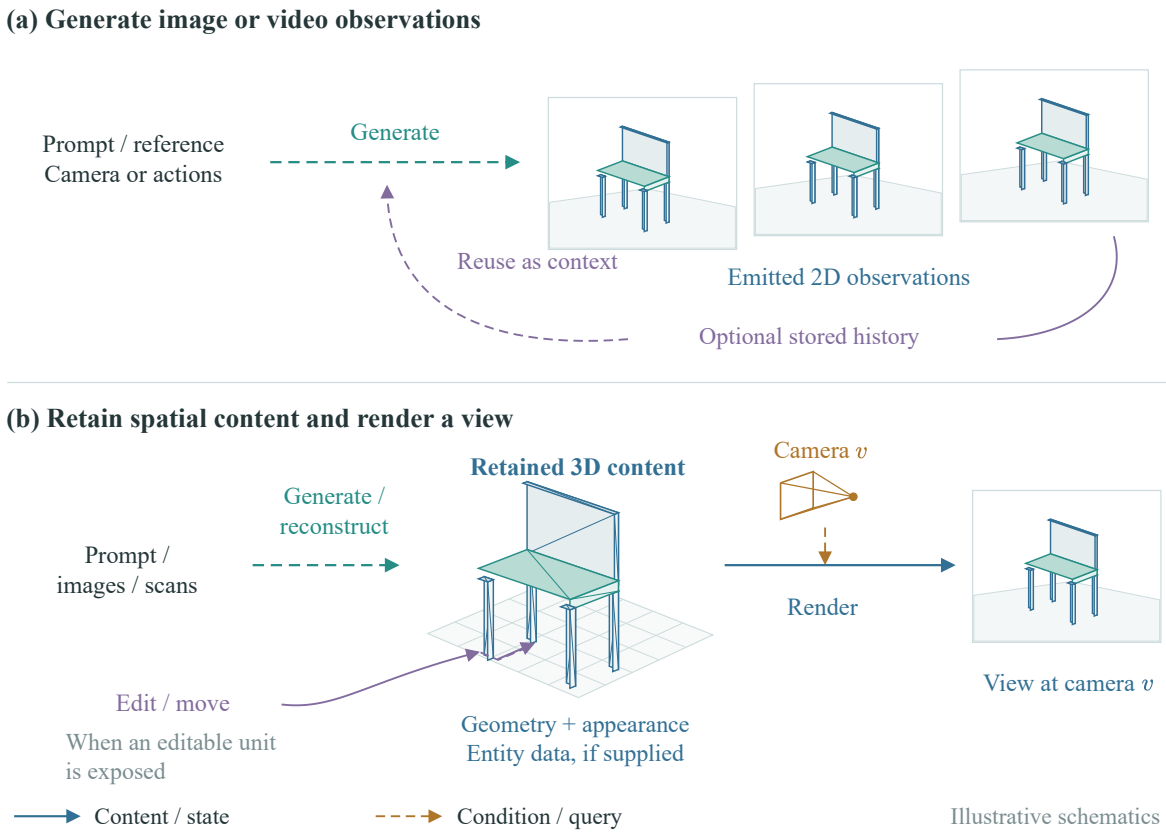



**Fig. 3** Visual-generation interfaces: (a) image/video observations with optional history reuse; (b) retained geometry and appearance queried by camera $v$ or modified through an exposed edit unit. Solid arrows carry content/state; dashed arrows carry conditions/queries. All elements are schematic.

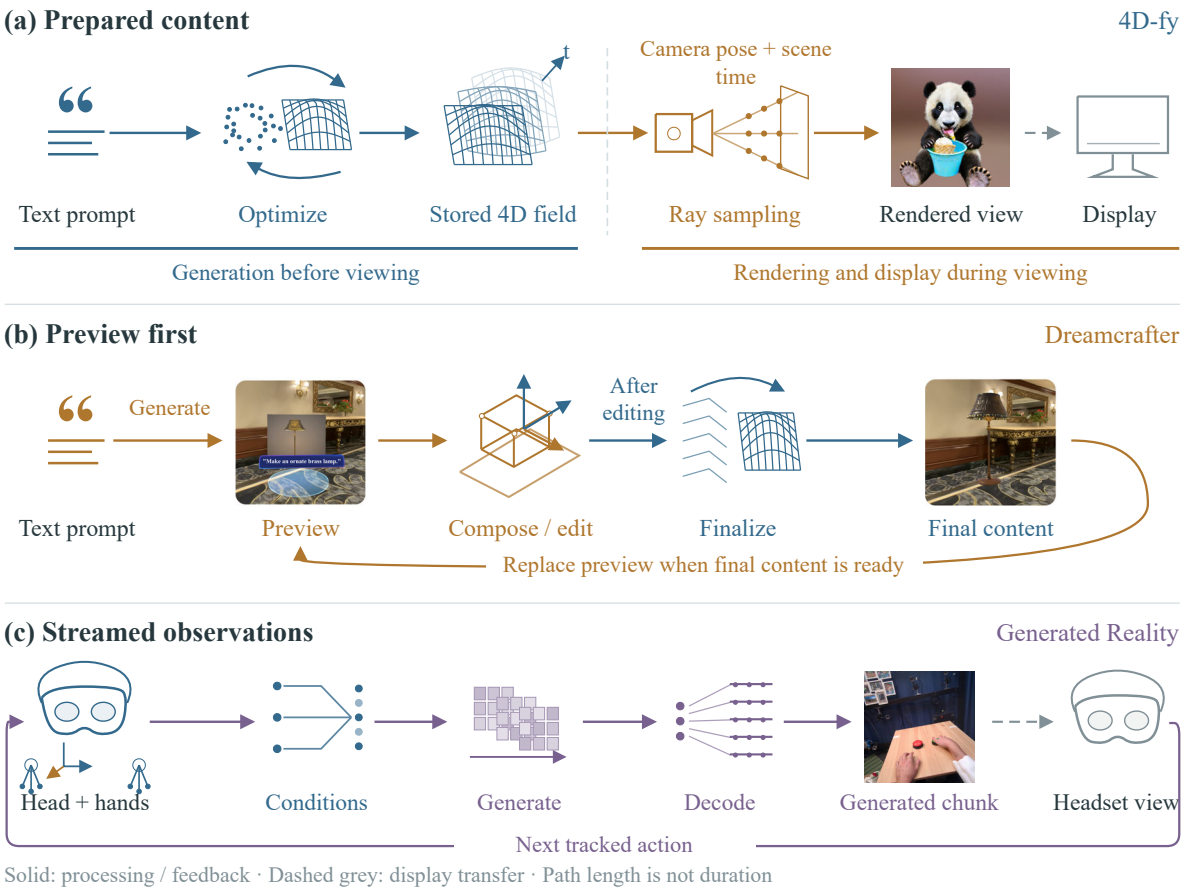



**Fig. 4** Generation relative to use: (a) prepared-content rendering; (b) preview interaction followed by final replacement; (c) generation/decoding during interaction. Solid arrows show processing and feedback; grey dashed arrows show display transfer. Schematics are not to time scale; scene time in (a) is physical time. Images from Bahmani et al. (2024); Vachha et al. (2025); Xie et al. (2026b).

### 2.1.2 Generative mechanisms and representations

The mechanism axis records how operators construct or change content. They may fit instance parameters to observations or a learned prior, sample a learned representation, or assemble assets from layouts. Other operators generate observations that may be fused into retained geometry. Learned models, procedural rules, and explicit constraints can participate in the same pipeline; Section 3 compares their conditions and update rules.

**Representation roles.** The internal representation is manipulated by the generator; the rendering representation supplies image formation; the delivered artifact is handed to another component. A shared file format does not establish that the same structure or operation survives every stage.

A point cloud used to condition generation and a mesh delivered for interaction serve different purposes, even when describing the same scene. The method and deployment comparisons therefore record conversion and system integration as explicit steps.

### 2.1.3 Operating stages and system configurations

The operating-stage axis distinguishes *offline generation* before viewing from *runtime rendering* of available content. *Asynchronous editing* keeps existing content or a proxy displayed, whereas *streaming generation* produces content as context or actions arrive. A system may combine these stages to schedule work across its components.

Figure 4 locates generation on the response paths explained in Section 2.2.3. Table 6 later compares source-reported costs at their stated endpoints.

A deployment configuration records the generator and version, asset processing, interaction components, device and compute arrangement, and task. It follows the implementation, including streams without conventional engines or exported assets.

### 2.1.4 XR capabilities and evidence targets

The four axes describe how a method works; evidence is examined separately along eight capability dimensions. They include spatial agreement (S), binocular viewing (B), temporal consistency and persistence (T), and conditioning and control (C). The remaining dimensions are interaction (I), delivery and engine use (D), operating cost (O), and XR deployment (X). Each observation identifies a property, its reference, and the responsible component, from model output to user task. Section 4 develops the evaluation targets; Figure 10 summarizes the evidence profiles.

## 2.2 XR viewing and interaction

Coordinate frames determine which view is requested, perception determines how paired views are experienced, and the interaction loop determines when an action becomes visible.

### 2.2.1 XR displays, viewpoints, and spatial coordinates

Extended reality (XR) places digital information in a spatial experience. Virtual reality (VR) presents a predominantly synthetic environment; augmented reality (AR) combines digital content with a physical view registered to three-dimensional coordinates. Mixed reality (MR) combines physical and virtual elements, with usage varying across research and commercial systems. The reality–virtuality continuum and subsequent MR research document these combinations and terminology (Milgram and Kishino 1994; Azuma 1997; Speicher et al. 2019). Applications can cross modes: MagicBook moves from a physical book to overlaid models and immersive scenes (Billinghurst et al. 2001).

A head-mounted display (HMD) presents images near the eyes while the system estimates the wearer's viewpoint. The renderer uses tracked head pose to update views, making a world-anchored object appear fixed as the viewer moves. This coupling of tracking and rendering was central to early head-mounted graphics (Sutherland 1968). Three degrees of freedom (3DoF) describe orientation around three axes. Six degrees of freedom (6DoF) add translation along three axes. Turning the head changes viewing direction; leaning sideways also changes surface visibility and foreground motion relative to the background. A panoramic image supports looking around from its capture center, whereas convincing translational exploration requires additional geometric or view-dependent information. Tracking technologies differ in precision, operating volume, latency, and sensitivity to occlusion or environmental conditions (Welch and Foxlin 2002).

Spatial content passes through object ($O$), world ($W$), headset, and eye ($E$) coordinate frames. Let $T_{ab}$ map coordinates from frame $b$ into frame $a$. A rigid pose and the resulting eye projection can be written as

$$\begin{aligned} T_{ab} &= \begin{bmatrix} R_{ab} & \mathbf{t}_{ab} \\ \mathbf{0}^{\mathsf{T}} & 1 \end{bmatrix} \in \mathrm{SE}(3), \quad \widetilde{\mathbf{X}}_E = T_{EW} T_{WO} \widetilde{\mathbf{X}}_O, \\ Z_E \widetilde{\mathbf{u}} &= K_E \mathbf{X}_E, \qquad\qquad \widetilde{\mathbf{u}} = (u_1, u_2, 1)^{\mathsf{T}}. \end{aligned} \tag{1}$$

Here $R_{ab}$ is a rotation, satisfying $R_{ab}^{\mathsf{T}} R_{ab} = I$ and $\det R_{ab} = 1$; $\mathbf{t}_{ab}$ is a translation. A tilde denotes homogeneous coordinates, and SE(3) is the set of these rigid transformations. In the ideal pinhole projection, $K_E$ contains the eye camera's intrinsic parameters, $Z_E > 0$ is forward depth, and $(u_1, u_2)$ are pixel coordinates. Each eye has its own calibrated offset from the tracked headset and its own projection parameters, jointly defining the view $v$. Rigid pose does not include scale: the object's dimensions must first agree with the world's units. A chair's placement can then align its support surface with the floor. The same transformation chain maps tracked hands and controller rays into the world for selection and contact tests. Figure 5(a) illustrates the frames and transformation chain.

Simultaneous localization and mapping (SLAM) estimates sensor trajectory while mapping the surroundings. Parallel Tracking and Mapping separates rapid tracking from slower map refinement for small AR workspaces. ORB-SLAM3 supports visual and visual–inertial localization and map reuse (Klein and Murray 2007; Campos et al. 2021). These maps locate devices relative to observed structure without defining generated objects' behavior. Registration aligns virtual and physical structures, accounting for tracking and calibration. Static and motion-dependent errors can make overlays drift or appear detached (Azuma and Bishop 1994; Holloway 1997).

Optical see-through displays transmit environmental light and add virtual imagery; video see-through displays combine camera images with rendered content. They differ in occlusion, brightness, camera-to-eye alignment, and delay (Azuma et al. 2001). Placing a virtual object behind a real desk requires aligned surface and depth estimates. Immersive VR controls the entire visible environment but still requires a consistent mapping from head movement to displayed views.

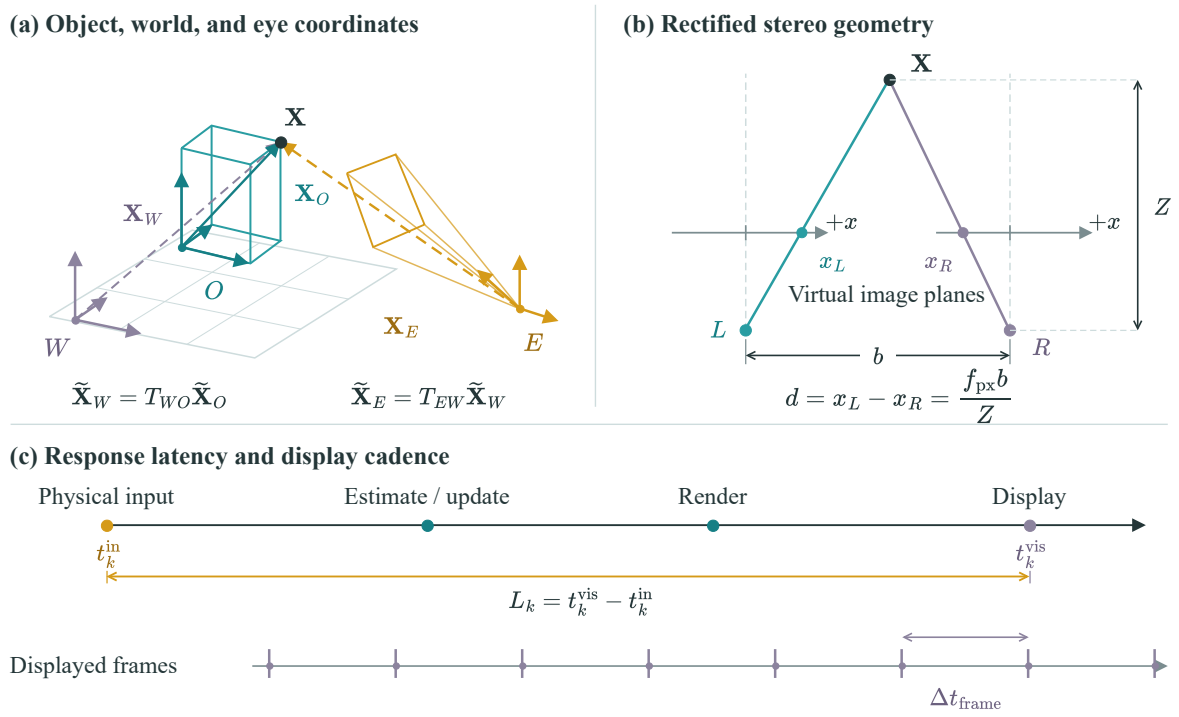



**Fig. 5** XR geometry and timing: (a) coordinate transformations; (b) rectified stereo projection; (c) response latency and display cadence.

### 2.2.2 Binocular perception, comfort, and embodiment

Humans recover depth from binocular disparity, motion parallax, occlusion, perspective, and changes in image sharpness. Binocular disparity compares the projections of the same scene point into the two eyes. In the simplified case of rectified pinhole cameras with parallel optical axes, a horizontal baseline, equal focal lengths, and aligned principal points,

$$d = x_L - x_R = \frac{f_{\text{px}} b}{Z}. \tag{2}$$

Here $x_L$ and $x_R$ are horizontal pixel coordinates, $f_{\text{px}}$ is focal length in pixels, and $b$ is camera separation. Forward depth $Z$ uses the same length units as $b$. Figure 5(b) shows this geometry: a nearer point produces a larger disparity. This relation applies to corresponding points visible in both images at the same scene time; general eye configurations require calibrated projections. Motion parallax instead concerns relative image movement as the observer moves. Stereoscopic presentation must coordinate these cues with viewing geometry and timing (Banks et al. 2012). Interpupillary distance (IPD), the distance between pupil centers, informs the eye separation used for rendering; a mismatch can alter perceived scale or depth. Generated stereo must also handle regions visible to only one eye, where paired correspondence is unavailable.

Vergence rotates the eyes toward a common fixation point; accommodation changes the eye's optical focus. In natural viewing, both normally correspond to the fixated object's distance. A conventional fixed-focus stereoscopic display can depict an object at a virtual distance that differs from the display's optical focal distance. This vergence–accommodation conflict (VAC) can impair binocular performance and cause visual fatigue. Its effects depend on viewing distance and the direction of the conflict (Hoffman et al. 2008; Shibata et al. 2011). Geometric stereo consistency and comfortable presentation are therefore related but distinct. Matching the two images does not change the headset's physical focal behavior.

Multiple focal planes and gaze-contingent adaptive focus address this mismatch through different optical and computational trade-offs (Akeley et al. 2004; Padmanaban et al. 2017). Their suitability depends on the viewing task and observer (Kramida 2016). Generation can influence disparity, depth ordering, temporal stability, and content placement; the display determines the resulting optical stimuli. Evaluation must therefore describe the headset and viewing geometry alongside the generated content, particularly for near-object manipulation.

Immersion describes system sensory and interaction capabilities, whereas presence concerns the experience of being in the represented environment (Slater and Wilbur 1997). Place illusion concerns being somewhere; plausibility concerns whether unfolding events feel real (Slater 2009). Presence is a perceptual and behavioral phenomenon, rather than a synonym for image realism (Sanchez-Vives and Slater 2005). A meta-analysis distinguishes contributions from tracking, stereoscopic presentation, field of view, and other system properties (Cummings and Bailenson 2016).

Embodiment concerns users' relationship with represented bodies, including self-location, movement agency, and body ownership (Kilteni et al. 2012). First-person virtual-body experiments show that body experience depends on presentation and coupling to the participant (Slater et al. 2010). Generated hands or avatars must therefore couple appearance and movement to user actions and nearby objects. Reviewed studies associate presence with cybersickness and suggest possible shared perceptual causes (Weech et al. 2019). Visual quality, bodily control, presence, and discomfort require separate measurements.

### 2.2.3 The interaction loop and its time scales

An XR application repeatedly senses the user, estimates pose, interprets input, updates scene state, renders eye views, and presents them. Motion-to-photon (MTP) latency measures the delay from physical movement to its displayed consequence, including sensing, computation, transport, scheduling, rendering, and display contributions (Friston and Steed 2014). For paired input and response events, distinguish

$$L_k = t_k^{\mathrm{vis}} - t_k^{\mathrm{in}}, \qquad f_{\mathrm{out}} = \frac{N-1}{t_N^{\mathrm{out}} - t_1^{\mathrm{out}}}. \tag{3}$$

The timestamps $t$ are wall-clock measurements. The first expression pairs input event $k$ with its corresponding visible response; the movement and response-detection rule must be specified. The second measures the average rate of $N > 1$ output events. Its inverse is the mean interval between outputs, not their response delay; Figure 5(c) separates these measurements. The output endpoint must also be named: generated frames, rendered frames, and display presentations can have different rates. Asynchronous time warp uses newer tracking information to transform an already rendered stereo pair near presentation (van Waveren 2016). Such reprojection reduces apparent viewpoint delay but does not execute a new object interaction or provide every surface revealed by a large viewpoint change.

Perceptual thresholds depend on the motion and protocol being tested. Mania et al. found latency discrimination largely unchanged across tested scene-complexity conditions for trained observers. Jerald and Whitton related latency thresholds to scene motion induced by head turns (Mania et al. 2004; Jerald and Whitton 2009). Controller-motion measurements further show that prediction produces different delays at movement onset and during predictable motion (Warburton et al. 2023).

Selection identifies objects; manipulation changes their position, orientation, or scale; locomotion changes the user's virtual location. Go-Go extends reach through nonlinear mapping, while remote-manipulation studies distinguish object acquisition from subsequent control (Poupyrev et al. 1996; Bowman and Hodges 1997). Body-relative techniques exploit proprioception, the sense of body and limb configuration (Mine et al. 1997). Redirected walking changes the physical-to-virtual movement mapping within experimentally studied perceptual limits (Steinicke et al. 2010). These mappings belong to the interface and inform interpretation of generative systems' control claims.

The instruction "make this chair taller" requires resolving an entity, preserving accepted placement, modifying geometry, and updating the display. A later grasp requires identifiable, selectable or collidable content that follows the intended manipulation. Collaborative use adds compatible object state and awareness of others' actions (Ens et al. 2019). Mini-Me adapts a remote collaborator's avatar to communicate gaze and gestures across AR and VR viewpoints (Piumsomboon et al. 2018).

Display and interaction fidelity can affect task performance and subjective experience (McMahan et al. 2012). Presence questionnaires require constructs and administration suited to the study question (Witmer and Singer 1998; Schubert et al. 2001). Comparisons across real and virtual settings further show that scores require interpretation within their protocol (Usoh et al. 2000). Visual output, retained application state, and participant task outcomes thus describe different stages of the interaction loop.

## 2.3 Spatial representations and rendering

We distinguish generating content from reconstructing observations, then introduce geometric representations and the rendering operations that turn them into images.

### 2.3.1 Generating and reconstructing spatial content

A visual generator may produce an image, a set of views, or spatial content renderable from new viewpoints. Walking around a stationary chair reveals occluded surfaces while the chair remains fixed. Rendering predicts an observation from a scene and camera; reconstruction infers scene properties from observations. Depth measurements can be fused into a shared volume (Curless and Levoy 1996), whereas a single photograph leaves hidden surfaces underdetermined. Learned image-conditioned methods supply additional information: pixelNeRF conditions a radiance field on features learned across scenes, while IBRNet combines features from available source views (Yu et al. 2021; Wang et al. 2021b).

We distinguish shared model parameters $\theta$ from parameters $\varphi$ fitted to one asset or scene. A trained decoder $D_\theta$ maps a latent code $z$ to content. Reconstruction may fit $\varphi$ against observations or infer content using a model learned across scenes (Lu et al. 2026a; Caldeira et al. 2026). A condition $c$ supplies a prompt, reference image, or other constraint. Reconstruction seeks an explanation of observations; generation samples content satisfying a request. Both can use learned priors, so the presence of a neural network does not distinguish these tasks.

Camera parameters $v$ contain intrinsics and extrinsics; physical time $\tau$ indexes scene evolution. In Eq. (5), changing $v$ at fixed $A_\varphi(\tau)$ changes viewpoint; advancing $\tau$ at fixed $v$ reveals temporal change. D-NeRF and K-Planes incorporate temporal variation (Pumarola et al. 2021; Fridovich-Keil et al. 2023), but time-indexed representations alone do not define responses to arbitrary new actions.

### 2.3.2 Geometry: surfaces, samples, and continuous fields

Geometry describes the spatial extent of matter or the location of surfaces. A triangle mesh stores vertex positions $V$ and triangular connectivity $F$; a point cloud retains positions without connectivity. Poisson surface reconstruction recovers a global surface from points and estimated orientations (Kazhdan et al. 2006). A voxel grid partitions space into cells; doubling resolution along all three axes multiplies the cell count by eight. Sparse storage reduces this burden when much of the volume is empty.

Implicit representations specify surfaces through functions queried at spatial locations $\mathbf{x}$. In a shared decoder parameterization with shape code $z$, the alternatives can be summarized as

$$\begin{aligned} M &= (V, F), \quad V = \{\mathbf{x}_j\}_{j=1}^{n_v}, \quad F \subseteq \{1, \dots, n_v\}^3, \\ S_z^{\mathrm{occ}} &= \{\mathbf{x} : o_\theta(\mathbf{x}, z) = \eta\}, \quad o_\theta(\mathbf{x}, z) \in [0, 1], \\ S_z^{\mathrm{sdf}} &= \{\mathbf{x} : f_\theta(\mathbf{x}, z) = 0\}. \end{aligned} \tag{4}$$

Each row of $F$ indexes a triangle in the mesh. An occupancy decoder $o_\theta$ estimates the probability of a point being inside; a chosen threshold $\eta \in (0, 1)$ defines its surface (Mescheder et al. 2019). For a closed surface, an exact signed distance function (SDF) has magnitude equal to the shortest Euclidean distance to that surface. Its sign is negative inside under the convention used here. DeepSDF learns an approximation $f_\theta$ together with shape-specific codes (Park et al. 2019). Its zero level set defines the predicted surface, but fitting a neural field does not guarantee exact distances everywhere. Occupancy values, distances, and mesh connectivity support different computations even when describing similar shapes.

The word "neural" describes a function's parameterization, not its physical quantity. Networks can predict occupancy or distance, with local features adding detail as in Convolutional Occupancy Networks (Peng et al. 2020). Fourier features and SIREN's periodic activations help coordinate networks represent rapidly changing signals (Tancik et al. 2020; Sitzmann et al. 2020).

Implicit surfaces can enter mesh pipelines through field sampling and triangle extraction. Marching Cubes constructs triangles where a chosen scalar value crosses grid cells (Lorensen and Cline 1987). Sphere tracing instead renders distance fields directly, advancing rays by distance-based steps whose valid bounds prevent surface crossing (Hart 1996).

A neural radiance field (NeRF) represents density $\sigma_\varphi(\mathbf{x}) \geq 0$ and direction-dependent RGB radiance $\mathbf{r}_\varphi(\mathbf{x}, \mathbf{d})$. Here $\mathbf{d}$ is a unit viewing direction and $\varphi$ contains the fitted scene parameters (Mildenhall et al. 2020). Density determines ray attenuation, rather than occupancy probability or signed distance; Eq. (6) gives its pixel contributions. However, density alone does not identify a unique watertight surface. NeuS and VolSDF incorporate signed-distance geometry into volume rendering to recover surfaces more directly from images (Wang et al. 2021a; Yariv et al. 2021). Photographic reconstruction quality and the quality of a collision mesh therefore remain distinct.

Gaussian splatting stores primitives with centers $\boldsymbol{\mu}_i$, covariance matrices $\Sigma_i$, opacities $\beta_i$, and view-dependent appearance (Kerbl et al. 2023); covariance specifies each Gaussian's extent and orientation.

Projected footprints are blended as in Eq. (7); the primitives need not form a connected surface mesh. Alternatives include oriented planar elements in 2D Gaussian Splatting, surface alignment and mesh extraction in SuGaR, and anchor-organized, view-adaptive Gaussians in Scaffold-GS (Huang et al. 2024a; Guédon and Lepetit 2024; Lu et al. 2024).

### 2.3.3 Appearance, rendering, and learning from images

Pixel color depends on geometry, material, lighting, visibility, and viewing direction (Huang et al. 2026a; Jiang et al. 2025a; Huang et al. 2025b; Jiang et al. 2025b). Texture maps attach spatially varying values, such as base color or roughness, to surfaces through texture coordinates; they may also contain baked shadows. A bidirectional reflectance distribution function describes how incident light is redistributed toward outgoing directions; Cook–Torrance provides a classic physically motivated model (Cook and Torrance 1982).

The rendering equation expresses the interaction between emitted light, incoming illumination, and surface reflection (Kajiya 1986). For this review, a compact interface is sufficient:

$$I_{v,\tau} = \mathcal{R}\big(A_\varphi(\tau), v, \ell\big). \tag{5}$$

Here $I_{v,\tau}$ is the rendered image, and $A_\varphi(\tau)$ specifies asset or scene geometry and appearance at physical time $\tau$. The camera is specified by $v$, illumination by $\ell$, and the rendering procedure by $\mathcal{R}$. For static content, this representation does not change with $\tau$. For radiance stored under fixed lighting, illumination may already be absorbed into $A_\varphi$ rather than exposed as a controllable input.

Rasterization projects surface primitives and determines pixel coverage, whereas ray-based rendering follows lines of sight. For a NeRF, sample a camera ray at positions $\mathbf{x}_i = \mathbf{o} + q_i\mathbf{d}$. Here $\mathbf{o}$ is its origin and $q_i$ is distance along the unit direction $\mathbf{d}$. Discrete volume rendering gives (Mildenhall et al. 2020)

$$\begin{aligned} \widehat{I}_v(\mathbf{u}) &= \sum_{i=1}^{N} T_i \alpha_i \mathbf{r}_\varphi(\mathbf{x}_i, \mathbf{d}) + T_{N+1} I_{\mathrm{bg}}, \\ \alpha_i &= 1 - \exp[-\sigma_\varphi(\mathbf{x}_i)\Delta_i], \qquad T_i = \prod_{j<i}(1-\alpha_j). \end{aligned} \tag{6}$$

Here $\Delta_i$ is the sample interval along the ray, $\alpha_i$ is that interval's opacity, and $T_i$ is the fraction transmitted through preceding intervals. Thus $T_i\alpha_i$ weights the sample's radiance. The residual transmittance $T_{N+1}$ weights the background color $I_{\mathrm{bg}}$; the empty product is one. Density has inverse-length units, so $\sigma_i\Delta_i$ is dimensionless. The weights couple visibility to pixel color.

Gaussian splatting uses a related compositing rule but computes opacity from projected primitives rather than ray-sampled density (Kerbl et al. 2023):

$$\begin{aligned} \widehat{I}_v(\mathbf{u}) &= \sum_{i=1}^{N_{\mathbf{u}}} T_i(\mathbf{u}) \alpha_i(\mathbf{u}) \mathbf{r}_i(\mathbf{d}_i) + T_{N_{\mathbf{u}}+1}(\mathbf{u}) I_{\mathrm{bg}}, \\ \alpha_i(\mathbf{u}) &= \beta_i G_i^{2D}(\mathbf{u}), \qquad T_i(\mathbf{u}) = \prod_{j<i}[1-\alpha_j(\mathbf{u})]. \end{aligned} \tag{7}$$

The $N_{\mathbf{u}}$ splats overlapping pixel $\mathbf{u}$ are indexed in the renderer's depth order. $G_i^{2D}$ is the unit-peak Gaussian footprint obtained by locally projecting its 3D covariance, without normalization to unit integral. Learned opacity is $\beta_i \in [0,1]$, and $\mathbf{r}_i(\mathbf{d}_i)$ is directional RGB appearance. These equations give the basic blending rule; practical rasterizers also filter, threshold, and approximate ordering.

Differentiable rendering estimates how pixels change with scene parameters, allowing an image loss to update a 3D representation. Neural 3D Mesh Renderer introduces approximate rasterization gradients, while IDR jointly estimates implicit geometry, appearance, and camera parameters (Kato et al. 2018; Yariv et al. 2020).

Inverse rendering seeks scene components that explain the observed appearance. NeRFactor separates albedo, reflectance, visibility, and illumination; nvdiffrec jointly extracts geometry, materials, and lighting (Zhang et al. 2021; Munkberg et al. 2022). Ref-NeRF structures view-dependent reflected radiance to better represent glossy surfaces (Verbin et al. 2022). Textures and materials require separate treatment because shading recomputed under changed lighting differs from colors stored under the original illumination.

Storage and sampling also influence rendering cost and image quality (Wang et al. 2026c,b). Plenoxels optimizes a sparse voxel radiance field without a neural network, whereas TensoRF factorizes a field into compact tensor components. Instant-NGP uses multiresolution hash features with small neural networks (Fridovich-Keil et al. 2022; Chen et al. 2022; Müller et al. 2022).

Pixels cover finite areas; ignoring their footprints can cause aliasing under camera movement or resolution changes. Mip-NeRF integrates features over conical ray regions, Mip-NeRF 360 extends to unbounded scenes, and Zip-NeRF combines scale-aware sampling with grid acceleration (Barron et al. 2021, 2022, 2023). Mip-Splatting handles analogous scale changes through spatial smoothing and image-space filtering (Yu et al. 2024b).

## 2.4 Generative learning, sampling, and state

Generative mechanisms determine how these representations are learned, sampled, and updated. Distribution learning supplies reusable priors, sampling produces content under conditions, and temporal models determine how history and actions influence subsequent outputs.

### 2.4.1 Learning visual distributions and latent representations

A visual generator models a distribution $p_\theta(Y \mid c)$ over outputs $Y$, given conditions $c$ and shared parameters $\theta$ learned from training data. The output may be an image, a video, or a representation rendered by another component. A wooden-chair prompt constrains category and material while leaving exact shape open; sampling produces one realization of that uncertainty. Conditions therefore restrict the set of possible outputs without necessarily specifying them completely (Huang et al. 2026b).

Many models generate through a *latent variable* $z$. A variational autoencoder (VAE) learns an encoder distribution $q_{\theta_E}(z \mid Y)$ and a probabilistic decoder $p_{\theta_D}(Y \mid z)$. Its standard unconditional objective maximizes the evidence lower bound (ELBO) (Kingma and Welling 2014; Rezende et al. 2014):

$$\begin{aligned}\mathcal{L}_{\mathrm{VAE}}(Y) = \mathbb{E}_{q_{\theta_E}(z|Y)}[\log p_{\theta_D}(Y \mid z)] \\ - \mathrm{KL}(q_{\theta_E}(z \mid Y) \,\|\, p(z)) \le \log p_{\theta_D}(Y).\end{aligned} \tag{8}$$

The first term rewards reconstruction, and the Kullback–Leibler divergence regularizes inferred codes against a chosen prior $p(z)$. The encoder and decoder parameters $\theta_E, \theta_D$ are shared across examples; they differ from the instance parameters $\varphi$ introduced earlier (Huang et al. 2023; Wu et al. 2025b). Sampling draws a code from the prior and decodes it. A latent variable need not compress the data: invertible normalizing flows such as Real NVP retain dimensionality while permitting exact density evaluation (Dinh et al. 2017). Latent diffusion instead compresses images with a perceptual autoencoder before learning a generative prior (Rombach et al. 2022). The prior controls which codes are sampled, while the decoder limits the details they can express.

A generative adversarial network (GAN) learns a generator $G_{\theta_G}$ through competition with a discriminator $C_{\theta_C}$ (Goodfellow et al. 2014). In the original conditional minimax formulation, both networks receive $c$ (Mirza and Osindero 2014):

$$\begin{aligned}\min_{\theta_G} \max_{\theta_C} \quad & \mathbb{E}_{(Y,c)\sim p_{\mathrm{data}}}[\log C_{\theta_C}(Y, c)] \\ & + \mathbb{E}_{c\sim p_{\mathrm{data}}(c),\, z\sim p(z)}[\log(1 - C_{\theta_C}(G_{\theta_G}(z, c), c))].\end{aligned} \tag{9}$$

The discriminator scores whether a sample comes from the training distribution, and the generator maps noise and a condition to an output. The discriminator supplies the training signal; inference uses the generator. StyleGAN additionally organizes variation across synthesis scales (Karras et al. 2019).

Conditions supplied to a generator include labels, text, images, and spatial maps. Contrastive Language–Image Pre-training (CLIP) learns aligned image and text representations, while Imagen uses a pretrained language model to encode prompts (Radford et al. 2021; Saharia et al. 2022). Camera parameters or depth provide further constraints when accepted by the generator. Discrete representations select entries from a learned codebook: VQ-VAE learns such codes, and VQGAN combines perceptual compression with a transformer prior (van den Oord et al. 2017; Esser et al. 2021). Tokens can encode image or video components without identifying named objects. Token histories therefore differ from XR object records.

### 2.4.2 Denoising, probability transport, and sampling

Denoising diffusion probabilistic models (DDPMs) learn to reverse a process that progressively corrupts data (Sohl-Dickstein et al. 2015; Ho et al. 2020). For a clean representation $z_0$, a noisy training input is

$$z_s = \alpha_s z_0 + \sigma_s \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \qquad s \in [0, 1]. \tag{10}$$

Here $s$ is diffusion time, $\alpha_s$ and $\sigma_s$ control signal and noise strength, and $I$ is the identity covariance. The clean endpoint has $\alpha_0 = 1$, $\sigma_0 = 0$; the terminal distribution is chosen to approximate standard Gaussian noise. A common training objective and a reverse step on a discrete DDPM sampling grid are

$$\begin{aligned} \mathcal{L}_{\text{den}}(\theta) &= \mathbb{E}_{z_0,c,s,\epsilon}\left[w(s)\|\epsilon_\theta(z_s, s, c) - \epsilon\|_2^2\right], \\ p_\theta(z_{s-\Delta s} \mid z_s, c) &= \mathcal{N}(\mu_\theta(z_s, s, c), \Sigma_\theta(z_s, s, c)), \\ \nabla_{z_s} \log q_s(z_s \mid c) &\approx -\epsilon_\theta(z_s, s, c)/\sigma_s, \qquad \sigma_s > 0. \end{aligned} \tag{11}$$

Training samples data–condition pairs, noise levels, and Gaussian noise; $w(s)$ weights noise levels and equals one in the simplified DDPM objective. The predicted noise parameterizes the reverse mean $\mu_\theta$; the covariance $\Sigma_\theta$ can be fixed or learned. Sampling starts near $s = 1$ and takes successive steps $\Delta s > 0$ toward $s = 0$. The last line links noise prediction to the *score* of the noisy data marginal $q_s$: a local gradient in representation space. Score-based models use it in reverse stochastic dynamics or an associated probability-flow ordinary differential equation (ODE) (Song et al. 2021b). The sampled clean representation is decoded as $Y = D_\theta(z_0)$; this decoder is the identity for pixel-space diffusion. Diffusion time $s$ is distinct from physical time $\tau$ in a video.

Flow matching trains a velocity field directly (Lipman et al. 2023). A linear-path example, closely related to rectified flow (Liu et al. 2023c), uses the same clean-to-noise orientation as Eq. 10:

$$\begin{aligned} \widetilde{z}_s &= (1 - s)z_0 + s\epsilon, \\ \mathcal{L}_{\text{CFM}}(\theta) &= \mathbb{E}_{z_0,c,\epsilon,s}\left[\|u_\theta(\widetilde{z}_s, s, c) - (\epsilon - z_0)\|_2^2\right], \\ \frac{\mathrm{d}z_s}{\mathrm{d}s} &= u_\theta(z_s, s, c), \qquad z_1 \sim \mathcal{N}(0, I), \quad s : 1 \longrightarrow 0. \end{aligned} \tag{12}$$

During training, a data code $z_0$ and independent Gaussian noise $\epsilon$ define a known interpolation and target velocity; $s$ is sampled uniformly. Conditional flow matching refers to these endpoint-conditioned training paths; external context $c$ additionally conditions the learned field. At inference, the ODE follows the fitted field from noise to data without access to the training endpoint $z_0$. Straight training paths need not yield straight sampled trajectories; rectification can improve numerical efficiency. This velocity changes a representation during synthesis rather than describing physical object motion.

The model and its sampler are separate choices. DDIM constructs alternative paths compatible with DDPM training, while DPM-Solver integrates the diffusion ODE with a dedicated higher-order method (Song et al. 2021a; Lu et al. 2022). Progressive distillation trains a student to replace successive teacher steps (Salimans and Ho 2022). Consistency models map trajectory points to a common clean endpoint and support one-step or multistep generation, with distillation or independent training (Song et al. 2023). These procedures change the sampling work that contributes to the response costs examined later.

Guidance modifies sampling to emphasize a condition: classifier guidance uses classifier gradients, whereas classifier-free guidance combines conditional and unconditional predictions (Dhariwal and Nichol 2021; Ho and Salimans 2022). Adjusting guidance strength trades diversity against fidelity to the supplied condition. Architecture is a separate choice: a diffusion transformer (DiT) operates on latent patches. Scaled rectified-flow transformers exchange information between text and image tokens (Peebles and Xie 2023; Esser et al. 2024).

### 2.4.3 Autoregression, temporal context, and world state

Autoregressive generation decomposes a joint distribution into ordered conditional predictions:

$$p_\theta(z_{1:N} \mid c) = \prod_{i=1}^{N} p_\theta(z_i \mid z_{<i}, c). \tag{13}$$

Here $N$ is sequence length, $z_i$ is token $i$, and $z_{<i}$ denotes preceding tokens. PixelRNN predicts pixels sequentially; transformers combine tokens through learned attention (van den Oord et al. 2016; Vaswani et al. 2017). This prediction order may follow spatial positions or physical time. MaskGIT instead refines masked token sets bidirectionally; continuous-token autoregressive models can train each conditional with diffusion (Chang et al. 2022; Li et al. 2024). Autoregression orders prediction units, while diffusion can model and sample each unit's conditional distribution, including frames or chunks in Section 3.3.

