WorldWeave: Growing Persistent Geometric Worlds for Video Generation

Yifan Huang[1] Lifan Jiang[1] Qingyue Hao[1] Cheng Chen[1]
Boxi Wu[2] Xiaoxue Ren[1] Xiaofei He[1] Dehai Zhao[1,†]
[1]Zhejiang University [2]Daerwen AI
[†]Corresponding author.

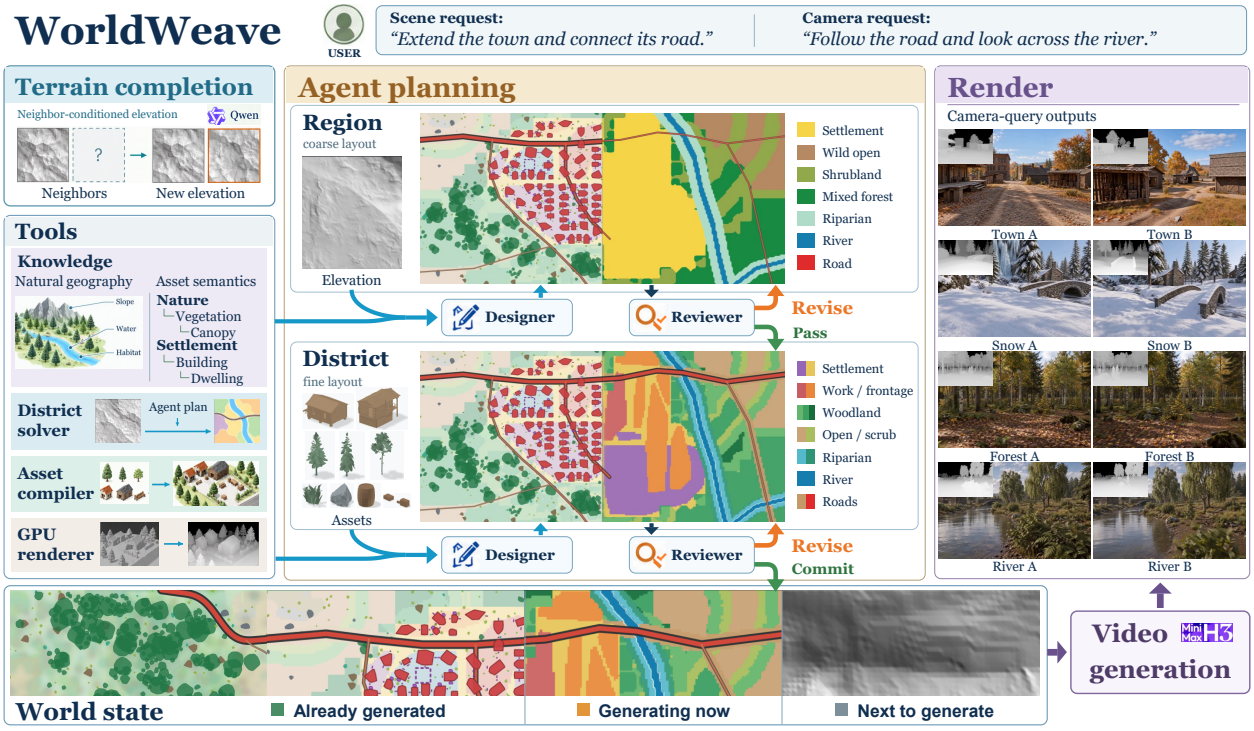

Figure 1: **Overview of WorldWeave.** Terrain completion extends metric terrain, agent planning organizes and validates coarse-to-fine world layouts, and rendering queries the persistent world state to guide video generation. Previously generated regions are preserved while new regions are progressively added.

ABSTRACT

Despite rapid progress, world models still lack explicit, persistent structural memory, making it difficult to preserve consistent world structure during continual scene expansion and cross-view revisits. To address this limitation, we present WORLDWEAVE, a world generation framework that decouples world-state maintenance from visual rendering. Specifically, WORLDWEAVE combines continual elevation-map generation with agent-guided scene organization and stitching to build an expandable explicit 3D world that incrementally extends structural memory while preserving existing structure. First, its terrain module uses diffusion-based image outpainting to generate continuous metric elevation maps under neighborhood conditioning and boundary constraints. Next, an agent integrates user intent, terrain evidence, and cross-region connectivity constraints to construct scenes through hierarchical semantic planning, deterministic geometry compilation, and local revision. Finally, during visual generation, planned camera trajectories query world geometry through a read-only interface, producing depth sequences that guide video synthesis without writing the generated results back into the world state. As a result, structural memory remains independent of short-window video generation, enabling continual expansion without predefined map boundaries and providing a consistent geometric basis for observations across trajectories and repeated visits.

1 Introduction

Generative world models have advanced rapidly in scene synthesis, view prediction and interactive environment generation, moving from local visual content toward explorable worlds (Duan et al., 2026; Chen et al., 2026; Huang et al., 2026; Zhu et al., 2026). Continual exploration requires more than realistic, temporally coherent observations: explored regions must persist outside the current view, new regions must connect to the existing environment, and revisits must encounter stable spatial structure. This requires structural memory beyond the current generation window, preserving world geometry, object identities and spatial relations as a common reference across time and viewpoints.

Despite improved visual generation, methods that rely primarily on images, video or finite observation histories still face challenges in maintaining explicit, persistent structural memory (Xiao et al., 2025; Ren et al., 2025; Yin et al., 2026). When observations implicitly carry world information, scene continuation depends on the generator re-inferring past content; local visual coherence does not directly establish lasting structural consistency. Expansion and revisitation expose two requirements: new regions must continue terrain, roads and waterways without altering established structure, and different observation trajectories must access the same world without reconstructing its geometric relations on each visit. A persistent world must therefore support both spatial addressing and relational composition: an extension must locate adjoining terrain, object support and continuing routes in a shared coordinate system. Persistence therefore governs both construction and observation.

We propose WORLDWEAVE, a world generation framework that decouples world-state maintenance from visual rendering. An explicit 3D world in a shared metric coordinate system serves as structural memory, separating incremental construction from observation generation. The state stores terrain, regional organization, asset identities and transforms, and boundary interfaces inherited by later regions. Each expansion constructs an independent candidate from the committed state; only validated additions create a new version, leaving existing structural records unchanged. Visual generation reads a specified version, allowing different camera trajectories to reuse the same geometry. World structure thus persists independently of the current video window, while construction can continue without predefined map boundaries.

WorldWeave first uses diffusion-based image outpainting to generate complete metric terrain chunks from committed neighbors. A multiscale residual codec referenced to neighboring terrain maps elevation to an image representation; height and first-derivative constraints confined to the new chunk join it continuously to the existing surface. An agent then integrates user intent, terrain evidence and inherited interfaces to plan region roles, functional districts and asset relations. Deterministic solvers and geometry compilers realize these plans as spatial layouts and infrastructure connections. Program checks and matched-view geometry previews guide bounded local revision, after which accepted candidates are atomically committed as new world versions. Planned camera trajectories query this state for depth-conditioned video synthesis, with no RGB write-back.

This separation also defines what must remain consistent across observations: the scene geometry and relations belong to the world, while appearance is synthesized for each viewing request. New construction can therefore inherit off-screen structure without recovering it from previous RGB frames. The same distinction supports controlled evaluation: camera motion should reveal new content while retaining the organization of previously observed regions. We therefore assess structural consistency and revisit memory alongside camera compliance and visual quality, connecting preservation of scene organization with the ability to explore it. Figure 1 illustrates how scene requests and camera requests operate on this shared spatial reference.

Our contributions are threefold:

- **Persistent structural-memory framework.** We decouple world-state maintenance from visual rendering, supporting incremental construction that preserves existing structure and shared geometric queries for observations across trajectories and revisits.
- **Continual metric terrain expansion.** We combine diffusion-based image outpainting, multiscale elevation-residual encoding, and new-side boundary constraints to generate complete metric terrain chunks and append them continuously.
- **Agent-guided scene construction.** We combine hierarchical semantic planning with deterministic geometry compilation, bounded local revision and validated commit to extend scenes following user intent, terrain conditions and inherited cross-region connections.

2 Related Work

Video world models and structural memory. Echo-WM combines metric camera control with geometry conditioning (Duan et al., 2026); Zing-0.5 uses text, keyboard control and causal caching (Chen et al., 2026). SolarWM-5B uses

camera-conditioned causal training (Huang et al., 2026), while SANA-WM uses hybrid linear attention for minute-long generation (Zhu et al., 2026).

WorldMem retrieves frames by their states (Xiao et al., 2025); ViewCrafter, WVD and Geometry-as-context condition view synthesis on geometry (Yu et al., 2025b; Zhang et al., 2025; Hu et al., 2026). GEN3C unprojects previous observations into a 3D cache and renders it to guide generation (Ren et al., 2025), while WorldStereo and Lyra 2.0 connect views through geometric memory or correspondences (Zhang et al., 2026b; Shen et al., 2026). Alaya-EVOKE-Turbo estimates geometry from generated frames and updates a camera-indexed world-state bank (Yin et al., 2026); AlayaWorld combines a 3D cache with compressed frame history (AlayaWorld Team et al., 2026). WorldWeave instead maintains memory through validated scene construction, with video generation reading committed geometry.

Persistent and expandable 3D worlds. Persistent Nature uses an extendable layout grid and camera-independent decoder (Chai et al., 2023). Text2Room, WonderJourney and WonderWorld grow scenes through view synthesis and geometric integration (Höllein et al., 2023; Yu et al., 2024, 2025a); ScenePainter tracks concept relations to limit semantic drift (Xia et al., 2025), and One2Scene establishes a shared Gaussian scaffold before generating novel views (Wang et al., 2026b). InfiniCube follows a construction-first approach, expanding semantic voxel worlds from HD maps and vehicle boxes and rendering guidance buffers for driving videos (Lu et al., 2025). XCube and LT3SD use sparse voxels and latent trees (Ren et al., 2024; Meng et al., 2025); BlockFusion, SceneFactor and NuiScene extend spatial blocks (Wu et al., 2024; Bokhovkin et al., 2025; Lee et al., 2025). We preserve committed structure and inherit terrain and scene interfaces.

Agent-guided incremental scene construction. Infinigen supplies procedural environments (Raistrick et al., 2023, 2024); 3D-GPT, SceneCraft and SceneX expose construction through language-driven programs (Sun et al., 2025; Hu et al., 2024; Zhou et al., 2025). Holodeck supports language- and vision-guided layouts (Yang et al., 2024; Bian et al., 2025); Scenethesis combines visual guidance with geometric constraints, while scene graphs and hierarchical motifs organize relations (Ling et al., 2026; Liu et al., 2025; Pun et al., 2026). WorldClaw builds region-aware terrain and places editable assets with rendering feedback (Guo et al., 2026). WorldGen imposes navigational structure to obtain traversable worlds (Wang et al., 2026a); Code2Worlds separates object generation from environmental orchestration and refines scene programs through feedback (Zhang et al., 2026a). Our planning inherits neighboring interfaces and uses deterministic compilation and bounded revision before committing additions.