Interaction conditions prediction on actions. Let $k$ index rollout steps, $o_k$ be an observation, and $a_k$ the action applied between steps $k-1$ and $k$. An observation generator can predict $p_\theta(o_k \mid \mathcal{H}_{k-1}, a_k, c)$ directly from a retained history $\mathcal{H}_{k-1}$ of observations and actions. For fixed external context $c$, suppressed below, a latent state-space model summarizes the past in $z_{k-1}^{\text{state}}$ and separates transition from observation generation:

$$\begin{aligned} z_k^{\text{state}} &\sim p_\theta(z_k^{\text{state}} \mid z_{k-1}^{\text{state}}, a_k), \\ o_k &\sim p_\theta(o_k \mid z_k^{\text{state}}). \end{aligned} \tag{14}$$

The state may contain deterministic recurrent memory and stochastic variables. PlaNet learns such transitions for planning from pixels, and Dreamer learns behavior through imagined latent trajectories (Hafner et al. 2019, 2020). The training target determines what the state is encouraged to retain (Huang 2022). MuZero predicts reward, policy, and value for planning, while V-JEPA 2 learns predictive embeddings and post-trains an action-conditioned variant for robotic planning (Schrittwieser et al. 2020; Assran et al. 2025).

Explicit application state exposes object identities, transforms, or geometry to queries and updates. A predictive latent state can support useful rollouts without exposing these quantities. Observation history can preserve visible context without specifying how an accepted object edit persists. Section 3.3.5 examines how these mechanisms store information and update it.

# 3 Methods: Assets, Scenes, and Dynamic Worlds

XR content generation involves three decisions: which structure an asset retains, how a scene organizes or extends it, and how a world responds to time and actions. We compare conditioning, representations, and update operators to explain their consequences for use. Table 1 catalogues methods in three domain-specific parts; Table 7 records documented XR integrations.

**Reading the catalogues.** Input tags can indicate alternative conditioning modes. Stage tags identify preparation before use (P), interactive authoring including asynchronous completion (A), or generation during an ongoing rollout (R). They describe workflow placement, not speed; rendering prepared content is separate. An arrow opens a verified official project or GitHub page; a dash indicates that no such page was verified.

The equations use the notation from Section 2: $c_{\text{txt}}$ and $c_{\text{img}}$ denote text and image conditions, and $k$ indexes construction or rollout steps. A distribution $p_\theta$ denotes the sampler-induced distribution without assuming a tractable likelihood. Operator chains describe retrieval, optimization, or geometric fusion. Camera queries render existing content; camera conditions guide its generation.

## 3.1 Generative Models for Visual Assets

Asset generation must infer missing shape and appearance while determining what can later be edited independently. We follow these decisions from 3D construction to materials, parts, and delivery. Let $A$ denote an asset and $A_\varphi$ its representation when fitted per instance; Table 1, Part A, catalogues the methods.

### 3.1.1 From image priors to native 3D generation

The first choice is where missing 3D structure is inferred. Image-guided optimization fits an asset through rendered views; view-based methods reconstruct it from generated observations; native 3D models sample spatial variables learned from 3D data.

**Representation priors and surface construction.** Early learned 3D priors differed in their geometric output. 3D-GAN generated volumetric shapes, latent point-cloud models included learned autoencoders and Gaussian mixtures, and PointFlow separated variation across shapes from variation among their points (Wu et al. 2016; Achlioptas et al. 2018; Yang et al. 2019). Text alignment requires a semantic condition or objective beyond these geometric distributions.

Surface parameterization and extraction connect these representations to mesh outputs. AtlasNet generates parametric surface patches (Groueix et al. 2018). DMTet combines deformable tetrahedral grids

with differentiable surface extraction; FlexiCubes adds control over isosurface geometry and connectivity during gradient-based optimization (Shen et al. 2021b, 2023). Training supervision can also differ from the generated output: GET3D learns textured mesh generation from images, whereas EG3D learns a hybrid representation for 3D-aware image synthesis (Gao et al. 2022; Chan et al. 2022).

**Semantic guidance and per-asset optimization.** Dream Fields optimizes rendered views toward a text description in CLIP's embedding space (Jain et al. 2022). Text2Mesh applies semantic guidance to the colors and local geometry of a supplied mesh, while CLIP-Mesh optimizes a surface together with texture and normal maps (Michel et al. 2022; Mohammad Khalid et al. 2022). A supplied mesh provides a starting structure, but local deformation means that its exact silhouette need not remain fixed.

DreamFusion keeps a pretrained image diffusion model fixed and optimizes an individual 3D representation through differentiable rendering (Poole et al. 2023). Let $I_v = \mathcal{R}(A_\varphi, v)$ denote a rendered image at camera $v$, with other rendering settings suppressed. A noisy image and the score distillation sampling (SDS) gradient are

$$\begin{aligned} I_{v,s} &= \alpha_s I_v + \sigma_s \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \\ g_{\mathrm{SDS}}(\varphi) &= \mathbb{E}_{v,s,\epsilon}\left[ w(s) \left( \frac{\partial I_v}{\partial \varphi} \right)^T \mathrm{sg}(\widehat{\epsilon}_\theta(I_{v,s}, c_{\mathrm{txt}}, s) - \epsilon) \right]. \end{aligned} \tag{15}$$

Here $s$ is diffusion time; $\alpha_s$ and $\sigma_s$ define the noise schedule, while $w(s)$ weights noise levels and absorbs schedule-dependent factors. The guided noise predictor $\widehat{\epsilon}_\theta$ supplies the image-space residual, which the stop-gradient operator sg holds fixed during differentiation. The update $\varphi \leftarrow \varphi - \eta g_{\mathrm{SDS}}$ therefore differentiates through the renderer but not the denoiser; $\eta$ is the step size. Equation (15) describes the prior-guided gradient; method-specific regularizers are handled separately. DreamFusion instantiates $A_\varphi$ as a density-and-albedo field rendered with explicit shading (Poole et al. 2023).

Magic3D first optimizes a sparse hash-grid field, then refines a textured mesh through differentiable rasterization (Lin et al. 2023). Its fine stage applies diffusion guidance to an encoded rendering $z_v = E_\theta(I_v)$, adding the image encoder $E_\theta$ to the gradient path through its Jacobian. ProlificDreamer models a distribution of 3D parameters in score distillation; DreamGaussian optimizes Gaussian primitives and subsequently converts them to a textured mesh (Wang et al. 2023b; Tang et al. 2024c).

**Generated views and reconstruction.** Camera-conditioned image priors provide evidence about unseen surfaces. Zero-1-to-3 predicts an object view from an image and relative camera change; SyncDreamer couples multiple generated views, and MVDream learns a multiview diffusion prior (Liu et al. 2023b, 2024b; Shi et al. 2024). Zero123++ similarly targets consistent image sets. Wonder3D and Unique3D additionally generate surface-normal cues that guide mesh reconstruction (Shi et al. 2023; Long et al. 2024; Wu et al. 2024b). Consistency in these intermediate predictions matters because a later stage must consolidate them into one surface.

Magic123 fits a per-object field and mesh using reference-view supervision with 2D and 3D diffusion priors (Qian et al. 2024). LRM instead maps an image through shared weights to three coordinate-aligned feature planes, forming a triplane radiance representation (Hong et al. 2024). One-2-3-45 and InstantMesh reconstruct surfaces or meshes from generated views using learned reconstructors (Liu et al. 2023a; Xu et al. 2024a). LGM and GRM reconstruct Gaussians from multiview images, including generated views (Tang et al. 2024a; Xu et al. 2024c). One-2-3-45++ uses the views to condition native 3D diffusion (Liu et al. 2024a).

**Native 3D distributions.** Other methods learn distributions directly over encoded spatial variables. LION uses global and point-structured latents; SDFusion encodes signed-distance geometry with multimodal conditions; MeshDiffusion models deformable tetrahedral grids (Zeng et al. 2022; Cheng et al. 2023; Liu et al. 2023d).

A learned sampler draws latent content while keeping shared model parameters fixed:

$$\begin{aligned} z &\sim p_\theta(z \mid c), \qquad A = D_\theta(z), \\ p_\theta(P, Z \mid c) &= p_\theta(P \mid c)\, p_\theta(Z \mid P, c) \qquad \text{(TRELLIS)}. \end{aligned} \tag{16}$$

The first line summarizes sampling and representation-specific decoding. For TRELLIS, $z = (P, Z)$ contains active voxel positions $P$ and their local features $Z$. Separate rectified-flow models generate sparse structure and then its features, conditioned on text or an image (Xiang et al. 2025). Different decoders map these structured latents to meshes, radiance fields, or Gaussians. Locality supports regional editing, while factorization permits detail variation with coarse structure held fixed. Learned sampling and decoding replace repeated fitting of $\varphi$, although sampling can still require multiple numerical steps.

Reconstruction metrics test the representation and decoder; generation metrics also test the sampler and condition adherence.

Other native models make different representation choices: Shap-E generates parameters of implicit functions, LN3Diff models latent neural fields, and 3DTopia-XL generates primitives encoding geometry and appearance (Jun and Nichol 2023; Lan et al. 2024; Chen et al. 2025c).

### 3.1.2 Geometry, appearance, and materials

**Separating shape from appearance.** Once geometry is generated, appearance must align with its surfaces and support the intended rendering controls. Let $M$ denote mesh geometry and $\mu$ its appearance attributes, ranging from a texture to explicit material channels. Geometry-first pipelines expose the dependency

$$p_\theta(M, \mu \mid c_{\text{img}}) = p_\theta(M \mid c_{\text{img}})\, p_\theta(\mu \mid M, c_{\text{img}}), \qquad A = (M, \mu). \tag{17}$$

This factorization separates the initial output from the quantities it constrains. In Hunyuan3D 2.0, image-conditioned shape latents decode to geometry. Rendered normal and position maps then condition multiview image synthesis, followed by texture baking (Zhao et al. 2025). The second factor thus summarizes several operations rather than a direct material decoder. Shape reconstruction, condition alignment, and texture assessments evaluate different components of this chain.

TRELLIS.2 represents local geometry and physically based rendering (PBR) attributes in O-Voxel, including base color, metallic value, roughness, and opacity (Xiang et al. 2026). Three stages generate sparse occupancy, geometry latents, and material latents. The material model conditions on the image and generated geometry latents. Separate shape and material autoencoders preserve spatial alignment. Unlike Hunyuan3D's multiview synthesis and baking, material generation operates directly in native 3D latent space. O-Voxel's flexible dual-grid surface representation also accommodates open and non-manifold geometry.

Both routes support appearance generation on supplied geometry, holding $M$ fixed while changing $\mu$ (Zhao et al. 2025; Xiang et al. 2026). Explicit material channels provide inputs for changing illumination. The delivered appearance must be assessed under the application's renderer and lighting.

Appearance representation also determines the available editing operation. Seed3D 2.0 combines coarse-to-fine geometry generation with a unified mixture-of-experts model for multiview albedo and metallic–roughness maps. Part decomposition and articulation construction are additional pipeline stages (Gu et al. 2026). MaterialApprentice retrieves expert construction traces, synthesizes a process, and compiles it into shader operations to produce an executable Blender material graph (Gupta et al. 2026). The former delivers explicit material channels; the latter retains the construction program for subsequent edits. Their evaluations accordingly emphasize rendered-asset preference and material-generation quality or expert editing effort.

**Texturing and relighting.** TEXTure and Text2Tex condition texture generation on supplied geometry and consolidate observations across viewpoints (Richardson et al. 2023; Chen et al. 2023a). Fantasia3D instead optimizes geometry and appearance with distinct components, including a physically based material representation (Chen et al. 2023b). The distinction is which quantities may change: painting an existing surface and synthesizing its geometry are different authoring operations.

A texture may also contain illumination that should change with the environment. Paint3D targets textures without embedded illumination through multiview generation and UV-space refinement, while SceneTex uses a multiresolution texture field and reference-based attention for consistent indoor-scene appearance (Zeng et al. 2024; Chen et al. 2024b). Relighting comparisons concern the renderer controls introduced through inverse rendering in Section 2.3.3 (Zhang et al. 2021; Munkberg et al. 2022).

### 3.1.3 Structured and compositional assets

**Generating addressable parts.** Editing individual components requires a different output unit from editing an object's appearance. PartCrafter generates a set of coordinated meshes from a red–green–blue (RGB) image and a requested part count $n$:

$$z_{1:n} \sim p_\theta(z_{1:n} \mid c_{\text{img}}, n), \qquad A = \{M_i = D_\theta(z_i)\}_{i=1}^{n}. \tag{18}$$

Each $z_i$ contains one part's tokens, decoded in a shared canonical coordinate system. The count $n$ sets the number of token groups, distinguished by learned part identities (Lin et al. 2025). Local attention processes each group, global attention exchanges information across groups, and image cross-attention

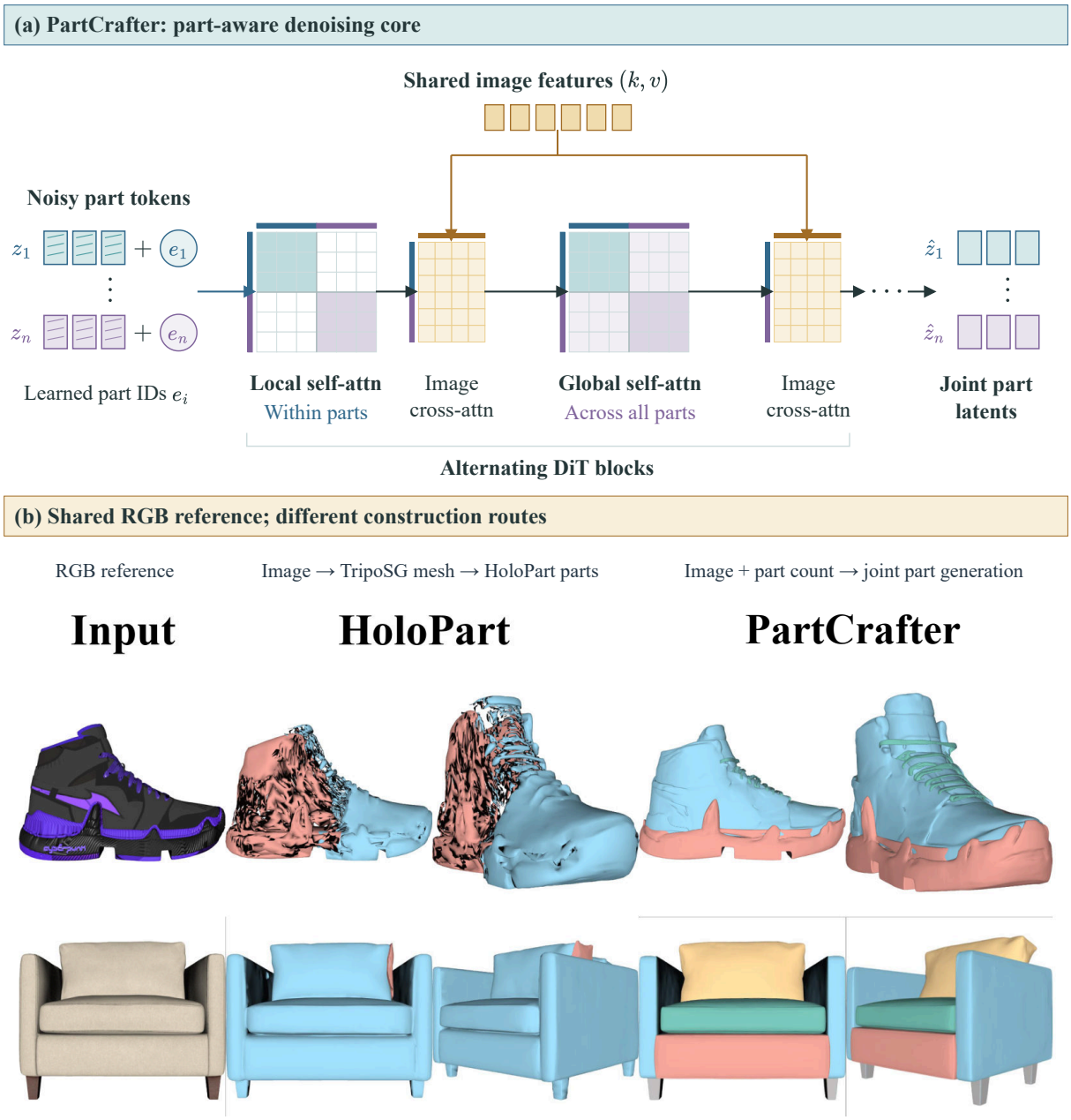



**Fig. 6** PartCrafter: (a) part-aware denoising; (b) TripoSG–HoloPart and PartCrafter outputs from a shared RGB image. PartCrafter also takes the requested part count. Images from Lin et al. (2025).

conditions both levels. All parts share flow time $s$ during training. The coupled denoising operator thus generates coordinated parts rather than independently sampling and assembling meshes.

Figure 6 connects joint latent generation to decomposition coherence: colors identify part surfaces in the shoe and armchair examples, allowing part structure to be inspected alongside shape quality.

PartCrafter evaluates reconstruction with Chamfer distance and F-score, and decomposition with pairwise voxel-overlap measures (Lin et al. 2025). Part selection and independent transformations remain inferred uses; joints, motion constraints, collision behavior, and persistent state require additional definitions and evidence.

TRELLIS edits regions within its structured representation, whereas PartCrafter delivers separate meshes (Xiang et al. 2025; Lin et al. 2025).

**Generation and decomposition as separate operations.** TripoSG uses rectified flow to generate shapes; HoloPart performs generative amodal part segmentation of an existing shape (Li et al. 2025e; Yang et al. 2025a). Their combined baseline in Table 3 generates the whole shape before inferring hidden part geometry, whereas PartCrafter jointly generates part structure. PartCrafter reports the comparison values for these complete routes.

**Mesh topology and semantic parts.** PolyGen models vertex and face sequences; MeshGPT generates mesh sequences from learned local features; MeshAnything converts an input shape into an economical mesh (Nash et al. 2020; Siddiqui et al. 2024; Chen et al. 2025b). These methods structure surface connectivity, while part-level methods define selectable or replaceable components. Animation-ready joints require a separate articulation stage.

Recent part generators introduce controls over resolution, semantic composition, and connectivity. FullPart gives each part a full-resolution grid and conditions generation on part bounding boxes. This permits box edits while retaining unchanged part latents (Ding et al. 2026). CubePart accepts an open-vocabulary schema of named parts, making semantic composition explicit beyond PartCrafter's requested count (Zhu et al. 2026). MeshMosaic generates local mesh patches autoregressively and assembles them through shared boundaries and global shape features. It addresses connectivity rather than semantic decomposition (Xu et al. 2026b). CubePart's qualitative engine interactions use downstream articulation scripts.

### 3.1.4 Conditioning, editing, and asset preparation

Conditioning and delivery turn these representational choices into authoring operations. Text specifies content, images anchor shape and appearance, and local representations constrain edit location. Magic3D demonstrates prompt-based re-optimization and image/style extensions; TRELLIS supports regional edits (Lin et al. 2023; Xiang et al. 2025). Hunyuan3D Studio adds image preparation for sketch-based generation and downstream animation tools around its shape and texture models (Zhao et al. 2025).

Evaluation should match the editing unit: condition adherence for whole assets, preservation of unselected regions for local edits, and part identity and object relations for part operations.

**Delivery and attribute preservation.** DreamFusion and Magic3D offer mesh exports or downloads (Poole et al. 2023; Lin et al. 2023). Official TRELLIS, Hunyuan3D 2.0, and TRELLIS.2 implementations document export paths, including binary glTF (GLB). PartCrafter exports separate parts, a merged object, and a part-to-file manifest (Xiang et al. 2025; Zhao et al. 2025; Xiang et al. 2026; Lin et al. 2025). Conversion can alter retained attributes: TRELLIS uses mesh processing and texture baking; TRELLIS.2's GLB export defaults to opaque materials. Transparency must therefore be enabled after import where relevant (Xiang et al. 2025, 2026). PartCrafter's separate files retain part identities (Lin et al. 2025).

### 3.1.5 Comparison and implications

Asset selection depends on both the source of missing structure and the required edit unit. Image-prior optimization can yield plausible views despite geometric ambiguity; learned spatial latents and coordinated part meshes expose different regional or component controls (Poole et al. 2023; Xiang et al. 2025; Lin et al. 2025). Delivery must preserve the materials, identities, and geometry required by those controls.

## 3.2 Generative Models for Spatial Scenes

Scene generation adds two problems to asset construction: organizing objects in space and completing the environment beyond observed views. Let $\mathcal{S}$ denote the retained scene. Its construction index $k$ differs from physical time $\tau$: traversing a static scene changes camera $v_k$ without requiring object dynamics. Table 1, Part B, catalogues assembly, completion, expansion, and memory methods.

### 3.2.1 Layout generation and asset assembly

Scene assembly separates spatial arrangement from the geometry placed within it. Procedural systems encode selected spatial rules: extended L-systems generate streets and buildings, while CGA shape grammars refine building masses into architectural detail (Parish and Müller 2001; Müller et al. 2006). Residential-layout generation combines learned architectural programs with floor-plan optimization (Merrell et al. 2010). These examples distinguish structural constraints from the decorative appearance of a scene.

Data-driven approaches learn object selection and placement from scenes. Example-based synthesis models plausible arrangements. Deep Convolutional Priors uses top-down room representations; GRAINS models recursive object groups and relationships (Fisher et al. 2012; Wang et al. 2018; Li et al. 2019). The representation determines whether context is expressed through spatial occupancy, object groups, or explicit relations.

Let $B$ describe the room boundary and $\mathcal{L}$ an assembly of assets $A_i$ with transforms $T_i$. A learned layout generator can operate on object attributes $o_i$ before assets are retrieved:

$$\begin{aligned} \mathcal{L} &= \{(A_i, T_i)\}_{i=1}^{N}, \qquad \mathcal{S} = \text{Assemble}(B, \mathcal{L}), \\ o_{n+1} &\sim p_\theta\big(o \mid \{o_j\}_{j=1}^{n}, B\big)\,. \end{aligned} \tag{19}$$

For ATISS, $o_i$ contains category, size, orientation, and position; room type and floor plan define the generation setting. The model predicts the next object's attributes from the existing object set and floor-plan features. A permutation-invariant context encoder and sampled object permutations during training avoid imposing a fixed ordering on room objects. Retrieved assets instantiate the resulting labeled boxes (Paschalidou et al. 2021). Geometry availability and layout probability are thus distinct parts of construction.

Scene graphs make those relations directly addressable. PlanIT generates a relation graph before spatial instantiation; Graph-to-3D jointly generates object shapes and layouts from a graph; EchoScene couples shape and layout denoising through information exchange (Wang et al. 2019; Dhamo et al. 2021; Zhai et al. 2024). Editing a graph specifies a change in objects or relations, while the generated geometry and transforms must still realize it spatially.

Language-conditioned systems provide an interface to these structural decisions. LayoutGPT uses in-context demonstrations to produce spatial plans; InstructScene links instructions to a semantic-graph prior and layout decoder (Feng et al. 2023; Lin and Mu 2024). Holodeck combines staged large language model (LLM) specifications, asset retrieval, and constraint solving (Yang et al. 2024d). Queries $Q = \text{LLM}_\theta(c_{\text{txt}})$ specify room geometry, object descriptions, and relations. Retrieval selects assets $A_i$; $\{T_i\} = \text{Solve}(B, \{A_i\}, Q)$ places them. The solver enforces containment and noncollision while seeking to satisfy soft relations, producing an AI2-THOR scene. Evaluation includes layout plausibility and navigation in NoveltyTHOR environments. Retained asset identities supply units for later authoring, while retrieval limits available shapes.

DiffuScene denoises unordered object attributes, while PhyScene adds collision, room-layout, and reachability guidance to scene synthesis (Tang et al. 2024b; Yang et al. 2024c). Ctrl-Room connects arrangement and appearance through a layout-conditioned panorama followed by 3D room reconstruction. Its scene representation supports individual furniture edits (Fang et al. 2025).

Agentic systems incorporate assessment into repeated scene construction. Scenethesis combines image-guided semantic placement with geometric optimization for contact, collision, and stability, followed by scene judging and replanning (Ling et al. 2026b). SAGE coordinates asset and layout tools with visual and physics critics, using simulation feedback to revise candidate scenes (Xia et al. 2026). Compared with Holodeck's staged constraint solving, these systems change where assessment feeds back into construction. Both evaluate scene plausibility and physical or embodied tasks under different protocols. Scenethesis includes reachability and walkability; SAGE reports short-horizon stability and policies trained from generated demonstrations.

### 3.2.2 Scene completion and spatial expansion

**Incremental surface construction.** View-based construction addresses coverage by generating newly revealed content in agreement with a retained surface. The main operation alternates observation synthesis with geometric fusion. With $m_k = 1$ for known projected pixels, a common update is

$$\begin{aligned} (\bar{I}_k, \bar{D}_k, m_k) &= \Pi(\mathcal{S}_{k-1}, v_k), \\ I_k &\sim p_\theta(I \mid \bar{I}_k, m_k, c_{\text{txt},k}), \\ \widehat{D}_k &= \text{AlignDepth}_k(I_k, \bar{D}_k, m_k), \\ \mathcal{S}_k &= \mathcal{U}_k(\mathcal{S}_{k-1}, I_k, \widehat{D}_k, m_k, v_k). \end{aligned} \tag{20}$$

Here $\Pi$ projects the retained representation to RGB, depth, and visibility; $\text{AlignDepth}_k$ combines depth prediction with geometric alignment; and $\mathcal{U}_k$ unprojects, filters, and merges new content. Projected RGB and visibility condition the image sampler; projected depth constrains depth completion or alignment in $\text{AlignDepth}_k$.

Text2Room instantiates $\mathcal{S}_k$ as a textured mesh $\mathcal{M}_k$. Its depth operator completes depth from the known rendered region, then fits scale and shift in disparity space. The update filters stretched triangles and stitches new faces to existing geometry (Höllein et al. 2023). A two-stage camera/prompt schedule

first constructs the room, then selects views exposing remaining holes. Mesh downloads and a browser viewer establish delivery and camera navigation; human judgments assess perceived structure and completeness. Poor viewpoints can leave stretched surfaces or holes; Poisson closure can oversmooth missed regions, and textures retain generated illumination (Höllein et al. 2023).

SceneScape retains mesh projection but changes the refinement operators. At each view, it fine-tunes the depth predictor against $\bar{D}_k$ over known pixels. It then adapts the image decoder to preserve projected appearance, regularizing new regions toward the original decoder output. Both networks reset to their pretrained weights at each step (Fridman et al. 2023). The image sampler in Equation (20) is therefore $p_{\theta_k}$, where $\theta_k$ includes the temporary decoder update. Depth adaptation occurs within $\text{AlignDepth}_k$; these network updates differ from fitting scene-content parameters $\varphi$. Mesh updates add newly visible content while retaining previously synthesized, possibly occluded surfaces. Camera recovery, reprojection/depth error, and reconstruction tests assess how these updates affect agreement across viewpoints. Incorrect retained surfaces can misguide later synthesis, and large sky–ground depth discontinuities remain difficult for the triangular mesh (Fridman et al. 2023).

Figure 7 illustrates why plausible frames and geometric agreement can diverge. The ablation evaluates 20 videos from 10 prompts. COLMAP provides a reconstruction-based check whose coverage depends on the generated views. Here, temporal continuity concerns camera traversal, not independently moving objects.

**Fields and coupled views.** A learned spatial field can replace incremental fusion state. SceneDreamer's core generator uses no external text or image condition. Scene noise $z_{\text{scene}}$ procedurally determines a bird's-eye-view height field $h$ and semantic field $m_{\text{sem}}$, while $z_{\text{style}}$ controls appearance. In shared notation, its construction is

$$\begin{aligned} (h, m_{\text{sem}}) &= \text{BEV}(z_{\text{scene}}), \qquad f_{\text{scene}} = E_\theta(h, m_{\text{sem}}), \\ f_{\mathbf{x}} &= H_\theta(\mathbf{x}, f_{\text{scene}}), \\ \mathcal{S}(\mathbf{x}) &= F_\theta(f_{\mathbf{x}}, m_{\text{sem}}(\mathbf{x}), z_{\text{style}}), \qquad I_v = \mathcal{R}(\mathcal{S}, v). \end{aligned} \tag{21}$$

Here $E_\theta$ encodes height and semantic fields into the scene feature $f_{\text{scene}}$; $\mathbf{x}$ is a queried 3D position. The generative hash grid $H_\theta$ indexes location- and scene-dependent features; $F_\theta$ supplies density and color for volumetric rendering (Chen et al. 2023c). Shared parameters are learned adversarially from unposed image collections. New cameras query a sampled scene rather than rebuilding it through Equation (20); style changes modify its appearance condition. The height field restricts vertically overlapping structures, including tunnels and caves. Reported trajectory tests therefore concern this class of static landscapes.

MVDiffusion couples generated observations by denoising multiple views with correspondence-aware attention. This replaces sampling each $I_k$ from a fused predecessor in Equation (20). Homographies connect panoramic crops; depth maps and camera poses supply correspondences in the geometry-conditioned configuration (Tang et al. 2023). Overlap-consistency evaluation tests agreement across views; panoramic angular coverage alone leaves stereo baseline and geometry for viewpoint translation unspecified.

**Panorama-to-scene construction.** HunyuanWorld 1.0 adds a panoramic intermediate, $I^{\text{pan}} \sim p_\theta(I^{\text{pan}} \mid c)$, followed by semantic decomposition and layer-wise reconstruction (HunyuanWorld Team et al. 2025). For text input, $c$ is an enhanced description. Image input supplies a projected image condition and an auxiliary scene description. The intermediate contains foreground, background, and, for outdoor scenes, sky layers. Predicted depths are aligned across overlapping regions before mesh construction. Foreground geometry comes from depth-based sheet warping or an additional image-to-3D stage with object placement. Compared with sequential mesh fusion, the state becomes a layered panorama, and reconstruction preserves separable object components. Reference-centered angular coverage still requires hidden surfaces to be completed or generated. The report describes engine integration and demonstrates object translation, rotation, and scaling; image-alignment measures evaluate the input condition. Its optional long-range extension uses Voyager's geometric memory, discussed next.

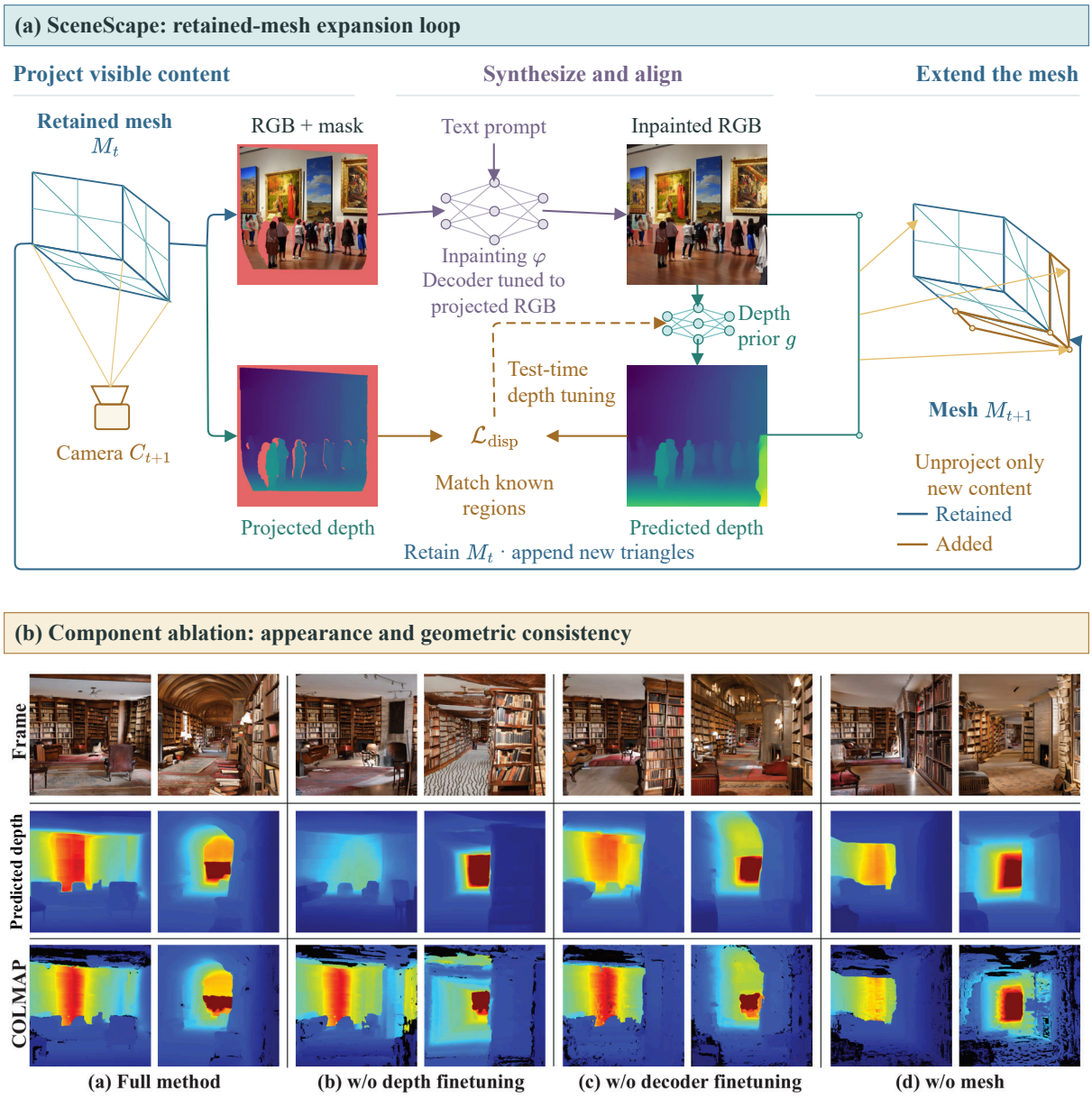



**Fig. 7** SceneScape: (a) mesh-guided view synthesis; (b) full model and ablations removing depth tuning, decoder tuning, or mesh retention, labeled (a)–(d). Rows show frames, predicted depth, and COLMAP depth; no shared depth scale or validity key is specified. Images from Fridman et al. (2023).

**Extending a scene across observations.** LucidDreamer expands Gaussian scenes with generated imagery; WonderJourney creates connected visual journeys; WonderWorld targets interactive scene generation from one image (Chung et al. 2025; Yu et al. 2024a, 2025a). Invisible Stitch uses depth inpainting to construct newly revealed regions (Engstler et al. 2025b). Overlap with retained surfaces constrains an extension, while unseen regions can admit several plausible completions.

Text2NeRF combines generated imagery, monocular depth, inpainting, and radiance-field updates for text-conditioned expansion (Zhang et al. 2024b). It uses a field rather than surface or Gaussian reconstruction to retain the aligned observations.

**Panoramic and layout-conditioned construction.** PanFusion couples panoramic and perspective diffusion branches to output a panorama (Zhang et al. 2024a). DreamScene360 reconstructs a Gaussian scene from a panorama, while Pano3DComposer generates and places object components from panoramic input (Zhou et al. 2024a; Qiu and Wu 2026). SceneCraft supplies geometry earlier: it renders a specified 3D semantic layout into multiview proxy maps, conditions diffusion on their semantics and depth, and fits a NeRF to the generated observations (Yang et al. 2024b). Its layout constrains synthesis before reconstruction; panorama-based routes infer geometry and placement from imagery.

Generation and reconstruction can enter at different stages of scene construction. I-Scene adapts a pretrained instance generator to scene RGB images and instance masks through spatial guidance, scene-context attention, and view-centric coordinates. It produces separately represented objects with shared spatial context (Ling et al. 2026a). SceneFrom3D starts from supplied outdoor geometry and per-object identity images. It schedules anchor views and interpolated video, then fits a Gaussian scene to the generated observations (Kim et al. 2026). I-Scene transfers an instance prior into scene organization; SceneFrom3D constrains observation synthesis with an existing layout. SceneFrom3D's component-substitution comparisons separately test generation, reconstruction, and view scheduling.

For temporal imagery, 360DVD generates motion-conditioned panoramic video and CubeComposer expands perspective video into high-resolution panoramic sequences (Wang et al. 2024b; Li et al. 2026a). Looking around a viewpoint uses angular coverage; translation additionally requires geometry or a novel-view mechanism.

SynCity combines pretrained image and 3D generators to condition each new tile on the surrounding world and fuse it with the scene (Engstler et al. 2025a).

### 3.2.3 Camera-controlled generation and geometric memory

**Geometry as generation memory.** Repeated expansion makes retained geometry a constraint on later synthesis. Voyager moves depth from a separate estimator into the jointly generated output. Let $\mathcal{P}_{k-1}$ be the accumulated colored point cloud, $\boldsymbol{v}_k$ a requested camera sequence, and $q_k$ its projected partial color-and-depth (RGB-D) observations and visibility masks. At clip level,

$$\begin{aligned} q_k &= \Pi(\mathcal{P}_{k-1}, \boldsymbol{v}_k), \\ (I_k, D_k) &\sim p_\theta(I, D \mid q_k, \omega_{k-1}), \\ \mathcal{P}_k &= \mathrm{Cull}(\mathcal{P}_{k-1} \cup \mathrm{Unproject}(I_k, D_k, \boldsymbol{v}_k)) \,. \end{aligned} \tag{22}$$

Here $\omega_{k-1}$ denotes overlapping-clip state used during sampling; the initial cache is unprojected from input image $c_{\mathrm{img}}$ with estimated depth (Huang et al. 2025c). Unlike RGB inpainting in Equation (20), the model jointly predicts aligned RGB and depth, adding projected depth as a geometric condition. Culling selects useful new points by visibility and orientation; overlapping-segment initialization and refinement smooth clip transitions. The cache supports subsequent synthesis, while generated observations and depths can also supervise three-dimensional Gaussian splatting (3DGS) reconstruction. Camera-control and spatial-consistency evaluations, including WorldScore's static subset, test path adherence and agreement among views (Huang et al. 2025c).

**Offline world construction.** HY-World 2.0 combines panoramic initialization, planned trajectories, memory-guided keyframe expansion, reconstruction, and scene composition offline (Team HY-World et al. 2026). Compared with Voyager, its WorldStereo 2.0 expansion component generates keyframes using both global point-cloud guidance and retrieved observations. Memory includes a panoramic point cloud and a bank $\mathcal{H}_k$ of image–camera pairs. Selected references are paired horizontally with target views; generated keyframes update $\mathcal{H}_k$ for later retrieval. Camera rays and embeddings provide further viewpoint control. WorldMirror 2.0 and depth alignment subsequently reconstruct this keyframe state, which is distinct from the final Gaussian representation.

Scene composition fits per-instance Gaussian parameters $\varphi$ to views and geometric supervision, then extracts a collision-proxy mesh through a truncated signed distance field. The output $Y = (\mathcal{S}_\varphi, \mathcal{M}_{\mathrm{collision}})$ thus separates rendering and collision roles. Evaluation likewise distinguishes text/image-conditioned world generation from reconstruction supplied with multiple views or video. Camera errors evaluate WorldStereo 2.0; single-image generative reconstruction is assessed against multiview-stereo pseudo-ground truth (Team HY-World et al. 2026). Navigation and stair-collision demonstrations use the subsequently constructed scene and proxy.

**View synthesis and persistent reconstruction.** Stable Virtual Camera (SEVA) generates camera-conditioned views without explicit reconstruction as its central representation (Zhou et al. 2025). ViewCrafter guides video diffusion with point clouds; See3D learns multiview generation from pose-free video with visual conditioning (Yu et al. 2025c; Ma et al. 2025). FlexWorld integrates generated content into a retained scene, whereas GEN3C uses a 3D cache for camera-controlled generation (Chen et al. 2025a; Ren et al. 2025).

### 3.2.4 Persistent representations and spatial functionality

Retained surfaces, addressable objects, and collision proxies expose different spatial operations. Editing methods make the update mechanism explicit.

**Editing retained representations.** Instruct-NeRF2NeRF updates a radiance field from instruction-edited observations. GaussianEditor couples semantic editing with Gaussians, and SceneFactor factors a latent 3D scene representation for controllable synthesis and modification (Haque et al. 2023; Chen et al. 2024d; Bokhovkin et al. 2025). Each retains content for subsequent rendering; later operations depend on whether the addressable unit is an appearance region, geometric component, scene entity, or simulator variable.

**Procedural generation and explicit constraints.** Infinigen generates natural-scene geometry and appearance procedurally; Infinigen Indoors adds indoor assets and constraint-based arrangement (Raistrick et al. 2023, 2024). Constraints specify compositions for a solver, while learned priors estimate regularities from examples; PhyScene shows how explicit requirements can also guide learned generation. Both routes require operation-specific assessment, such as keeping passages clear or supporting placed objects.