Terrain generation and boundary continuity. Hydrology-based modeling organizes terrain around drainage (Génevaux et al., 2013). Learned approaches include sketch-conditioned diffusion with upscaling (Lochner et al., 2023), joint height–texture synthesis with separate latent encoders (Higo et al., 2025) and text-conditioned generation trained on geospatial data (Borne-Pons et al., 2025). InfiniteDiffusion supports seed-consistent random access through overlapping queries (Goslin, 2026). Our fixed-order expansion instead generates metric chunks against committed neighbors and joins them through new-side height and derivative constraints.

3 Method

3.1 Persistent Structural Memory

WorldWeave starts from a seed terrain chunk and maintains an explicit world in a shared metric coordinate system. This structural memory has committed version $M_k = (\mathcal{T}_k, \mathcal{P}_k, \mathcal{I}_k, \mathcal{J}_k)$ after k accepted additions: $\mathcal{T}_k$ stores elevation and boundary derivatives, $\mathcal{P}_k$ stores regional organization and object relations, $\mathcal{I}_k$ stores asset mesh references, identities and world transforms, and $\mathcal{J}_k$ stores exposed boundary interfaces. These interfaces specify terrain attachment and the positions and types of road or water connections inherited by new regions.

An expansion request specifies intent u and frontier chunk q. Construction reads M_k and forms an independent candidate ΔM_q through terrain expansion and hierarchical scene compilation (Figure 2), using the committed terrain and inherited interfaces. Validation and commit (Section 3.4) determine whether the candidate becomes a new version. A camera query selects an existing committed version; changing the observation does not change its structural records. The state links semantic and geometric descriptions through stable instance identities. Regional plans specify which parts of the terrain serve each function, while instance transforms place the corresponding assets in the same metric frame. Exposed interfaces transfer this organization across chunk boundaries: a new region receives both the neighboring surface and the road or water connections it must extend. This gives local construction a consistent global reference as the map grows.

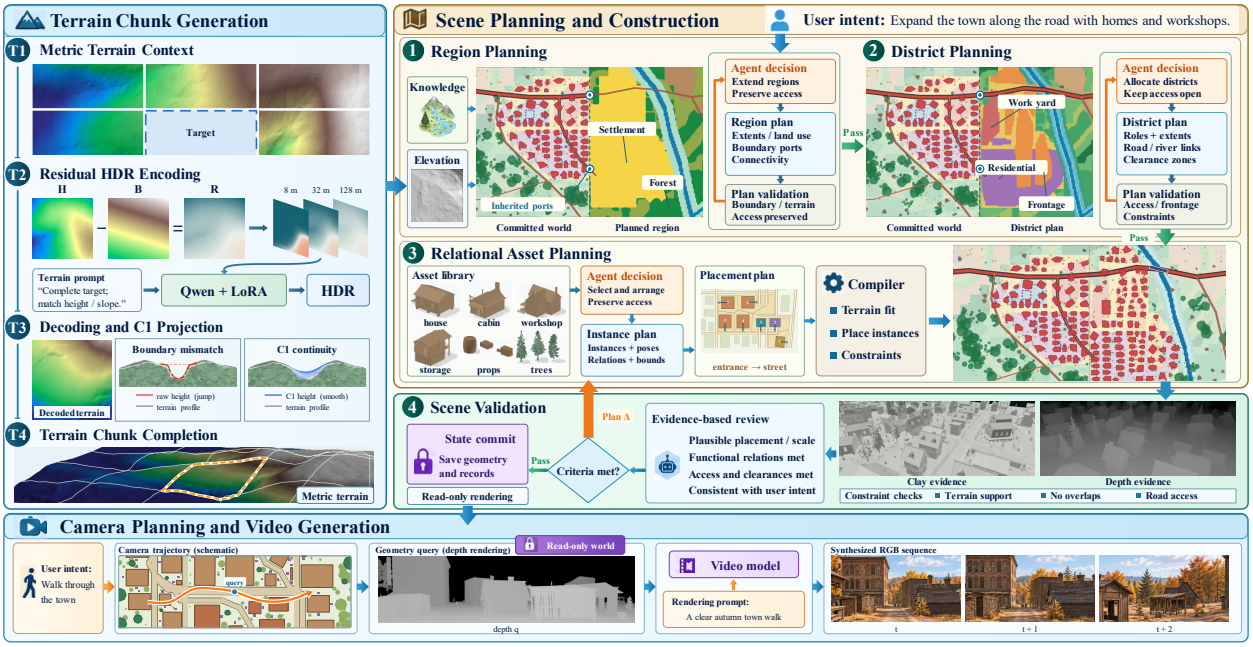

Figure 2: **WorldWeave pipeline.** Neighbor-conditioned terrain generation uses residual HDR encoding and new-side joining. The agent plans regions, districts and asset relations, while deterministic tools compile geometry and provide structural and visual evidence for revision. Validated additions extend persistent world state; camera trajectories query its geometry to condition video generation without writing RGB back into the world.