### 3.2.5 Comparison and implications

Scene assembly and incremental completion fail for different reasons: unavailable assets or incompatible layouts constrain the former, while incorrect depth and accumulated surface errors constrain the latter (Yang et al. 2024d; Fridman et al. 2023). Object rearrangement requires addressable entities; visual walkthroughs require coherent surface coverage.

## 3.3 Dynamic Content and Action-Conditioned Worlds

Static scene construction determines what exists; dynamic generation also determines what changes and what drives those changes. Prepared 3D dynamics expose camera/time queries, while controlled rollouts predict content from actions and history. Table 1, Part C, catalogues these methods.

### 3.3.1 Video priors for dynamic generation

**From image synthesis to temporal prediction.** Independently generated frames can flicker, change identity, or break motion continuity. Video Diffusion Models adds temporal modeling and joint image–video training (Ho et al. 2022b). Make-A-Video learns appearance and language associations from paired text–image data and motion from uncaptioned videos (Singer et al. 2023a). Video LDM learns temporal alignment in an image generator's compressed representation (Blattmann et al. 2023b). These approaches transfer image priors while using additional architecture and video training to connect observations across time.

**Temporal resolution and the organization of synthesis.** Imagen Video distributes spatial detail and frame rate across a base generator and spatial/temporal super-resolution cascade (Ho et al. 2022a). Lumiere processes the full temporal extent at several space–time scales, generating a full-frame-rate, low-resolution clip within its denoising network (Bar-Tal et al. 2024). The full clip is refined through iterative denoising. Training organization also differs: Stable Video Diffusion separates image pretraining, video pretraining, and high-quality video finetuning (Blattmann et al. 2023a). VideoCrafter2 studies spatial/temporal coupling, then improves appearance with high-quality images while retaining learned motion (Chen et al. 2024c).

**Large video backbones and downstream capabilities.** HunyuanVideo spans architecture, training, data processing, and execution infrastructure (Kong et al. 2024). Wan combines a video VAE with a flow-matching transformer for generation and editing (Wan Team et al. 2025). CogVideoX introduces spatiotemporal compression and expert transformer components for text–video integration (Yang et al. 2025b). Movie Gen extends foundation-model training to video synthesis, personalization, instruction-based editing, and synchronized audio (Polyak et al. 2024). These backbones supply visual and motion priors for downstream adaptation. That adaptation must determine how generated edits remain available to later application views and interactions.

**Discrete video representations and temporal ordering.** VideoGPT combines a video VQ-VAE with a transformer over discrete space–time codes (Yan et al. 2021). CogVideo transfers a text-to-image transformer and trains hierarchically across frame rates (Hong et al. 2023). Phenaki combines a temporally causal tokenizer for variable-length video with masked-transformer generation under changing text prompts (Villegas et al. 2023). Tokenizer causality limits which frames contribute to

a code; code-generation order is a separate choice. MAGVIT uses masked video-token modeling for multiple synthesis tasks (Yu et al. 2023), while VideoPoet represents multimodal inputs and outputs autoregressively (Kondratyuk et al. 2024).

**Hybrid autoregressive and denoising models.** NOVA predicts frames causally while modeling spatial sets in a continuous representation within each frame (Deng et al. 2025). Lumos-1 combines a language-model-style architecture with masked discrete diffusion, using bidirectional within-frame and causal between-frame attention (Yuan et al. 2026a). MAGI-1 denoises successive video chunks with noise levels organized for causal temporal generation (Sand.ai et al. 2025). These ordered temporal processes differ in token type, prediction unit, and objective. Streaming assessment must identify when each unit becomes available, which previous outputs remain in context, and whether a new instruction can affect the next unit.

### 3.3.2 Generating time-varying 3D content

To view generated motion from different positions, one route fits a time-indexed 3D representation $\mathcal{S}_\varphi(\tau)$ before playback. A renderer then separates the scene-time query from the camera query:

$$I(v, \tau) = \mathcal{R}(\mathcal{S}_\varphi(\tau), v)\,. \tag{23}$$

Varying $v$ at fixed $\tau$ produces different views of the same modeled state; varying $\tau$ at fixed $v$ exposes its temporal evolution. These queries allow spatial agreement and motion to be evaluated separately. Methods differ in how $\mathcal{S}_\varphi(\tau)$ is parameterized and which priors or reference observations supervise $\varphi$. A time coordinate selects a point in the prepared animation; it does not itself supply an action-dependent transition.

4D-fy instantiates $\mathcal{S}_\varphi(\tau)$ as a radiance field with static and dynamic hash-grid features (Bahmani et al. 2024). In common notation, $(\sigma, \mathbf{r}) = f_{\varphi_{\text{MLP}}}(h_{\varphi_{\text{stat}}}(\mathbf{x}) + h_{\varphi_{\text{dyn}}}(\mathbf{x}, \tau))$ gives density and color at position $\mathbf{x}$; the inspected model omits view-dependent color. Compared with Eq. (15), guidance is staged among multiview SDS, image-based variational score distillation (VSD), and video SDS. Camera extrinsics additionally condition the multiview prior; rendered frame sequences supply the video prior. These supervision operators alternate updates rather than forming a fixed sum of identical SDS losses. Dynamic features are unfrozen for video-guided updates in the last stage. Human evaluation separates appearance, perceived 3D structure, and motion. Against MAV3D, 41% of motion-quality preferences favored 4D-fy, with no statistically significant difference reported for that criterion.

Figure 8 illustrates how guidance and field representation affect shape and detail for two prompts. Its time/viewpoint axes distinguish spatial from temporal agreement. The sampled stills support inspection of appearance and geometry. Motion quality comes from human judgments of the animation.

**Deforming a static asset.** DreamGaussian4D instead writes the dynamic representation as $\mathcal{S}_\varphi(\tau) = \mathcal{W}_\varphi(\mathcal{S}_0, \tau)$, where $\mathcal{S}_0$ is a static Gaussian asset and $\mathcal{W}_\varphi$ predicts changes in position, rotation, and scale using HexPlane features (Ren et al. 2023). Its additional supervision is a driving video $\{I_k^{\text{drive}}\}_{k=1}^K$ at a reference camera $v_{\text{ref}}$:

$$\mathcal{L}_{\text{ref}}(\varphi) = \frac{1}{K}\sum_{k=1}^{K} \left\| \mathcal{R}(\mathcal{W}_\varphi(\mathcal{S}_0, \tau_k), v_{\text{ref}}) - I_k^{\text{drive}} \right\|_2^2 . \tag{24}$$

The video may be supplied or generated from the input image. Reference-view fitting is combined with 3D-aware image SDS at novel views; subsequent video-to-video refinement improves mesh-sequence textures. Whereas 4D-fy uses a video-prior gradient, the driving video supplies explicit framewise deformation targets. Evaluation covers video-to-4D novel-view reconstruction and video quality; demonstrations compose textured mesh sequences in Blender.

**From dynamic reconstruction to generation.** D-NeRF models deformation relative to a canonical radiance field, and K-Planes factorizes spatiotemporal fields into lower-dimensional planes (Pumarola et al. 2021; Fridovich-Keil et al. 2023). They supply representation and reconstruction mechanisms that later generative systems can use; neither an additional time coordinate nor a deformable field by itself specifies a generative prior. MAV3D learns a dynamic scene from text and video priors, while Consistent4D reconstructs a dynamic object from monocular video using consistency across generated observations (Singer et al. 2023b; Jiang et al. 2024). This separates the source of motion supervision from the data structure used to replay that motion.

Align Your Gaussians combines dynamic Gaussians and diffusion priors for text-to-4D generation; SV4D generates jointly consistent frames and views before fitting dynamic 3D content (Ling et al. 2024a;

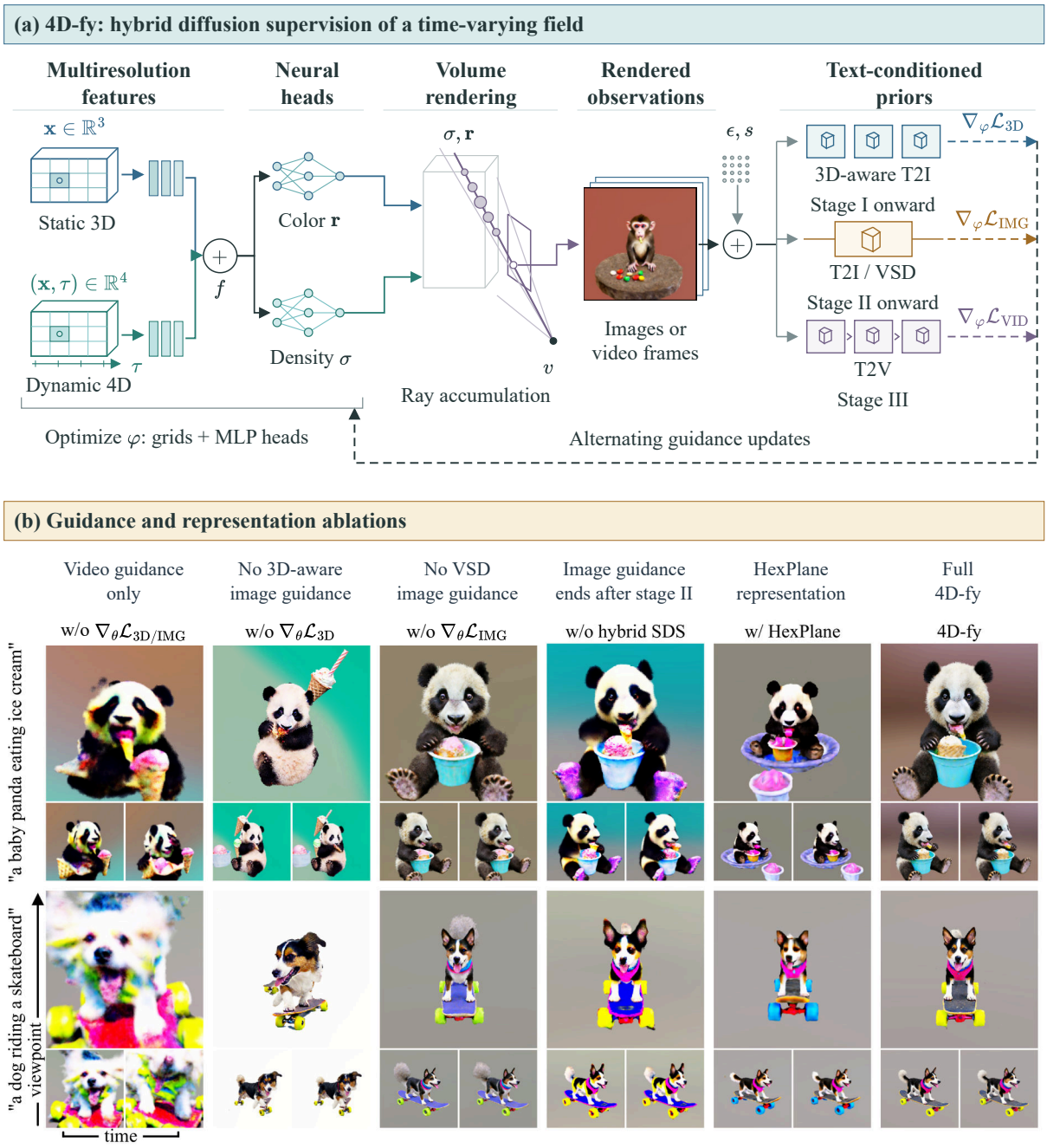



**Fig. 8** 4D-fy: (a) time-varying field optimization; (b) video-only guidance, no 3D-aware image guidance, no VSD image guidance, image guidance ending after stage II, HexPlane, and full-model results. Within each block, horizontal offsets vary time; vertical offsets vary viewpoint. Instance parameters: $\theta$ in (b), $\varphi$ in (a). Images from Bahmani et al. (2024).

Xie et al. 2025). The former emphasizes a deformable explicit rendering representation; the latter concentrates consistency modeling in multiview video generation. Full-4D likewise synthesizes time–view observations, using geometric reprojection during video generation and flow-matching distillation when fitting deformable Gaussians (Chen et al. 2026a). Its comparisons combine delivered 4D scenes with generated-video baselines used when reconstruction is unavailable or fails.

### 3.3.3 Separating camera motion from object dynamics

The same camera/time separation must be resolved when generated observations supply the reconstruction targets. DimensionX models these factors through two desired observation families, expressed using Eq. (23) as

$$\mathcal{V}_{\mathrm{S}} = \{I(v_k, \tau_0)\}_{k=1}^{K}, \qquad \mathcal{V}_{\mathrm{T}} = \{I(v_0, \tau_k)\}_{k=1}^{K}. \tag{25}$$

The S-Director learns camera variation at fixed scene time; the T-Director learns temporal variation at a fixed camera (Sun et al. 2025). In the inspected preprint, separate low-rank adaptations learn the two factors. A switch during diffusion time $s$ applies spatial guidance before temporal guidance. The guidance switch alters the denoising operator, rather than simply appending physical time $\tau$ to the input.

Compared with direct optimization of $\mathcal{S}_\varphi(\tau)$, DimensionX first generates spatially and temporally varying observations from an image. Trajectory-specific S-Directors control the view sequence; the sparse-view route additionally conditions on an ending image. Identity-preserving denoising and temporal refinement connect generated views before 3D or 4D reconstruction (Sun et al. 2025). Novel-view reconstruction metrics and subject-consistency assessment test different components. The reconstructed output is a Gaussian representation.

**Independent controls for viewpoint and motion.** MotionCtrl controls camera and object motion within video generation; CameraCtrl adapts video diffusion to supplied camera trajectories (Wang et al. 2024d; He et al. 2025a). CAT4D conditions multiview video on camera and time, then recovers a deformable scene (Wu et al. 2025a). These mechanisms distinguish walking around a stationary person from observing that person's motion with a fixed camera. In XR, tracked user motion continuously changes viewpoint, while object animation or manipulation supplies a separate source of change.

**Reference appearance and motion conditions.** Tune-A-Video adapts image diffusion to a single text–video example and guides structure through inversion (Wu et al. 2023). AnimateDiff learns a motion module transferable to personalized image models derived from the same base (Guo et al. 2024); DynamiCrafter conditions video on contextual and direct image information (Xing et al. 2024). VideoComposer adds textual, spatial, and temporal conditions, including motion vectors (Wang et al. 2023a). Distinguishing object from camera motion still requires separate conditioning and evaluation; an animated reference alone leaves the decomposition ambiguous (Xu et al. 2026a; Hu et al. 2026).

### 3.3.4 Action-conditioned video generation

Action-conditioned models sample the next observation from the current input and available history. Let $V_k$ be the next frame or chunk, $a_k$ its action inputs, $\mathcal{H}_{k-1}$ the retained history with any associated metadata, and $\mathcal{C}_k$ the context selected for prediction. A common rollout interface is

$$\begin{aligned} V_k &\sim p_\theta(V_k \mid \mathcal{C}_k, a_k, c), \\ \mathcal{H}_k &= \mathrm{Update}(\mathcal{H}_{k-1}, V_k, a_k). \end{aligned} \tag{26}$$

Selection and update may retain observations, latent features, or attention caches. They need not preserve all past information or constitute a complete Markov state. In contrast to Eq. (23), each new action affects the sampling of subsequent content rather than only a query to a prepared representation.

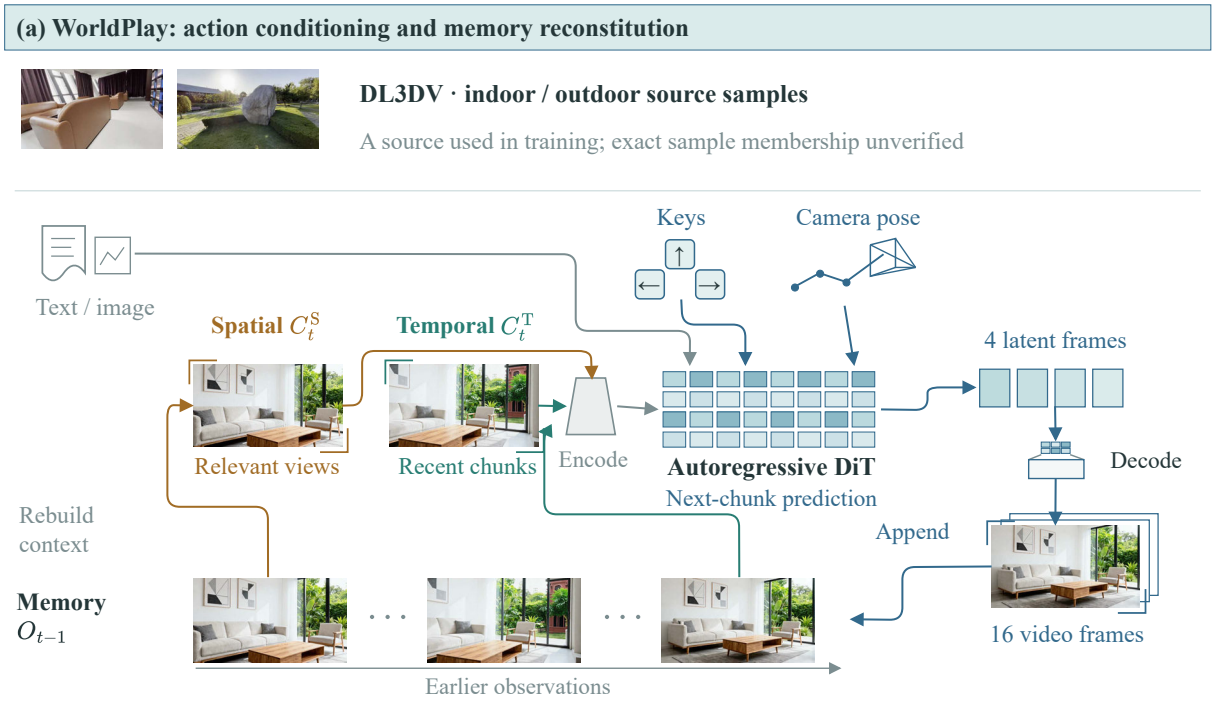
Matrix-Game 2.0 instantiates $a_k$ with frame-level keyboard and mouse signals and initializes generation from an image; its inspected architecture removes the text branch (He et al. 2025b). Mouse signals enter through latent concatenation and temporal self-attention, while keyboard embeddings enter through cross-attention. Its fixed-length latent/action cache restricts $\mathcal{C}_k$ to recent context. Self-Forcing trains the few-step autoregressive student on its own generated histories. The reported temporal, motion, and input-adherence tests assess this configuration; scene/object scores remain perceptual video measures.

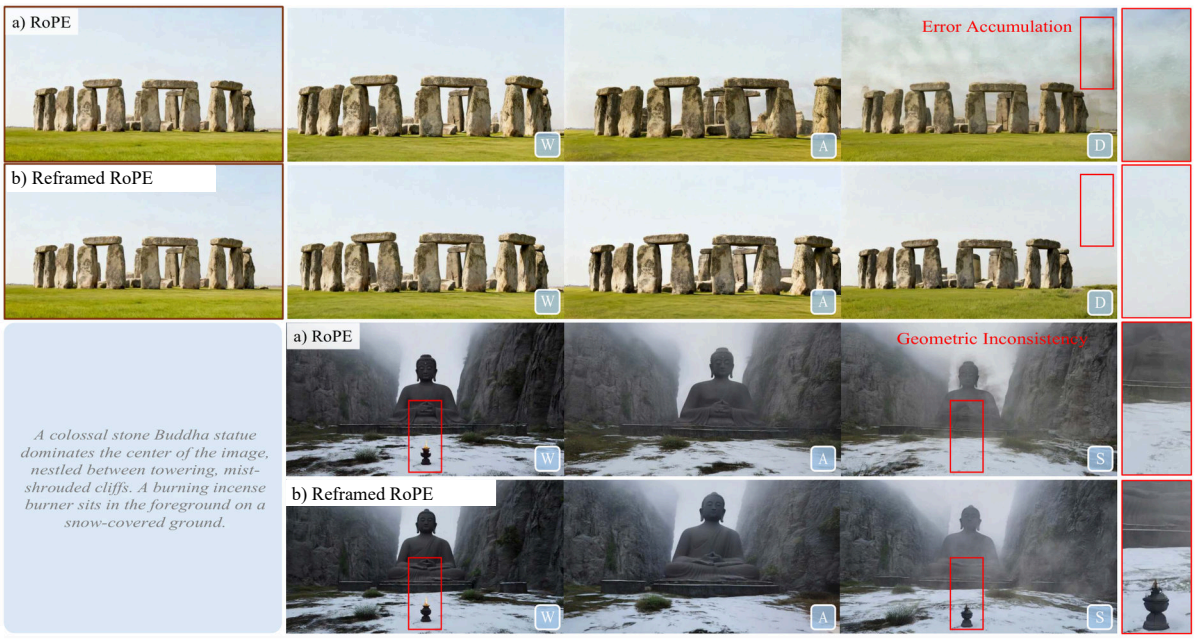
WorldPlay uses dual action conditioning $a_k = (a_k^{\mathrm{key}}, v_k)$, combining discrete actions with explicit camera parameters (Sun et al. 2026a). Key embeddings enter the time embedding; camera intrinsics and extrinsics enter an additional projective positional-encoding attention path. Poses also support retrieval of historical observations. Context Forcing aligns the teacher's and student's available memory during distillation, without adding a user input. Evaluation distinguishes trajectory-based camera-control measurements from navigation and text-triggered event demonstrations.

**Interface-level demonstrations.** Google DeepMind's Genie 3 research release is an industrial counterpart. Its official recordings show navigation, revisiting, weather changes, and new objects, and the release describes agent-provided navigation actions (Parker-Holder and Fruchter 2025). The release documents interface behavior, with limited architectural and experimental detail for internal-memory comparison.

Generated Reality supplies the tracked-input configuration examined in Section 5 (Xie et al. 2026b).

**Action interfaces and learned dynamics.** Genie learns latent actions from video without explicit action labels. GameNGen instead conditions the next game observation on recorded game actions and recent history (Bruce et al. 2024; Valevski et al. 2025). This distinguishes an inferred action encoding from a supplied command vocabulary. DIAMOND also conditions visual prediction on actions, but studies the




Within each pair: top Standard RoPE; bottom Reframed RoPE.

b1 Stone scene | Keys W → A → D | Readout: error accumulation

b2 Statue scene | Keys W → A → S | Readout: geometric inconsistency

**Fig. 9** WorldPlay: (a) action conditioning and spatial/temporal history; (b) rotary positional embedding (RoPE) comparison. The stone scene (b1) and statue scene (b2) each pair standard RoPE above with reframed RoPE below, labeled a)/b) internally. Key inputs are W, A, D for b1 and W, A, S for b2. Red boxes mark accumulated errors; right-edge magnifications show geometric inconsistencies. Dataset images from Ling et al. (2024b); generated images from Sun et al. (2026a).

resulting model as an environment for learning a control policy (Alonso et al. 2024). Action encoding and the purpose of a rollout are therefore separate choices. An XR integration must map its tracked pose, hand event, or authoring command to the accepted action space.

Hunyuan-GameCraft combines controls with hybrid historical conditioning; GameCraft-2 supports instruction-following interaction and introduces InterBench (Li et al. 2025b; Tang et al. 2025). Image-to-video configurations of DynamiCrafter, EasyAnimate, Allegro, and CogVideoX are also world-generation benchmark comparators (Xing et al. 2024; Xu et al. 2024b; Zhou et al. 2024b; Yang et al. 2025b). WorldScore's scores apply to the camera and action conditions of the evaluated image-to-video checkpoints.

**Noise levels as a representation of available history.** Unlike joint clip denoising, interactive continuation has an observed past and uncertain future. Diffusion Forcing trains with independent noise levels per sequence token, retaining cleaner context while denoising future tokens (Chen et al. 2024a). Noise levels express generation uncertainty, not elapsed physical time. History-Guided Video Diffusion accepts variable-length history through its transformer and objective and uses that history for sampling guidance (Song et al. 2025). Both make the amount and reliability of context central to prediction.

**Causal generation and training on generated histories.** CausVid adapts bidirectional video diffusion into a causal generator and distills sampling into few steps (Yin et al. 2025). Causal attention enables caching as the sequence grows, reducing work before display. Self Forcing instead addresses the mismatch between clean training histories and imperfect generated histories. It rolls out and supervises generated sequences during training (Huang et al. 2025d). Architecture and distillation reduce dependencies and sampling work; generated-context training addresses rollout errors. Throughput must therefore be read alongside training context and the evaluated quality horizon.

**Extending the generation horizon.** FreeNoise reschedules initial noise and uses windowed temporal attention for longer sequences, including changing text conditions (Qiu et al. 2024). FIFO-Diffusion queues frames at different noise levels, emitting completed frames and adding new noise (Kim et al. 2024). StreamingT2V conditions chunks on recent features and an appearance reference from the initial chunk (Henschel et al. 2025). Correlated noise, recent context, and persistent appearance can retain different information about earlier events. LongLive combines causal generation, streaming long-video training, an attention sink, and cache refresh when prompts change (Yang et al. 2026). Revisiting locations and restoring changed objects test spatial recall and retained modifications.

**Turning controls into predicted consequences.** GameFactory learns keyboard/mouse control separately from game-style adaptation (Yu et al. 2025b); Pandora combines language modeling and video generation to predict visual states from free-text actions (Xiang et al. 2024). iVideoGPT places observations, actions, and rewards in a common token sequence for prediction, planning, and reinforcement learning (Wu et al. 2024a). Vid2World adapts pretrained video diffusion through causalization and action guidance (Huang et al. 2026c). Responses must be interpreted at their action granularity: language commands and calibrated controller displacements carry different control semantics.

### 3.3.5 Memory and persistent interaction

**Retrieving visual history.** A recent-only context can lose access to locations left earlier in a rollout. Spatial retrieval addresses this by selecting relevant past observations for Eq. (26). For Matrix-Game 3.0 and WorldPlay, this selection can be described as

$$\begin{aligned} \mathcal{C}_k^{\mathrm{T}} &= \mathrm{Recent}(\mathcal{H}_{k-1}), \qquad \mathcal{C}_k^{\mathrm{S}} = \mathrm{Retrieve}(\mathcal{H}_{k-1}; v_k), \\ \mathcal{C}_k &= \mathrm{Merge}(\mathcal{C}_k^{\mathrm{T}}, \mathcal{C}_k^{\mathrm{S}}). \end{aligned} \tag{27}$$

Here Merge denotes model-specific encoding and attention, not necessarily set union. Matrix-Game 3.0 retrieves frames using camera and field-of-view relations and encodes their relative camera information. It jointly attends to retrieved memory, recent context, and current predictions (Wang et al. 2026g). Error-aware training perturbs historical conditions with prediction residuals. Rollout is autoregressive across segments, while the student retains bidirectional attention within each prediction segment. The retrieval examples test recovery after camera movement and occlusion; the speed ablation tests a separate property.

WorldPlay instantiates $\mathcal{C}_k^{\mathrm{T}}$ with recent chunks and $\mathcal{C}_k^{\mathrm{S}}$ with non-adjacent frames selected by field-of-view overlap and camera distance (Sun et al. 2026a). Temporal reframing assigns retrieved observations nearby positional indices instead of retaining unbounded absolute gaps. It changes the encoding of selected conditions, not their physical capture times. Figure 9 pairs the selection/append loop with

the standard-versus-reframed rotary positional embedding (RoPE) ablation. The round-trip protocol compares returning views with earlier generated views, testing visual recall of the established environment.

Relative to WorldPlay's frame retrieval, AnchorWeave selects local point-cloud memories by viewpoint coverage and reconciles them through a multi-anchor controller (Wang et al. 2026f). The geometry serves as generation context (Section 3.2.3). ActWorld changes what history is prioritized: an interaction-aware router retains transitions, while persistent event/object tokens preserve anchors after frames leave the buffer (Xiong et al. 2026). Geometric coverage and interaction relevance therefore select different evidence for a later observation.

**Memory allocation and cost.** Additional history can preserve useful context while reintroducing generated errors. Matrix-Game 2.0's qualitative cache comparison shows earlier artifacts with nine rather than six latent frames (He et al. 2025b). WorldPlay's fixed-budget comparison favors different spatial/temporal allocations under different consistency and control metrics (Sun et al. 2026a). Memory selection therefore trades recent continuity against access to older views. Implementation also affects throughput: Matrix-Game 3.0's asynchronous ablation uses eight graphics processing units (GPUs) for diffusion inference and one for video decoding. This configuration also applies 75% pruning of the video autoencoder. Removing GPU retrieval optimization reduces throughput from about 40 to 6.60 frames per second (FPS) (Wang et al. 2026g).

**Shared state and generated observations.** MultiGen adds explicit geometric state to the history-conditioned observation model (Po et al. 2026). Let $\Lambda$ be its static level map, $\mathbf{p}_{k-1}$ the current player pose, and $r_{k-1}$ its disparity readout. Its observation/dynamics split can be summarized as

$$\begin{aligned} r_{k-1} &= \mathrm{RayDisp}(\Lambda, \mathbf{p}_{k-1}), \\ I_k &\sim p_\theta(I_k \mid \mathcal{C}_k, r_{k-1}, a_k), \\ \mathbf{p}_k &= \mathbf{p}_{k-1} \oplus \Delta_\theta(\mathbf{p}_{k-1}, a_k, r_{k-1}, f_k). \end{aligned} \tag{28}$$

Here $\mathcal{C}_k$ contains recent frames ending at $I_{k-1}$; $a_k$ drives the transition to frame $k$. The readout is concatenated with context frames, and action embeddings enter the observation network through cross-attention. The dynamics module also uses features $f_k$ from denoising $I_k$; $\oplus$ applies the predicted pose increment and wraps the angle. Generation is therefore coupled to an explicit state update, beyond adding another history frame. The map is editable before rollout and fixed during it; player states update online and are shared. Simulator-referenced rollouts evaluate observation fidelity; the single-map multiplayer test uses a vision-language model to judge opponent presence.

Solaris jointly denoises two Minecraft views under separate player actions, coupling their streams through cross-player self-attention (Savva et al. 2026). This differs from MultiGen's pose updates against an authored map. Building and cross-view tests with two players extend multiplayer evidence beyond opponent presence.

**Predictive state and its intended use.** DreamerV3 learns state for control: its training and normalization support behavior learning across diverse tasks (Hafner et al. 2025). As distinguished in Section 2.4.3, state sufficient for action selection need not retain every display detail; a detailed frame need not recover earlier object interactions.

**View memory and recurrent context.** VMem retrieves spatially relevant views through a surfel index (Li et al. 2025c). Video World Models with Long-term Spatial Memory retains geometrically grounded information; Long-Context State-Space Video World Models extends temporal context with recurrent state-space processing (Wu et al. 2025c; Po et al. 2025). ReWorld combines a bounded key–value cache with a pose-indexed landmark bank for revisits (Chen et al. 2026b). Geometric indices, chronological caches, and recurrent summaries provide different access to earlier observations.

**Retaining changes in a scene.** WorldMem associates frames with poses and timestamps. It demonstrates object placement and tests temporal conditioning on placing-event videos (Xiao et al. 2025). PERSIST learns an evolving latent 3D environment specified or edited in three dimensions (Garcin et al. 2026). FloWM structures memory into velocity channels transformed by self-motion and object flow, with tests in simplified dynamic worlds (Lillemark et al. 2026). LiveWorld instead keeps entity monitors active outside the observer's view and uses their evolved states on return (Duan et al. 2026). These approaches distinguish recalling changed observations from predicting changes during absence.

Programmable World Model separates state transitions from visual generation. External rules update entity attributes and relations, which the engine compiles into generator conditions (Huang et al. 2026d). These transitions are programmed rather than learned as latent dynamics. WorldDirector augments

appearance retrieval with planned 3D object trajectories and identity-specific conditions for disappearance and re-entry (Wang et al. 2026d). Trajectory orchestration alone does not establish physical simulation. From Pixels to States supplies an action–state–observation vocabulary for this comparison (Li et al. 2026b). Section 6.3 distinguishes accepting a change from retaining it after departure and return.

### 3.3.6 Comparison and implications

Prepared dynamics, history-conditioned observations, and explicit state updates support different tasks: animated exhibits, visual exploration, and observers linked by shared state. MultiGen, for example, reports that textures and small objects absent from its map are not explicitly preserved on revisiting (Po et al. 2026). Evaluation should measure retention of the task-relevant variable.

**Table 1**: Catalogue of visual-asset, spatial-scene, and temporal generative methods.

**Input:** **T** text; **I** image; **V** video; **P** camera/pose; **G** geometry; **L** layout/graph/constraints; **A** action; **S** sketch; **M** motion/flow; **N** latent/random seed; **U** unconditional. $n$ requested part count (PartCrafter).
**Stage:** P preparation; A interactive authoring; R ongoing generation/prediction. **ID:** A assets; B scenes; C temporal methods.
**Reading:** Inputs: configuration-specific combinations. Representation/state: internal or retained content. Output: delivered content. GS Gaussian splats; NR not established in inspected material; ↗ official project/GitHub. Mechanism keys appear at the start of each domain.

| ID | Method / year | Input | Mechanism | Representation / state | Output | Stage | ↗ |
|---|---|---|---|---|---|---|---|
| **A. Assets — Visual-asset generation and editing** | | | | | | | |
| **Mechanism (A):** GAN adversarial; GMM Gaussian mixture; NF normalizing flow; DEC surface decoder; GEN generator family unspecified; CLIP CLIP guidance; OPT instance optimization; SDS score distillation sampling; VSD variational score distillation; SD score distillation family; Diff. diffusion; FM flow matching; RF rectified flow; REC reconstruction/fusion; SEG segmentation; AR autoregressive; INP inpainting; LLM language model; PROC procedural. | | | | | | | |
| **A. Assets — Learned geometry and 3D-aware generation** | | | | | | | |
| A1 | **3D-GAN** (2016) | **N** | **GAN** | Voxel grid | Voxel shapes | P | – |
| A2 | **Latent-GMM** (2018) | **N** | **GMM** | Point-cloud latent | Point clouds | P | – |
| A3 | **PointFlow** (2019) | **N** | **NF** | Shape/point distributions | Point clouds | P | – |
| A4 | **AtlasNet** (2018) | **I G** | **DEC** | Parametric surface patches | Surface meshes | P | – |
| A5 | **DMTet** (2021b) | **G** | **GEN** | Deformable tetrahedral SDF | Meshes | P | ↗ |
| A6 | **GET3D** (2022) | **N** | **GAN** | Geometry/texture fields | Textured meshes | P | – |
| A7 | **EG3D** (2022) | **N P** | **GAN** | Triplane radiance field | Rendered images | P | – |
| **A. Assets — Optimization through image priors** | | | | | | | |
| A8 | **Dream Fields** (2022) | **T** | **CLIP+ OPT** | Radiance field | Radiance field | P | – |
| A9 | **Text2Mesh** (2022) | **T G** | **CLIP+ OPT** | Mesh/color/displacement | Edited mesh | P | – |
| A10 | **CLIP-Mesh** (2022) | **T** | **CLIP+ OPT** | Mesh/texture/normal maps | Textured mesh | P | – |
| A11 | **DreamFusion** (2023) | **T** | **SDS** | Density/albedo field | Mesh/renders | P | ↗ |
| A12 | **Magic3D** (2023) | **T** | **SDS** | Hash-grid/mesh | Textured mesh | P | ↗ |
| A13 | **ProlificDreamer** (2023b) | **T** | **VSD** | 3D parameter particles | Textured mesh | P | – |
| A14 | **DreamGaussian** (2024c) | **T I** | **SD** | 3D Gaussians | Textured mesh | P | – |
| **A. Assets — View synthesis and feed-forward reconstruction** | | | | | | | |
| A15 | **Zero-1-to-3** (2023b) | **I P** | **Diff.** | Image latents | Novel-view images | P | – |
| A16 | **SyncDreamer** (2024b) | **I** | **Diff.** | Coupled image latents | Multiview images | P | – |

Continued on next page

**Table 1**: Catalogue of visual-asset, spatial-scene, and temporal generative methods. (continued).

**ID:** A assets; B scenes; C temporal methods. **Stage:** P preparation; A interactive authoring; R ongoing generation/prediction. Input and reading key on p. 30; mechanism keys on pp. 30, 32, and 34. ↗ official project/GitHub.

| ID | Method / year | Input | Mechanism | Representation / state | Output | Stage | ↗ |
|---|---|---|---|---|---|---|---|
| A17 | **MVDream** (2024) | T P | Diff. | Multiview image latents | Multiview images | P | – |
| A18 | **Zero123++** (2023) | I | Diff. | Multiview image latents | Multiview images | P | – |
| A19 | **Wonder3D** (2024) | I | Diff. | Multiview color/normals | Textured mesh | P | – |
| A20 | **Unique3D** (2024b) | I | Diff.+REC | Multiview color/normals | Textured mesh | P | – |
| A21 | **Magic123** (2024) | I | Diff.+OPT | Radiance-field/mesh | Textured mesh | P | – |
| A22 | **One-2-3-45** (2023a) | I | Diff.+REC | Multiview/SDF | Textured mesh | P | – |
| A23 | **One-2-3-45++** (2024a) | I | Diff. | 3D diffusion latents | Textured mesh | P | – |
| A24 | **LRM** (2024) | I | REC | Triplane radiance field | Radiance field | P | – |
| A25 | **InstantMesh** (2024a) | I | Diff.+REC | Triplane/isosurface | Textured mesh | P | – |
| A26 | **LGM** (2024a) | I | REC | 3D Gaussians | Gaussian asset | P | – |
| A27 | **GRM** (2024c) | I | REC | Pixel-aligned Gaussians | Gaussian asset | P | – |
| **A. Assets — Native 3D and shape/appearance generation** | | | | | | | |
| A28 | **LION** (2022) | N | Diff. | Global/point latents | Point clouds | P | – |
| A29 | **SDFusion** (2023) | T I G | Diff. | SDF latent grid | Meshes | P | – |
| A30 | **MeshDiffusion** (2023d) | N | Diff. | Deformable tetrahedral SDF | Meshes | P | – |
| A31 | **TRELLIS** (2025) | T I | RF | Sparse structured latent | Mesh/field/ Gaussians | P | ↗ |
| A32 | **Shap-E** (2023) | T I | Diff. | Implicit-function parameters | Mesh/radiance field | P | – |
| A33 | **LN3Diff** (2024) | T I | Diff. | 3D-aware latent | Neural field | P | – |
| A34 | **3DTopia-XL** (2025c) | T I | Diff. | Geometry/material primitives | PBR mesh | P | – |
| A35 | **Hunyuan3D 2.0** (2025) | I | FM+Diff. | Shape latents/mesh | Textured mesh | P | ↗ |
| A36 | **TRELLIS.2** (2026) | I | RF | O-Voxel latents | PBR mesh | P | ↗ |
| A37 | **Fantasia3D** (2023b) | T | SD | Geometry/BRDF fields | PBR mesh | P | – |
| A38 | **TripoSG** (2025e) | I | RF | SDF shape latents | Mesh | P | – |
| A39 | **Seed3D 2.0** (2026) | I G | Diff. | Shape vecsets /PBR latents | PBR mesh /optional URDF | P | ↗ |
| **A. Assets — Textures and materials** | | | | | | | |
| A40 | **TEXTure** (2023) | T G | Diff. | Mesh/texture atlas | Textured mesh | P | – |
| A41 | **Text2Tex** (2023a) | T G | Diff.+INP | Mesh/texture atlas | Textured mesh | P | – |
| A42 | **Paint3D** (2024) | T I G | Diff. | Mesh/UV texture | Textured mesh | P | – |
| A43 | **SceneTex** (2024b) | T G | Diff.+OPT | Multiresolution texture field | Textured scene | P | – |
| A44 | **Material Apprentice** (2026) | T I | LLM+ PROC | Process traces /shader graph | Procedural material graph | P/A | ↗ |
| **A. Assets — Parts and explicit mesh structure** | | | | | | | |
| A45 | **PartCrafter** (2025) | I +$n$ | FM | Part-indexed latent sets | Separate part meshes | P | ↗ |
| A46 | **HoloPart** (2025a) | G | SEG+Diff. | Part latents | Completed part meshes | P | – |
| A47 | **PolyGen** (2020) | I G | AR | Vertex/face sequences | Polygon meshes | P | – |
| A48 | **MeshGPT** (2024) | U | AR | Quantized mesh tokens | Triangle meshes | P | – |
| A49 | **MeshAnything** (2025b) | G | AR | Mesh tokens | Triangle meshes | P | – |
| A50 | **FullPart** (2026) | T I L | FM | Per-part grids /boxes | Textured part meshes | P/A | ↗ |
| A51 | **CubePart** (2026) | T L G | FM | Named-part latent sets | Named part meshes | P | ↗ |