3.2 Continual Metric Terrain Expansion

A candidate terrain chunk contains 512×512 elevation samples at 0.5 m spacing, giving a nominal 256×256 m footprint. Its context comprises one to three committed axial neighbors: one side, two opposite sides, two adjacent sides, or three sides. We generate a complete target in one image-outpainting call and repeat this operation in a fixed expansion order, using previously committed outputs as subsequent context.

To encode metric height within the image model's dynamic range, we fit a reference surface $b_q(x, y)$ from valid neighboring boundary bands. The unknown target does not enter this fit. For elevation h_q, the residual r_q is encoded through three monotone channels:

$$r_q = h_q - b_q, \qquad c_{q,j} = \tfrac{1}{2} + \tfrac{1}{2}\tanh(r_q/s_j), \qquad (s_1, s_2, s_3) = (8, 32, 128)\,\text{m}. \tag{1}$$

Each encoded sample is repeated in a 2×2 image block before VAE encoding, adding redundancy without changing the metric sampling grid. The reference surface carries the elevation level and broad trend supplied by the context, leaving the residual channels to describe local relief. Small channel scales allocate greater sensitivity to subtle variations, whereas larger scales retain information over stronger relief. The same reference is restored after decoding, so neighboring chunks remain expressed in world units even though generation takes place in image space.

We adapt Qwen-Image-Edit-2509 (Wu et al., 2025) with rank-32 LoRA (Hu et al., 2021), freezing its VAE and text conditioner. Encoded neighbors and a fixed completion prompt condition generation. We arrange the context on a spatial canvas, retaining neighbor directions around a neutral target slot. This exposes opposite or adjacent boundary conditions in one layout, allowing the target to be completed jointly against all attached sides. Let $\mathcal{V}_q$ index valid target loss locations, $\ell_q(x)$ denote the local diffusion loss, and $d_q(x)$ the metric distance to the nearest attached edge. We use

$$\begin{aligned} w_q(x) &= 1 + \tfrac{3}{2}[1 + \cos(\pi \min\{d_q(x)/a, 1\})]\,, \\ \mathcal{L}_q &= \frac{\sum_{x \in \mathcal{V}_q} w_q(x)\,\ell_q(x)}{\sum_{x \in \mathcal{V}_q} w_q(x)}, \qquad a = 16\,\text{m}. \end{aligned} \tag{2}$$

Weights are evaluated on the loss grid. The nearest-edge distance implements the maximum over intersecting edge bands; normalization keeps loss scales comparable across targets.

After VAE decoding, each channel gives a residual estimate $\hat{r}_{q,j} = s_j \operatorname{atanh}(2\hat{c}_{q,j} - 1)$, with channel values clipped inside $(0, 1)$. We combine the repeated estimates using inverse-sensitivity weighting, downweighting near-saturated channels, and refine the result with five robust projection iterations onto the three-channel encoding curve. Adding b_q

recovers the metric prediction $\hat{h}_q = b_q + \hat{r}_q$ on the original sampling grid. For each attached side $e \in \mathcal{E}_q$, the neighbor supplies height g_e and directional derivative v_e on the world-coordinate interface Γ_e, using a shared normal n_e. The joined surface satisfies

$$\begin{aligned} h_q^+ &= \hat{h}_q + \delta_q, \qquad \operatorname{supp}(\delta_q) \subseteq \mathcal{B}_q, \\ h_q^+\big|_{\Gamma_e} &= g_e, \qquad \partial_{n_e} h_q^+\big|_{\Gamma_e} = v_e, \qquad e \in \mathcal{E}_q. \end{aligned} \tag{3}$$

Here $\mathcal{B}_q$ includes the new-side height-transition band and connector cells between existing sample-center grids. New connector cells are bicubic Hermite patches whose old-facing edges inherit boundary heights and derivatives (Figure 3); slope correction uses an independent taper. Corrections vanish at their inner support boundaries, with $\delta_q = \partial_n \delta_q = 0$. Committed geometry is preserved, and modification magnitude is reported separately from continuity (Appendix D). Connector cells bridge the gap between the old and new sample-center grids, providing a shared boundary for subsequent terrain and scene construction. Height correction adjusts the offset between chunks, whereas slope correction controls the approach to the inherited boundary slope. Separate support widths let the surface satisfy both conditions while confining derivative-induced displacement near the seam.

3.3 Hierarchical Scene Planning and Compilation

Deterministic terrain analysis extracts slope, local relief, drainage and buildable-area evidence from the accepted elevation field. The agent combines this evidence with u, inherited interfaces $\mathcal{J}_k$, and available asset capabilities to assign broad region roles, such as settlement, woodland or river corridor. A region solver converts these semantic assignments into spatial extents. After region review (Section 3.4), the agent specifies district roles, adjacency and access requirements. A deterministic solver resolves their extents and infrastructure under inherited road and water interfaces, producing functional or ecological subdivisions with the access, clearance and terrain conditions required by downstream assets.

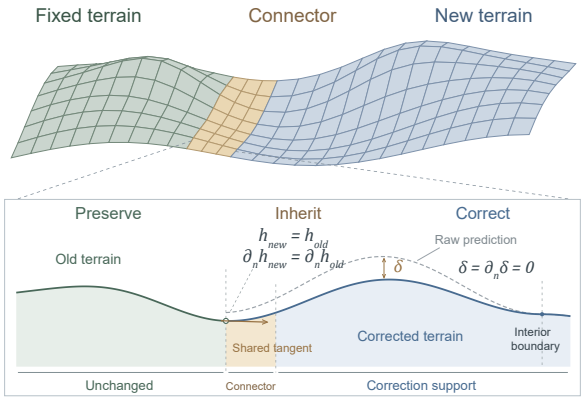

Figure 3: New-side terrain joining. Old samples remain fixed; newly owned connector cells inherit boundary height and normal derivative. The correction vanishes at the interior support boundary. Schematic.