Continued on next page

**Table 1**: Catalogue of visual-asset, spatial-scene, and temporal generative methods. (continued).

**ID:** A assets; B scenes; C temporal methods. **Stage:** P preparation; A interactive authoring; R ongoing generation/prediction. Input and reading key on p. 30; mechanism keys on pp. 30, 32, and 34. ↗ official project/GitHub.

| ID | Method / year | Input | Mechanism | Representation / state | Output | Stage | ↗ |
|---|---|---|---|---|---|---|---|
| A52 | **MeshMosaic** (2026b) | G | AR | Local mesh sequences / boundaries | Triangle mesh | P | ↗ |
| **B. Scenes — Spatial-scene generation, assembly, and expansion** | | | | | | | |
| **Mechanism (B):** GAN adversarial; GEN generator family unspecified; OPT instance optimization; Diff. diffusion; RF rectified flow; REC reconstruction/fusion; AR autoregressive; PROC procedural; PROB probabilistic scene sampling; SEQ sequential placement; VAE variational autoencoder; GRAPH graph synthesis; LLM language model; ASM retrieval/assembly; INP inpainting; ALIGN spatial alignment; EDIT scene editing. | | | | | | | |
| **B. Scenes — Layouts and scene assembly** | | | | | | | |
| B1 | **CityEngine** (2001) | L | PROC | Road/building geometry | City meshes | P | – |
| B2 | **CGA shape** (2006) | G L | PROC | Hierarchical shape rules | Building geometry | P | – |
| B3 | **Residential layouts** (2010) | L | OPT | Architectural program/ layout | Building model | P | – |
| B4 | **Example-based synthesis** (2012) | L | PROB | Object/relationship models | Arranged scene | P | – |
| B5 | **DeepSynth** (2018) | L | SEQ | Top-down scene images | Arranged scene | P | – |
| B6 | **GRAINS** (2019) | N | VAE | Scene hierarchy | Arranged scene | P | – |
| B7 | **ATISS** (2021) | L | AR | Unordered object attributes | Layout/retrieved assets | P | – |
| B8 | **PlanIT** (2019) | L | GRAPH+ ASM | Relation graph | Arranged scene | P | – |
| B9 | **Graph-to-3D** (2021) | L | VAE | Graph/shape/layout latents | Shapes +layout | P | – |
| B10 | **EchoScene** (2024) | L | Diff. | Graph/shape/layout latents | Shapes +layout | P | – |
| B11 | **LayoutGPT** (2023) | T | LLM | Structured layout | Object boxes | P | – |
| B12 | **InstructScene** (2024) | T | Diff. | Semantic-graph/layout latents | Layout/retrieved assets | P | – |
| B13 | **Holodeck** (2024d) | T | LLM+ ASM | Objects +transforms | Simulator scene | P | ↗ |
| B14 | **DiffuScene** (2024b) | T L | Diff. | Unordered object attributes | Layout/retrieved assets | P | – |
| B15 | **PhyScene** (2024c) | L | Diff. | Object layouts | Arranged scene | P | – |
| B16 | **Infinigen** (2023) | N | PROC | Procedural scene graph | Mesh/material scene | P | – |
| B17 | **Infinigen Indoors** (2024) | L N | PROC+ ASM | Procedural assets/ constraints | Mesh/material scene | P | – |
| B18 | **SAGE** (2026) | T I | LLM+ GEN+ OPT | Objects /layout /physics attributes | Simulator scene | P | ↗ |
| B19 | **Scenethesis** (2026b) | T | LLM+ ASM+ OPT | Asset poses /geometric constraints | Assembled scene | P | ↗ |
| **B. Scenes — Reconstruction and incremental geometry** | | | | | | | |
| B20 | **Ctrl-Room** (2025) | T L | Diff.+REC | Layout/panorama/mesh | Room mesh | P | – |
| B21 | **Text2Room** (2023) | T P | INP+REC | Retained textured mesh | Textured mesh | P | ↗ |
| B22 | **SceneScape** (2023) | T P | INP+REC | Retained textured mesh | Mesh/views | P | ↗ |
| B23 | **HunyuanWorld 1.0** (2025) | T I | GEN+ REC | Layered panorama/ meshes | Object-level meshes | P | ↗ |
| B24 | **LucidDreamer** (2025) | T I P | Diff.+OPT | Point-cloud/Gaussians | Gaussian scene | P | – |
| B25 | **WonderJourney** (2024a) | T I | LLM+ GEN | Connected point-cloud scenes | Scene journey | P | – |

Continued on next page

**Table 1**: Catalogue of visual-asset, spatial-scene, and temporal generative methods. (continued).

**ID:** A assets; B scenes; C temporal methods. **Stage:** P preparation; A interactive authoring; R ongoing generation/prediction. Input and reading key on p. 30; mechanism keys on pp. 30, 32, and 34. ↗ official project/GitHub.

| ID | Method / year | Input | Mechanism | Representation / state | Output | Stage | ↗ |
|---|---|---|---|---|---|---|---|
| B26 | **WonderWorld** (2025a) | T I P | Diff. | Layered Gaussian surfels | Gaussian scene | A | – |
| B27 | **Invisible Stitch** (2025b) | I P | INP+OPT | Point-cloud/Gaussians | Gaussian scene | P | – |
| B28 | **Text2NeRF** (2024b) | T P | INP+OPT | Radiance field | NeRF scene | P | – |
| B29 | **DreamScene360** (2024a) | T | GEN+ REC | Point-cloud/Gaussians | Gaussian scene | P | – |
| B30 | **Pano3DComposer** (2026) | I | GEN+ ALIGN | Objects +transforms | Compositional 3D scene | P | – |
| B31 | **I-Scene** (2026a) | I | RF | Instance latents /scene context | Object-separated scene | P | ↗ |
| **B. Scenes — Scene fields, views, and panoramas** | | | | | | | |
| B32 | **SceneDreamer** (2023c) | N | GAN | BEV/generative hash-grid | Rendered views | P | ↗ |
| B33 | **MVDiffusion** (2023) | T G P | Diff. | Multiview image latents | Panorama/multiview images | P | – |
| B34 | **PanFusion** (2024a) | T L | Diff. | Panorama/perspective latents | Panoramic image | P | – |
| B35 | **SceneCraft** (2024b) | T L | Diff.+REC | Semantic/depth proxies/ NeRF | NeRF scene | P | – |
| B36 | **360DVD** (2024b) | T M | Diff. | Panoramic video latents | Panoramic video | P | – |
| B37 | **CubeComposer** (2026a) | V | AR | Cube-face video latents | Panoramic video | P | – |
| **B. Scenes — Spatial expansion, control, and memory** | | | | | | | |
| B38 | **SynCity** (2025a) | T L | GEN | Structured latents/ Gaussians | Gaussian world | P | – |
| B39 | **Voyager** (2025c) | I P | Diff. | Colored point-cloud cache | RGB-D video | P | ↗ |
| B40 | **HY-World 2.0** (2026) | T I V P | GEN+ REC | Point-cloud/keyframes/ Gaussians | Gaussians/collision mesh | P | ↗ |
| B41 | **SEVA** (2025) | I P | Diff. | Multiview image latents | Novel-view images | P | – |
| B42 | **ViewCrafter** (2025c) | I P | Diff. | Point-cloud guidance | Novel-view video | P | – |
| B43 | **See3D** (2025) | I | Diff. | Visual-condition latents | Multiview images | P | – |
| B44 | **FlexWorld** (2025a) | I P | Diff.+REC | Persistent Gaussian scene | Gaussian scene/ video | P | – |
| B45 | **GEN3C** (2025) | I V P | Diff. | 3D point-cloud cache | Camera-controlled video | P | – |
| B46 | **SceneFrom3D** (2026) | G I | Diff.+REC | Geometry /scheduled views /GS | Gaussian scene | P | ↗ |
| **B. Scenes — Scene editing** | | | | | | | |
| B47 | **Instruct-NeRF2NeRF** (2023) | T G | EDIT+ OPT | Radiance field | Edited NeRF | P | – |
| B48 | **GaussianEditor** (2024d) | T G | Diff.+ EDIT | Semantic Gaussian scene | Edited Gaussians | A | ↗ |
| B49 | **SceneFactor** (2025) | T L | Diff. | Semantic/geometric latents | 3D scene | P | – |

Continued on next page

**Table 1**: Catalogue of visual-asset, spatial-scene, and temporal generative methods. (continued).

**ID:** A assets; B scenes; C temporal methods. **Stage:** P preparation; A interactive authoring; R ongoing generation/prediction. Input and reading key on p. 30; mechanism keys on pp. 30, 32, and 34. ↗ official project/GitHub.

| ID | Method / year | Input | Mechanism | Representation / state | Output | Stage | ↗ |
|---|---|---|---|---|---|---|---|
| **C. Temporal — Video, time-varying 3D, and interactive-world models** | | | | | | | |
| **Mechanism (C):** GEN generator family unspecified; OPT instance optimization; SD score distillation family; Diff. diffusion; FM flow matching; REC reconstruction/fusion; AR autoregressive; MTM masked tokens; dDiff. discrete diffusion; CTRL conditioning module; DIST generator distillation; DYN learned or programmed state/pose updates; RSSM recurrent stochastic state; SSM state-space sequence model. | | | | | | | |
| **C. Temporal — Video generation and temporal priors** | | | | | | | |
| C1 | **Video Diffusion** (2022b) | T V N | Diff. | NR | Video | P | – |
| C2 | **Make-A-Video** (2023a) | T | Diff. | NR | Video | P | – |
| C3 | **Video LDM** (2023b) | T | Diff. | Video latents | Video | P | – |
| C4 | **Imagen Video** (2022a) | T | Diff. | NR | Video | P | – |
| C5 | **Lumiere** (2024) | T I V | Diff. | NR | Video | P | – |
| C6 | **Stable Video Diff.** (2023a) | T I | Diff. | Video latents | Video | P | – |
| C7 | **VideoCrafter2** (2024c) | T | Diff. | NR | Video | P | – |
| C8 | **HunyuanVideo** (2024) | T | FM | Video latents | Video | P | – |
| C9 | **Wan** (2025) | T I V | FM | Video latents | Video | P | – |
| C10 | **CogVideoX** (2025b) | T I | Diff. | Video latents | Video | P | – |
| C11 | **Movie Gen** (2024) | T I V | FM | Video latents | Video | P | – |
| C12 | **VideoGPT** (2021) | I A U | AR | Video tokens | Video | P | – |
| C13 | **CogVideo** (2023) | T | AR | NR | Video | P | – |
| C14 | **Phenaki** (2023) | T | MTM | Video tokens | Video | P | – |
| C15 | **MAGVIT** (2023) | I V U | MTM | Video tokens | Video | P | – |
| C16 | **VideoPoet** (2024) | T I V | AR | Multimodal tokens | Video | P | – |
| C17 | **NOVA** (2025) | T | AR | Visual tokens | Video | P | – |
| C18 | **Lumos-1** (2026a) | T I | AR+dDiff. | Video tokens | Video | P | – |
| C19 | **MAGI-1** (2025) | T I | AR+Diff. | NR | Video | R | – |
| C20 | **EasyAnimate (V5 InP)** (2024b) | T I | Diff. | NR | Video | P | – |
| C21 | **Allegro (TI2V)** (2024b) | T I | NR | NR | Video | P | – |
| **C. Temporal — Time-varying 3D content** | | | | | | | |
| C22 | **4D-fy** (2024) | T | SD | Dynamic field | 4D scene | P | ↗ |
| C23 | **DreamGaussian4D** (2023) | I V | OPT | Deformable GS | Animated 3D asset | P | ↗ |
| C24 | **MAV3D** (2023b) | T | SD | Dynamic field | 4D scene | P | – |
| C25 | **Consistent4D** (2024) | V | Diff.+REC | Dynamic field | 4D object | P | – |
| C26 | **Align Your Gaussians** (2024a) | T | SD | Deformable GS | 4D animation | P | – |
| C27 | **Stable Video 4D** (2025) | V | Diff.+REC | Dynamic field | 4D object | P | – |
| C28 | **DimensionX** (2025) | I P | Diff.+REC | Space-time GS | 3D/4D scene | P | ↗ |
| C29 | **CAT4D** (2025a) | V P | Diff.+REC | Deformable GS | 4D scene | P | – |
| C30 | **Full-4D** (2026a) | V P | Diff.+REC | Multiview video / deforming GS | Dynamic GS /views | P | – |

Continued on next page

**Table 1**: Catalogue of visual-asset, spatial-scene, and temporal generative methods. (continued).

**ID:** A assets; B scenes; C temporal methods. **Stage:** P preparation; A interactive authoring; R ongoing generation/prediction. Input and reading key on p. 30; mechanism keys on pp. 30, 32, and 34. ↗ official project/GitHub.

| ID | Method / year | Input | Mechanism | Representation / state | Output | Stage | ↗ |
|---|---|---|---|---|---|---|---|
| **C. Temporal — Controlled and reference-guided video** | | | | | | | |
| C31 | **MotionCtrl** (2024d) | PM | CTRL | NR | Video | P | – |
| C32 | **CameraCtrl** (2025a) | TP | Diff. | NR | Video | P | – |
| C33 | **Tune-A-Video** (2023) | TV | Diff. | NR | Video | P | – |
| C34 | **AnimateDiff** (2024) | T | Diff. | NR | Video | P | – |
| C35 | **DynamiCrafter** (2024) | I | Diff. | Image-context features | Video | P | – |
| C36 | **VideoComposer** (2023a) | TVSM | Diff. | Space-time features | Video | P | – |
| **C. Temporal — Action-conditioned worlds and predictive models** | | | | | | | |
| C37 | **Matrix-Game 2.0** (2025b) | IA | AR+Diff. | Recent KV cache | Video | R | ↗ |
| C38 | **Genie 3** (2025) | TA | NR | NR | Video | R | ↗ |
| C39 | **Genie** (2024) | TISA | AR | Video/action tokens | Video | R | – |
| C40 | **GameNGen** (2025) | VA | AR+Diff. | Recent frame history | Video | R | – |
| C41 | **DIAMOND** (2024) | VA | AR+Diff. | Recent frame history | Predicted video | R | ↗ |
| C42 | **Hunyuan-GameCraft** (2025b) | AP | AR+Diff. | Historical context | Video | R | – |
| C43 | **Hunyuan-GameCraft-2** (2025) | TIA | GEN | NR | Video | R | – |
| C44 | **GameFactory** (2025b) | A | AR+Diff. | NR | Video | R | – |
| C45 | **Pandora** (2024) | TA | AR+Diff. | NR | Video | R | – |
| C46 | **iVideoGPT** (2024a) | VA | AR | Obs./action/reward | Predicted video | R | – |
| C47 | **Vid2World** (2026c) | VA | AR+Diff. | NR | Predicted video | R | – |
| C48 | **DreamerV3** (2025) | IA | RSSM | Recurrent/stochastic state | Latent/reward predictions | R | – |
| C49 | **Solaris** (2026) | VA | AR+Diff. | Per-player video latents /KV cache | Multiplayer videos | R | ↗ |
| **C. Temporal — Spatial memory and persistent state** | | | | | | | |
| C50 | **WorldPlay** (2026a) | TIPA | Diff. | Pose/frame memory | Video | R | ↗ |
| C51 | **Matrix-Game 3.0** (2026g) | IPA | AR+Diff. | Pose/frame memory | Video | R | ↗ |
| C52 | **MultiGen** (2026) | GPA | Diff.+DYN | Map/player poses | Video +poses | R | ↗ |
| C53 | **VMem** (2025c) | IP | AR+Diff. | Surfel-indexed views | Video | R | ↗ |
| C54 | **Long-term Spatial Memory** (2025c) | TVP | AR+Diff. | Point-map/frame memory | Video | R | – |
| C55 | **State-Space Video WM** (2025) | A | SSM+Diff. | Recurrent state | Video | R | – |
| C56 | **ReWorld** (2026b) | AP | Diff. | KV/landmark memory | Video | R | – |
| C57 | **WorldMem** (2025) | VPA | Diff. | Frame/pose/time memory | Video | R | – |
| C58 | **PERSIST** (2026) | IAG | DYN | Latent 3D scene | Video | A/R | – |
| C59 | **FloWM** (2026) | VA | DYN | Latent spatial map / velocity channels | Predicted video | R | ↗ |
| C60 | **LiveWorld** (2026) | TIP | Diff.+DYN | Static geometry /entity states | Camera-conditioned video | R | ↗ |
| C61 | **AnchorWeave** (2026f) | TIVP | Diff. | Retrieved local point clouds /poses | Camera-controlled video | R | ↗ |

Continued on next page

**Table 1**: Catalogue of visual-asset, spatial-scene, and temporal generative methods. (continued).

**ID:** A assets; B scenes; C temporal methods. **Stage:** P preparation; A interactive authoring; R ongoing generation/prediction. Input and reading key on p. 30; mechanism keys on pp. 30, 32, and 34. ↗ official project/GitHub.

| ID | Method / year | Input | Mechanism | Representation / state | Output | Stage | ↗ |
|---|---|---|---|---|---|---|---|
| C62 | **ActWorld** (2026) | T V P A | AR+Diff. | Ranked history /event / identity tokens | Action-conditioned video | R | – |
| C63 | **Programmable World Model** (2026d) | T I P A | Diff.+DYN | Entity IDs /3D boxes / rules | State-conditioned video | A/R | ↗ |
| C64 | **WorldDirector** (2026d) | T I P | Diff.+ CTRL | 3D boxes /appearance / memory | Controlled video | R | ↗ |
| **C. Temporal — Long-form and streaming generation** | | | | | | | |
| C65 | **Diffusion Forcing** (2024a) | NR | Diff. | Sequence tokens | Predicted sequences | P/R | – |
| C66 | **History-Guided VD** (2025) | V | Diff. | Frame history | Video | P/R | – |
| C67 | **CausVid** (2025) | T I V | Diff.+ DIST | KV cache | Video | R | – |
| C68 | **Self Forcing** (2025d) | T | AR+Diff. | Rolling KV cache | Video | R | – |
| C69 | **FreeNoise** (2024) | T | Diff. | Correlated noise windows | Video | P | – |
| C70 | **FIFO-Diffusion** (2024) | T | Diff. | Latent frame queue | Video | R | – |
| C71 | **StreamingT2V** (2025) | T | AR+Diff. | Context/appearance features | Video | R | – |
| C72 | **LongLive** (2026) | T | AR | Refreshed KV cache | Video | R | – |

# 4 Data and Evaluation

Evaluation connects model capabilities to XR task requirements through data roles, measured endpoints, quality, and operating costs. Table 2 summarizes supporting resources; Figure 10 maps capability coverage in the detailed evidence subset. These results inform the system analysis and task-based selection in Section 5.

## 4.1 Data sources and their roles

Object collections supervise geometry and appearance, scene observations support spatial reconstruction, and executable environments expose actions and outcomes. WorldPlay uses DL3DV among its training sources, while Generated Reality uses HOT3D for egocentric hand–object interaction (Sun et al. 2026a; Xie et al. 2026b). HOT3D provides synchronized egocentric views with tracked hand, object, and camera poses (Banerjee et al. 2025). Figures 9 and 13 illustrate the datasets' capture contexts.

Test data define the reference a result can recover. SceneScape evaluates 22 indoor RealEstate10K videos subsampled to 25 frames. Under the same prompt, SceneScape receives camera poses and GEN-1 receives RGB frames (Fridman et al. 2023). Holodeck's artist-designed NoveltyTHOR environments instead support navigation evaluation (Yang et al. 2024d).

Table 2 describes public resources and their scales; Table 8 records the source-specific experimental subsets and workloads. Counts retain their original units, including objects, trajectories, and user ratings.

**Geometry, appearance, and their capture.** ShapeNet, Objaverse, and Objaverse-XL provide object models with different category coverage and annotations. ABO adds product-centric images, models, and benchmarks (Chang et al. 2015; Deitke et al. 2023b,a; Collins et al. 2022). Google Scanned Objects supplies textured household models, while Common Objects in 3D (CO3D) couples object-centric videos with reconstructed geometry (Downs et al. 2022; Reizenstein et al. 2021). Filtering, rendering, licensing, and split decisions determine the subset used for shape and appearance learning; resource totals therefore differ from effective training size.

**Scene observations and reconstructed structure.** 3D-FRONT supplies furnished room structure, DL3DV-10K supplies scene observations, and RealEstate10K supports posed video-based view synthesis (Fu et al. 2021; Ling et al. 2024b; Zhou et al. 2018). SUN RGB-D provides color images, depth measurements, and scene annotations. The Stanford Large-Scale 3D Indoor Spaces (S3DIS) dataset

annotates building interiors. ScanNet links camera trajectories, RGB-D observations, reconstructed surfaces, and semantics (Song et al. 2015; Armeni et al. 2016; Dai et al. 2017). Matterport3D and Replica emphasize reconstructed indoor environments, while ScanNet++ and ARKitScenes cover different capture devices and geometry/appearance conditions (Chang et al. 2017; Straub et al. 2019; Yeshwanth et al. 2023; Baruch et al. 2021). To distinguish view agreement from recovery of unobserved structure, evaluation must identify observed versus completed surfaces and the reference type: imagery, geometry, or annotation.

**Executable environments and task state.** Habitat-Matterport 3D (HM3D) provides reconstructed spaces that can support simulation. Habitat and Gibson expose environments for embodied navigation (Ramakrishnan et al. 2021; Savva et al. 2019; Xia et al. 2018). AI2-THOR adds object actions, ProcTHOR scales scene construction, and Habitat 2.0 supports rearrangement (Kolve et al. 2017; Deitke et al. 2022; Szot et al. 2021). SAPIEN supplies articulated objects and physics-based manipulation. iGibson 1.0 combines interactive scenes with demonstration tools, and iGibson 2.0 adds object-state changes and a VR interface (Xiang et al. 2020; Shen et al. 2021a; Li et al. 2022). BEHAVIOR and BEHAVIOR-1K organize such interactions into household goals and activities (Srivastava et al. 2022; Li et al. 2023). An evaluation can vary appearance, layout, object properties, or available actions while retaining a defined task outcome.

**Recorded activity and transfer.** Ego4D and Ego-Exo4D document everyday and skilled actions from first-person or paired first- and third-person views (Grauman et al. 2022, 2024). RoboTHOR studies navigation across simulated and physical environments, while ManiSkill2 evaluates manipulation across objects and task families (Deitke et al. 2020; Gu et al. 2023). Recorded actions document what occurred; testing an unrecorded intervention additionally requires an executable environment or a separate experimental protocol.

**Table 2**: Datasets, simulators, and evaluation benchmarks used in the review.

**Signals:** R RGB; D depth; P camera/object/hand pose; G geometry/layout; S semantics/boxes; A actions. A check marks a selected documented signal; a dash marks a signal unlisted in the inspected description. **Origin:** R real capture; S synthetic/simulated; M mixed. For simulators, origin refers to generated observations. Scale: dataset size or simulator/benchmark workload, in the cited release's units; NR not established here. Mat. materials; Meta. metadata; Tex. texture; Phys. physics; Pref. preference; Anim. animation.

| # | Resource / year | Domain | Origin | R | D | P | G | S | A | Other signals | Reported scale |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | | | | | | | | | | | |
| 1 | **ShapeNet** (2015) | assets | S | – | – | – | ✓ | ✓ | – | – | 220K classified models |
| 2 | **Objaverse 1.0** (2023b) | assets | NR | – | – | – | ✓ | – | – | Text/Anim | 800K+models |
| 3 | **Objaverse-XL** (2023a) | assets | M | – | – | – | ✓ | – | – | Meta. | >10M objects |
| 4 | **ABO** (2022) | assets | M | ✓ | – | – | ✓ | – | – | Meta. | NR |
| 5 | **Google Scanned Objects** (2022) | assets | R | – | – | – | ✓ | – | – | Tex. | >1,000 objects |
| 6 | **3D-FRONT** (2021) | scenes | S | – | – | – | ✓ | ✓ | – | Tex. | 18,968 rooms |
| 7 | **SUN RGB-D** (2015) | scenes | R | ✓ | ✓ | – | – | ✓ | – | – | 10,335 images |
| 8 | **S3DIS** (2016) | scenes | R | – | – | – | ✓ | ✓ | – | – | >6,000 $m^2$ |
| 9 | **ScanNet** (2017) | scenes | R | ✓ | ✓ | ✓ | ✓ | ✓ | – | – | 1,513 scenes |
| 10 | **Matterport3D** (2017) | scenes | R | ✓ | ✓ | ✓ | ✓ | ✓ | – | – | 90 buildings |
| 11 | **Replica** (2019) | scenes | R | – | – | – | ✓ | ✓ | – | Tex. | 18 reconstructions |
| 12 | **ScanNet++** (2023) | scenes | R | ✓ | ✓ | – | ✓ | ✓ | – | – | 460 scenes |
| 13 | **ARKitScenes** (2021) | scenes | R | ✓ | ✓ | – | – | ✓ | – | – | NR |
| 14 | **HM3D** (2021) | scenes | R | – | – | – | ✓ | – | – | Tex. | 1,000 reconstructions |
| 15 | **DL3DV-10K** (2024b) | video | R | ✓ | – | – | – | – | – | – | NR |
| 16 | **RealEstate10K** (2018) | video | R | ✓ | – | ✓ | – | – | – | – | ∼80,000 clips |
| 17 | **CO3D** (2021) | assets | R | ✓ | – | ✓ | ✓ | – | – | – | ∼19,000 videos |
| 18 | **Ego4D** (2022) | actions | R | ✓ | – | – | – | – | ✓ | – | 3,670 hours |
| 19 | **Ego-Exo4D** (2024) | actions | R | ✓ | – | ✓ | – | – | ✓ | IMU | 1,286 hours |
| 20 | **HOT3D** (2025) | actions | R | ✓ | – | ✓ | – | – | – | – | NR |
| 21 | **SCOPE** (2025) | scenes | M | – | – | – | – | – | – | Stereo/Pref. | 2,400 samples |
| **Simulator** | | | | | | | | | | | |
| 22 | **Habitat** (2019) | scenes | S | ✓ | ✓ | – | – | – | ✓ | – | NR |
| 23 | **Gibson** (2018) | scenes | S | ✓ | – | – | ✓ | – | ✓ | – | 572 buildings |
| 24 | **AI2-THOR** (2017) | actions | S | ✓ | – | – | – | – | ✓ | State | NR |
| 25 | **ProcTHOR** (2022) | scenes | S | ✓ | – | – | ✓ | – | ✓ | – | 10,000 sampled houses |
| 26 | **SAPIEN** (2020) | actions | S | – | – | – | ✓ | – | ✓ | Phys. | NR |
| 27 | **Habitat 2.0** (2021) | actions | S | ✓ | – | – | – | – | ✓ | Phys. | NR |
| 28 | **iGibson 1.0** (2021a) | actions | S | ✓ | ✓ | – | – | ✓ | ✓ | – | 15 home scenes |
| 29 | **iGibson 2.0** (2022) | actions | S | – | – | – | – | – | ✓ | State/Demo. | NR |
| 30 | **RoboTHOR** (2020) | actions | M | ✓ | – | – | – | – | ✓ | – | NR |
| **Benchmark** | | | | | | | | | | | |
| 31 | **BEHAVIOR** (2022) | actions | S | – | – | – | – | – | ✓ | Goal/State/Demo. | 100 activities |
| 32 | **BEHAVIOR-1K** (2023) | actions | S | – | – | – | – | – | ✓ | Goal/State/Phys. | 1,000 activities |
| 33 | **ManiSkill2** (2023) | actions | S | ✓ | ✓ | – | ✓ | – | ✓ | – | 20 task families |
| 34 | **NoveltyTHOR** (2024d) | scenes | S | – | – | – | ✓ | – | ✓ | Goal | 10 test scenes |
| 35 | **VBench** (2024b) | video | S | – | – | – | – | – | – | T/Video/Rating | NR |
| 36 | **WorldScore** (2025) | video | NR | – | – | ✓ | – | – | – | T/I/Motion | 3,000 specifications |
| 37 | **4DWorldBench** (2026b) | video | NR | – | – | – | – | – | – | T/I/Video/Rating | NR |

## 4.2 Evaluation targets, metrics, and evidence

**Geometry, perception, and operation.** Text2Room's judgments concern perceived scene structure; TRELLIS.2's reconstruction tests concern reference geometry (Höllein et al. 2023; Xiang et al. 2026). CoGen3D's imported-object manipulation illustrates evaluation at the system level (Jiang et al. 2026).

**Camera motion and scene dynamics.** Spatial evaluation separates viewpoint change from change within the world. WorldScore measures camera adherence, static-world agreement estimated from DROID-SLAM depth and co-visible-pixel reprojection, and requested in-scene motion under a fixed camera (Duan et al. 2025).

**Validation of model-based judgments.** Automated judgments require calibration against the endpoint they intend to measure. 4DWorldBench assesses perceptual quality, condition alignment, physical realism, and 4D consistency under text, image, and video conditions. It uses model-based and

language/vision-language-model judgments (Lu et al. 2026b). Against ten participants' ratings of 100 videos, its initial style-consistency ablation increases Pearson correlation from 0.383 to 0.545; Spearman correlation remains 0.457. Linear association with human ratings improves, while rank agreement is unchanged.

**Stereoscopic viewing experience.** Perceptual evaluation depends on both generated content and its display conditions. The Stereoscopic Content Preference Evaluation (SCOPE) dataset records static-stereo preferences on Apple Vision Pro; iSQoE predicts quality of experience from these annotations (Tamir et al. 2025). A separate comparison uses Apple Vision Pro, Meta Quest Pro, anaglyph viewing, and alternating monocular monitor views. These conditions distinguish geometric proxies, desktop preferences, and headset judgments. The resulting evidence concerns static-stereo viewing.

**Feature-based and distributional measures.** CLIP aligns image and text representations (Radford et al. 2021); TRELLIS compares rendered views with an input image, whereas Text2Room compares them with text (Xiang et al. 2025; Höllein et al. 2023). LPIPS approximates image perceptual similarity using calibrated deep features (Zhang et al. 2018); SSIM compares structure, luminance, and contrast under its signal model (Wang et al. 2004). Semantic correspondence can therefore improve while less of a reference view is preserved.

Fréchet Inception Distance compares fitted feature distributions; FVD extends this assessment to video features (Huang and Bors 2020, 2021, 2022b,a) (Heusel et al. 2017; Unterthiner et al. 2018). Because Inception-v3 and DINOv2 represent different learned visual spaces (Szegedy et al. 2016; Oquab et al. 2024), Table 3 retains separate Inception, DINO, and point-feature distances. VBench characterizes generated video across multiple quality dimensions (Huang et al. 2024b).

**Measures should follow the user-facing claim.** Presence questionnaires describe subjective experience (Witmer and Singer 1998; Schubert et al. 2001). Schwind et al. found unchanged mean presence scores for questionnaires administered inside versus outside VR, although response variance and interruptions differed (Schwind et al. 2019). Studies should therefore record the rating procedure, preceding task, and display conditions. The Simulator Sickness Questionnaire measures symptoms; a review of head-mounted VR studies associates sickness with participant, content, and exposure factors (Kennedy et al. 1993; Saredakis et al. 2020). Presence ratings should accordingly be accompanied by exposure duration, withdrawals, and discomfort.

**Workload, usability, and successful use.** The NASA Task Load Index measures perceived workload dimensions; IBM usability questionnaires assess satisfaction with system usability (Hart and Staveland 1988; Lewis 1995). These ratings and time savings should accompany task criteria, including correction accuracy, multimodal reference resolution, or coordinated manipulation. Task criteria distinguish completion of the intended operation from acceptance of an incorrect result.

**Reading the evidence profiles.** The categories describe evidence types and do not rank capabilities. E denotes a measured result for a named property and setup, including mixed or negative findings; D denotes a demonstrated implemented function. C denotes a source claim without a located targeted test, and I denotes a mechanism-based inference. NR means no evidence was located in the inspected source, not that the capability is absent; NA denotes a property outside the selected output's scope. Temporal entries distinguish evolution, ordered views, revisiting, and retained modifications. For operating cost, E requires a comparison, profile, configuration experiment, or repeated/averaged measurement; an isolated approximate timing remains C.

## 4.3 Source-reported quantitative comparisons

Tables 3–5 compare assets, scenes, and dynamic worlds under eight source protocols. Each block retains comparison methods, inputs, test sets, metric directions, and reported precision so that trade-offs remain interpretable within the original task.

**Assets: appearance and part structure.** TRELLIS-L improves appearance- and shape-distribution scores over the listed Toys4k baselines. Across all three datasets, PartCrafter has lower Chamfer distance and inter-part overlap and higher F-score than TripoSG–HoloPart (Xiang et al. 2025; Lin et al. 2025). Overlap measures generated-part intersection, rather than segmentation agreement with ground truth.

**Scenes: camera adherence and visual quality.** Text2Room combines image scores with perceived completeness; Voyager measures fidelity to held-out observations. WorldScore compares complete configurations with different control interfaces. Its displayed 3D methods receive precise camera poses and score 84.60–94.01 on camera control; the displayed video methods lack that input and score 22.39–38.27. The contrast combines representation and control-interface differences. Hailuo leads object control

## Detailed extraction subset

31 records × 8 dimensions. E/ = evidence ID; catalogue IDs are separate.

| ID | Method / system | Temporal target | Output properties | | | Control/action | | Delivery/use | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Spatial | Binocular | Temporal | Control | Interaction | Delivery | Cost | XR device |
| **Assets** | | | | | | | | | | |
| E/A01 | DreamFusion | | D | NR | NA | E | NR | D | E | NR |
| E/A02 | Magic3D | | D | NR | NA | D | NR | D | E | NR |
| E/A03 | TRELLIS | | E | NR | NA | D | NR | D | NR | NR |
| E/A04 | Hunyuan3D 2.0 [a] | Studio motion | E | NR | D | E | NR | D | NR | NR |
| E/A05 | TRELLIS.2 [b] | | E | NR | NA | D | NR | D | E | NR |
| E/A06 | PartCrafter | | E | NR | NA | D | I | D | E | NR |
| **Spatial scenes and multiview generation** | | | | | | | | | | |
| E/SCE01 | Text2Room | | E | NR | NA | E | D | D | C | C |
| E/SCE02 | SceneScape | Ordered views | E | NR | E | E | NR | I | E | NR |
| E/SCE03 | SceneDreamer | Ordered views | E | NR | E | D | NR | NR | C | NR |
| E/SCE04 | Holodeck | | E | NR | NA | E | E | D | C | NR |
| E/SCE05 | HunyuanWorld 1.0 | | D | NR | NA | E | D | D | NR | C |
| E/SCE06 | HY-World 2.0 | View memory | E | NR | D | E | D | I | E | C |
| E/D04 | Voyager | World-cache reuse | E | NR | D | E | NR | NR | NR | C |
| E/P01 | MVDiffusion | | E | NR | NA | E | NR | D | NR | C |
| **Dynamic content and action-conditioned worlds** | | | | | | | | | | |
| E/D01 | 4D-fy | Motion quality | E | NR | E | E | NR | NR | E | NR |
| E/D02 | DreamGaussian4D | Video / motion | E | NR | E | D | NR | D | E | C |
| E/D03 | DimensionX | Subject consistency | E | NR | E | D | NR | NR | NR | NR |
| E/D05 | Matrix-Game 2.0 | 597-frame video | NR | NR | E | E | D | NR | E | NR |
| E/D06 | Matrix-Game 3.0 | Minute memory | D | NR | D | D | D | NR | E | C |
| E/D07 | WorldPlay [d] | Revisit recall | E | NR | E | E | D | NR | E | NR |
| E/D08 | Genie 3 | Minute coherence | D | NR | D | D | D | NR | C | NR |
| E/D09 | MultiGen | Reference fidelity | E | NR | E | D | E | NR | C | NR |
| **Stereo video** | | | | | | | | | | |
| E/P02 | EPIC | Rigid tracks | E | E | E | D | NR | D | NR | C |
| **XR systems and displayed-content studies** | | | | | | | | | | |
| E/X01 | VRCopilot | | D | NR | NR | E | E | D | NR | E |
| E/X02 | MS2Mesh-XR | | D | NR | NR | D | D | D | E | D |
| E/X03 | Dreamcrafter | | D | NR | NR | E | E | D | C | E |
| E/X04 | CoGen3D | | D | NR | NR | D | E | D | E | E |
| E/X05 | Generated Reality [c] | Short hand tasks | E | NR | D | E | E | D | C | E |
| E/X06 | ImaginateAR (phone AR) | State/asset reuse | E | NR | D | E | E | D | E | E |
| E/X07 | Heritage visualization | | E | NR | NR | D | D | D | E | E |
| E/X08 | Panoramic stimulus validation | | E | NR | NA | E | NR | D | NR | E |

| | | | | | |
|---|---|---|---|---|---|
| E | Targeted evaluation | D | Implementation / demo | C | Author claim |
| I | Analytical inference | NR | Not located | NA | Not applicable |

E includes mixed/negative findings. NR is an inspection boundary; NA is task-specific.

[a] Temporal: Studio animation. [b] Cost: decoder measurement; full generation is a claim.

[c] Spatial: offline 14B; deployed XR: 5B. [d] WorldPlay: arXiv v2.

Index of inspected configurations, not a jointly tested system; categories are non-ordinal.

**Fig. 10** Evidence profiles for 31 methods and systems, with property, configuration and temporal-target keys. E/ IDs identify companion records described in Appendix A.

and CogVideoX-I2V leads style consistency. Camera adherence and visual instruction fulfillment therefore remain separate selection targets. Table 4 preserves the benchmark scores and evaluation settings (Höllein et al. 2023; Huang et al. 2025c; Duan et al. 2025).

**Dynamics: reconstruction, return, and shared observations.** DreamGaussian4D evaluates novel-view fidelity and offline generation time for prepared dynamic content. WorldPlay's short-trajectory PSNR is 21.92 dB versus Gen3C's 21.68 dB; on long revisits, the scores are 18.94 and 15.37 dB. The long protocol references the initial generated pass, so the two protocols assess different reconstruction targets. MultiGen raises opponent-presence accuracy from the split-screen baseline's 65.31% to 75.38% on one Doom map. Its rollouts demonstrate player removal and reintroduction in shared memory, with opponent presence as the quantitative endpoint (Ren et al. 2023; Sun et al. 2026a; Po et al. 2026). MultiGen's baselines include IP-Adapter, which introduces image prompting through separate image and text attention (Ye et al. 2023). Section 6.3 adds an edit-and-return test for retained object changes.

**Table 3** Reported asset-generation results under two source protocols.

**A:** Single-image input; 1,250 Toys4k objects; one asset/input. FD/KD: Fréchet/kernel distance. FD-point uses PointNet++ features of depth-derived point clouds. CLIP and both KD columns use the source's ×100 scale.

**A. Whole-object generation** — Xiang et al. (2025) (arXiv v1)

| Method | CLIP↑ | Inception-v3 | | DINOv2 | | FD-point↓ |
|---|---|---|---|---|---|---|
| | | FD↓ | KD↓ | FD↓ | KD↓ | |
| Shap-E | 82.11 | 34.72 | 0.87 | 465.74 | 62.72 | 8.20 |
| LGM | 83.97 | 26.31 | 0.48 | 322.71 | 38.27 | 15.90 |
| InstantMesh | 84.43 | 20.22 | 0.30 | 264.36 | 25.99 | 9.63 |
| 3DTopia-XL† | 78.45 | 37.68 | 1.20 | 437.37 | 53.24 | 18.21 |
| LN3Diff | 82.74 | 26.61 | 0.68 | 357.93 | 50.72 | 7.86 |
| TRELLIS-L | **85.77** | **9.35** | **0.02** | **67.21** | **0.72** | **2.03** |

**B:** RGB image; PartCrafter also takes part count $N$. CD: L2 Chamfer distance; F: F-score at 0.1; IoU: inter-part overlap on a $64^3$ grid.

**B. Part-level object generation** — Lin et al. (2025)

| Method | Objaverse | | | ShapeNet | | | ABO | | |
|---|---|---|---|---|---|---|---|---|---|
| | CD↓ | F↑ | IoU↓ | CD↓ | F↑ | IoU↓ | CD↓ | F↑ | IoU↓ |
| TripoSG | 0.3104 | 0.5940 | — | 0.3751 | 0.5050 | — | 0.2017 | 0.7096 | — |
| TripoSG* | 0.1821 | 0.7115 | — | 0.3301 | 0.5589 | — | 0.1503 | 0.7723 | — |
| TripoSG → HoloPart | 0.1916 | 0.6916 | 0.0443 | 0.3511 | 0.5498 | 0.1107 | 0.1338 | 0.8093 | 0.0449 |
| PartCrafter | 0.1726 | 0.7472 | 0.0359 | 0.3205 | 0.5668 | 0.0293 | 0.1047 | 0.8617 | 0.0243 |

†3DTopia-XL's five appearance/alignment scores use shaded PBR-mesh renders. TRELLIS appearance uses Gaussian renders; geometry uses meshes. *TripoSG uses $512N$ tokens, matching PartCrafter. Monolithic TripoSG part IoU is unreported (—). About 2K test samples overall; per-dataset counts unspecified. Four-part examples on one H20 GPU: TripoSG → HoloPart 18 min; PartCrafter 34 s. Bold: block-A column optima; block B retains separate token/decomposition configurations.

## 4.4 Cross-method qualitative comparisons

Figures 11 and 12 complement source-reported scores with shared-input comparisons. Both plates retain the source papers' selected results and comparison conditions.

**Assets: appearance and geometric detail.** Figure 11 compares six image-conditioned generators on shared excavator and chest inputs. The excavator reveals differences in boom, cabin, and track continuity; the chest reveals differences in spikes, metal trim, and the front ornament. Several outputs retain overall colors while simplifying or distorting these structures; TRELLIS preserves more local detail in these examples. Alongside Table 3, the views show why appearance agreement and geometric completeness require separate inspection.

**Scenes: content retention under a viewpoint change.** Figure 12 compares Voyager, FlexWorld, SEVA, See3D, and ViewCrafter on four shared examples. The staircase reveals displaced surfaces and distorted furniture; the bathroom reveals lost or altered small objects. Alongside Table 4, they show

**Table 4** Scene-generation results under three source protocols. Arrows give preferred directions; bold marks within-block optima.

**(a) Text-to-room generation.** Höllein et al. (2023). Text input; 60 novel views/room; 61 raters. CS: CLIP score; IS: Inception score; PQ: perceptual quality; 3DS: perceived 3D completeness.

| Method / pipeline | CS↑ | IS↑ | PQ↑ (1–5) | 3DS↑ (1–5) |
|---|---|---|---|---|
| PureClipNeRF | 24.06 | 1.26 | 2.34 | 2.38 |
| Outpainting | 23.10 | 1.60 | 2.90 | 2.58 |
| Text2Light + mesh fusion | 25.99 | 2.21 | 2.82 | 2.97 |
| Blockade + mesh fusion | 26.29 | 2.13 | 3.35 | 3.36 |
| Text2Room | **28.02** | **2.31** | **4.01** | **4.19** |

**(b) Image-conditioned novel-view synthesis.** Huang et al. (2025c) (arXiv v1). 150 RealEstate10K test clips; image and target trajectory; source-estimated camera/depth conditions.

| Method | PSNR↑ (dB) | SSIM↑ | LPIPS↓ |
|---|---|---|---|
| SEVA | 16.648 | 0.613 | 0.349 |
| ViewCrafter | 16.512 | 0.636 | 0.332 |
| See3D | 18.189 | 0.694 | 0.290 |
| FlexWorld | 18.278 | 0.693 | 0.281 |
| Voyager | **18.751** | **0.715** | **0.277** |

**(c) WorldScore-Static.** Duan et al. (2025) (ICCV 2025 snapshot). 2,000 static specifications; seven normalized scores (0–100). First six rows: 3D scene methods; last six: video methods.

| | Controllability | | | Quality | | | |
|---|---|---|---|---|---|---|---|
| Method | Camera ctrl.↑ | Object ctrl.↑ | Content align.↑ | 3D consist.↑ | Photo. consist.↑ | Style consist.↑ | Subject. quality↑ |
| SceneScape | 84.99 | 47.44 | 28.64 | 76.54 | 62.88 | 21.85 | 32.75 |
| Text2Room | **94.01** | 38.93 | 50.79 | 88.71 | 88.36 | 37.23 | 36.69 |
| LucidDreamer | 88.93 | 41.18 | **75.00** | **90.37** | **90.20** | 48.10 | 58.99 |
| WonderJourney | 84.60 | 37.10 | 35.54 | 80.60 | 79.03 | 62.82 | **66.56** |
| InvisibleStitch | 93.20 | 36.51 | 29.53 | 88.51 | 89.19 | 32.37 | 58.50 |
| WonderWorld | 92.98 | 51.76 | 71.25 | 86.87 | 85.56 | 70.57 | 49.81 |
| Gen-3 | 29.47 | 62.92 | 50.49 | 68.31 | 87.09 | 62.82 | 63.85 |
| Hailuo | 22.39 | **69.56** | 73.53 | 67.18 | 62.82 | 54.91 | 52.44 |
| DynamiCrafter | 25.15 | 47.36 | 25.00 | 72.90 | 60.95 | 78.85 | 54.40 |
| EasyAnimate | 26.72 | 54.50 | 50.76 | 67.29 | 47.35 | 73.05 | 50.31 |
| Allegro | 24.84 | 57.47 | 51.48 | 70.50 | 69.89 | 65.60 | 47.41 |
| CogVideoX-I2V | 38.27 | 40.07 | 36.73 | 86.21 | 88.12 | **83.22** | 62.44 |

**Evaluation settings.** (a) Panorama baselines use Text2Room's mesh fusion; scene count and resolution NR. (b) Resolution and frame sampling NR; scores evaluate video synthesis, not 3DGS reconstruction. (c) All six 3D rows receive precise camera poses; all six video rows lack this input. Other inputs, resolution, duration and compute also differ. Subjective quality uses an automatic ensemble.

that plausible room appearance need not preserve the reference scene's structure or details. Methods retain their native conditioning representations, and the displayed target frames support single-view comparisons (Huang et al. 2025c).

## 4.5 Costs of generation, rendering, and interaction

**Timing endpoints.** Table 6 separates preparation, component execution, and generation during use by timed stage, output, and hardware. WorldPlay's progressive decoding targets earlier feedback, whereas its speed table measures throughput (Sun et al. 2026a). Input-to-display latency also depends on buffering, scheduling, transfer, and presentation.

**Table 5** Dynamic-content and interactive-world performance. Arrows give preferred directions; bold marks within-block optima.

**(a) Video-to-4D novel-view fidelity.** Ren et al. (2023). Seven Consistent4D videos; four novel evaluation views/video. Full DG4D: 32 frames/four optimization views; Fast: 16 frames/one optimization view.

| Method / configuration | LPIPS↓ | CLIP↑ | FVD↓ | Gen. time↓ (offline) |
|---|---|---|---|---|
| D-NeRF | 0.51 | 0.68 | 2327.83 | — |
| K-planes | 0.38 | 0.72 | 2295.68 | — |
| Zero123 | 0.15 | 0.90 | 1571.60 | — |
| Consistent4D | 0.16 | 0.87 | 1133.44 | 2 hrs |
| DreamGaussian4D (Fast) | 0.13 | 0.91 | 775.90 | 5 mins |
| DreamGaussian4D | **0.12** | **0.92** | **729.74** | 15 mins |

**(b) Camera control and revisit consistency.** Sun et al. (2026a). 600 total cases; separate short/long counts NR.

| | Short: 61 frames | | Long: ≥ 250 frames | | | |
|---|---|---|---|---|---|---|
| Method | PSNR↑ (dB) | LPIPS↓ | PSNR↑ (dB) | LPIPS↓ | $R_{\text{dist}}$ ↓ | $T_{\text{dist}}$ ↓ |
| CameraCtrl | 17.93 | 0.298 | 10.09 | 0.549 | 0.733 | 1.117 |
| SEVA | 19.84 | 0.313 | 10.51 | 0.517 | 0.721 | 1.893 |
| ViewCrafter | 19.91 | 0.327 | 9.32 | 0.661 | 1.573 | 3.051 |
| Gen3C | 21.68 | 0.278 | 15.37 | 0.483 | 0.357 | 0.979 |
| VMem | 19.97 | 0.316 | 12.77 | 0.542 | 0.748 | 1.547 |
| Matrix-Game-2.0 | 17.26 | 0.383 | 9.57 | 0.631 | 2.125 | 2.742 |
| GameCraft | 21.05 | 0.341 | 10.09 | 0.614 | 2.497 | 3.291 |
| WorldPlay (full) | **21.92** | **0.247** | **18.94** | **0.371** | **0.332** | **0.797** |

**(c) Two-player opponent-presence consistency.** Po et al. (2026). One Doom map; matched action streams; VLM judgments of generated/reference views.

| Method | Accuracy↑ | Precision↑ | Recall↑ |
|---|---|---|---|
| IP-Adapter | 62.12% | 84.62% | 40.73% |
| ControlNet | 60.71% | 82.16% | 39.20% |
| Split-screen | 65.31% | 86.86% | 44.59% |
| MultiGen | **75.38%** | **88.12%** | **65.07%** |

**Evaluation settings.** (a) Times measure content optimization; baseline hardware unspecified. (b) Short: observed-frame reference; long: return-path frames against the initial generated pass. ViPE estimates generated poses; both trajectories are first-frame-relative. $R_{\text{dist}}$: angular distance (source unit unspecified); $T_{\text{dist}}$: squared translation distance after furthest-frame scale normalization, not metres. Common resolution unspecified. (c) Opponent-visible is positive; aggregation across trajectories/players is unspecified. Test count, rollout length, frame resolution and exact VLM version NR. —: not reported.

**Component time and complete response.** TRELLIS.2's A100 experiment measures shape-decoder time; its H100 figures describe full generation (Xiang et al. 2026). Generated Reality times generation and decoding; the display interval is unreported. Section 6.4 specifies complete-response measurement.

**Time to usable content.** The relevant endpoint is the first state that supports the intended operation. In the heritage study's three-pair comparison, raw generation takes approximately 3–4 minutes. Basic structural editability requires 4–5 hours of total post-processing, including 1–2 hours of close-up visual refinement (Liu et al. 2026). These measurements characterize a small hosted-service preparation workflow.

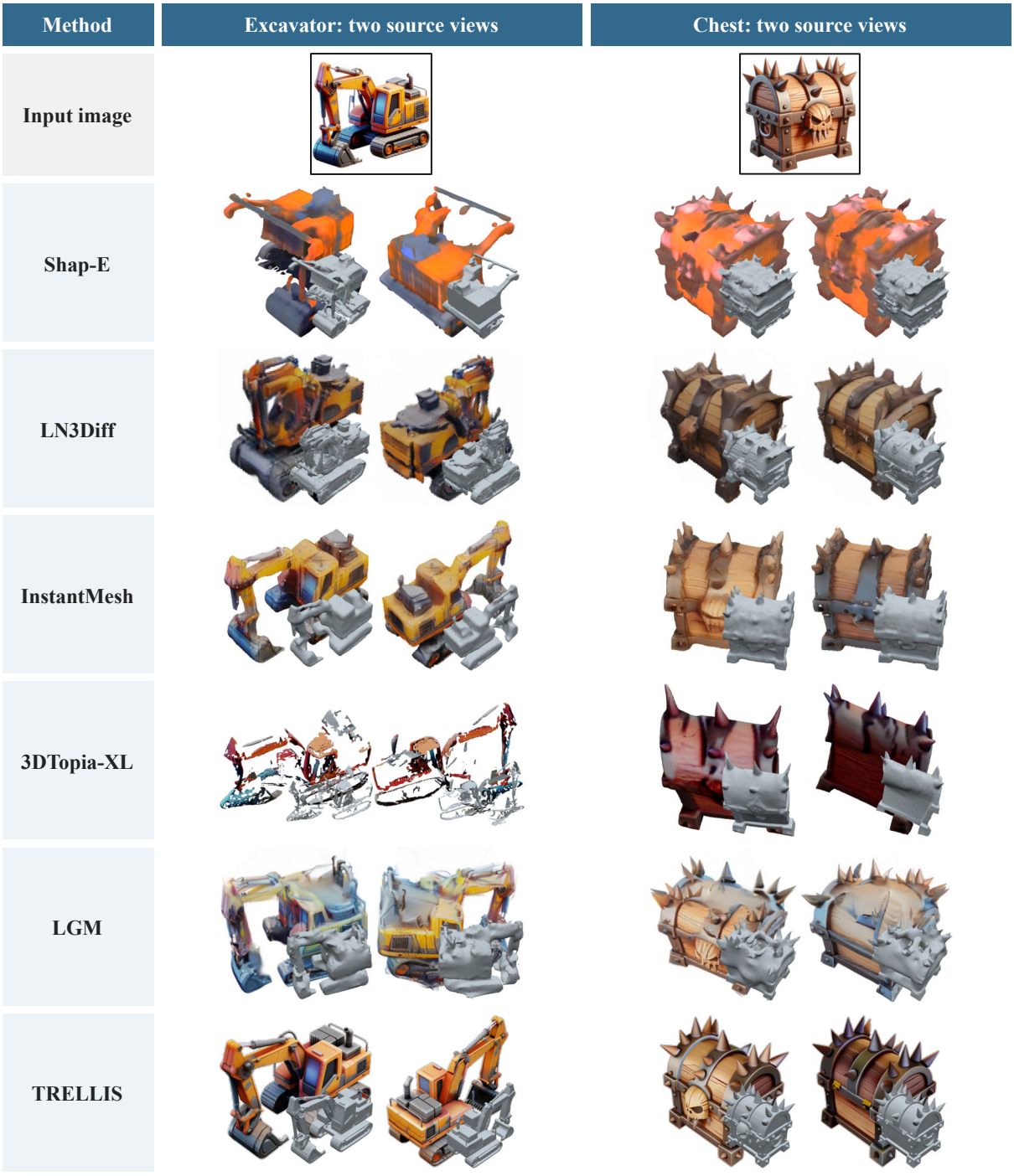



**Fig. 11** Image-conditioned 3D asset generation from shared inputs (top row). Each method retains two views and grey geometry insets. Images from Xiang et al. (2025).

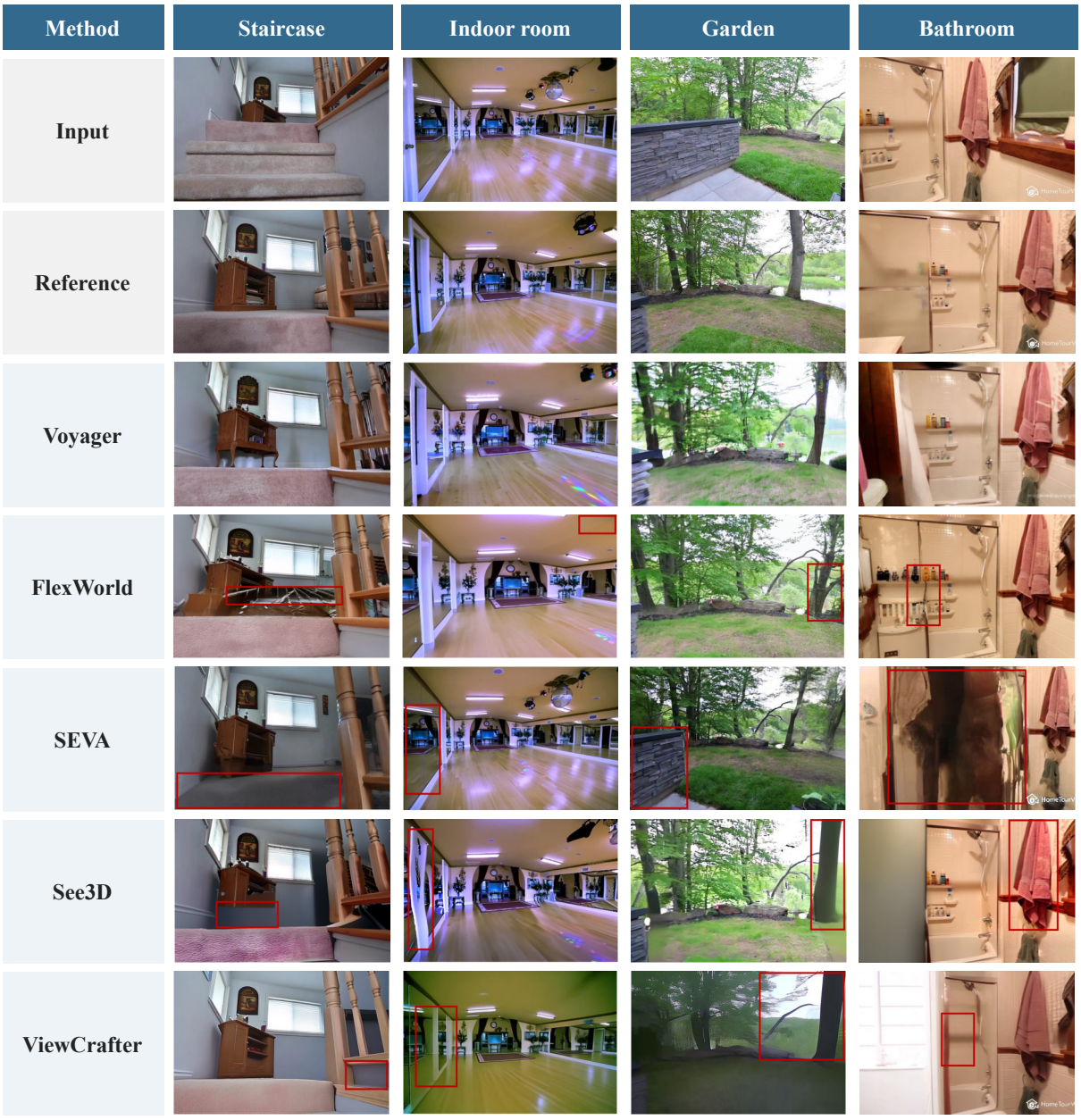



**Fig. 12** Camera-controlled scene generation across four examples. Rows show the starting input, observed target-view reference, and five methods at corresponding target views. Red boxes mark selected discrepancies. Images from Huang et al. (2025c).

**Table 6** Reported operating costs by stage, hardware, and output setting.

**Units:** gen. FPS generated frames/s; b batch size. **Evidence:** E timing evaluation; C isolated timing report. Both are source-reported (Section 4.2).

| # | Method | Hardware | Timed stage | Setting | Value | Unit | Ev. |
|---|---|---|---|---|---|---|---|
| **Preparation before use** | | | | | | | |
| 1 | Magic3D (2023) | 8× A100 | Optimization | 5k/3k steps; b32 | ≈ 40 | min | E |
| 2 | TRELLIS.2 (2026) | 1× H100 | Textured asset | $512^3$ | ≈ 3 | s | C |
| 3 | TRELLIS.2 | 1× H100 | Textured asset | $1024^3$ | ≈ 17 | s | C |
| 4 | TRELLIS.2 | 1× H100 | Textured asset | $1536^3$ | ≈ 60 | s | C |
| 5 | Text2Room (2023) | 1× RTX3090 | Scene build | 20 paths × 10 frames | ≈ 50 | min | C |
| 6 | 4D-fy (2024) | A100; ∼ 80 GB | Optimization | 2 + 2 + 19 h | ≈ 23 | h | C |
| 7 | Heritage (2026) | Hosted; NR | Raw generation | 3 paired workflows | ≈ 3–4 | min | E |
| 8 | Heritage | Hosted; NR | Post-processing | Structural editability | 4–5 | h | E |
| 9 | Heritage | Hosted; NR | Refinement | Included in 4–5 h | 1–2 | h | E |
| **Component execution** | | | | | | | |
| 10 | TRELLIS.2 | 1× A100 | Decoder | $512^3$ | 0.077 | s | E |
| 11 | TRELLIS.2 | 1× A100 | Decoder | $1024^3$ | 0.301 | s | E |
| 12 | 4D-fy | A100 | Dynamic render | Resolution NR | 71 | ms/ frame | E |
| 13 | 4D-fy | A100 | Static render | Resolution NR | 68 | ms/ frame | E |
| **Generation or preview during use** | | | | | | | |
| 14 | Dreamcrafter (2025) | GPU NR | Object preview | Quest 3; resolution NR | ≈ 15 | s | C |
| 15 | WorldPlay (2026a) | 8× H800 | Generation/ decode | 720p | 24 | gen. FPS | E |
| 16 | Generated Reality (2026b) | 1× H100 | Generation/ decode | 5B; 12-frame chunks | 11 | gen. FPS | C |
| 17 | Generated Reality | 1× H100 | Bottleneck delay | 5B; 12-frame chunks | ≈ 1.4 | s | C |

TRELLIS.2 reports shape-decoder time; 4D-fy renders prepared content; its 68 ms static comparator is MVDream-based. Generated Reality uses the deployed 5B model; delay bounds/display interval are unspecified. NR: Dreamcrafter final-asset time; WorldPlay first-frame delay; separate rendering time except 4D-fy; instrumented input-to-display delay throughout.

# 5 XR Systems, Workflows, and Model Selection

Table 7 links documented XR systems to their models, artifacts, platforms, tasks, and evaluations. We follow these configurations from generation through preparation and interaction to the measured user outcome.

## 5.1 Documented model-to-system links

**Assembly and asset-delivery pipelines.** Layout generation enters XR through object structure retained by the application. VRCopilot integrates ATISS, which autoregressively predicts furniture categories and bounding boxes from a room and partial layout. Unity instantiates furniture from an existing library and exposes it for manipulation (Paschalidou et al. 2021; Zhang et al. 2024c). The model supplies selection and arrangement; the library and application supply meshes and operations.

Image-to-mesh systems extend this handoff to geometry generated from user input. MS2Mesh-XR combines sketch and voice input, ControlNet 1.1 image generation, CRM reconstruction, and runtime OBJ import into Unity (Tong et al. 2025). ControlNet supplies spatial conditioning to image diffusion (Qi et al. 2024) (Zhang et al. 2023); CRM generates orthographic views and reconstructs a textured mesh through a convolutional triplane representation and FlexiCubes (Wang et al. 2024c). CoGen3D connects FLUX.1 images and Hunyuan3D-2 to textured GLB delivery in Unity (Jiang et al. 2026). Its desktop co-design workflow, hardware, and VR study document deployment beyond the Hunyuan3D 2.0 model evaluation (Zhao et al. 2025). Dreamcrafter combines captured Gaussian splats, meshes, and replaceable generative modules, including LGM or GRM (Vachha et al. 2025).

**Generated observations in the display loop.** Generated Reality specializes Eq. (26) for tracked head and hand inputs, then streams generated observations through Unity to Quest 3 (Xie et al. 2026b).

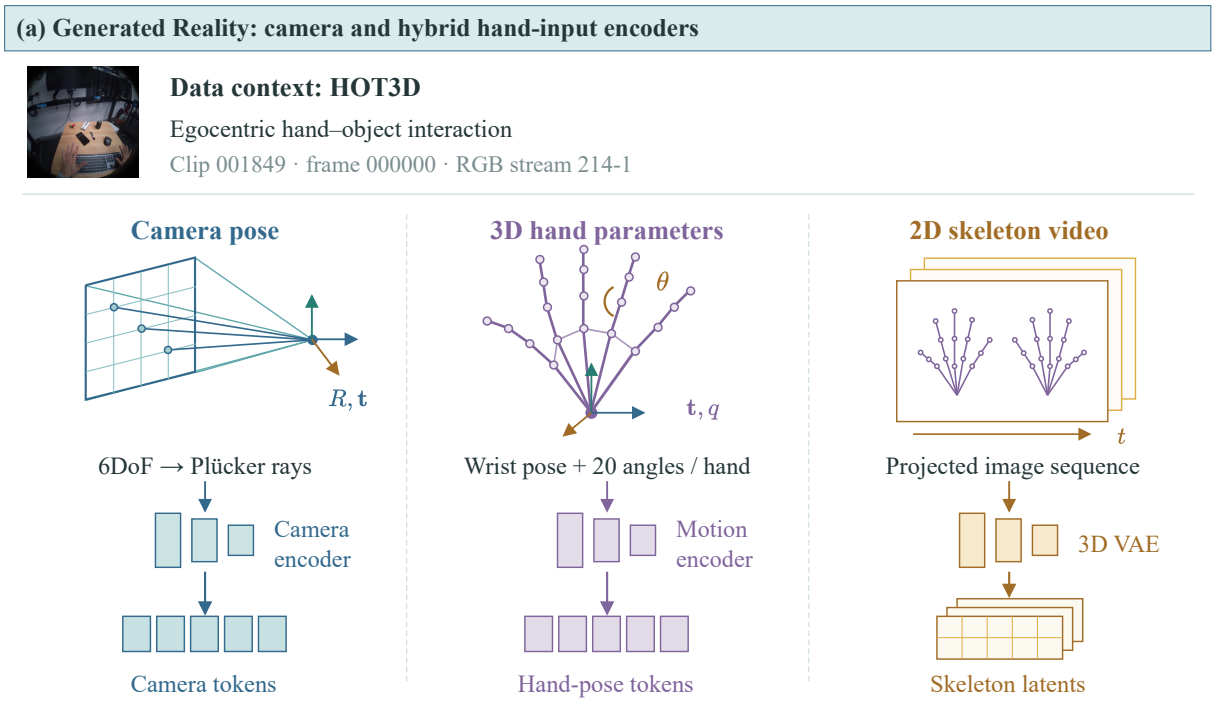



**Fig. 13** Generated Reality conditioning and visual responses. (a) Camera, hand-parameter, and skeleton encoders, with a HOT3D image (Banerjee et al. 2025). (b) Button pressing, wheel turning, and jar opening in the 5B headset study, compared with a head-pose/text baseline without tracked hands. Images from Xie et al. (2026b).

Relative to camera-conditioned generation, it adds hand parameters $H_k$ and projected skeleton latents $z_k^{\text{skel}}$. In the source's conditioning construction, the transformer input is

$$X_k = \text{Patchify}([z_k, z_k^{\text{skel}}]_{\text{ch}}) + E_{\text{hand}}(H_k) + E_{\text{cam}}(P_k). \tag{29}$$

Here $z_k$ is the video-latent input, $[\cdot]_{\text{ch}}$ concatenates channels, and $P_k$ contains Plücker embeddings of tracked camera rays. Each hand's $H_k$ includes wrist translation/quaternion and 20 joint angles; learned encoders match the patch-token shape. These 3D parameters complement the aligned 2D skeleton branch. Figure 13 shows encoder inputs and visual-task sequences from the deployed 5B model, distinct from the 14B conditioning experiments.

**XR as an authoring interface.** Some systems use XR to direct generation for a separate output medium. VideoCraft uses MR spatial-layer editing to steer concept videos viewed elsewhere; its evaluation concerns authoring control (Li et al. 2025a). Context-aware AR generation recommends objects from scene images using a vision–language model; Say It, See It compares speech-driven generation pipelines within AR (Behravan et al. 2025; Xiu et al. 2025). Their findings concern semantic or aesthetic suitability. Gesture2Prop explores joint speech and hand-gesture conditioning for prop creation in an adjunct demonstration (Yao et al. 2025).

## 5.2 Where operational capabilities are introduced

Operational capabilities often enter the pipeline after visual content is generated. After GLB import, CoGen3D sets scale/orientation heuristically and adds manipulation components for resizing and repositioning (Jiang et al. 2026). Conversion determines whether materials, hierarchy, and physical properties survive delivery. Dreamcrafter associates Gaussian appearance with meshes for interaction and collision; boundary agreement affects selection and contact precision (Vachha et al. 2025). A grab component supplies rigid manipulation, whereas articulation and deformation require further structure.

Part-level interaction requires both segmented geometry and rules for its use. HOICraft uses vision-language-model assistance to recommend parts and hand–object interactions after segmentation, labels, and motion constraints are supplied (Lee et al. 2026). In its 12-designer study, fewer exploration attempts did not yield a statistically significant decision-speed improvement. PartInteractor instead exposes generated part hierarchies for selection, removal, and recombination in XR (Jiang and Li 2026). This supports component editing, while maintaining semantic part identity during later interaction remains an implementation issue.

**Generating behavior through orchestration.** LLMR creates and modifies interactive worlds through generated code, while ImoGenXR combines asset retrieval with generated behaviors in immersive authoring (De La Torre et al. 2024; Limbert and Karnik 2025). Executable behavior comes from orchestration and author control beyond the visual generator's output. GenAssist retrieves models and documentation to generate and check executable scene programs; its experiments use a browser viewer (Srinidhi et al. 2026). CreepyCoCreator studies AI representation and edit presentation through Wizard-of-Oz co-creation. Its object additions and replacements are system-selected and artist-created (Rasch et al. 2025). Its findings concern interaction and expectations within this simulated co-creation workflow.

## 5.3 Authoring and interaction workflows

**Staging decisions before final generation.** Authoring systems expose intermediate content so users can guide generation before final delivery. VRCopilot uses reviewable wireframes in manual, scaffolded, and automatic modes (Zhang et al. 2024c). Dreamcrafter records online-preview annotations for broker-executed generation (Vachha et al. 2025); CoGen3D uses asynchronous jobs (Jiang et al. 2026). MS2Mesh-XR places image selection before reconstruction and import (Tong et al. 2025). These workflows differ in when users commit decisions and how those decisions transfer to the final asset. Section 4.5 distinguishes the corresponding preview, delivery, and display intervals.

**Spatial input and editable representations.** The representation determines what a spatial gesture can change. GaussianShopVR combines immersive selection and drawing with Gaussian-based editing and object generation; DreamSpace propagates panoramic appearance across captured geometry (Shen et al. 2025; Yang et al. 2024a). Generative terrain authoring maps mid-air hand contours to a learned terrain representation (Hu et al. 2024). These interfaces expose different editable units: points, surfaces, objects, or scene regions. Local appearance changes, articulation, and nearby object structure consequently require distinct editing support.

**Situated creation and physical context.** Dream Mesh connects speech-driven generation to mesh import and manipulation in MR (Weng et al. 2024). Roomify transforms captured rooms while preserving approximate object footprints, major contact surfaces, and high-level functional roles. Editing Reality explores in-situ speech, sketching, and direct manipulation, informed by a formative workshop and expert review (Wang et al. 2026e; Weng et al. 2026). FusionProtor uses component-level physical-to-virtual transitions for mixed prototypes and subsequent simulation (Zhang et al. 2025).

**Generation during collaboration.** Generated content can be the shared design object or the setting for collaboration. Thing2Reality explores avatar decoration, collaborative furniture layout, and toy ideation/pitching through shared two-dimensional content and three-dimensional objects (Hu et al. 2025a). Furniture layout emphasizes spatial manipulation, while ideation also uses a two-dimensional whiteboard. SpaceBlender compares blended familiar spaces with Generic3D and Text2Room during

collaborative sticky-note clustering, without manipulating generated room surfaces (Numan et al. 2024). These studies address the awareness, communication, and coordination concerns of collaborative MR (Ens et al. 2019).

## 5.4 VR, AR, and MR integration conditions

**Alignment with physical space.** Tracked input and environmental registration address different problems in XR integration. MS2Mesh-XR supplies drawing, speech, and manipulation interfaces and demonstrates manual resizing and positioning of MR furniture (Tong et al. 2025). Its inspected implementation and use-case sections provide no registration-error, real-object occlusion, or lighting-adaptation measurements. ImaginateAR represents pre-scanned surroundings with labeled bounding boxes and a scene graph. A visual positioning system localizes the mobile device, and the graph guides asset placement (Lee et al. 2025). Sensing and localization supply registration. Reported evaluations concern components, technical performance, and authoring.

**Viewing geometry.** The display pipeline determines the spatial queries supported by VR content. Images, panoramic backgrounds, and meshes rendered from two eye cameras produce different depth behavior even when physical-surface registration is unnecessary. Dreamcrafter's Magic Camera produces a stylized snapshot (Vachha et al. 2025). Extending such content to head translation requires binocular consistency and disocclusion to be checked in the resulting view-generation pipeline.

EPIC addresses binocular panoramic video through left/right view generation, spherical-epipolar-preference refinement, and omnidirectional-stereo preparation (Mandal et al. 2026). Before binocular super-resolution, preparation re-renders both eyes from the left view using depth scaled against the right view. Geometric scores assess stereo, temporal, and cross-time cross-eye residuals on rigid-scene correspondences, excluding independently moving tracks and invalid low-parallax pairs. Refinement improves stereo residuals but worsens cross-time cross-eye residuals on this subset.

## 5.5 Headset demonstrations and user tasks

**Agency and authoring experience.** User studies reveal how integration choices affect control over generated content. VRCopilot reports a 14-participant study and a second with 15 different Quest 2 participants (Zhang et al. 2024c). In the second, bedroom-authoring agency is highest for manual creation, followed by scaffolded and automatic creation. This ordering links authoring structure to perceived control; circulation and window access add practical layout constraints.

Dreamcrafter's seven-person exploratory study uses computer-linked Quest 3 for holiday scenes (Vachha et al. 2025). Participants report greater control from sculpting before stylization, uncertain final appearance, and lost size information in image previews. Low-fidelity mesh previews for prompted objects were added after the study. CoGen3D studies 60 design participants and an independently recruited validation group of 60 in affective VR scenes (Jiang et al. 2026). Satisfaction is higher for generated images than resulting 3D assets.

**Visual action response.** Generated Reality evaluates tracked-hand tasks through blinded judgments of generated recordings and perceived control (Xie et al. 2026b). Hand conditioning improves task accuracy over a camera/text baseline lacking tracked hands. The measured endpoint is perceived and judged visual response. MS2Mesh-XR contributes VR/MR demonstrations and stage timings (Tong et al. 2025).

**Application-level presentation.** ImmerseGen prepares lightweight textured scenes for headset rendering, while CinemaWorld generates film-linked spatial augmentations before playback (Yuan et al. 2026b; Ihara et al. 2026). ShadAR injects generated shaders into monoscopic passthrough augmentation during use. It measures code-generation delay separately from frame-update throughput (Mei et al. 2026).

Heritage visualization combines hosted generation, manual inspection, and Unity/PICO 4 presentation (Liu et al. 2026). Questionnaires and interviews cover four painting-derived scene fragments, nine experts, and 30 non-experts. Experts contribute either on-site VR or remote first-person video judgments. Topology and preparation effort are assessed separately on three AI/manual object pairs.

Han and Jeong compare static panoramas generated with Skybox AI and image processing against SketchUp/Enscape references, using seven Quest 3 experts (Han and Jeong 2026). Image-property and presentation-quality judgments provide device-based evidence for static viewing.

**Table 7**: Catalogue of XR authoring systems and generated-content presentation methods.

**Input:** T text; I image; V video; P head/pose; G geometry/scene context; H hand; S sketch; Sp speech; C controller/touch/manipulation. Input combinations are configuration-specific.
**Eval.:** U user study; T technical test; D demonstration; X exploratory study; F formative study; E expert assessment; R ratings; WoZ Wizard-of-Oz study.
GS Gaussian splats; ODS omnidirectional stereo; HOI human–object interaction. ↗ official project/GitHub.

| # | System / year | Input | Model / component | Artifact | Platform | Task | Eval. | ↗ |
|---|---|---|---|---|---|---|---|---|
| **Spatial authoring and situated creation** | | | | | | | | |
| 1 | **VRCopilot** (2024c) | SpSC | ATISS | Layout+library objects | Quest 2 | Layout | U | ↗ |
| 2 | **Dreamcrafter** (2025) | SpGC | Modular gen. | Images/GS/ meshes | Quest 3+PC | Preview | X | ↗ |
| 3 | **GaussianShopVR** (2025) | GS | 3DGS editing | Edited/new GS | VR | 3D editing | U | ↗ |
| 4 | **DreamSpace** (2024a) | TG | Panoramic diffusion | Textured scene mesh | HMD VR | Stylization | T/D | ↗ |
| 5 | **Generative terrain** (2024) | HS | CGAN | Height-map terrain | Quest 3 | Terrain | U | – |
| 6 | **Roomify** (2026e) | VTI | gpt-image-1/ Tripo2.5 | Stylized environment | Quest 3 | Room edit | U | – |
| 7 | **ImaginateAR** (2025) | TCG | Graph-guided gen. | Placed meshes | Phone AR | Situated | T/U | ↗ |
| 8 | **Context-aware 3D** (2025) | I | LLaVA+3D generator | 3D objects | AR headset | Context | T/R | – |
| 9 | **Dream Mesh** (2024) | Sp | DreamFusion/ ThreeStudio | Imported mesh | Quest Pro | Objects | D | – |
| 10 | **Editing Reality** (2026) | SpSCG | Modular gen. | Edited scene elements | MR | Co-creation | F/E | ↗ |
| 11 | **ImmerseGen** (2026b) | T | GPT-4o /SDXL-ControlNet / layered diff. | Alpha-textured meshes | PICO 4 Ultra /XR2 Gen 2 | Immersive worlds | T/U | ↗ |
| **Object creation, prototyping, and collaboration** | | | | | | | | |
| 12 | **MS2Mesh-XR** (2025) | SpSH | ControlNet1.1/ CRM | OBJ mesh | Quest 3 | Objects | D/T | ↗ |
| 13 | **CoGen3D** (2026) | T | FLUX.1/ Hunyuan3D-2 | GLB mesh | Desktop→ Quest 3 | Co-design | U | – |
| 14 | **Say It, See It** (2025) | Sp | Modular 3D pipelines | 3D objects | Quest 3 | Speech-to-3D | U/T | – |
| 15 | **Gesture2Prop** (2025) | SpH | ControlNet+3D reconstruction | 3D props | Quest 3 | Props | D | – |
| 16 | **FusionProtor** (2025) | ITC | SDXL-1.0/ Rodin | 3D components | Apple Vision Pro | Prototyping | T/U | – |
| 17 | **Thing2Reality** (2025a) | IVS | Image-to-3D | Views/3D Gaussians | Quest 3 | Discussion | U | – |
| 18 | **SpaceBlender** (2024) | I | ControlNet-Layout | Shared scene mesh | Quest 3 | Blending | U | – |
| 19 | **CreepyCoCreator** (2025) | C | Scripted alternatives | Co-created object | Quest Pro | Co-creation | WoZ | – |
| 20 | **PartInteractor** (2026) | SpSIHG | GPT-4o /GPT-Image-1 / OmniPart | Part-structured GLB | Quest 3 +PC +APIs | Part editing | U/T | ↗ |
| **Video and behavior authoring** | | | | | | | | |
| 21 | **VideoCraft** (2025a) | VG | Video-to-video | Concept video | MR | Video edit | U/E | – |
| 22 | **LLMR** (2024) | TG | GPT-4 | Scene code | Desktop/AR/ VR | Worlds | T/U | ↗ |
| 23 | **ImoGenXR** (2025) | SpC | GPT-4o+ retrieval | Models+ behavior scripts | VR | Behavior | U | – |
| 24 | **HOICraft** (2026) | GH | GPT-4o | HOI mappings | Quest 3+PC | HOI | U | – |
| 25 | **GenAssist** (2026) | TG | GPT-4o +doc./ model RAG | Scene / interaction code | ARENA / browser / WebXR | XR programs | T/ U/R | ↗ |

Continued on next page

**Table 7**: Catalogue of XR authoring systems and generated-content presentation methods. (continued).

**Input:** T text; I image; V video; P head/pose; G geometry/scene context; H hand; S sketch; Sp speech; C controller/touch/manipulation. Input combinations are configuration-specific.
**Eval.:** U user study; T technical test; D demonstration; X exploratory study; F formative study; E expert assessment; R ratings; WoZ Wizard-of-Oz study.
GS Gaussian splats; ODS omnidirectional stereo; HOI human–object interaction. ↗ official project/GitHub.

| # | System / year | Input | Model / component | Artifact | Platform | Task | Eval. | ↗ |
|---|---|---|---|---|---|---|---|---|
| **Presentation and action response** | | | | | | | | |
| 26 | **Generated Reality** (2026b) | T P H | 5B video diffusion | Video | Quest 3 | Action response | U/T | ↗ |
| 27 | **EPIC** (2026) | T I | Stereo video diffusion | ODS stereo video | VR-target | Stereo | T | – |
| 28 | **Heritage prototyping** (2026) | I | Hunyuan hosted models | Assets+scene fragments | PICO 4/video | Heritage | T/U | – |
| 29 | **Panoramic stimuli** (2026) | T I | Skybox AI | Static panoramas | Quest 3 | Stimuli | T/E | – |
| 30 | **CinemaWorld** (2026) | V G | Gemini-2.5-pro / gpt-image-1 / InstantMesh | Timed meshes / effects | Quest 3 / Unity | Film augmentation | F/T/ U/E | ↗ |
| 31 | **ShadAR** (2026) | Sp I | GPT-5.2 / YOLO26 / Whisper | HLSL /altered passthrough | Quest 3 + RTX 5090 | Visual augmentation | T/X | – |

## 5.6 Task requirements and model selection

Paradigm-level requirements such as efficiency and consistency narrow the choice of generation route (Xie et al. 2026a). Within a route, selection depends on the exact operation, the state or artifact it changes, the component implementing that change, and the endpoint tested. The following comparisons show how this evidence changes a choice.

**Modify a part, preserve contact, and return.** Consider moving a chair back independently, maintaining the intended contact/collision behavior, and finding the accepted change after leaving and returning. CoGen3D demonstrates whole-object manipulation through Unity components applied to Hunyuan3D-2 assets (Jiang et al. 2026). PartCrafter supplies separate meshes; its geometric scores assess decomposition (Lin et al. 2025). PartInteractor provides direct XR component-editing evidence; HOICraft configures behavior on prepared parts and constraints (Jiang and Li 2026; Lee et al. 2026). These are alternative starting points with different preparation costs. The task requires stable part identity, an implemented update, compatible collision geometry, and stored accepted state. An integrated pipeline would need to combine these components and test the complete sequence; Figure 14 specifies the retention test.