Within each district, the agent turns its functional requirements into an asset-relation plan. The plan assigns primary, companion and service roles, specifies shared spaces and orientation, and connects buildings and vegetation groups to the district's roads and ecological zones. The typed catalog supplies each asset's dimensions, support anchors, entrance direction and required context. These attributes connect semantic choices to the spatial constraints inherited from regional and district planning. The hierarchy passes explicit commitments between scales: region boundaries delimit admissible land use, district connections reserve routes through those regions, and asset relations specify how individual instances participate in that organization. The deterministic compiler resolves the plan into stable instance identities and metric transforms under the spatial and asset contracts in Appendix E. It fits assets to supporting surfaces, realizes infrastructure connections and vegetation groupings, and preserves access corridors and collision clearance. For n_q candidate instances with required geometric relations $\mathcal{R}_q$, acceptance requires

$$\mathcal{I}_q = \{(a_i, T_i, \rho_i, d_i)\}_{i=1}^{n_q}, \qquad g_r(\mathcal{I}_q, h_q^+; M_k) = 1 \quad (r \in \mathcal{R}_q). \tag{4}$$

Here a_i identifies catalog geometry, T_i its world transform, ρ_i its semantic role and d_i its district. Each predicate g_r checks a catalog-supported relation against the candidate and committed context. Failed predicates retain instance and relation identifiers, locating the plan entries to revise. The plan, geometry and checks form the scene portion of ΔM_q.

3.4 Bounded Revision and Validated Commit

Region review checks inherited connections, coverage and terrain feasibility before district refinement. After compilation, program checks evaluate support, collision and connectivity in parallel with matched-view GPU geometry and depth previews. The tests establish geometric validity; matched viewpoints show whether the compiled arrangement realizes the intended regional organization. The agent receives both for the same candidate, allowing a revision to address the responsible functional area or obstructed connection while retaining unaffected content. Accepted regions then expose their outer interfaces as context for the next expansion, linking each local planning cycle to continued world

construction. Review evidence is bound to the candidate revision: after an edit, program reports and matched-view previews are regenerated for that revision. This keeps acceptance tied to the geometry that will be committed.

Let $\pi_q^{(t)}$ be the candidate plan after t revisions and K the revision budget. A failed review returns a structured delta $\delta\pi_q^{(t)}$ targeting the responsible region, district or asset relation:

$$\pi_q^{(t+1)} = \mathrm{Edit}(\pi_q^{(t)}, \delta\pi_q^{(t)}), \qquad \Delta M_q^{(t+1)} = \mathrm{Compile}(\pi_q^{(t+1)}; M_k), \qquad t < K. \tag{5}$$

Edit updates the indicated plan entries; Compile recomputes affected geometry with M_k read-only. Reviews use fixed acceptance criteria and reject candidates that exhaust the revision budget.

When all geometric checks and agent reviews pass, the candidate is atomically published as a new world version. Let Ω_k denote the terrain, semantic-plan, instance and boundary-geometry records already committed in M_k. The update appends accepted records while preserving those fields:

$$M_{k+1} = M_k \oplus \Delta M_q, \qquad M_{k+1}\big|_{\Omega_k} = M_k\big|_{\Omega_k}. \tag{6}$$

Here $\oplus$ denotes the validated addition, including new exposed interfaces for subsequent expansion. The equality preserves the terrain, organization and instance records reused by future expansions and observations. Visibility is computed from the complete geometry of the selected version.

3.5 Read-only Queries for Video Generation

Observation generation selects a committed world version independently of expansion. A viewing request specifies desired motion and targets, either supplied directly by the user or sampled by the agent from the user's intent. The camera planner converts this request into intrinsics, poses and timing $C_{1:T}$, then checks the trajectory for clearance and the required viewpoints. GPU queries resolve nearest visible intersections across terrain and asset meshes, returning metric depth with its convention, validity mask and calibration. Per-pixel hit identities associate queried surfaces with persistent asset records. They bind named camera targets to world geometry when assessing visibility along a trajectory. The resulting RGB video visualizes this world version.

As the camera moves, visible surfaces are recomputed from this common geometry, providing spatially coordinated conditions throughout the trajectory. For video generator F, let A_F adapt depths from query Q. Its supported optional style input s_F comprises text, reference images, or both:

$$D_{1:T} = Q(M_k, C_{1:T}), \qquad V_{1:T} = F\big(A_F(D_{1:T}), s_F\big). \tag{7}$$

A surface point X retains its world coordinates in the selected version. For two views in which it is visible, the homogeneous image coordinates obey

$$\widetilde{x}_t(X) \sim K_t R_t^\top (X - c_t), \qquad \widetilde{x}_s(X) \sim K_s R_s^\top (X - c_s). \tag{8}$$

Here K_t, R_t and c_t denote intrinsics, camera-to-world rotation and camera center. Both projections refer to the same terrain or asset surface, supplying a shared geometric correspondence through viewpoint changes and revisits. The depth sequence encodes how camera motion changes visibility, including occlusions and the boundaries of terrain and assets.