**Explore a generated room with tracked head motion.** A prepared scene with local rendering exposes camera queries independently of later generation. Text2Room supplies a surface for walkthroughs; Holodeck adds addressable scene objects, and HY-World 2.0 supplies a separate collision proxy (Höllein et al. 2023; Yang et al. 2024d; Team HY-World et al. 2026). These structures support geometric queries; binocular viewing requires assessment in the headset configuration. WorldPlay instead generates views from camera/action conditions and retrieved history, with quantitative camera and revisit evidence. Generated Reality supplies a 5B tracked-head/hand deployment and short visual-task evaluation (Sun et al. 2026a; Xie et al. 2026b). These routes permit generated visual responses, with reported timings for generation throughput or decoding delay. The choice trades explicit queryable geometry and local viewpoint updates against ongoing synthesis. Selection depends on supported view changes, disocclusion, stereo behavior, and the complete response path.

**What changes the choice.** Rough composition can use a preview before geometry is complete; precise placement depends on the final boundary and scale, as Dreamcrafter's findings illustrate (Section 5.5). An edit-and-return requirement instead makes the supported update and retained variable decisive. WorldMem and PERSIST address changed content within their respective settings (Xiao et al. 2025; Garcin et al. 2026); those settings must match the requested edit before their persistence evidence transfers. Table 6 then constrains when the required output becomes available.

# 6 Challenges and Future Directions

We propose five matched evaluation designs for transferring model-level evidence to XR operations. Each holds a relevant task or content reference fixed while varying a stage in the pipeline.

## 6.1 Binocular consistency under head motion

Transferring stereo consistency from prescribed views to tracked head motion requires evaluating newly requested binocular views. StereoDiffusion generates stereo image pairs using diffusion, and StereoCrafter converts monocular video to stereoscopic video (Wang et al. 2024a; Zhao et al. 2024). StereoWorld adds geometry-aware monocular-to-stereo generation; Stereo World Model targets camera-guided stereo synthesis (Xing et al. 2026; Sun et al. 2026b). This transfer adds camera control and response timing to fixed-pair and stereo-sequence evaluation.

A paired study could test whether offline consistency predicts binocular viewing quality for the same content. The study would link findings from cross-view, static-stereo, and device-based panorama evaluations (Tang et al. 2023; Tamir et al. 2025; Han and Jeong 2026). Link each asset to its displayed version and record viewing queries, rendering configuration, stereo baseline, depth scale, display settings, and correction/reprojection. Use separate applicable conditions for fixed stereo pairs, panoramic rotation, and translated-view generation. Include any reconstruction or renderer used for translation in the pipeline evaluation.

Within each condition, compare geometric errors and artifacts with task performance and participant judgments. Retain disagreements, such as improved appearance with unchanged binocular error, when interpreting the results. Vary rendering and display settings to separate representation from presentation effects, and validate dynamic content separately.

## 6.2 Consistency between previews and final assets

In preview-based authoring, users accept size, placement, and appearance decisions before final geometry becomes available. Dreamcrafter's selected preview images can condition subsequent 3D generation (Vachha et al. 2025). Evaluation should determine which decisions survive replacement and how the interface or generator can preserve them.

Retain the preview, user operations, and final asset for each request. A presentation-only experiment could replay requests with the same final asset and matched delivery times, counterbalancing preview format and order. A complementary live experiment could randomize preview interfaces across comparable requests while fixing the task and generator configuration. Preserve selected-preview conditioning where used, and record the assets generated from those choices. Replay isolates displayed-proxy effects, whereas live evaluation tests the complete preview-to-final workflow.

Compare three-dimensional proxies through occupied volumes, relative scale, contact regions, and boundaries in common coordinates. For two-dimensional previews, use projected outlines and the placement constraints they convey. Measure constraint satisfaction, revised decisions, and total correction time after replacement. The generator and authoring system take on different responsibilities when preserving bounding structure, transferring edits, or delaying geometry-dependent operations.

## 6.3 Persistent and shared object state

Persistent interaction requires a successful update to survive absence and remain consistent across observers. Different variables are available in WorldPlay's visual history, MultiGen's shared player state, and Holodeck/HY-World 2.0 simulation structures (Sun et al. 2026a; Po et al. 2026; Yang et al. 2024d; Team HY-World et al. 2026). MultiGen's level map remains fixed during rollout, limiting the updates that can be tested. LiveWorld evaluates off-screen events using scripts and VLM judgments. Programmable World Model checks engine-driven entity counts and death events, although event success is not tied to a specified entity (Duan et al. 2026; Huang et al. 2026d). An XR test should identify the supported action, changed variable, and storage component. Choose an object edit, a supported transition, or an unchanged-scene revisit according to the interface.

For a supported edit, modify an object, move until it is out of view, and return through a different route. Report update success before departure, retention among successful updates, and overall completion separately. A matched unmodified revisit provides the reference for testing visual recall. Assess identity, pose, appearance, and action-relevant state separately: a visually closed door need not have closed collision or traversal state. Test the action's expected visible consequence, and compare engine state with generated observations when that state is accessible.

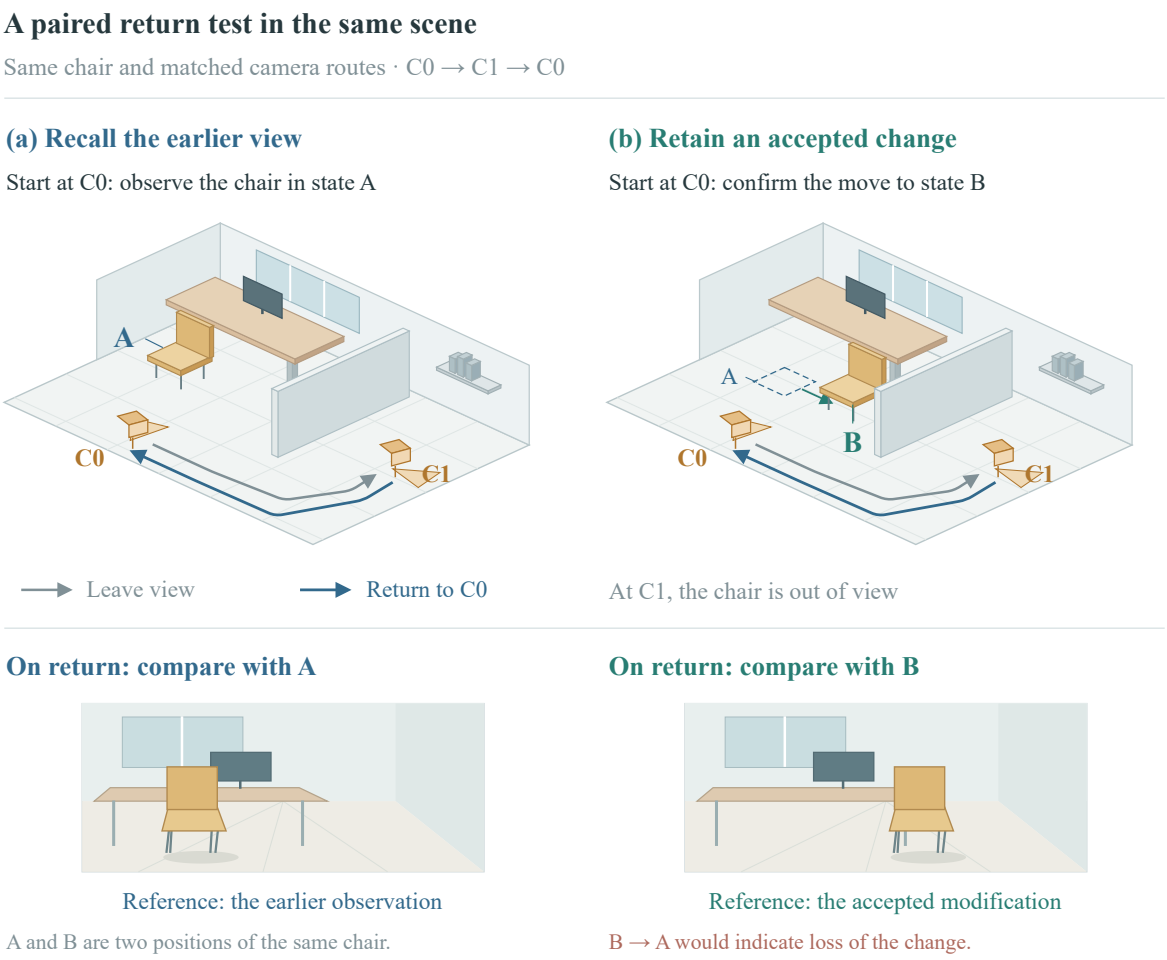



**Fig. 14** Proposed recall and retention tests on matched C0–C1–C0 routes, with the chair hidden at C1. Return references are (a) original position A and (b) accepted edited position B. Update success and retention are scored separately. All scenes are schematic.

Figure 14 pairs the same scene and return viewpoint with references to the earlier observation or accepted modification.

Vary the return interval, route, occlusion, and repeated operations. Trace failures to action interpretation, storage, retrieval, or rendering to identify which retained information the task needs.

A collaborative test would inspect the same accepted update from multiple viewpoints before introducing delayed or conflicting inputs. MultiGen's opponent-presence evaluation and Solaris's two-player Minecraft tasks already test cross-observer behavior (Po et al. 2026; Savva et al. 2026). For a specified update, compare stored state, generated observations, and outcomes. This would distinguish disagreement about the update from a mismatch in its presentation.

Control memory budget and history quality independently (Section 3.3.5); test extended absence separately from short action response.

## 6.4 Response latency and system scheduling

Response scheduling connects model execution to the time when a user sees a useful change. Generated Reality reports generation/decoding delay (Xie et al. 2026b). A complete response trace would also timestamp input receipt, queuing/scheduling, memory retrieval, transfer, rendering, and display. Record applicable loading/import stages, dependencies, and overlap alongside the generation stage.

With display and task fixed, vary input frequency, scene complexity, hardware allocation, and network conditions. Measure time to the first meaningful response and to a usable final result. Also record displayed-content age and the frequency of interrupted or superseded requests. Report both average and tail delays, tracing head-motion updates separately when a local scene keeps rendering during generation.

These traces support task-specific comparisons of synchronous generation, preview-based authoring, and local rendering with asynchronous updates. A scheduling ablation should fix the input trace, model/output settings, hardware, and network while varying scheduling or presentation. Report visual quality and task completion alongside latency, since provisional responses may support fewer operations than completed results.

### 6.5 Physical-space alignment and adaptation

Physical-space adaptation requires generation, placement, and display to preserve constraints from supporting surfaces, room boundaries, foreground occluders, and illumination. ImaginateAR supplies context through a pre-scanned graph and localization pipeline (Lee et al. 2025). The question is how well each constraint survives across these stages and after the physical context changes.

A controlled task could generate and place a constrained asset, then revisit it after viewpoint or environmental changes. Measure support-relative placement error, reference-relative scale error, occlusion-order violations, and task success. Record whether spatial inputs come from manual specification, sensing, or inference. Fixing the asset while varying alignment can help isolate registration effects. Fixing registration while varying the asset tests preservation of the required geometry.

Vary illumination and partial occlusion to test adaptation to environmental changes (Lin et al. 2026c,a; Huang et al. 2025a; Lin et al. 2026b; Malyugina et al. 2025). Evaluate geometric fit, appearance adaptation, and interaction outcomes separately to locate failures in generation, alignment, or presentation.

## 7 Conclusion

This review connects visual generative models to XR use by following the path from generated output to a tested operation. Common mathematical descriptions explain how conditioning inputs, representations, and update mechanisms differ across assets, scenes, and dynamic worlds. The evidence profiles and deployment traces then identify the components and configurations supporting the reported capabilities.

The comparisons reveal a recurring distinction between producing content and maintaining the structures needed to operate on it. Separate part meshes expose a possible editing unit; manipulation and retained edits additionally depend on the update and storage components supplied by the complete configuration. Prepared scenes permit local viewpoint queries, whereas observation generators synthesize new views from actions and history. These routes place different demands on geometry, memory, and response scheduling. Accordingly, selection depends on the editable unit, the state that must survive, and the time at which a useful response becomes visible.

The five proposed matched evaluations are designed to test how model-level evidence transfers through viewing, authoring, state updates, scheduling, and physical-space adaptation. They separate successful updates from later retention and component speed from the response experienced by a user. Progress should therefore be assessed through the viewing, editing, and interaction tasks supported by a complete XR configuration.

## A Review records and evaluation details

**Scope and selection.** This is a critical narrative review based on purposive retrieval, primary-source checks, and citation-led follow-up. Selection sought complementary generation mechanisms, output structures, and documented XR operations. Foundational reconstruction, rendering, and learning papers supply technical context; they are distinguished from the 204 catalogued methods/systems and 37 resources. The final update used targeted retrieval through 23 October 2026. Search-result and full-exclusion counts were not retained.

**Detailed evidence subset.** The 31 coded methods and systems form a deliberately selected explanatory subset spanning assets, scenes, dynamics, and XR integration. Eight dimensions yield 248 evidence cells; three benchmark records and nine neighboring-survey records provide context but no additional cells. Coding records the named property, inspected source, source location, and qualification associated with each code. Source material includes technical papers, selected official implementation files, and the Genie 3 official description. Static code inspection documents export paths; execution and device use require separate runtime evidence. Codes were developed through author-side, AI-assisted extraction and source checks. Independent duplicate human screening and intercoder agreement are not documented. Later catalogue additions, including WorldMem, PERSIST, and Solaris, inform the task synthesis and agenda outside this coded subset. The map supports case comparisons; field-wide capability prevalence lies outside this purposive sample's scope.

**Traceability and interpretation.** Evidence supplement E includes a browsable record index and a catalogue crosswalk. Its E/A01-style IDs distinguish evidence records from catalogue IDs. The accompanying `evidence-matrix.csv` links each cell to its note, source, and location; `extraction-records.json` retains source versions and inspected configurations. `search-log.txt` records the retained queries and follow-ups, and the quantitative evidence files record source protocols and result strings. Codes describe evidence for a specified property and configuration; comparisons therefore use the associated cell notes. Quantitative blocks retain their own inputs, references, workloads, and precision, with no pooled ranking. Technical editions are recorded at the point of extraction: Table 3A uses TRELLIS arXiv v1, whereas WorldPlay's coded profile uses arXiv v2 and Table 5 uses its PMLR edition.

Table 8 lists the experimental subsets and workloads underlying the quantitative comparisons in Section 4.3.

**Table 8** Data resources and workloads used in the quantitative comparisons.

T: training resource; E: evaluation context. Counts reflect paper-specific curation; evaluation rows correspond to Tables 3–5.

| ID | Dataset / context | Role | Scale | Unit | Used by / source |
|---|---|---|---|---|---|
| **Training resources** | | | | | |
| T1 | TRELLIS curated union | Asset training | 500,777 | 3D objects | TRELLIS (Xiang et al. 2025) |
| T2 | Objaverse / ShapeNet / ABO, part-labeled | Part training | ≈50,000 / ≈300,000 | objects / parts | PartCrafter (Lin et al. 2025) |
| T3 | RealEstate / DL3DV / UE mixture | Video training | >100,000 | curated clips | Voyager (Huang et al. 2025c) |
| T4 | Sekai / DL3DV / UE / game mixture | World training | ≈320,000 | 30–40 s clips | WorldPlay (Sun et al. 2026a) |
| **Evaluation protocols** | | | | | |
| E1 | Toys4k subset | Image-to-3D | 1,250 | test objects | TRELLIS (Xiang et al. 2025) |
| E2 | Objaverse / ShapeNet / ABO[a] | Part generation | ≈2,000[a] | overall test samples | PartCrafter (Lin et al. 2025) |
| E3 | Text2Room room prompts[b] | Room quality | 60 / 61 | views per room / raters | Text2Room (Höllein et al. 2023) |
| E4 | RealEstate10K test subset | Novel views | 150 | test clips | Voyager (Huang et al. 2025c) |
| E5 | WorldScore | World evaluation | 3,000 | specifications[c] | WorldScore (Duan et al. 2025) |
| E6 | Consistent4D subset[d] | Video-to-4D | 7 × 32 | videos × frames | DG4D (Ren et al. 2023) |
| E7 | DL3DV / game / AI-image test suite | Control / revisit | 600 | test cases | WorldPlay (Sun et al. 2026a) |
| E8 | ViZDoom multiplayer context[e] | Opponent presence | 1 / 2 | map / test players | MultiGen (Po et al. 2026) |

[a] About 2K overall; per-dataset counts and object/scene allocation unspecified. T2 excludes the separate 30K monolithic regularization pool. [b] Room count NR; 61 raters, 1,098 datapoints. [c] 2,000 static and 1,000 dynamic specifications; Table 4(c) uses static scores. [d] Four novel evaluation views/video; DG4D-Fast uses 16 input frames. [e] One map, two test players; test-trajectory count NR. The >10-million-frame training collection is separate. T4 counts curated clips; E7 source-specific test counts are NR.

# Declarations

## Funding

No funding was received for this work.

## Author Contributions

G.H. was responsible for the conception and design of the review, literature curation and synthesis, development of the analytical framework, preparation of figures and tables, and drafting and revision of the manuscript. G.H. takes responsibility for the manuscript's content.

## Data Availability

The review-level data, including catalogue metadata, XR capability annotations, source-linked extraction records, and source-reported quantitative comparison data, are provided in the accompanying Evidence supplement E. The underlying publications are cited in the manuscript and linked in the supplementary records.

## Use of Generative AI Tools

OpenAI's ChatGPT and Codex tools assisted literature retrieval and source inspection, evidence organization and comparison, manuscript drafting and revision, figure and table preparation, and code generation for document preparation and LaTeX compilation checks. The author takes responsibility for the accuracy, interpretation, citations, and final content of the manuscript.

## Conflicts of Interest

The author declares no conflicts of interest.

## References


Achlioptas P, Diamanti O, Mitliagkas I, Guibas L (2018) Learning Representations and Generative Models for 3D Point Clouds. In: ICML, pp 40–49

Akeley K, Watt SJ, Girshick AR, Banks MS (2004) A stereo display prototype with multiple focal distances. ACM Trans Graph 23(3):804–813. https://doi.org/10.1145/1015706.1015804

Alonso E, Jelley A, Micheli V, Kanervisto A, Storkey A, Pearce T, Fleuret F (2024) Diffusion for World Modeling: Visual Details Matter in Atari. In: NeurIPS

Armeni I, Sener O, Zamir AR, Jiang H, Brilakis I, Fischer M, Savarese S (2016) 3D Semantic Parsing of Large-Scale Indoor Spaces. In: CVPR, pp 1534–1543, https://doi.org/10.1109/cvpr.2016.170

Assran M, Bardes A, Fan D, Garrido Q, Howes R, Komeili M, Muckley M, Rizvi A, Roberts C, Sinha K, Zholus A, Arnaud S, Gejji A, Martin A, Hogan FR, Dugas D, Bojanowski P, Khalidov V, Labatut P, Massa F, Szafraniec M, Krishnakumar K, Li Y, Ma X, Chandar S, Meier F, LeCun Y, Rabbat M, Ballas N (2025) V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. arXiv preprint arXiv:2506.09985

Azuma R, Bishop G (1994) Improving static and dynamic registration in an optical see-through HMD. In: SIGGRAPH, pp 197–204, https://doi.org/10.1145/192161.192199

Azuma R, Baillot Y, Behringer R, Feiner S, Julier S, MacIntyre B (2001) Recent advances in augmented reality. IEEE Comput Graph Appl 21(6):34–47. https://doi.org/10.1109/38.963459

Azuma RT (1997) A Survey of Augmented Reality. Presence (Camb) 6(4):355–385. https://doi.org/10.1162/pres.1997.6.4.355

Bahmani S, Skorokhodov I, Rong V, Wetzstein G, Guibas L, Wonka P, Tulyakov S, Park JJ, Tagliasacchi A, Lindell DB (2024) 4D-fy: Text-to-4D Generation Using Hybrid Score Distillation Sampling. In: CVPR, pp 7996–8006, https://doi.org/10.1109/cvpr52733.2024.00764

Banerjee P, Shkodrani S, Moulon P, Hampali S, Han S, Zhang F, Zhang L, Fountain J, Miller E, Basol S, Newcombe R, Wang R, Engel JJ, Hodan T (2025) HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. In: CVPR, pp 7061–7071, https://doi.org/10.1109/CVPR52734.2025.00662

Banks MS, Read JCA, Allison RS, Watt SJ (2012) Stereoscopy and the Human Visual System. SMPTE Motion Imaging Journal 121(4):24–43. https://doi.org/10.5594/j18173

Bar-Tal O, Chefer H, Tov O, Herrmann C, Paiss R, Zada S, Ephrat A, Hur J, Liu G, Raj A, Li Y, Rubinstein M, Michaeli T, Wang O, Sun D, Dekel T, Mosseri I (2024) Lumiere: A Space-Time Diffusion Model for Video Generation. In: SIGGRAPH Asia Conference Papers, pp 1–11, https://doi.org/10.1145/3680528.3687614

Barron JT, Mildenhall B, Tancik M, Hedman P, Martin-Brualla R, Srinivasan PP (2021) Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In: ICCV, https://doi.org/10.1109/iccv48922.2021.00580

Barron JT, Mildenhall B, Verbin D, Srinivasan PP, Hedman P (2022) Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In: CVPR, https://doi.org/10.1109/cvpr52688.2022.00539

Barron JT, Mildenhall B, Verbin D, Srinivasan PP, Hedman P (2023) Zip-NeRF: Anti-Aliased Grid-Based Neural Radiance Fields. In: ICCV, https://doi.org/10.1109/iccv51070.2023.01804

Baruch G, Chen Z, Dehghan A, Dimry T, Feigin Y, Fu P, Gebauer T, Joffe B, Kurz D, Schwartz A, Shulman E (2021) ARKitScenes: A Diverse Real-World Dataset For 3D Indoor Scene Understanding Using Mobile RGB-D Data. In: NeurIPS Datasets and Benchmarks Track

Behravan M, Matković K, Gračanin D (2025) Generative AI for Context-Aware 3D Object Creation Using Vision-Language Models in Augmented Reality. In: IEEE AIxVR, https://doi.org/10.1109/AIxVR63409.2025.00018

Billinghurst M, Kato H, Poupyrev I (2001) The MagicBook: a transitional AR interface. Comput Graph 25(5):745–753. https://doi.org/10.1016/s0097-8493(01)00117-0

Blattmann A, Dockhorn T, Kulal S, Mendelevitch D, Kilian M, Lorenz D, Levi Y, English Z, Voleti V, Letts A, Jampani V, Rombach R (2023a) Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. arXiv preprint arXiv:2311.15127

Blattmann A, Rombach R, Ling H, Dockhorn T, Kim SW, Fidler S, Kreis K (2023b) Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. In: CVPR, https://doi.org/10.1109/cvpr52729.2023.02161

Bokhovkin A, Meng Q, Tulsiani S, Dai A (2025) SceneFactor: Factored Latent 3D Diffusion for Controllable 3D Scene Generation. In: CVPR, https://doi.org/10.1109/cvpr52734.2025.00067

Bowman DA, Hodges LF (1997) An evaluation of techniques for grabbing and manipulating remote objects in immersive virtual environments. In: SI3D, pp 35–ff., https://doi.org/10.1145/253284.253301

Bruce J, Dennis MD, Edwards A, Parker-Holder J, Shi Y, Hughes E, Lai M, Mavalankar A, Steigerwald R, Apps C, Aytar Y, Bechtle SME, Behbahani F, Chan SC, Heess N, Gonzalez L, Osindero S, Ozair S, Reed S, Zhang J, Zolna K, Clune J, Freitas ND, Singh S, Rocktäschel T (2024) Genie: Generative Interactive Environments. In: ICML

Caldeira H, Wang H, Huang G, Cai S, Fu R, Anantrasirichai N (2026) Geometry beneath the Waves: Dense Priors for Sparse-View Underwater 3D Gaussian Splatting. In: SIGGRAPH Asia 2026 Posters, accepted; technical version inspected: arXiv v2, 30 September 2026, arXiv:2609.18737v2

Campos C, Elvira R, Rodriguez JJG, M. Montiel JM, D. Tardos J (2021) ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM. IEEE Trans Robot 37(6):1874–1890. https://doi.org/10.1109/tro.2021.3075644

Chan ER, Lin CZ, Chan MA, Nagano K, Pan B, Mello SD, Gallo O, Guibas LJ, Tremblay J, Khamis S, Karras T, Wetzstein G (2022) Efficient Geometry-Aware 3D Generative Adversarial Networks. In: CVPR, https://doi.org/10.1109/cvpr52688.2022.01565

Chang A, Dai A, Funkhouser T, Halber M, Nießner M, Savva M, Song S, Zeng A, Zhang Y (2017) Matterport3D: Learning from RGB-D Data in Indoor Environments. In: 3DV, https://doi.org/10.1109/3dv.2017.00081

Chang AX, Funkhouser T, Guibas L, Hanrahan P, Huang Q, Li Z, Savarese S, Savva M, Song S, Su H, Xiao J, Yi L, Yu F (2015) ShapeNet: An Information-Rich 3D Model Repository. ArXiv preprint, arXiv:1512.03012

Chang H, Zhang H, Jiang L, Liu C, Freeman WT (2022) MaskGIT: Masked Generative Image Transformer. In: CVPR, pp 11315–11325, https://doi.org/10.1109/cvpr52688.2022.01103

Chen A, Xu Z, Geiger A, Yu J, Su H (2022) TensoRF: Tensorial Radiance Fields. In: ECCV, https://doi.org/10.1007/978-3-031-19824-3_20

Chen B, Martí Monsó D, Du Y, Simchowitz M, Tedrake R, Sitzmann V (2024a) Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion. In: NeurIPS, arXiv:2407.01392

Chen DZ, Siddiqui Y, Lee HY, Tulyakov S, Nießner M (2023a) Text2Tex: Text-driven Texture Synthesis via Diffusion Models. In: ICCV, https://doi.org/10.1109/iccv51070.2023.01701

Chen DZ, Li H, Lee HY, Tulyakov S, Nießner M (2024b) SceneTex: High-Quality Texture Synthesis for Indoor Scenes via Diffusion Priors. In: CVPR, https://doi.org/10.1109/cvpr52733.2024.01992

Chen H, Zhang Y, Cun X, Xia M, Wang X, Weng C, Shan Y (2024c) VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models. In: CVPR, pp 7310–7320, https://doi.org/10.1109/cvpr52733.2024.00698

Chen L, Zhou Z, Zhao M, Wang Y, Zhang G, Huang W, Sun H, Wen JR, Li C (2025a) FlexWorld: Progressively Expanding 3D Scenes for Flexible-View Exploration. In: NeurIPS, https://doi.org/10.52202/085713-5732

Chen R, Chen Y, Jiao N, Jia K (2023b) Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. In: ICCV, https://doi.org/10.1109/iccv51070.2023.02033

Chen T, Hao K, Chen Y, Cheng Z, Xie R, Song L, Huang H, Zhang C, Li X (2026a) Full-4D: Generating Full-Scope 4D Scenes from a Single-View Video. arXiv preprint https://doi.org/10.48550/arXiv.2605.25500, Version v1

Chen Y, Chen Z, Zhang C, Wang F, Yang X, Wang Y, Cai Z, Yang L, Liu H, Lin G (2024d) GaussianEditor: Swift and Controllable 3D Editing with Gaussian Splatting. In: CVPR, https://doi.org/10.1109/cvpr52733.2024.02029

Chen Y, He T, Huang D, Ye W, Chen S, Tang J, Cai Z, Yang L, Yu G, Lin G, Zhang C (2025b) MeshAnything: Artist-Created Mesh Generation with Autoregressive Transformers. In: ICLR

Chen Z, Wang G, Liu Z (2023c) SceneDreamer: Unbounded 3D Scene Generation From 2D Image Collections. IEEE Trans Pattern Anal Mach Intell 45(12):15562–15576. https://doi.org/10.1109/TPAMI.2023.3321857

Chen Z, Tang J, Dong Y, Cao Z, Hong F, Lan Y, Wang T, Xie H, Wu T, Saito S, Pan L, Lin D, Liu Z (2025c) 3DTopia-XL: Scaling High-quality 3D Asset Generation via Primitive Diffusion. In: CVPR, https://doi.org/10.1109/cvpr52734.2025.02475

Chen Z, Wang L, Shen G, Yan D, Yang S, Xu T, Du Y, Wang W, Gui T, Huang L, Chen Y (2026b) ReWorld: An Interactive World Model with Long-Horizon Memory. ArXiv preprint, arXiv:2608.23565

Cheng YC, Lee HY, Tulyakov S, Schwing AG, Gui LY (2023) SDFusion: Multimodal 3D Shape Completion, Reconstruction, and Generation. In: CVPR, https://doi.org/10.1109/cvpr52729.2023.00433

Chung J, Lee S, Nam H, Lee J, Lee KM (2025) LucidDreamer: Domain-Free Generation of 3D Gaussian Splatting Scenes. IEEE Trans Vis Comput Graph 31(12). https://doi.org/10.1109/TVCG.2025.3611489

Collins J, Goel S, Deng K, Luthra A, Xu L, Gundogdu E, Zhang X, Yago Vicente TF, Dideriksen T, Arora H, Guillaumin M, Malik J (2022) ABO: Dataset and Benchmarks for Real-World 3D Object Understanding. In: CVPR, https://doi.org/10.1109/cvpr52688.2022.02045

Cook RL, Torrance KE (1982) A Reflectance Model for Computer Graphics. ACM Trans Graph 1(1):7–24. https://doi.org/10.1145/357290.357293

Cummings JJ, Bailenson JN (2016) How Immersive Is Enough? A Meta-Analysis of the Effect of Immersive Technology on User Presence. Media Psychol 19(2):272–309. https://doi.org/10.1080/15213269.2015.1015740

Curless B, Levoy M (1996) A volumetric method for building complex models from range images. In: SIGGRAPH, pp 303–312, https://doi.org/10.1145/237170.237269

Dai A, Chang AX, Savva M, Halber M, Funkhouser T, Niessner M (2017) ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In: CVPR, pp 5828–5839, https://doi.org/10.1109/cvpr.2017.261

De La Torre F, Fang CM, Huang H, Banburski-Fahey A, Fernandez JA, Lanier J (2024) LLMR: Real-time Prompting of Interactive Worlds using Large Language Models. In: CHI, https://doi.org/10.1145/3613904.3642579

Deitke M, Han W, Herrasti A, Kembhavi A, Kolve E, Mottaghi R, Salvador J, Schwenk D, VanderBilt E, Wallingford M, Weihs L, Yatskar M, Farhadi A (2020) RoboTHOR: An Open Simulation-to-Real Embodied AI Platform. In: CVPR, pp 3164–3174, https://doi.org/10.1109/cvpr42600.2020.00323

Deitke M, VanderBilt E, Herrasti A, Weihs L, Ehsani K, Salvador J, Han W, Kolve E, Kembhavi A, Mottaghi R (2022) ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. In: NeurIPS, pp 5982–5994

Deitke M, Liu R, Wallingford M, Ngo H, Michel O, Kusupati A, Fan A, Laforte C, Voleti V, Gadre SY, VanderBilt E, Kembhavi A, Vondrick C, Gkioxari G, Ehsani K, Schmidt L, Farhadi A (2023a) Objaverse-XL: A Universe of 10M+ 3D Objects. In: NeurIPS Datasets and Benchmarks Track, https://doi.org/10.52202/075280-1554

Deitke M, Schwenk D, Salvador J, Weihs L, Michel O, VanderBilt E, Schmidt L, Ehsani K, Kembhavi A, Farhadi A (2023b) Objaverse: A Universe of Annotated 3D Objects. In: CVPR, https://doi.org/10.1109/cvpr52729.2023.01263

Deng H, Pan T, Diao H, Luo Z, Cui Y, Lu H, Shan S, Qi Y, Wang X (2025) Autoregressive Video Generation without Vector Quantization. In: ICLR, arXiv:2412.14169

Dhamo H, Manhardt F, Navab N, Tombari F (2021) Graph-to-3D: End-to-End Generation and Manipulation of 3D Scenes Using Scene Graphs. In: ICCV, https://doi.org/10.1109/iccv48922.2021.01604

Dhariwal P, Nichol A (2021) Diffusion Models Beat GANs on Image Synthesis. In: NeurIPS, pp 8780–8794

Ding L, Dong S, Li Y, Gao C, Chen X, Han R, Kuang Y, Zhang H, Huang B, Huang Z, Wang Z, Xu D, Xue T (2026) FullPart: Generating each 3D Part at Full Resolution. In: ICLR

Dinh L, Sohl-Dickstein J, Bengio S (2017) Density estimation using Real NVP. In: ICLR, arXiv:1605.08803

Downs L, Francis A, Koenig N, Kinman B, Hickman R, Reymann K, McHugh TB, Vanhoucke V (2022) Google Scanned Objects: A High-Quality Dataset of 3D Scanned Household Items. In: IEEE ICRA, https://doi.org/10.1109/icra46639.2022.9811809

Duan H, Yu HX, Chen S, Fei-Fei L, Wu J (2025) WorldScore: A Unified Evaluation Benchmark for World Generation. In: ICCV, pp 27713–27724, https://doi.org/10.1109/iccv51701.2025.02573

Duan Z, Xia J, Zhang Z, Zhang W, Zhou G, Gou C, He Y, Chen F, Zhang X, Liu L (2026) LiveWorld: Simulating Out-of-Sight Dynamics in Generative Video World Models. arXiv preprint https://doi.org/10.48550/arXiv.2603.07145, Version v2

Engstler P, Shtedritski A, Laina I, Rupprecht C, Vedaldi A (2025a) SynCity: Training-Free Generation of 3D Worlds. In: ICCV, https://doi.org/10.1109/iccv51701.2025.02561

Engstler P, Vedaldi A, Laina I, Rupprecht C (2025b) Invisible Stitch: Generating Smooth 3D Scenes with Depth Inpainting. In: 3DV, pp 457–468, https://doi.org/10.1109/3DV66043.2025.00047

Ens B, Lanir J, Tang A, Bateman S, Lee G, Piumsomboon T, Billinghurst M (2019) Revisiting collaboration through mixed reality: The evolution of groupware. Int J Hum Comput Stud 131:81–98. https://doi.org/10.1016/j.ijhcs.2019.05.011

Esser P, Rombach R, Ommer B (2021) Taming Transformers for High-Resolution Image Synthesis. In: CVPR, pp 12873–12883, https://doi.org/10.1109/cvpr46437.2021.01268

Esser P, Kulal S, Blattmann A, Entezari R, Müller J, Saini H, Levi Y, Lorenz D, Sauer A, Boesel F, Podell D, Dockhorn T, English Z, Lacey K, Goodwin A, Marek Y, Rombach R (2024) Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In: ICML, pp 12606–12633

Fang C, Dong Y, Luo K, Hu X, Shrestha R, Tan P (2025) Ctrl-Room: Controllable Text-to-3D Room Meshes Generation with Layout Constraints. In: 3DV, pp 692–701, https://doi.org/10.1109/3DV66043.2025.00069

Feng W, Zhu W, Fu TJ, Jampani V, Akula A, He X, Basu S, Wang XE, Wang WY (2023) LayoutGPT: Compositional Visual Planning and Generation with Large Language Models. In: NeurIPS, https://doi.org/10.52202/075280-0802

Fisher M, Ritchie D, Savva M, Funkhouser T, Hanrahan P (2012) Example-based synthesis of 3D object arrangements. ACM Trans Graph 31(6):1–11. https://doi.org/10.1145/2366145.2366154

Fridman R, Abecasis A, Kasten Y, Dekel T (2023) SceneScape: Text-Driven Consistent Scene Generation. In: NeurIPS, vol 36. Curran Associates, Inc., pp 39897–39914, https://doi.org/10.52202/075280-1734

Fridovich-Keil S, Yu A, Tancik M, Chen Q, Recht B, Kanazawa A (2022) Plenoxels: Radiance Fields Without Neural Networks. In: CVPR, https://doi.org/10.1109/cvpr52688.2022.00542

Fridovich-Keil S, Meanti G, Warburg FR, Recht B, Kanazawa A (2023) K-Planes: Explicit Radiance Fields in Space, Time, and Appearance. In: CVPR, https://doi.org/10.1109/cvpr52729.2023.01201

Friston S, Steed A (2014) Measuring Latency in Virtual Environments. IEEE Trans Vis Comput Graph 20(4):616–625. https://doi.org/10.1109/tvcg.2014.30

Fu H, Cai B, Gao L, Zhang LX, Wang J, Li C, Zeng Q, Sun C, Jia R, Zhao B, Zhang H (2021) 3D-FRONT: 3D Furnished Rooms With layOuts and semaNTics. In: ICCV, https://doi.org/10.1109/iccv48922.2021.01075

Gao J, Shen T, Wang Z, Chen W, Yin K, Li D, Litany O, Gojcic Z, Fidler S (2022) GET3D: A Generative Model of High Quality 3D Textured Shapes Learned from Images. In: NeurIPS

Garcin S, Walker T, Mcdonagh S, Pearce T, Bilen H, He T, Wang K, Bian J (2026) Beyond Pixel Histories: World Models with Persistent 3D State. In: ICML

Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative Adversarial Nets. In: NIPS

Grauman K, Westbury A, Byrne E, Chavis Z, Furnari A, Girdhar R, Hamburger J, Jiang H, Liu M, Liu X, Martin M, Nagarajan T, Radosavovic I, Ramakrishnan SK, Ryan F, Sharma J, Wray M, Xu M, Xu EZ, Zhao C, Bansal S, Batra D, Cartillier V, Crane S, Do T, Doulaty M, Erapalli A, Feichtenhofer C, Fragomeni A, Fu Q, Gebreselasie A, González C, Hillis J, Huang X, Huang Y, Jia W, Khoo W, Kolář J, Kottur S, Kumar A, Landini F, Li C, Li Y, Li Z, Mangalam K, Modhugu R, Munro J, Murrell T, Nishiyasu T, Price W, Ruiz P, Ramazanova M, Sari L, Somasundaram K, Southerland A, Sugano Y, Tao R, Vo M, Wang Y, Wu X, Yagi T, Zhao Z, Zhu Y, Arbeláez P, Crandall D, Damen D, Farinella GM, Fuegen C, Ghanem B, Ithapu VK, Jawahar CV, Joo H, Kitani K, Li H, Newcombe R, Oliva A, Park HS, Rehg JM, Sato Y, Shi J, Shou MZ, Torralba A, Torresani L, Yan M, Malik J (2022) Ego4D: Around the World in 3,000 Hours of Egocentric Video. In: CVPR, pp 18995–19012, https://doi.org/10.1109/cvpr52688.2022.01842

Grauman K, Westbury A, Torresani L, Kitani K, Malik J, Afouras T, Ashutosh K, Baiyya V, Bansal S, Boote B, Byrne E, Chavis Z, Chen J, Cheng F, Chu FJ, Crane S, Dasgupta A, Dong J, Escobar M, Forigua C, Gebreselasie A, Haresh S, Huang J, Islam MM, Jain S, Khirodkar R, Kukreja D, Liang KJ, Liu JW, Majumder S, Mao Y, Martin M, Mavroudi E, Nagarajan T, Ragusa F, Ramakrishnan SK, Seminara L, Somayazulu A, Song Y, Su S, Xue Z, Zhang E, Zhang J, Castillo A, Chen C, Fu X, Furuta R, Gonzalez C, Gupta P, Hu J, Huang Y, Huang Y, Khoo W, Kumar A, Kuo R, Lakhavani S, Liu M, Luo M, Luo Z, Meredith B, Miller A, Oguntola O, Pan X, Peng P, Pramanick S, Ramazanova M, Ryan F, Shan W, Somasundaram K, Song C, Southerland A, Tateno M, Wang H, Wang Y, Yagi T,

Yan M, Yang X, Yu Z, Zha SC, Zhao C, Zhao Z, Zhu Z, Zhuo J, Arbelaez P, Bertasius G, Damen D, Engel J, Farinella GM, Furnari A, Ghanem B, Hoffman J, Jawahar C, Newcombe R, Park HS, Rehg JM, Sato Y, Savva M, Shi J, Shou MZ, Wray M (2024) Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In: CVPR, pp 19383–19400, https://doi.org/10.1109/cvpr52733.2024.01834

Groueix T, Fisher M, Kim VG, Russell BC, Aubry M (2018) A Papier-Mâché Approach to Learning 3D Surface Generation. In: CVPR, https://doi.org/10.1109/cvpr.2018.00030

Gu D, Lin J, Liu G, Liu J, Ma S, Shi G, Wang J, Wang Q, Wu Q, Xu Z, Yi X, Yu Z, Zhang J, Zheng Z, Zhu Y, Chen R, Guo H, Guo X, Han M, Han X, Li X, Liang Y, Lou W, Lu J, Luo G, Qin M, Wang S, Wang Y (2026) Seed3D 2.0: Advancing High-Fidelity Simulation-Ready 3D Content Generation. https://doi.org/10.48550/arXiv.2605.13862

Gu J, Xiang F, Li X, Ling Z, Liu X, Mu T, Tang Y, Tao S, Wei X, Yao Y, Yuan X, Xie P, Huang Z, Chen R, Su H (2023) ManiSkill2: A Unified Benchmark for Generalizable Manipulation Skills. In: ICLR

Guo Y, Yang C, Rao A, Liang Z, Wang Y, Qiao Y, Agrawala M, Lin D, Dai B (2024) AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In: ICLR, arXiv:2307.04725

Gupta K, Joshi G, Chen YR, Jain S, Mehta I, Chandraker M (2026) Reflecting Process Expertise in Procedural Material Generation. In: Computer Vision – ECCV 2026, Lecture Notes in Computer Science, vol 17046. Springer, pp 318–336, https://doi.org/10.1007/978-3-032-37281-9_19

Guédon A, Lepetit V (2024) SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering. In: CVPR, https://doi.org/10.1109/cvpr52733.2024.00512

Hafner D, Lillicrap T, Fischer I, Villegas R, Ha D, Lee H, Davidson J (2019) Learning Latent Dynamics for Planning from Pixels. In: ICML, arXiv:1811.04551

Hafner D, Lillicrap T, Ba J, Norouzi M (2020) Dream to Control: Learning Behaviors by Latent Imagination. In: ICLR, arXiv:1912.01603

Hafner D, Pasukonis J, Ba J, Lillicrap T (2025) Mastering diverse control tasks through world models. Nature 640:647–653. https://doi.org/10.1038/s41586-025-08744-2

Han Y, Jeong J (2026) Leveraging generative AI for high-fidelity 360° spatial images: Methodological validation for use as experimental stimuli. Buildings 16(9):1679. https://doi.org/10.3390/buildings16091679

Haque A, Tancik M, Efros AA, Holynski A, Kanazawa A (2023) Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions. In: ICCV, https://doi.org/10.1109/iccv51070.2023.01808

Hart JC (1996) Sphere tracing: a geometric method for the antialiased ray tracing of implicit surfaces. Vis Comput 12(10):527–545. https://doi.org/10.1007/s003710050084

Hart SG, Staveland LE (1988) Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: Hancock PA, Meshkati N (eds) Human Mental Workload, Advances in Psychology, vol 52. North-Holland, p 139–183, https://doi.org/10.1016/s0166-4115(08)62386-9

He H, Xu Y, Guo Y, Wetzstein G, DAI B, Li H, Yang C (2025a) CameraCtrl: Enabling Camera Control for Video Diffusion Models. In: ICLR

He X, Peng C, Liu Z, Wang B, Zhang Y, Cui Q, Kang F, Jiang B, An M, Ren Y, Xu B, Guo HX, Gong K, Wu C, Li W, Song X, Liu Y, Li E, Zhou Y (2025b) Matrix-game 2.0: An open-source real-time and streaming interactive world model. https://doi.org/10.48550/arXiv.2508.13009, Version v2, revised 24 November 2025

Henschel R, Khachatryan L, Poghosyan H, Hayrapetyan D, Tadevosyan V, Wang Z, Navasardyan S, Shi H (2025) StreamingT2V: Consistent, Dynamic, and Extendable Long Video Generation from Text. In: CVPR, pp 2568–2577, https://doi.org/10.1109/cvpr52734.2025.00245

Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: NIPS

Ho J, Salimans T (2022) Classifier-Free Diffusion Guidance. arXiv preprint arXiv:2207.12598

Ho J, Jain A, Abbeel P (2020) Denoising Diffusion Probabilistic Models. In: NeurIPS, arXiv:2006.11239

Ho J, Chan W, Saharia C, Whang J, Gao R, Gritsenko A, Kingma DP, Poole B, Norouzi M, Fleet DJ, Salimans T (2022a) Imagen Video: High Definition Video Generation with Diffusion Models. arXiv preprint arXiv:2210.02303

Ho J, Salimans T, Gritsenko A, Chan W, Norouzi M, Fleet DJ (2022b) Video Diffusion Models. In: NeurIPS, pp 8633–8646, https://doi.org/10.52202/068431-0628

Hoffman DM, Girshick AR, Akeley K, Banks MS (2008) Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. J Vis 8(3):33. https://doi.org/10.1167/8.3.33

Höllein L, Cao A, Owens A, Johnson J, Nießner M (2023) Text2Room: Extracting Textured 3D Meshes from 2D Text-to-Image Models. In: ICCV, pp 7909–7920, https://doi.org/10.1109/iccv51070.2023.00727

Holloway RL (1997) Registration Error Analysis for Augmented Reality. Presence (Camb) 6(4):413–432. https://doi.org/10.1162/pres.1997.6.4.413

Hong W, Ding M, Zheng W, Liu X, Tang J (2023) CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. In: ICLR, arXiv:2205.15868

Hong Y, Zhang K, Gu J, Bi S, Zhou Y, Liu D, Liu F, Sunkavalli K, Bui T, Tan H (2024) LRM: Large Reconstruction Model for Single Image to 3D. In: ICLR

Hu B, Qi Z, Huang G, Xu Z, Zhang R, Ye C, Zhou J, Li X, Wang J (2026) Identity-consistent video generation under large facial-angle variations. https://doi.org/10.48550/arXiv.2603.21299

Hu E, Li M, Hong J, Qian X, Olwal A, Kim D, Heo S, Du R (2025a) Thing2Reality: Enabling Spontaneous Creation of 3D Objects from 2D Content using Generative AI in XR Meetings. In: UIST, https://doi.org/10.1145/3746059.3747621

Hu K, Jin Y, Zhou H, Du L, Liu J, Liu X (2025b) Generative AI for Immersive Video: Recent Advances and Future Opportunities. In: IJCAI, pp 10464–10472, https://doi.org/10.24963/ijcai.2025/1162

Hu Y, Wang K, Shao Y, Plass J, Wang Z, Perlin K (2024) Generative Terrain Authoring with Mid-air Hand Sketching in Virtual Reality. In: ACM VRST, https://doi.org/10.1145/3641825.3687736

Huang B, Yu Z, Chen A, Geiger A, Gao S (2024a) 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. In: SIGGRAPH Conference Papers, https://doi.org/10.1145/3641519.3657428

Huang G (2022) Spatio-temporal modeling for action recognition in videos. PhD thesis, University of York, URL https://etheses.whiterose.ac.uk/id/eprint/31262/

Huang G, Bors AG (2020) Learning spatio-temporal representations with temporal squeeze pooling. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 2103–2107, https://doi.org/10.1109/ICASSP40776.2020.9054200

Huang G, Bors AG (2021) Region-based non-local operation for video classification. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, pp 10010–10017, https://doi.org/10.1109/ICPR48806.2021.9411997

Huang G, Bors AG (2022a) BQN: Busy-quiet net enabled by motion band-pass module for action recognition. IEEE Trans Image Process 31:4966–4979. https://doi.org/10.1109/TIP.2022.3189810

Huang G, Bors AG (2022b) Busy-quiet video disentangling for video classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp 1341–1350, https://doi.org/10.1109/WACV51458.2022.00083

Huang G, Fu H, Bors AG (2023) Masked image residual learning for scaling deeper vision transformers. In: NeurIPS, vol 36. Curran Associates, Inc., pp 57570–57582, https://doi.org/10.52202/075280-2511

Huang G, Lin R, Li Y, Bull DR, Anantrasirichai N (2025a) BVI-Mamba: Video enhancement using a visual state-space model for low-light and underwater environments. In: Markopoulos P, Ouyang B, Sklivanitis G (eds) Machine Learning from Challenging Data 2025, Proceedings of SPIE, vol 13460. SPIE, p 1346009, https://doi.org/10.1117/12.3053998

Huang G, Wang H, Qi Z, Lu W, Bull D, Anantrasirichai N (2025b) From Restoration to Reconstruction: Rethinking 3D Gaussian Splatting for Underwater Scenes. arXiv preprint https://doi.org/10.48550/arXiv.2509.17789, Version v1

Huang G, Wang H, Seymour B, Kovacs E, Ellerbroc J, Blackham D, Anantrasirichai N (2026a) Visual enhancement and 3D representation for underwater scenes: a review. Artif Intell Rev 59(10):205. https://doi.org/10.1007/s10462-026-11597-4

Huang G, Yang Q, Lin R, Qi Z, Bull D, Anantrasirichai N (2026b) Bayesian neural networks for one-to-many mapping in image enhancement. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 5004–5012, https://doi.org/10.1609/aaai.v40i7.37413

Huang S, Wu J, Zhou Q, Miao S, Long M (2026c) Vid2World: Crafting Video Diffusion Models to Interactive World Models. In: ICLR, arXiv:2505.14357

Huang T, Zheng W, Wang T, Liu Y, Wang Z, Wu J, Jiang J, Li H, Lau RWH, Zuo W, Guo C (2025c) Voyager: Long-Range and World-Consistent Video Diffusion for Explorable 3D Scene Generation. ACM Trans Graph 44(6):1–15. https://doi.org/10.1145/3763330, article 245

Huang X, Li Z, He G, Zhou M, Shechtman E (2025d) Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. In: NeurIPS, arXiv:2506.08009

Huang Z, He Y, Yu J, Zhang F, Si C, Jiang Y, Zhang Y, Wu T, Jin Q, Chanpaisit N, Wang Y, Chen X, Wang L, Lin D, Qiao Y, Liu Z (2024b) VBench: Comprehensive Benchmark Suite for Video Generative Models. In: CVPR, https://doi.org/10.1109/cvpr52733.2024.02060

Huang ZH, Lin G, Lin J, Huang YC, Yu R, Niu M, Yang S, Liu YL, Chuang YY, Zhang K, Wang Z (2026d) Programmable World Model. arXiv preprint https://doi.org/10.48550/arXiv.2609.10540, Version v2

HunyuanWorld Team, Wang Z, Liu Y, Wu J, Gu Z, Wang H, Zuo X, Huang T, Li W, Zhang S, Lian Y, Tsai Y, Wang L, Liu S, Jiang P, Yang X, Guo D, Tang Y, Mao X, Yu J, Yu J, Zhang J, Chen M, Dong L, Jia Y, Zhang C, Tan Y, Zhang H, Ye Z, He P, Wu R, Chen M, Li Z, Qin W, Wang L, Sun Y, Niu L, Yuan X, Yang X, He Y, Xiao J, Tao Y, Zhu J, Xue J, Liu K, Zhao C, Wu X, Liu T, Chen P, Wang D, Liu Y, Linus, Jiang J, Wang T, Guo C (2025) HunyuanWorld 1.0: Generating Immersive, Explorable, and Interactive 3D Worlds from Words or Pixels. https://doi.org/10.48550/arXiv.2507.21809, technical report, Version v1, dated 2025-07-29

Ihara K, Lee D, Abe M, Jo HY, Suzuki R (2026) CinemaWorld: Generative Augmented Reality with LLMs and 3D Scene Generation for Movie Augmentation. https://doi.org/10.48550/arXiv.2603.08060, preprint Version v1, 9 March 2026; accepted to UIST 2026

Jain A, Mildenhall B, Barron JT, Abbeel P, Poole B (2022) Zero-Shot Text-Guided Object Generation With Dream Fields. In: CVPR, https://doi.org/10.1109/cvpr52688.2022.00094

Jerald J, Whitton MC (2009) Relating Scene-Motion Thresholds to Latency Thresholds for Head-Mounted Displays. In: IEEE VR, pp 211–218, https://doi.org/10.1109/vr.2009.4811025

Jiang J, Li B (2026) PartInteractor: Intent-Driven Part-Aware 3D Authoring for Continuous Co-Creation in XR. https://doi.org/10.48550/arXiv.2608.01335, preprint Version v1, 2 August 2026; accepted to UIST 2026

Jiang W, He W, Tan Z, Kuang Z, Yu D, Hasegawa S, Mayer S, Sarsenbayeva Z (2026) CoGen3D: An Agentic Human-AI Co-Design Pipeline for 3D Asset Generation for Virtual Reality. https://doi.org/10.48550/arXiv.2607.03731, Version v1, dated 2026-07-04

Jiang Y, Zhang L, Gao J, Hu W, Yao Y (2024) Consistent4D: Consistent 360° Dynamic Object Generation from Monocular Video. In: ICLR

Jiang Z, Wang H, Huang G, Seymour B, Anantrasirichai N (2025a) RUSplatting: Robust 3D Gaussian Splatting for Sparse-View Underwater Scene Reconstruction. In: Proceedings of the British Machine Vision Conference

Jiang Z, Wang H, Huang G, Seymour B, Anantrasirichai N (2025b) Semantic-guided Gaussian Splatting for High-Fidelity Underwater Scene Reconstruction. arXiv preprint https://doi.org/10.48550/arXiv.2509.00800, technical version inspected: Version v3, 21 April 2026; initial preprint 2025

Jun H, Nichol A (2023) Shap-E: Generating Conditional 3D Implicit Functions. ArXiv preprint, arXiv:2305.02463

Kajiya JT (1986) The rendering equation. SIGGRAPH Comput Graph 20(4):143–150. https://doi.org/10.1145/15886.15902

Karras T, Laine S, Aila T (2019) A Style-Based Generator Architecture for Generative Adversarial Networks. In: CVPR, https://doi.org/10.1109/cvpr.2019.00453

Kato H, Ushiku Y, Harada T (2018) Neural 3D Mesh Renderer. In: CVPR, https://doi.org/10.1109/cvpr.2018.00411

Kazhdan M, Bolitho M, Hoppe H (2006) Poisson Surface Reconstruction. In: SGP, pp 61–70, https://doi.org/10.2312/SGP/SGP06/061-070

Kennedy RS, Lane NE, Berbaum KS, Lilienthal MG (1993) Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. Int J Aviat Psychol 3(3):203–220. https://doi.org/10.1207/s15327108ijap0303_3

Kerbl B, Kopanas G, Leimkühler T, Drettakis G (2023) 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Trans Graph 42(4). https://doi.org/10.1145/3592433

Kilteni K, Groten R, Slater M (2012) The Sense of Embodiment in Virtual Reality. Presence (Camb) 21(4):373–387. https://doi.org/10.1162/pres_a_00124

Kim G, Park J, Ryu N, Liu D, Cho S (2026) SceneFrom3D: Geometry-Conditioned Outdoor 3D Scene Generation via View Scheduling with Object-Level Control. https://doi.org/10.48550/arXiv.2607.04540

Kim J, Kang J, Choi J, Han B (2024) FIFO-Diffusion: Generating Infinite Videos from Text without Training. In: NeurIPS, arXiv:2405.11473

Kingma DP, Welling M (2014) Auto-Encoding Variational Bayes. In: ICLR

Klein G, Murray D (2007) Parallel Tracking and Mapping for Small AR Workspaces. In: IEEE ISMAR, pp 1–10, https://doi.org/10.1109/ismar.2007.4538852

Kolve E, Mottaghi R, Gordon D, Zhu Y, Gupta A, Farhadi A (2017) AI2-THOR: An Interactive 3D Environment for Visual AI. arXiv:1712.05474v1

Kondratyuk D, Yu L, Gu X, Lezama J, Huang J, Schindler G, Hornung R, Birodkar V, Yan J, Chiu MC, Somandepalli K, Akbari H, Alon Y, Cheng Y, Dillon J, Gupta A, Hahn M, Hauth A, Hendon D, Martinez A, Minnen D, Sirotenko M, Sohn K, Yang X, Adam H, Yang MH, Essa I, Wang H, Ross DA, Seybold B, Jiang L (2024) VideoPoet: A Large Language Model for Zero-Shot Video Generation. In: ICML, arXiv:2312.14125

Kong L, Yang Y, Mei J, Liu Y, Liang A, Zhu D, Lu D, Yin W, Hu X, Jia M, Deng J, Zhang K, Wu Y, Yan T, Gao S, Wang S, Li L, Pan L, Liu Y, Zhu J, Ooi WT, Hoi SCH, Liu Z (2025) 3D and 4D World Modeling: A Survey. https://doi.org/10.48550/arXiv.2509.07996, Version v4, revised 20 July 2026

Kong W, Tian Q, Zhang Z, Min R, Dai Z, Zhou J, Xiong J, Li X, Wu B, Zhang J, Wu K, Lin Q, Yuan J, Long Y, Wang A, Wang A, Li C, Huang D, Yang F, Tan H, Wang H, Song J, Bai J, Wu J, Xue J, Wang J, Wang K, Liu M, Li P, Li S, Wang W, Yu W, Deng X, Li Y, Chen Y, Cui Y, Peng Y, Yu Z, He Z, Xu Z, Zhou Z, Xu Z, Tao Y, Lu Q, Liu S, Zhou D, Wang H, Yang Y, Wang D, Liu Y, Jiang J, Zhong C (2024) HunyuanVideo: A Systematic Framework For Large Video Generative Models. arXiv preprint arXiv:2412.03603

Kramida G (2016) Resolving the Vergence-Accommodation Conflict in Head-Mounted Displays. IEEE Trans Vis Comput Graph 22(7):1912–1931. https://doi.org/10.1109/tvcg.2015.2473855

Lan Y, Hong F, Yang S, Zhou S, Meng X, Dai B, Pan X, Loy CC (2024) LN3Diff: Scalable Latent Neural Fields Diffusion for Speedy 3D Generation. In: ECCV, https://doi.org/10.1007/978-3-031-73235-5_7, Inspected arXiv:2403.12019v2

Lee D, Sun Q, Yoon SH (2026) HOICraft: In-situ VLM-based authoring tool for part-level hand-object interaction design in VR. In: CHI. Association for Computing Machinery, pp 1–25, https://doi.org/10.1145/3772318.3790823

Lee J, Aleotti F, Mazala D, Garcia-Hernando G, Vicente S, Johnston OJ, Kraus-Liang I, Powierza J, Shin D, Froehlich JE, Brostow G, Van Brummelen J (2025) ImaginateAR: AI-assisted in-situ authoring in augmented reality. In: UIST. Association for Computing Machinery, UIST '25, https://doi.org/10.1145/3746059.3747635

Lewis JR (1995) IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. Int J Hum Comput Interact 7(1):57–78. https://doi.org/10.1080/10447319509526110

Li B, Yuan LP, Wang Z (2025a) VideoCraft: A Mixed Reality-empowered Video Generation Workflow with Spatial Layer Editing for Concept Video Creation. In: UIST, https://doi.org/10.1145/3746059.3747606

Li C, Xia F, Martín-Martín R, Lingelbach M, Srivastava S, Shen B, Vainio KE, Gokmen C, Dharan G, Jain T, Kurenkov A, Liu K, Gweon H, Wu J, Fei-Fei L, Savarese S (2022) iGibson 2.0: Object-Centric Simulation for Robot Learning of Everyday Household Tasks. In: CoRL

Li C, Zhang R, Wong J, Gokmen C, Srivastava S, Martín-Martín R, Wang C, Levine G, Lingelbach M, Sun J, Anvari M, Hwang M, Sharma M, Aydin A, Bansal D, Hunter S, Kim KY, Lou A, Matthews CR, Villa-Renteria I, Tang JH, Tang C, Xia F, Savarese S, Gweon H, Liu K, Wu J, Fei-Fei L (2023) BEHAVIOR-1K: A Benchmark for Embodied AI with 1,000 Everyday Activities and Realistic Simulation. In: CoRL

Li J, Tang J, Xu Z, Wu L, Zhou Y, Shao S, Yu T, Cao Z, Lu Q (2025b) Hunyuan-GameCraft: High-dynamic Interactive Game Video Generation with Hybrid History Condition. ArXiv preprint, arXiv:2506.17201

Li L, Wang G, Li X, Zhang Z, Dou Q, Gu J, Xue T, Shan Y (2026a) CubeComposer: Spatio-Temporal Autoregressive 4K 360° Video Generation from Perspective Video. In: CVPR

Li M, Patil AG, Xu K, Chaudhuri S, Khan O, Shamir A, Tu C, Chen B, Cohen-Or D, Zhang H (2019) GRAINS: Generative Recursive Autoencoders for INdoor Scenes. ACM Trans Graph 38(2). https:

//doi.org/10.1145/3303766

Li R, Torr P, Vedaldi A, Jakab T (2025c) VMem: Consistent Interactive Video Scene Generation with Surfel-Indexed View Memory. In: ICCV, https://doi.org/10.1109/iccv51701.2025.02383

Li T, Tian Y, Li H, Deng M, He K (2024) Autoregressive Image Generation without Vector Quantization. In: NeurIPS, arXiv:2406.11838

Li X, He X, Zhang L, Wu M, Li X, Liu Y (2025d) A Comprehensive Survey on World Models for Embodied AI. Version 3, revised June 2026, arXiv:2510.16732v3

Li Y, Zou ZX, Liu Z, Wang D, Liang Y, Yu Z, Liu X, Guo YC, Liang D, Ouyang W, Cao YP (2025e) TripoSG: High-Fidelity 3D Shape Synthesis using Large-Scale Rectified Flow Models. ArXiv preprint, arXiv:2502.06608

Li Z, Meng Z, Shi S, Zhai M, Tan J, Li C, Zhang K (2026b) From Pixels to States: Rethinking Interactive World Models as Game Engines. ArXiv preprint, arXiv:2607.14076

Lillemark H, Huang B, Zhan F, Du Y, Keller TA (2026) Flow Equivariant World Models: Structured Memory for Dynamic Environments. In: ICML, pp 73968–74002

Limbert G, Karnik A (2025) ImoGenXR: Investigating the Impact on Creativity Support in a Generative-AI Assisted Immersive Authoring Workflow. In: ICVR, https://doi.org/10.1109/ICVR66534.2025.11172549

Lin C, Mu Y (2024) InstructScene: Instruction-Driven 3D Indoor Scene Synthesis with Semantic Graph Prior. In: ICLR

Lin CH, Gao J, Tang L, Takikawa T, Zeng X, Huang X, Kreis K, Fidler S, Liu MY, Lin TY (2023) Magic3D: High-Resolution Text-to-3D Content Creation. In: CVPR, pp 300–309, https://doi.org/10.1109/cvpr52729.2023.00037

Lin R, Huang G, Anantrasirichai N (2026a) Dynamic weight-based temporal aggregation for low-light video enhancement under extreme noise. In: 2026 IEEE International Conference on Image Processing (ICIP). IEEE, https://doi.org/10.1109/ICIP61757.2026.11630319

Lin R, Huang G, Bull D, Anantrasirichai N (2026b) PixIE: Prompted pixel-space low-light image enhancement. https://doi.org/10.48550/arXiv.2605.23531

Lin R, Huang G, Lin J, Sun Q, Malyugina A, Bull DR, Anantrasirichai N (2026c) BVI-RLV: A fully registered dataset for low-light video enhancement. IEEE Transactions on Multimedia Accepted/In press, 3 October 2026, arXiv:2407.03535v3

Lin Y, Lin C, Pan P, Yan H, Feng Y, Mu Y, Fragkiadaki K (2025) PartCrafter: Structured 3D Mesh Generation via Compositional Latent Diffusion Transformers. In: NeurIPS, vol 38. Curran Associates, Inc., pp 35387–35415, https://doi.org/10.52202/085713-1189

Ling H, Kim SW, Torralba A, Fidler S, Kreis K (2024a) Align Your Gaussians: Text-to-4D with Dynamic 3D Gaussians and Composed Diffusion Models. In: CVPR, https://doi.org/10.1109/cvpr52733.2024.00819

Ling L, Sheng Y, Tu Z, Zhao W, Xin C, Wan K, Yu L, Guo Q, Yu Z, Lu Y, Li X, Sun X, Ashok R, Mukherjee A, Kang H, Kong X, Hua G, Zhang T, Benes B, Bera A (2024b) DL3DV-10K: A Large-Scale Scene Dataset for Deep Learning-based 3D Vision. In: CVPR, https://doi.org/10.1109/cvpr52733.2024.02092

Ling L, Ge Y, Sheng Y, Bera A (2026a) I-Scene: 3D Instance Models are Implicit Generalizable Spatial Learners. In: CVPR, pp 26974–26983

Ling L, Lin CH, Lin TY, Ding Y, Zeng Y, Sheng Y, Ge Y, Liu MY, Bera A, Li Z (2026b) Scenethesis: A Language and Vision Agentic Framework for 3D Scene Generation. In: ICLR

Lipman Y, Chen RTQ, Ben-Hamu H, Nickel M, Le M (2023) Flow Matching for Generative Modeling. In: ICLR, arXiv:2210.02747

Liu J, Li R, Zhao H (2026) Generative AI-assisted visualization prototyping for cultural heritage: A computational framework from 2D planes to 3D immersive scenes. Buildings 16(16):3319. https://doi.org/10.3390/buildings16163319

Liu M, Xu C, Jin H, Chen L, T MV, Xu Z, Su H (2023a) One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization. In: NeurIPS

Liu M, Shi R, Chen L, Zhang Z, Xu C, Wei X, Chen H, Zeng C, Gu J, Su H (2024a) One-2-3-45++: Fast Single Image to 3D Objects with Consistent Multi-View Generation and 3D Diffusion. In: CVPR, https://doi.org/10.1109/cvpr52733.2024.00960

Liu R, Wu R, Van Hoorick B, Tokmakov P, Zakharov S, Vondrick C (2023b) Zero-1-to-3: Zero-shot One Image to 3D Object. In: ICCV, https://doi.org/10.1109/iccv51070.2023.00853

Liu X, Gong C, Liu Q (2023c) Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In: ICLR, arXiv:2209.03003

Liu Y, Lin C, Zeng Z, Long X, Liu L, Komura T, Wang W (2024b) SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. In: ICLR

Liu Z, Feng Y, Black MJ, Nowrouzezahrai D, Paull L, Liu W (2023d) MeshDiffusion: Score-based Generative 3D Mesh Modeling. In: ICLR

Long X, Guo YC, Lin C, Liu Y, Dou Z, Liu L, Ma Y, Zhang SH, Habermann M, Theobalt C, Wang W (2024) Wonder3D: Single Image to 3D using Cross-Domain Diffusion. In: CVPR, https://doi.org/10.1109/cvpr52733.2024.00951

Lorensen WE, Cline HE (1987) Marching cubes: A high resolution 3D surface construction algorithm. SIGGRAPH Comput Graph 21(4):163–169. https://doi.org/10.1145/37402.37422

Lu C, Zhou Y, Bao F, Chen J, Li C, Zhu J (2022) DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. In: NeurIPS, arXiv:2206.00927

Lu T, Yu M, Xu L, Xiangli Y, Wang L, Lin D, Dai B (2024) Scaffold-GS: Structured 3D Gaussians for View-Adaptive Rendering. In: CVPR, https://doi.org/10.1109/cvpr52733.2024.01952

Lu W, Chen H, Yi A, Huang G, Chung YY, Hu K, Wang Z (2026a) Learning Fine-Grained Geometry for Sparse-View Splatting via Cascade Depth Loss. In: IEEE International Conference on Multimedia and Expo, technical version inspected: arXiv v2, 2 April 2026, arXiv:2505.22279v2

Lu Y, Luo W, Tu P, Li H, Zhu H, Yu Z, Wang X, Chen X, Peng X, Li X, Chen Z (2026b) 4DWorldBench: A Comprehensive Evaluation Framework for 3D/4D World Generation Models. In: CVPR, pp 34322–34332

Ma B, Gao H, Deng H, Luo Z, Huang T, Tang L, Wang X (2025) You See it, You Got it: Learning 3D Creation on Pose-Free Videos at Scale. In: CVPR, https://doi.org/10.1109/cvpr52734.2025.00194

Malyugina A, Huang G, Ruiz E, Leslie B, Anantrasirichai N (2025) Marine snow removal using internally generated pseudo ground truth. In: 2025 33rd European Signal Processing Conference (EUSIPCO). IEEE, pp 900–904, https://doi.org/10.23919/EUSIPCO63237.2025.11226718

Mandal D, Fu D, Miller J, Villareal W, Peng X, Chakravarthula P (2026) EPIC: Epipolar-consistent 360-degree immersive stereo video generation. https://doi.org/10.48550/arXiv.2609.38689, preprint, Version v1, dated 2026-09-30

Mania K, Adelstein BD, Ellis SR, Hill MI (2004) Perceptual sensitivity to head tracking latency in virtual environments with varying degrees of scene complexity. In: Proceedings of the 1st Symposium on Applied perception in graphics and visualization, pp 39–47, https://doi.org/10.1145/1012551.1012559

McMahan RP, Bowman DA, Zielinski DJ, Brady RB (2012) Evaluating Display Fidelity and Interaction Fidelity in a Virtual Reality Game. IEEE Trans Vis Comput Graph 18(4):626–633. https://doi.org/10.1109/tvcg.2012.43

Mei Y, Hartenstein S, Wombacher J, Mueller F, Gugenheimer J (2026) ShadAR: Real-time Prompting On-demand Visual Augmentation in Augmented Reality. https://doi.org/10.48550/arXiv.2610.19998, preprint Version v1, 14 October 2026; accepted to UIST 2026

Merrell P, Schkufza E, Koltun V (2010) Computer-generated residential building layouts. ACM Trans Graph 29(6):1–12. https://doi.org/10.1145/1882261.1866203

Mescheder L, Oechsle M, Niemeyer M, Nowozin S, Geiger A (2019) Occupancy Networks: Learning 3D Reconstruction in Function Space. In: CVPR, https://doi.org/10.1109/cvpr.2019.00459

Michel O, Bar-On R, Liu R, Benaim S, Hanocka R (2022) Text2Mesh: Text-Driven Neural Stylization for Meshes. In: CVPR, https://doi.org/10.1109/cvpr52688.2022.01313

Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R (2020) NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In: ECCV, https://doi.org/10.1007/978-3-030-58452-8_24

Milgram P, Kishino F (1994) A Taxonomy of Mixed Reality Visual Displays. IEICE Trans Inf & Syst E77-D(12):1321–1329

Mine MR, Brooks FPJr., Séquin CH (1997) Moving Objects in Space: Exploiting Proprioception in Virtual-Environment Interaction. In: SIGGRAPH, pp 19–26, https://doi.org/10.1145/258734.258747

Mirza M, Osindero S (2014) Conditional Generative Adversarial Nets. arXiv preprint arXiv:1411.1784

Mohammad Khalid N, Xie T, Belilovsky E, Popa T (2022) CLIP-Mesh: Generating textured meshes from text using pretrained image-text models. In: SIGGRAPH Asia Conference Papers, pp 25:1–25:8, https://doi.org/10.1145/3550469.3555392

Munkberg J, Hasselgren J, Shen T, Gao J, Chen W, Evans A, Müller T, Fidler S (2022) Extracting Triangular 3D Models, Materials, and Lighting From Images. In: CVPR, https://doi.org/10.1109/cvpr52688.2022.00810

Müller P, Wonka P, Haegler S, Ulmer A, Gool LV (2006) Procedural modeling of buildings. ACM Trans Graph 25(3):614–623. https://doi.org/10.1145/1141911.1141931

Müller T, Evans A, Schied C, Keller A (2022) Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. ACM Trans Graph 41(4):102:1–102:15. https://doi.org/10.1145/3528223.3530127

Nash C, Ganin Y, Eslami SMA, Battaglia P (2020) PolyGen: An Autoregressive Generative Model of 3D Meshes. In: ICML, pp 7220–7229

Nguyen T, Simoff SJ, Qu Z, Adam R, Catchpoole DR, Kennedy PJ, Nguyen QV (2026) Generative artificial intelligence and human artificial intelligence interaction in immersive environments: Research and development directions. Int J Hum Comput Interact https://doi.org/10.1080/10447318.2026.2719170, published online 26 August 2026

Ning X, Zhuo Y, Wang X, Sio CID, Lee LH (2026) When Generative Artificial Intelligence Meets Extended Reality: A Systematic Review. Int J Hum Comput Interact 42(10):6870–6890. https://doi.org/10.1080/10447318.2025.2565392, first published online 30 September 2025

Numan N, Rajaram S, Kumaravel BT, Marquardt N, Wilson AD (2024) SpaceBlender: Creating Context-Rich Collaborative Spaces Through Generative 3D Scene Blending. In: UIST, https://doi.org/10.1145/3654777.3676361

van den Oord A, Kalchbrenner N, Kavukcuoglu K (2016) Pixel Recurrent Neural Networks. In: ICML, arXiv:1601.06759

van den Oord A, Vinyals O, Kavukcuoglu K (2017) Neural Discrete Representation Learning. In: NIPS, arXiv:1711.00937

Oquab M, Darcet T, Moutakanni T, Vo HV, Szafraniec M, Khalidov V, Fernandez P, Haziza D, Massa F, El-Nouby A, Assran M, Ballas N, Galuba W, Howes R, Huang PY, Li SW, Misra I, Rabbat M, Sharma V, Synnaeve G, Xu H, Jegou H, Mairal J, Labatut P, Joulin A, Bojanowski P (2024) DINOv2: Learning Robust Visual Features without Supervision. Transact Mach Learn Res

Padmanaban N, Konrad R, Stramer T, Cooper EA, Wetzstein G (2017) Optimizing virtual reality for all users through gaze-contingent and adaptive focus displays. Proc Natl Acad Sci U S A 114(9):2183–2188. https://doi.org/10.1073/pnas.1617251114

Parish YIH, Müller P (2001) Procedural modeling of cities. In: SIGGRAPH, pp 301–308, https://doi.org/10.1145/383259.383292

Park JJ, Florence P, Straub J, Newcombe R, Lovegrove S (2019) DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In: CVPR, https://doi.org/10.1109/cvpr.2019.00025

Parker-Holder J, Fruchter S (2025) Genie 3: A new frontier for world models. Google DeepMind research blog, URL https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/, published 5 August 2025; accessed 21 October 2026

Paschalidou D, Kar A, Shugrina M, Kreis K, Geiger A, Fidler S (2021) ATISS: Autoregressive Transformers for Indoor Scene Synthesis. In: NeurIPS, vol 34. Curran Associates, Inc., pp 12013–12026

Peebles W, Xie S (2023) Scalable Diffusion Models with Transformers. In: ICCV, pp 4195–4205, https://doi.org/10.1109/iccv51070.2023.00387

Peng S, Niemeyer M, Mescheder L, Pollefeys M, Geiger A (2020) Convolutional Occupancy Networks. In: ECCV, https://doi.org/10.1007/978-3-030-58580-8_31

Piumsomboon T, Lee GA, Hart JD, Ens B, Lindeman RW, Thomas BH, Billinghurst M (2018) Mini-Me: An Adaptive Avatar for Mixed Reality Remote Collaboration. In: CHI, pp 1–13, https://doi.org/10.1145/3173574.3173620

Po R, Nitzan Y, Zhang R, Chen B, Dao T, Shechtman E, Wetzstein G, Huang X (2025) Long-Context State-Space Video World Models. In: ICCV, https://doi.org/10.1109/iccv51701.2025.00817

Po R, Zhang DJ, Hertz A, Wetzstein G, Wadhwa N, Ruiz N (2026) MultiGen: Level-design for editable multiplayer worlds in diffusion game engines. https://doi.org/10.48550/arXiv.2603.06679, Version v2, 30 March 2026

Polyak A, Zohar A, Brown A, Tjandra A, Sinha A, Lee A, Vyas A, Shi B, Ma CY, Chuang CY, Yan D, Choudhary D, Wang D, Sethi G, Pang G, Ma H, Misra I, Hou J, Wang J, Jagadeesh K, Li K, Zhang L, Singh M, Williamson M, Le M, Yu M, Singh MK, Zhang P, Vajda P, Duval Q, Girdhar R, Sumbaly R, Rambhatla SS, Tsai S, Azadi S, Datta S, Chen S, Bell S, Ramaswamy S, Sheynin S, Bhattacharya S, Motwani S, Xu T, Li T, Hou T, Hsu WN, Yin X, Dai X, Taigman Y, Luo Y, Liu YC, Wu YC, Zhao Y, Kirstain Y, He Z, He Z, Pumarola A, Thabet A, Sanakoyeu A, Mallya A, Guo B, Araya B, Kerr B, Wood C, Liu C, Peng C, Vengertsev D, Schonfeld E, Blanchard E, Juefei-Xu F, Nord F, Liang J, Hoffman J, Kohler J, Fire K, Sivakumar K, Chen L, Yu L, Gao L, Georgopoulos M, Moritz R, Sampson SK, Li S, Parmeggiani S, Fine S, Fowler T, Petrovic V, Du Y (2024) Movie Gen: A Cast of Media Foundation Models. arXiv preprint arXiv:2410.13720

Poole B, Jain A, Barron JT, Mildenhall B (2023) DreamFusion: Text-to-3D using 2D Diffusion. In: ICLR

Poupyrev I, Billinghurst M, Weghorst S, Ichikawa T (1996) The go-go interaction technique. In: UIST, pp 79–80, https://doi.org/10.1145/237091.237102

Pumarola A, Corona E, Pons-Moll G, Moreno-Noguer F (2021) D-NeRF: Neural Radiance Fields for Dynamic Scenes. In: CVPR, https://doi.org/10.1109/cvpr46437.2021.01018

Qi Z, Huang G, Liu C, Ye F (2024) Layered rendering diffusion model for controllable zero-shot image synthesis. In: ECCV, Lecture Notes in Computer Science, vol 15124. Springer, pp 426–443, https://doi.org/10.1007/978-3-031-72848-8_25, first online 29 November 2024; print edition 2025

Qian G, Mai J, Hamdi A, Ren J, Siarohin A, Li B, Lee HY, Skorokhodov I, Wonka P, Tulyakov S, Ghanem B (2024) Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors. In: ICLR

Qiu H, Xia M, Zhang Y, He Y, Wang X, Shan Y, Liu Z (2024) FreeNoise: Tuning-Free Longer Video Diffusion via Noise Rescheduling. In: ICLR, arXiv:2310.15169

Qiu Z, Wu A (2026) Pano3DComposer: Feed-Forward Compositional 3D Scene Generation from Single Panoramic Image. In: CVPR

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I (2021) Learning Transferable Visual Models From Natural Language Supervision. In: ICML, pp 8748–8763

Rahimi F, Sadeghi-Niaraki A, Choi SM (2025) Generative AI Meets Virtual Reality: A Comprehensive Survey on Applications, Challenges, and Future Direction. IEEE Access 13:94893–94909. https://doi.org/10.1109/ACCESS.2025.3574779

Raistrick A, Lipson L, Ma Z, Mei L, Wang M, Zuo Y, Kayan K, Wen H, Han B, Wang Y, Newell A, Law H, Goyal A, Yang K, Deng J (2023) Infinite Photorealistic Worlds using Procedural Generation. In: CVPR, https://doi.org/10.1109/cvpr52729.2023.01215

Raistrick A, Mei L, Kayan K, Yan D, Zuo Y, Han B, Wen H, Parakh M, Alexandropoulos S, Lipson L, Ma Z, Deng J (2024) Infinigen Indoors: Photorealistic Indoor Scenes using Procedural Generation. In: CVPR, https://doi.org/10.1109/cvpr52733.2024.02058

Ramakrishnan SK, Gokaslan A, Wijmans E, Maksymets O, Clegg A, Turner J, Undersander E, Galuba W, Westbury A, Chang A, Savva M, Zhao Y, Batra D (2021) Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI. In: NeurIPS Datasets and Benchmarks Track

Rasch J, Töws J, Hirzle T, Müller F, Schmitz M (2025) CreepyCoCreator? Investigating AI Representation Modes for 3D Object Co-Creation in Virtual Reality. In: CHI, https://doi.org/10.1145/3706598.3713720

Reizenstein J, Shapovalov R, Henzler P, Sbordone L, Labatut P, Novotny D (2021) Common Objects in 3D: Large-Scale Learning and Evaluation of Real-Life 3D Category Reconstruction. In: ICCV, pp 10901–10911, https://doi.org/10.1109/iccv48922.2021.01072

Ren J, Pan L, Tang J, Zhang C, Cao A, Zeng G, Liu Z (2023) DreamGaussian4D: Generative 4D Gaussian Splatting. https://doi.org/10.48550/arXiv.2312.17142, Version v3, dated 2024-06-10

Ren X, Shen T, Huang J, Ling H, Lu Y, Nimier-David M, Müller T, Keller A, Fidler S, Gao J (2025) GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control. In: CVPR, pp 6121–6132, https://doi.org/10.1109/cvpr52734.2025.00574

Rezende DJ, Mohamed S, Wierstra D (2014) Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In: ICML, arXiv:1401.4082