The adapter converts metric depth to the generator's required encoding, resolution and frame rate while preserving camera calibration. Available style references control appearance, while depth supplies the spatial structure along the requested trajectory. Generated RGB never writes back into world state. Appendix F specifies revision, publication and observation.

4 Experiments

4.1 Experimental Setup

Benchmark. We evaluate 120 tasks covering forward exploration, viewpoint changes, and occlusion or off-screen return. Terrain relief, vegetation biomes (e.g., forests and grasslands), and settlement layouts (e.g., villages and towns) vary across scenes. Seasonal appearance and lighting are varied as rendering attributes. Videos are evaluated over 15 seconds at 12 fps.

Baselines. We compare with six open-source world models: SANA-WM, Zing, SolarWM-5B, AlayaWorld, Alaya-EVOKE-Turbo and JoyAI-Echo-1.5 (WM) (Zhu et al., 2026; Chen et al., 2026; Huang et al., 2026; AlayaWorld Team et al., 2026; Yin et al., 2026; Duan et al., 2026), and four video baselines: Seedance 2.0 (closed-source), Wan3.0, Kling 3.0 and our base generator MiniMax-H3. Since world models commonly require a first frame and camera controls, we provide them with the same first frame as the WorldWeave-generated video and the camera trajectory used to produce its depth input. All models receive the same task and scene-description prompt. WorldWeave additionally receives depth queried from the constructed world.

Figure 4: World-grounded video generation across four scenes. Each row follows a viewing trajectory from left to right, showing revisitation or continued exploration. Insets show depth guidance, and numbered markers identify corresponding scene elements across views.

Metrics. We assess visual quality with IQ and AQ from VBench (Huang et al., 2024), and revisit memory and camera compliance with SMC and Cam. We jointly calibrate smoothness and geometric consistency for realized motion, yielding MN-MS, MC-GeCo (Gu et al., 2026), MC-MEt3R (Asim et al., 2025) and MC-GeoCon (Xue et al., 2026), to reduce systematic scoring biases across different camera-motion speeds. Appendix B defines these metrics and their motion calibration.

4.2 Main Quantitative Results

Table 1 compares visual quality, structural consistency and camera control. WorldWeave ranks fifth in IQ and second in AQ, while leading MN-MS, MC-GeCo, SMC, Cam, MC-MEt3R and MC-GeoCon. Its strongest gains therefore concern maintaining an explorable world: even with greater mean image-space motion than six comparison methods, WorldWeave retains leading structural-consistency and camera-compliance scores alongside competitive appearance and smooth motion. Figure 4 illustrates the corresponding video sequences across revisitation and exploration tasks.

The Base comparison examines the benefit of world-grounded conditioning within the same video-model family. Relative to MiniMax-H3, WorldWeave improves all eight measures. The largest gains are in SMC (+426.7%) and Cam (+42.2%), together with a 32.2% reduction in MC-GeoCon error. IQ and AQ also improve, supporting both spatial organization and appearance.

4.3 Structural Consistency and Camera Compliance

Geometric consistency. WorldWeave achieves the lowest MC-GeCo and MC-MEt3R errors, 0.0573 and 0.122, improving over MiniMax-H3 by 9.6% and 12.5%, respectively. The lower motion-calibrated errors indicate more coherent geometry and reprojected features across generated views. MC-GeoCon corroborates this result through geometric correspondences. These results support transferring spatial organization from shared depth guidance into generated video.

Memory across revisits. WorldWeave obtains the highest SMC score, 2.30 compared with 2.21 for the runner-up Echo-WM, while MiniMax-H3 scores 0.438. This improvement strengthens memory over disappearance and return. When a target leaves the view or becomes occluded, later observations must recover its identity and spatial relations. Reusing a committed world version provides the same target geometry and surrounding layout on return, reducing the need to reconstruct these relations from a limited visual history. Appendix G provides matched occlusion and return sequences in farmstead, town and woodland scenes.

Camera compliance. WorldWeave also ranks first in Cam at 72.3, ahead of Seedance 2.0 by 2.49 points. Relative to MiniMax-H3's 50.8, this is a 42.2% increase in score, reflecting stronger adherence to the requested translations,

Table 1: World-video comparison. The first six methods are open-source world models; the remaining baselines are video models. Bold/underline indicate best/second-best scores; yellow highlights the three largest relative gains over MiniMax-H3 (Base), with signed percentage changes.

Method	Visual quality		Memory and camera control		Structural consistency			
	Imaging quality↑	Aesthetic quality↑	Structural memory↑	Camera compliance↑	MN-MS↑	MC-GeCo↓	MC-MEt3R↓	MC-GeoCon↓
SANA-WM	0.739	0.619	0.328	55.5	0.973	0.113	0.203	0.189
Zing	0.763	0.685	0.989	55.8	0.974	0.0698	<u>0.129</u>	0.142
SolarWM-5B	0.755	0.604	0.370	54.7	0.956	0.0786	0.166	0.217
AlayaWorld	0.784	0.641	1.51	56.4	0.896	0.926	0.239	0.295
EVOKE-Turbo	0.784	**0.718**	1.12	61.6	0.961	0.105	0.182	0.249
Echo-WM	**0.791**	0.655	<u>2.21</u>	63.5	0.968	<u>0.0604</u>	0.135	0.135
Seedance 2.0	0.717	0.617	2.14	<u>69.8</u>	<u>0.980</u>	0.101	0.170	0.307
Wan3.0	<u>0.786</u>	0.702	1.63	59.1	0.938	0.146	0.184	<u>0.131</u>
Kling 3.0	0.739	0.588	2.09	59.3	0.978	0.0745	0.155	0.163
MiniMax-H3 (*Base*)	0.725	0.617	0.438	50.8	0.979	0.0633	0.140	0.178
WorldWeave	0.781	<u>0.704</u>	**2.30** (+426.7%)	**72.3** (+42.2%)	**0.981**	**0.0573**	**0.122**	**0.121** (-32.2%)

turns and returns. On the 120 matched tasks, its mean optical-flow amplitude exceeds those of SANA-WM, EVOKE-Turbo, AlayaWorld, Seedance 2.0, Wan3.0 and Kling 3.0 by 8.1–24.1%. Even with this greater realized motion, WorldWeave maintains leading structural-consistency scores, supporting exploration with coherent changes in viewpoint. Appendix B.2 contrasts favorable raw scores with texture degradation or limited motion.

4.4 Ablation Studies

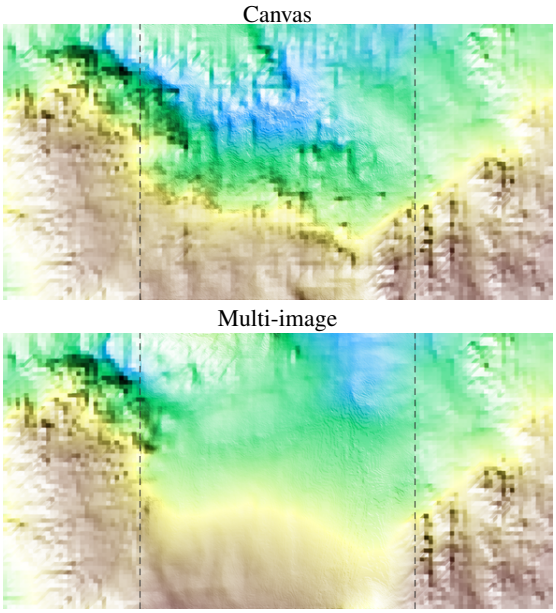

Figure 5: Opposite-neighbor completion with canvas (top) and multi-image inputs (bottom). Additional cases: Appendix A.2.

We evaluate elevation encoding, terrain conditioning and joining, followed by planning-model substitution and video-generation seed variation (Appendix A).

Metric elevation encoding. Using residual rather than absolute grayscale reduces mean VAE round-trip RMSE from 2.182 to 0.469 m (Table 2(a)). Multiscale residual channels further reduce it to 0.158 m, and the full codec reaches 0.136 m, a further 14.0% reduction. Referencing neighboring terrain removes the large elevation offset, while the channel scales retain sensitivity across different relief levels. These reductions demonstrate improved elevation-codec fidelity: context referencing and multiscale channels jointly preserve metric height through the frozen image VAE.

Spatially arranged terrain conditioning. Canvas conditioning yields lower boundary-height gaps in all four neighbor configurations (Table 2(b)), reducing them by 65.4–76.1% relative to independent multi-image references. The remaining gaps are 0.264–0.368 m for nominally 256 m chunks. Slope continuity also improves in the two- and three-neighbor settings, although the single-neighbor case favors multi-image input. This pattern supports arranging neighboring terrain in its spatial context, especially when several boundaries must be satisfied together. Figure 5 shows the opposing-neighbor case; Appendix A.2 compares all configurations. Training configurations and evaluation details are specified in Appendix A.1.

Continuous terrain joining. Table 2(c) compares four settings on 256 matched targets. Height-only joining closes the height gap but leaves slope residual 0.359; full derivative inheritance reduces both residuals below numerical tolerance. Independent slope taper preserves continuity while reducing mean whole-target modification from 0.2088 to 0.0825 m (60.5%). All joining variants connect every target and preserve valid existing heights and derivatives, supporting separate height and slope transition widths (Appendix D.4).

Table 2: Terrain ablations of (a) elevation encoding, (b) input organization and (c) terrain joining. Δh denotes whole-target RMS modification.

(a) Elevation encoding

Reference subtraction	Multiscale HDR	Spatial repetition	RMSE (m)↓
–	–	–	2.182
–	✓	–	1.935
✓	–	–	0.469
✓	✓	–	0.158
✓	✓	✓	**0.136**

(b) Terrain input

Input	Height gap (m)↓		Slope gap (m/m)↓	
	Multi	Canvas	Multi	Canvas
One	0.763	**0.264**	**0.240**	0.278
Opposite	1.394	**0.333**	0.373	**0.338**
L-shaped	1.076	**0.296**	0.310	**0.275**
Three	1.214	**0.368**	0.361	**0.328**

(c) Terrain joining

Height correction	Derivative inheritance	Slope taper	$H_c \downarrow$ (m)	$S_c \downarrow$ (m/m)	Δh (m)
–	–	–	0.319	0.638	0
✓	–	–	0	0.359	0.0773
✓	✓	–	0	0	0.2088
✓	✓	✓	0	0	**0.0825**

Planning model. Replacing the planner changes settlement density and infrastructure while retaining the shared terrain and cross-region organization. Both planners continue inherited roads: the alternative favors woodland and small settlement groups, while the default extends a denser street network. Appendix A.3 compares full layouts with identical terrain and the same existing region.