Richardson E, Metzer G, Alaluf Y, Giryes R, Cohen-Or D (2023) TEXTure: Text-Guided Texturing of 3D Shapes. In: SIGGRAPH Conference Papers, https://doi.org/10.1145/3588432.3591503

Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-Resolution Image Synthesis with Latent Diffusion Models. In: CVPR, https://doi.org/10.1109/cvpr52688.2022.01042

Saharia C, Chan W, Saxena S, Li L, Whang J, Denton EL, Ghasemipour K, Gontijo Lopes R, Karagol Ayan B, Salimans T, Ho J, Fleet D, Norouzi M (2022) Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In: NeurIPS, pp 36479–36494, https://doi.org/10.52202

/068431-2643

Salimans T, Ho J (2022) Progressive Distillation for Fast Sampling of Diffusion Models. In: ICLR, arXiv:2202.00512

Sanchez-Vives MV, Slater M (2005) From presence to consciousness through virtual reality. Nat Rev Neurosci 6(4):332–339. https://doi.org/10.1038/nrn1651

Sand.ai, Teng H, Jia H, Sun L, Li L, Li M, Tang M, Han S, Zhang T, Zhang WQ, Luo W, Kang X, Sun Y, Cao Y, Huang Y, Lin Y, Fang Y, Tao Z, Zhang Z, Wang Z, Liu Z, Shi D, Su G, Sun H, Pan H, Wang J, Sheng J, Cui M, Hu M, Yan M, Yin S, Zhang S, Liu T, Yin X, Yang X, Song X, Hu X, Zhang Y, Li Y (2025) MAGI-1: Autoregressive Video Generation at Scale. arXiv preprint arXiv:2505.13211

Saredakis D, Szpak A, Birckhead B, Keage HAD, Rizzo A, Loetscher T (2020) Factors Associated With Virtual Reality Sickness in Head-Mounted Displays: A Systematic Review and Meta-Analysis. Front Hum Neurosci 14. https://doi.org/10.3389/fnhum.2020.00096

Savva G, Michel O, Lu D, Waiwitlikhit S, Meehan T, Mishra D, Poddar S, Lu J, Xie S (2026) Solaris: Building a Multiplayer Video World Model in Minecraft. arXiv preprint https://doi.org/10.48550/arXiv.2602.22208, Version v2

Savva M, Kadian A, Maksymets O, Zhao Y, Wijmans E, Jain B, Straub J, Liu J, Koltun V, Malik J, Parikh D, Batra D (2019) Habitat: A Platform for Embodied AI Research. In: ICCV, pp 9339–9347, https://doi.org/10.1109/iccv.2019.00943

Schrittwieser J, Antonoglou I, Hubert T, Simonyan K, Sifre L, Schmitt S, Guez A, Lockhart E, Hassabis D, Graepel T, Lillicrap T, Silver D (2020) Mastering Atari, Go, chess and shogi by planning with a learned model. Nature 588:604–609. https://doi.org/10.1038/s41586-020-03051-4

Schubert T, Friedmann F, Regenbrecht H (2001) The Experience of Presence: Factor Analytic Insights. Presence (Camb) 10(3):266–281. https://doi.org/10.1162/105474601300343603

Schwind V, Knierim P, Haas N, Henze N (2019) Using Presence Questionnaires in Virtual Reality. In: CHI, pp 1–12, https://doi.org/10.1145/3290605.3300590

Shang Y, Tang Y, Zhang X, Wang S, Yan Y, Zhang H, Zheng Z, Zhao J, Feng J, Gao C, Chen J, Li Y (2025) A Survey of Embodied World Models. URL https://fi.ee.tsinghua.edu.cn/public/publications/0940dda4-af15-11f0-9d60-0242ac120002.pdf, author-hosted preprint; supplied twelve-author edition

Shen B, Xia F, Li C, Martin-Martin R, Fan L, Wang G, Perez-D'Arpino C, Buch S, Srivastava S, Tchapmi L, Tchapmi M, Vainio K, Wong J, Fei-Fei L, Savarese S (2021a) iGibson 1.0: a Simulation Environment for Interactive Tasks in Large Realistic Scenes. In: IEEE/RSJ IROS, pp 7520–7527, https://doi.org/10.1109/iros51168.2021.9636667

Shen T, Gao J, Yin K, Liu MY, Fidler S (2021b) Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis. In: NeurIPS

Shen T, Munkberg J, Hasselgren J, Yin K, Wang Z, Chen W, Gojcic Z, Fidler S, Sharp N, Gao J (2023) Flexible Isosurface Extraction for Gradient-Based Mesh Optimization. ACM Trans Graph 42(4):37:1–37:16. https://doi.org/10.1145/3592430

Shen Y, Li B, Huang J, Yip D, Wang Z (2025) GaussianShopVR: Facilitating Immersive 3D Authoring Using Gaussian Splatting in VR. In: UIST, https://doi.org/10.1145/3746059.3747803

Shi R, Chen H, Zhang Z, Liu M, Xu C, Wei X, Chen L, Zeng C, Su H (2023) Zero123++: a Single Image to Consistent Multi-view Diffusion Base Model. arXiv preprint arXiv:2310.15110, arXiv:2310.15110

Shi Y, Wang P, Ye J, Mai L, Li K, Yang X (2024) MVDream: Multi-view Diffusion for 3D Generation. In: ICLR

Shibata T, Kim J, Hoffman DM, Banks MS (2011) The zone of comfort: Predicting visual discomfort with stereo displays. J Vis 11(8):11–11. https://doi.org/10.1167/11.8.11

Siddiqui Y, Alliegro A, Artemov A, Tommasi T, Sirigatti D, Rosov V, Dai A, Nießner M (2024) MeshGPT: Generating Triangle Meshes with Decoder-Only Transformers. In: CVPR, https://doi.org/10.1109/cvpr52733.2024.01855

Singer U, Polyak A, Hayes T, Yin X, An J, Zhang S, Hu Q, Yang H, Ashual O, Gafni O, Parikh D, Gupta S, Taigman Y (2023a) Make-A-Video: Text-to-Video Generation without Text-Video Data. In: ICLR, arXiv:2209.14792

Singer U, Sheynin S, Polyak A, Ashual O, Makarov I, Kokkinos F, Goyal N, Vedaldi A, Parikh D, Johnson J, Taigman Y (2023b) Text-To-4D Dynamic Scene Generation. In: ICML

Sitzmann V, Martel JNP, Bergman AW, Lindell DB, Wetzstein G (2020) Implicit Neural Representations with Periodic Activation Functions. In: NeurIPS

Slater M (2009) Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Philos Trans R Soc Lond B Biol Sci 364(1535):3549–3557. https://doi.org/10.1098/rstb.2009.0138

Slater M, Wilbur S (1997) A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. Presence (Camb) 6(6):603–616. https://doi.org/10.1162/pres.1997.6.6.603

Slater M, Spanlang B, Sanchez-Vives MV, Blanke O (2010) First Person Experience of Body Transfer in Virtual Reality. PLoS One 5(5):e10564. https://doi.org/10.1371/journal.pone.0010564

Sohl-Dickstein J, Weiss EA, Maheswaranathan N, Ganguli S (2015) Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In: ICML, arXiv:1503.03585

Song J, Meng C, Ermon S (2021a) Denoising Diffusion Implicit Models. In: ICLR, arXiv:2010.02502

Song K, Chen B, Simchowitz M, Du Y, Tedrake R, Sitzmann V (2025) History-Guided Video Diffusion. In: ICML, arXiv:2502.06764

Song S, Lichtenberg SP, Xiao J (2015) SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In: CVPR, pp 567–576, https://doi.org/10.1109/cvpr.2015.7298655

Song Y, Sohl-Dickstein J, Kingma DP, Kumar A, Ermon S, Poole B (2021b) Score-Based Generative Modeling through Stochastic Differential Equations. In: ICLR, arXiv:2011.13456

Song Y, Dhariwal P, Chen M, Sutskever I (2023) Consistency Models. In: ICML, arXiv:2303.01469

Speicher M, Hall BD, Nebeling M (2019) What is Mixed Reality? In: CHI, pp 1–15, https://doi.org/10.1145/3290605.3300767

Srinidhi S, Singh A, Lu E, Rowe A (2026) GenAssist: Interactive Prompt-Driven XR Program Generation. In: 2026 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). IEEE, pp 615–625, https://doi.org/10.1109/VR67842.2026.00081

Srivastava S, Li C, Lingelbach M, Martín-Martín R, Xia F, Vainio KE, Lian Z, Gokmen C, Buch S, Liu K, Savarese S, Gweon H, Wu J, Fei-Fei L (2022) BEHAVIOR: Benchmark for Everyday Household Activities in Virtual, Interactive, and Ecological Environments. In: CoRL

Steinicke F, Bruder G, Jerald J, Frenz H, Lappe M (2010) Estimation of Detection Thresholds for Redirected Walking Techniques. IEEE Trans Vis Comput Graph 16(1):17–27. https://doi.org/10.1109/tvcg.2009.62

Straub J, Whelan T, Ma L, Chen Y, Wijmans E, Green S, Engel JJ, Mur-Artal R, Ren C, Verma S, Clarkson A, Yan M, Budge B, Yan Y, Pan X, Yon J, Zou Y, Leon K, Carter N, Briales J, Gillingham

T, Mueggler E, Pesqueira L, Savva M, Batra D, Strasdat HM, De Nardi R, Goesele M, Lovegrove S, Newcombe R (2019) The Replica Dataset: A Digital Replica of Indoor Spaces. arXiv:1906.05797

Sun W, Chen S, Liu F, Chen Z, Duan Y, Zhu J, Zhang J, Wang Y (2025) DimensionX: Create Any 3D and 4D Scenes from a Single Image with Decoupled Video Diffusion. In: ICCV, pp 13695–13706, https://doi.org/10.1109/iccv51701.2025.01271, technical analysis: arXiv:2411.04928v1 (2024)

Sun W, Zhang H, Wang H, Wu J, Wang Z, Wang Z, Wang Y, Zhang J, Wang T, Guo C (2026a) WorldPlay: Towards long-term geometric consistency for real-time interactive world modeling. In: ICML, Proceedings of Machine Learning Research, vol 306. PMLR, pp 117501–117523

Sun YT, Huang Z, Niu Y, Ma L, Cao YP, Ma Y, Qi X (2026b) Stereo World Model: Camera-Guided Stereo Video Generation. In: CVPR

Sutherland IE (1968) A head-mounted three dimensional display. In: Proceedings of the December 9-11, 1968, fall joint computer conference, part I on - AFIPS '68 (Fall, part I), pp 757–764, https://doi.org/10.1145/1476589.1476686

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the Inception Architecture for Computer Vision. In: CVPR, https://doi.org/10.1109/cvpr.2016.308

Szot A, Clegg A, Undersander E, Wijmans E, Zhao Y, Turner J, Maestre N, Mukadam M, Chaplot DS, Maksymets O, Gokaslan A, Vondruš V, Dharur S, Meier F, Galuba W, Chang A, Kira Z, Koltun V, Malik J, Savva M, Batra D (2021) Habitat 2.0: Training Home Assistants to Rearrange their Habitat. In: NeurIPS, pp 251–266

Taha K (2026) Generative AI for multimodal content: a survey with empirical and experimental evaluations. Artif Intell Rev 59. https://doi.org/10.1007/s10462-026-11525-6, article 140

Tamir NY, Amir S, Itzhaky R, Atia N, Sundaram S, Fu S, Sokolovsky R, Isola P, Dekel T, Zhang R, Farber M (2025) What Makes for a Good Stereoscopic Image? In: CVPR Workshops, pp 261–272, https://doi.org/10.1109/cvprw67362.2025.00031

Tancik M, Srinivasan PP, Mildenhall B, Fridovich-Keil S, Raghavan N, Singhal U, Ramamoorthi R, Barron JT, Ng R (2020) Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. In: NeurIPS

Tang J, Chen Z, Chen X, Wang T, Zeng G, Liu Z (2024a) LGM: Large Multi-View Gaussian Model for High-Resolution 3D Content Creation. In: ECCV, https://doi.org/10.1007/978-3-031-73235-5_1

Tang J, Nie Y, Markhasin L, Dai A, Thies J, Nießner M (2024b) DiffuScene: Denoising Diffusion Models for Generative Indoor Scene Synthesis. In: CVPR, https://doi.org/10.1109/cvpr52733.2024.01938

Tang J, Ren J, Zhou H, Liu Z, Zeng G (2024c) DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. In: ICLR

Tang J, Liu J, Li J, Wu L, Yang H, Zhao P, Gong S, Yuan X, Shao S, Zhang L, Lu Q (2025) Hunyuan-GameCraft-2: Instruction-following Interactive Game World Model. ArXiv preprint, arXiv:2511.23429

Tang S, Zhang F, Chen J, Wang P, Furukawa Y (2023) MVDiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In: NeurIPS, vol 36. Curran Associates, Inc., pp 51202–51233, https://doi.org/10.52202/075280-2229

Team HY-World, Cao C, Zuo X, Wang Z, Zhang Y, Wu J, Liu Z, Gong Y, Liu Y, Yuan B, Zhang C, Li C, Guo D, Yang F, Zhang H, Cao H, Zhu J, Lin J, Xiao J, Zhang J, Yu J, Wang L, Wang L, Wang L, Linus, Chen M, He P, Zhao P, Chen Q, Chen R, Shao R, Liu S, Qin W, Niu X, Yuan X, Sun Y, Tang Y, Sun Y, Lian Y, Tan Y, Liu Y, Yin Y, Min Z, Wang T, Guo C (2026) HY-World 2.0: A Multi-Modal World Model for Reconstructing, Generating, and Simulating 3D Worlds. https://doi.org/10.48550/arXiv.2604.14268, technical report, Version v1, dated 2026-04-15

Tong Y, Qiu Y, Li R, Qiu S, Heng PA (2025) MS2Mesh-XR: Multi-Modal Sketch-to-Mesh Generation in XR Environments. In: IEEE AIxVR. IEEE, pp 272–276, https://doi.org/10.1109/AIxVR63409.2025.00052

Unterthiner T, van Steenkiste S, Kurach K, Marinier R, Michalski M, Gelly S (2018) Towards Accurate Generative Models of Video: A New Metric & Challenges. ArXiv preprint, arXiv:1812.01717

Usoh M, Catena E, Arman S, Slater M (2000) Using Presence Questionnaires in Reality. Presence (Camb) 9(5):497–503. https://doi.org/10.1162/105474600566989

Vachha C, Kang Y, Dive Z, Chidambaram A, Gupta A, Jun E, Hartmann B (2025) Dreamcrafter: Immersive Editing of 3D Radiance Fields Through Flexible, Generative Inputs and Outputs. In: CHI, https://doi.org/10.1145/3706598.3714312

Valevski D, Leviathan Y, Arar M, Fruchter S (2025) Diffusion Models Are Real-Time Game Engines. In: ICLR

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention Is All You Need. In: NIPS, arXiv:1706.03762

Verbin D, Hedman P, Mildenhall B, Zickler T, Barron JT, Srinivasan PP (2022) Ref-NeRF: Structured View-Dependent Appearance for Neural Radiance Fields. In: CVPR, https://doi.org/10.1109/cvpr52688.2022.00541

Villegas R, Babaeizadeh M, Kindermans PJ, Moraldo H, Zhang H, Saffar MT, Castro S, Kunze J, Erhan D (2023) Phenaki: Variable Length Video Generation from Open Domain Textual Descriptions. In: ICLR, arXiv:2210.02399

Wan Team, Wang A, Ai B, Wen B, Mao C, Xie CW, Chen D, Yu F, Zhao H, Yang J, Zeng J, Wang J, Zhang J, Zhou J, Wang J, Chen J, Zhu K, Zhao K, Yan K, Huang L, Feng M, Zhang N, Li P, Wu P, Chu R, Feng R, Zhang S, Sun S, Fang T, Wang T, Gui T, Weng T, Shen T, Lin W, Wang W, Wang W, Zhou W, Wang W, Shen W, Yu W, Shi X, Huang X, Xu X, Kou Y, Lv Y, Li Y, Liu Y, Wang Y, Zhang Y, Huang Y, Li Y, Wu Y, Liu Y, Pan Y, Zheng Y, Hong Y, Shi Y, Feng Y, Jiang Z, Han Z, Wu ZF, Liu Z (2025) Wan: Open and Advanced Large-Scale Video Generative Models. arXiv preprint arXiv:2503.20314

Wang D, Qiu L, Cai L, Jiang Y, Fan M, Wang Z (2026a) Visual Generative Artificial Intelligence (GenAI) in Mixed Reality (MR): A Survey of Techniques, Interaction Designs, and Evaluations. Int J Hum Comput Interact https://doi.org/10.1080/10447318.2026.2713822, published online 18 August 2026

Wang H, Cai S, Azzarelli A, Jiang Z, Huang G, Khoo ET, Seymour B, Zhang F, Bull D, Anantrasirichai N (2026b) OceanXL: Large-scale Underwater 3D Gaussian Splatting via Block Partitioning and Adaptive Pruning. In: SIGGRAPH Asia 2026 Conference Papers, accepted; technical version inspected: arXiv v1, 24 September 2026, arXiv:2609.29985v1

Wang H, Huang G, Zhang F, Bull D, Anantrasirichai N (2026c) Prune Wisely, Reconstruct Sharply: Compact 3D Gaussian Splatting via Adaptive Pruning and Difference-of-Gaussian Primitives. In: CVPR, pp 11716–11725

Wang H, Ouyang H, Wang Q, Wang W, Bai Q, Cheng KL, Yu Y, Li Y, Meng Y, Liu Z, Zeng Y, Shen Y, Chen Q (2026d) WorldDirector: Building Controllable World Simulators with Persistent Dynamic Memory. arXiv preprint https://doi.org/10.48550/arXiv.2607.02517, Version v1

Wang K, Savva M, Chang AX, Ritchie D (2018) Deep convolutional priors for indoor scene synthesis. ACM Trans Graph 37(4):1–14. https://doi.org/10.1145/3197517.3201362

Wang K, Lin YA, Weissmann B, Savva M, Chang AX, Ritchie D (2019) PlanIT: Planning and Instantiating Indoor Scenes with Relation Graph and Spatial Prior Networks. ACM Trans Graph 38(4):1–15. https://doi.org/10.1145/3306346.3322941

Wang L, Frisvad JR, Jensen MB, Bigdeli SA (2024a) StereoDiffusion: Training-Free Stereo Image Generation Using Latent Diffusion Models. In: CVPR Workshops (GCV), https://doi.org/10.1109/CVPRW63382.2024.00737

Wang P, Liu L, Liu Y, Theobalt C, Komura T, Wang W (2021a) NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. In: NeurIPS

Wang Q, Wang Z, Genova K, Srinivasan PP, Zhou H, Barron JT, Martin-Brualla R, Snavely N, Funkhouser T (2021b) IBRNet: Learning Multi-View Image-Based Rendering. In: CVPR, https://doi.org/10.1109/cvpr46437.2021.00466

Wang Q, Li W, Mou C, Cheng X, Zhang J (2024b) 360DVD: Controllable Panorama Video Generation with 360-Degree Video Diffusion Model. In: CVPR, https://doi.org/10.1109/cvpr52733.2024.00660

Wang X, Yuan H, Zhang S, Chen D, Wang J, Zhang Y, Shen Y, Zhao D, Zhou J (2023a) VideoComposer: Compositional Video Synthesis with Motion Controllability. In: NeurIPS, pp 7594–7611, https://doi.org/10.52202/075280-0334

Wang X, Cen Q, Bi W, Ma Y, Yi X, Xiao R, Fu X, Li H (2026e) Roomify: Spatially-Grounded Style Transformation for Immersive Virtual Environments. In: CHI, https://doi.org/10.1145/3772318.3791803

Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Trans Image Process 13(4):600–612. https://doi.org/10.1109/TIP.2003.819861

Wang Z, Lu C, Wang Y, Bao F, Li C, Su H, Zhu J (2023b) ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. In: NeurIPS

Wang Z, Wang Y, Chen Y, Xiang C, Chen S, Yu D, Li C, Su H, Zhu J (2024c) CRM: Single Image to 3D Textured Mesh with Convolutional Reconstruction Model. In: ECCV. Springer Nature Switzerland, pp 57–74, https://doi.org/10.1007/978-3-031-72751-1_4

Wang Z, Yuan Z, Wang X, Li Y, Chen T, Xia M, Luo P, Shan Y (2024d) MotionCtrl: A Unified and Flexible Motion Controller for Video Generation. In: SIGGRAPH Conference Papers, https://doi.org/10.1145/3641519.3657518

Wang Z, Lin H, Yoon J, Cho J, Zhang Y, Bansal M (2026f) AnchorWeave: World-Consistent Video Generation with Retrieved Local Spatial Memories. In: Computer Vision – ECCV 2026, pp 203–224, https://doi.org/10.1007/978-3-032-37356-4_12

Wang Z, Liu Z, Li J, Huang K, Xu B, Kang F, An M, Wang P, Jiang B, Wei Y, Xietian Y, Pei J, Hu L, Jiang B, Xue H, Wang Z, Sun H, Li W, Ouyang W, He X, Liu Y, Li Y, Zhou Y (2026g) Matrix-Game 3.0: Real-Time and Streaming Interactive World Model with Long-Horizon Memory. https://doi.org/10.48550/arXiv.2604.08995, Version v1, posted 10 April 2026

Warburton M, Mon-Williams M, Mushtaq F, Morehead JR (2023) Measuring motion-to-photon latency for sensorimotor experiments with virtual reality systems. Behav Res Methods 55(7):3658–3678. https://doi.org/10.3758/s13428-022-01983-5

van Waveren JMP (2016) The asynchronous time warp for virtual reality on consumer hardware. In: ACM VRST, pp 37–46, https://doi.org/10.1145/2993369.2993375

Weech S, Kenny S, Barnett-Cowan M (2019) Presence and Cybersickness in Virtual Reality Are Negatively Related: A Review. Front Psychol 10. https://doi.org/10.3389/fpsyg.2019.00158

Welch G, Foxlin E (2002) Motion tracking: no silver bullet, but a respectable arsenal. IEEE Comput Graph Appl 22(6):24–38. https://doi.org/10.1109/mcg.2002.1046626

Weng SCC, Chiou YM, Do EYL (2024) Dream Mesh: A Speech-to-3D Model Generative Pipeline in Mixed Reality. In: IEEE AIxVR, https://doi.org/10.1109/AIxVR59861.2024.00059

Weng SCC, Ma SY, Reinig S, Kadam P, Zhao AY, Banić A, Suzuki R, Do EYL (2026) Editing Reality: Designing In-Situ Co-Creation with Generative AI in Mixed Reality. In: DIS, https://doi.org/10.1145/3800645.3813087

Witmer BG, Singer MJ (1998) Measuring Presence in Virtual Environments: A Presence Questionnaire. Presence (Camb) 7(3):225–240. https://doi.org/10.1162/105474698565686

Wu J, Zhang C, Xue T, Freeman WT, Tenenbaum JB (2016) Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling. In: NIPS

Wu J, Yin S, Feng N, He X, Li D, Hao J, Long M (2024a) iVideoGPT: Interactive VideoGPTs are Scalable World Models. In: NeurIPS, arXiv:2405.15223

Wu JZ, Ge Y, Wang X, Lei SW, Gu Y, Shi Y, Hsu W, Shan Y, Qie X, Shou MZ (2023) Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. In: ICCV, pp 7623–7633, https://doi.org/10.1109/iccv51070.2023.00701

Wu K, Liu F, Cai Z, Yan R, Wang H, Hu Y, Duan Y, Ma K (2024b) Unique3D: High-Quality and Efficient 3D Mesh Generation from a Single Image. In: NeurIPS, https://doi.org/10.52202/079017-3974

Wu R, Gao R, Poole B, Trevithick A, Zheng C, Barron JT, Holynski A (2025a) CAT4D: Create Anything in 4D with Multi-View Video Diffusion Models. In: CVPR, https://doi.org/10.1109/cvpr52734.2025.02427

Wu R, Ye F, Liu Q, Huang G, Guo J, Hu R (2025b) Incrementally learning multiple diverse data domains via multi-source dynamic expansion model. https://doi.org/10.48550/arXiv.2501.08878

Wu T, Yang S, Po R, Xu Y, Liu Z, Lin D, Wetzstein G (2025c) Video World Models with Long-term Spatial Memory. In: NeurIPS

Xia F, Zamir AR, He Z, Sax A, Malik J, Savarese S (2018) Gibson Env: Real-World Perception for Embodied Agents. In: CVPR, pp 9068–9079, https://doi.org/10.1109/cvpr.2018.00945

Xia H, Li X, Li Z, Ma Q, Xu J, Liu MY, Cui Y, Lin TY, Ma WC, Wang S, Song S, Wei F (2026) SAGE: Scalable Agentic 3D Scene Generation for Embodied AI. In: CVPR, pp 22358–22368

Xiang F, Qin Y, Mo K, Xia Y, Zhu H, Liu F, Liu M, Jiang H, Yuan Y, Wang H, Yi L, Chang AX, Guibas LJ, Su H (2020) SAPIEN: A SimulAted Part-Based Interactive ENvironment. In: CVPR, pp 11097–11107, https://doi.org/10.1109/cvpr42600.2020.01111

Xiang J, Liu G, Gu Y, Gao Q, Ning Y, Zha Y, Feng Z, Tao T, Hao S, Shi Y, Liu Z, Xing EP, Hu Z (2024) Pandora: Towards General World Model with Natural Language Actions and Video States. arXiv preprint arXiv:2406.09455

Xiang J, Lv Z, Xu S, Deng Y, Wang R, Zhang B, Chen D, Tong X, Yang J (2025) Structured 3D Latents for Scalable and Versatile 3D Generation. In: CVPR, pp 21469–21480, https://doi.org/10.1109/cvpr52734.2025.02000

Xiang J, Chen X, Xu S, Wang R, Lv Z, Deng Y, Zhu H, Dong Y, Zhao H, Yuan NJ, Yang J (2026) Native and Compact Structured Latents for 3D Generation. In: CVPR, pp 14419–14429

Xiao Z, Lan Y, Zhou Y, Ouyang W, Yang S, Zeng Y, Pan X (2025) WorldMem: Long-term Consistent World Simulation with Memory. In: NeurIPS, Inspected arXiv:2504.12369v2

Xie H, Wen B, Chen Z, Hong F, Liu Z (2026a) 3D scene generation: A survey. Int J Comput Vis 134. https://doi.org/10.1007/s11263-026-02999-6, article 418

Xie L, Sun LC, Neall A, Wu T, Cai S, Wetzstein G (2026b) Generated Reality: Human-centric world simulation using interactive video generation with hand and camera control. In: CVPR Findings, https://doi.org/10.1109/cvprf71909.2026.00387, technical version inspected: 20 February 2026; arXiv:2602.18422v1

Xie Y, Yao CH, Voleti V, Jiang H, Jampani V (2025) SV4D: Dynamic 3D Content Generation with Multi-Frame and Multi-View Consistency. In: ICLR

Xing J, Xia M, Zhang Y, Chen H, Yu W, Liu H, Liu G, Wang X, Shan Y, Wong TT (2024) DynamiCrafter: Animating Open-Domain Images with Video Diffusion Priors. In: ECCV, https://doi.org/10.1007/978-3-031-72952-2_23

Xing K, Li L, Yin Y, Liang H, Luo G, Fang C, Wang J, Plataniotis KN, Jin X, Zhao Y, Wei Y (2026) StereoWorld: Geometry-Aware Monocular-to-Stereo Video Generation. In: CVPR

Xiong Z, Song Y, Kang H, Yan Q, Jiang L, Yang J, Fu Z, Fotiadis S, Wang A, Liu Z, Liu B, Yang Y, Lu X, Jacobs N (2026) ActWorld: From Explorable to Interactive World Model via Action-Aware Memory. arXiv preprint https://doi.org/10.48550/arXiv.2606.17730, Version v1

Xiu Y, Chilukuri J, Sen S, Gorlatova M (2025) Say It, See It: A Systematic Evaluation on Speech-Based 3D Content Generation Methods in Augmented Reality. In: IEEE ISMAR Adjunct (UNAI workshop), https://doi.org/10.1109/ISMAR-Adjunct68609.2025.00069

Xu J, Cheng W, Gao Y, Wang X, Gao S, Shan Y (2024a) InstantMesh: Efficient 3D Mesh Generation from a Single Image with Sparse-view Large Reconstruction Models. ArXiv preprint, arXiv:2404.07191

Xu J, Zou X, Huang K, Chen Y, Liu B, Cheng M, Shi X, Huang J (2024b) EasyAnimate: A High-Performance Long Video Generation Method based on Transformer Architecture. https://doi.org/10.48550/arXiv.2405.18991, arXiv preprint; Version v1

Xu J, Shao S, Bai L, Zhou Z, Huang G, Anantrasirichai N, Xie Z (2026a) Golden anchor for video generation. https://doi.org/10.48550/arXiv.2610.13217

Xu R, Xue T, Dong Q, Wan L, Zhu Z, Li P, Dou Z, Lin C, Xin S, Liu Y, Wang W, Komura T (2026b) MeshMosaic: Scaling Artist Mesh Generation via Local-to-Global Assembly. In: CVPR, pp 20003–20013

Xu Y, Shi Z, Wang Y, Chen H, Yang C, Peng S, Shen Y, Wetzstein G (2024c) GRM: Large Gaussian Reconstruction Model for Efficient 3D Reconstruction and Generation. In: ECCV, https://doi.org/10.1007/978-3-031-72633-0_1

Yan W, Zhang Y, Abbeel P, Srinivas A (2021) VideoGPT: Video Generation using VQ-VAE and Transformers. arXiv preprint arXiv:2104.10157

Yang B, Dong W, Ma L, Hu W, Liu X, Cui Z, Ma Y (2024a) DreamSpace: Dreaming Your Room Space with Text-Driven Panoramic Texture Propagation. In: IEEE VR, https://doi.org/10.1109/VR58804.2024.00085

Yang G, Huang X, Hao Z, Liu MY, Belongie S, Hariharan B (2019) PointFlow: 3D Point Cloud Generation with Continuous Normalizing Flows. In: ICCV, https://doi.org/10.1109/iccv.2019.00464

Yang S, Huang W, Chu R, Xiao Y, Zhao Y, Wang X, Li M, Xie E, Chen Y, Lu Y, Han S, Chen Y (2026) LongLive: Real-time Interactive Long Video Generation. In: ICLR, pp 90037–90059

Yang X, Man Y, Chen JK, Wang YX (2024b) SceneCraft: Layout-Guided 3D Scene Generation. In: NeurIPS

Yang Y, Jia B, Zhi P, Huang S (2024c) PhyScene: Physically Interactable 3D Scene Synthesis for Embodied AI. In: CVPR, https://doi.org/10.1109/cvpr52733.2024.01539

Yang Y, Sun FY, Weihs L, VanderBilt E, Herrasti A, Han W, Wu J, Haber N, Krishna R, Liu L, Callison-Burch C, Yatskar M, Kembhavi A, Clark C (2024d) Holodeck: Language Guided Generation of 3D Embodied AI Environments. In: CVPR, pp 16227–16237, https://doi.org/10.1109/cvpr52733.2024.01536

Yang Y, Guo YC, Huang Y, Zou ZX, Yu Z, Li Y, Cao YP, Liu X (2025a) HoloPart: Generative 3D Part Amodal Segmentation. ArXiv preprint, arXiv:2504.07943

Yang Z, Teng J, Zheng W, Ding M, Huang S, Xu J, Yang Y, Hong W, Zhang X, Feng G, Yin D, Zhang Y, Wang W, Cheng Y, Xu B, Gu X, Dong Y, Tang J (2025b) CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In: ICLR

Yao Z, Yao X, Sun Q, Xiong H, Lau KW, Mi H (2025) Gesture2Prop: Conditional VR Prop Generation from Hand Gestures and Speech. In: UIST Adjunct, https://doi.org/10.1145/3746058.3759017

Yariv L, Kasten Y, Moran D, Galun M, Atzmon M, Basri R, Lipman Y (2020) Multiview Neural Surface Reconstruction by Disentangling Geometry and Appearance. In: NeurIPS

Yariv L, Gu J, Kasten Y, Lipman Y (2021) Volume Rendering of Neural Implicit Surfaces. In: NeurIPS

Ye H, Zhang J, Liu S, Han X, Yang W (2023) IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. ArXiv preprint, arXiv:2308.06721

Ye T, Mao Y, Liao M, Liu J, Guo C, Du D, Shou Q, Zhu F, Guo S (2026) 3D generation for embodied AI and robotic simulation: A survey. https://doi.org/10.48550/arXiv.2604.26509, Version v3, 8 May 2026

Yeshwanth C, Liu YC, Nießner M, Dai A (2023) ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes. In: ICCV, pp 12–22, https://doi.org/10.1109/iccv51070.2023.00008

Yin T, Zhang Q, Zhang R, Freeman WT, Durand F, Shechtman E, Huang X (2025) From Slow Bidirectional to Fast Autoregressive Video Diffusion Models. In: CVPR, https://doi.org/10.1109/cvpr52734.2025.02138

Yu A, Ye V, Tancik M, Kanazawa A (2021) pixelNeRF: Neural Radiance Fields From One or Few Images. In: CVPR, https://doi.org/10.1109/cvpr46437.2021.00455

Yu HX, Duan H, Hur J, Sargent K, Rubinstein M, Freeman WT, Cole F, Sun D, Snavely N, Wu J, Herrmann C (2024a) WonderJourney: Going from Anywhere to Everywhere. In: CVPR, pp 6658–6667, https://doi.org/10.1109/cvpr52733.2024.00636

Yu HX, Duan H, Herrmann C, Freeman WT, Wu J (2025a) WonderWorld: Interactive 3D Scene Generation from a Single Image. In: CVPR, https://doi.org/10.1109/cvpr52734.2025.00555

Yu J, Qin Y, Wang X, Wan P, Zhang D, Liu X (2025b) GameFactory: Creating New Games with Generative Interactive Videos. In: ICCV, https://doi.org/10.1109/iccv51701.2025.01078

Yu L, Cheng Y, Sohn K, Lezama J, Zhang H, Chang H, Hauptmann AG, Yang MH, Hao Y, Essa I, Jiang L (2023) MAGVIT: Masked Generative Video Transformer. In: CVPR, https://doi.org/10.1109/cvpr52729.2023.01008

Yu W, Xing J, Yuan L, Hu W, Li X, Huang Z, Gao X, Wong TT, Shan Y, Tian Y (2025c) ViewCrafter: Taming Video Diffusion Models for High-fidelity Novel View Synthesis. IEEE Trans Pattern Anal Mach Intell https://doi.org/10.1109/TPAMI.2025.3613256

Yu Z, Chen A, Huang B, Sattler T, Geiger A (2024b) Mip-Splatting: Alias-free 3D Gaussian Splatting. In: CVPR, https://doi.org/10.1109/cvpr52733.2024.01839

Yuan H, Chen W, Cen J, Yu H, Liang J, Chang S, Lin Z, Feng T, Liu P, Xing J, Luo H, Tang J, Wang F, Yang Y (2026a) Lumos-1: On Autoregressive Video Generation with Discrete Diffusion from a Unified Model Perspective. In: ICLR, arXiv:2507.08801

Yuan J, Yang B, Wang K, Pan P, Ma L, Zhang X, Liu X, Cui Z, Ma Y (2026b) ImmerseGen: Agent-Guided Immersive World Generation with Alpha-Textured Proxies. IEEE Trans Vis Comput Graph 32(5):3530–3540. https://doi.org/10.1109/TVCG.2026.3679097

Zeng X, Vahdat A, Williams F, Gojcic Z, Litany O, Fidler S, Kreis K (2022) LION: Latent Point Diffusion Models for 3D Shape Generation. In: NeurIPS

Zeng X, Chen X, Qi Z, Liu W, Zhao Z, Wang Z, Fu B, Liu Y, Yu G (2024) Paint3D: Paint Anything 3D with Lighting-Less Texture Diffusion Models. In: CVPR, https://doi.org/10.1109/cvpr52733.2024.00407

Zhai G, Örnek EP, Chen DZ, Liao R, Di Y, Navab N, Tombari F, Busam B (2024) EchoScene: Indoor Scene Generation via Information Echo over Scene Graph Diffusion. In: ECCV, https://doi.org/10.1007/978-3-031-72664-4_10

Zhang C, Wu Q, Gambardella CC, Huang X, Phung D, Ouyang W, Cai J (2024a) Taming Stable Diffusion for Text to 360° Panorama Image Generation. In: CVPR, https://doi.org/10.1109/cvpr52733.2024.00607

Zhang H, Chen P, Xie X, Jiang Z, Wu Y, Li Z, Chen X, Sun L (2025) FusionProtor: A Mixed-Prototype Tool for Component-level Physical-to-Virtual 3D Transition and Simulation. In: CHI, https://doi.org/10.1145/3706598.3713686

Zhang J, Li X, Wan Z, Wang C, Liao J (2024b) Text2NeRF: Text-Driven 3D Scene Generation with Neural Radiance Fields. IEEE Trans Vis Comput Graph 30(12):7749–7762. https://doi.org/10.1109/TVCG.2024.3361502

Zhang L, Rao A, Agrawala M (2023) Adding Conditional Control to Text-to-Image Diffusion Models. In: ICCV, pp 3836–3847, https://doi.org/10.1109/iccv51070.2023.00355

Zhang L, Pan J, Gettig J, Oney S, Guo A (2024c) VRCopilot: Authoring 3D Layouts with Generative AI Models in VR. In: UIST. Association for Computing Machinery, pp 1–13, https://doi.org/10.1145/3654777.3676451

Zhang R, Isola P, Efros AA, Shechtman E, Wang O (2018) The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: CVPR, https://doi.org/10.1109/CVPR.2018.00068

Zhang X, Srinivasan PP, Deng B, Debevec P, Freeman WT, Barron JT (2021) NeRFactor: Neural Factorization of Shape and Reflectance Under an Unknown Illumination. ACM Trans Graph 40(6). https://doi.org/10.1145/3478513.3480496

Zhao S, Hu W, Cun X, Zhang Y, Li X, Kong Z, Gao X, Niu M, Shan Y (2024) StereoCrafter: Diffusion-based Generation of Long and High-fidelity Stereoscopic 3D from Monocular Videos. ArXiv preprint, arXiv:2409.07447

Zhao Z, Lai Z, Lin Q, Zhao Y, Liu H, Yang S, Feng Y, Yang M, Zhang S, Yang X, Shi H, Liu S, Wu J, Lian Y, Yang F, Tang R, He Z, Wang X, Liu J, Zuo X, Chen Z, Lei B, Weng H, Xu J, Zhu Y, Liu X, Xu L, Hu C, Huang T, Wang L, Zhang J, Chen M, Dong L, Jia Y, Cai Y, Yu J, Tang Y, Zhang H, Ye Z, He P, Wu R, Zhang C, Tan Y, Xiao J, Tao Y, Zhu J, Xue J, Liu K, Zhao C, Wu X, Hu Z, Qin L, Peng J, Li Z, Chen M, Zhang X, Niu L, Wang P, Wang Y, Kuang H, Fan Z, Zheng X, Zhuang W, He Y, Liu T, Yang Y, Wang D, Liu Y, Jiang J, Huang J, Guo C (2025) Hunyuan3D 2.0: Scaling Diffusion Models for High Resolution Textured 3D Assets Generation. https://doi.org/10.48550/arXiv.2501.12202, Version v1, dated 2025-01-21

Zhou J, Gao H, Voleti V, Vasishta A, Yao CH, Boss M, Torr P, Rupprecht C, Jampani V (2025) Stable Virtual Camera: Generative View Synthesis with Diffusion Models. In: ICCV, pp 12405–12414, https://doi.org/10.1109/ICCV51701.2025.01153

Zhou S, Fan Z, Xu D, Chang H, Chari P, Bharadwaj T, You S, Wang Z, Kadambi A (2024a) DreamScene360: Unconstrained Text-to-3D Scene Generation with Panoramic Gaussian Splatting. In: ECCV, https://doi.org/10.1007/978-3-031-72658-3_19

Zhou T, Tucker R, Flynn J, Fyffe G, Snavely N (2018) Stereo Magnification: Learning View Synthesis Using Multiplane Images. ACM Trans Graph 37. https://doi.org/10.1145/3197517.3201323

Zhou Y, Wang Q, Cai Y, Yang H (2024b) Allegro: Open the Black Box of Commercial-Level Video Generation Model. https://doi.org/10.48550/arXiv.2410.15458, arXiv preprint

Zhu Y, Deng K, Fauconnier JP, Navarro I, Li D, Pun A, Zhang Y, Zhuang P, Sun X, Agrawala M, Bhat K, Zhou T (2026) CubePart: An Open-Vocabulary Part-Controllable 3D Generator. In: SIGGRAPH Conference Papers, https://doi.org/10.1145/3799902.3811117