Random-seed sensitivity. Across five fixed scene–trajectory pairs and five generation seeds, the seed-wise mean IQ, AQ and MN-MS have coefficients of variation below 0.4%; those of the three geometric metrics range from 3.1% to 6.3%. Thus, the aggregate computational measures remain comparatively stable as the noise realization changes. Event-based SMC and Cam vary more strongly. Appendix A.4 provides per-task scores, seed-wise means and dispersion, including the observed black-sky case under fixed geometry, first frame, prompt and depth.

4.5 User Study

	SANA WM	Zing	Solar WM	Alaya World	EVOKE Turbo	Echo WM	**Ours**
Camera consistency	2.65	2.76	1.99	3.23	2.84	3.92	**4.57**
Object consistency	1.52	2.88	1.49	3.06	2.52	3.42	**4.73**
Structure/ texture	1.61	2.85	1.84	2.68	2.64	3.32	**4.52**

Table 3: User ratings (1–5; higher is better).

We conduct a blinded study of WorldWeave and six world models on 20 selected scene–trajectory cases (140 videos). Participants score camera, object, and structure/texture consistency from 1 to 5. Each video receives three or four ratings, totaling 436 evaluations. Table 3 reports video means averaged equally across cases (Appendix C). WorldWeave leads with 4.57, 4.73 and 4.52, compared with 3.92, 3.42 and 3.32 for Echo-WM. The gains over Echo-WM are 0.65, 1.31 and 1.20 points.

5 Conclusion

We presented WorldWeave, a world generation framework that decouples persistent structural memory from visual synthesis. Continual metric terrain expansion and agent-guided scene compilation extend an explicit world while preserving previously committed structure. Hierarchical planning connects user intent, terrain evidence and inherited interfaces; deterministic compilation, bounded revision and validated publication turn these decisions into reusable geometry. Read-only camera queries then provide depth conditions for video generation without writing synthesized observations back into world state. Across the evaluated methods, WorldWeave leads the structural-consistency and camera-compliance measures while retaining competitive visual quality. Human ratings on the 20 selected cases also favor WorldWeave in camera, object and structure/texture consistency. These results support maintaining a shared geometric world independently of individual video-generation windows, including observations with greater realized motion than most baselines.

Limitations and future work. Our current focus is static geometry, and depth provides guidance rather than a hard rendering constraint. Future work will address dynamic interactions, long-horizon appearance memory and adaptive spatial detail for larger worlds.

AI Use Statement

Generative AI tools assisted with drafting sections of the manuscript, organizing and polishing the writing, and literature retrieval and discovery, including identifying potentially relevant related work. The authors are responsible for checking AI-assisted text and references against the underlying sources, and for the accuracy, originality and integrity of the final manuscript.

Ethics Statement

Our evaluation uses virtual environments and generated videos. Participants in the user study provide informed consent and evaluate anonymized method outputs. The terrain data and geometry assets are drawn from authored virtual environments, not presented as surveyed real-world geography. Third-party assets, pretrained models and services remain subject to their respective licenses and terms; redistribution requires the corresponding permissions. As with other visual generation systems, outputs could be used to create misleading depictions. Applications should clearly disclose synthetic content and preserve provenance, and should not treat generated worlds as verified representations of real locations.

Reproducibility Statement

Section 3 defines the world representation, incremental construction and read-only observation interfaces; Section 4 specifies the evaluation setting and reports the principal comparisons. Appendices D–F detail terrain encoding and joining, scene compilation, revision and geometric queries. Appendix B defines the metrics and Appendix A.1 specifies ablation protocols; Appendices A.2 and A.4 provide matched terrain visualizations and per-task seed results. Appendix C specifies user-study selection and score aggregation.

References

AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Mingliang Zhai, Ruicong Liu, Xiaojie Xu, Xuangeng Chu, Zhen Li, Zhengyuan Lin, Zhixiang Wang, Zian Meng, and Zihui Gao. AlayaWorld: Interactive Long-Horizon World Modeling - Full Technical Report (v1.1), 2026. URL https://arxiv.org/abs/2608.13492. Published: arXiv preprint.

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. MEt3R: Measuring Multi-View Consistency in Generated Images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6034–6044. IEEE, 2025. URL https://arxiv.org/abs/2501.06336.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. *arXiv preprint arXiv:2301.08243*, 2023. URL https://arxiv.org/abs/2301.08243.

Mahmoud Assran, Adrien Bardes, David Fan, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. URL https://arxiv.org/abs/2506.09985.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024. URL https://arxiv.org/abs/2404.08471.

Zixuan Bian, Ruohan Ren, Yue Yang, and Chris Callison-Burch. HOLODECK 2.0: Vision-Language-Guided 3D World Generation with Editing. *arXiv preprint arXiv:2508.05899*, 2025. URL https://arxiv.org/abs/2508.05899.

Aleksey Bokhovkin, Quan Meng, Shubham Tulsiani, and Angela Dai. SceneFactor: Factored Latent 3D Diffusion for Controllable 3D Scene Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 628–639, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Bokhovkin_SceneFactor_Factored_Latent_3D_Diffusion_for_Controllable_3D_Scene_Generation_CVPR_2025_paper.html.

Paul Borne-Pons, Mikolaj Czerkawski, Rosalie Martin, and Romain Rouffet. MESA: Text-Driven Terrain Generation Using Latent Diffusion and Global Copernicus Data. *arXiv preprint arXiv:2504.07210*, 2025. URL https://arxiv.org/abs/2504.07210.

Lucy Chai, Richard Tucker, Zhengqi Li, Phillip Isola, and Noah Snavely. Persistent Nature: A Generative Model of Unbounded 3D Worlds. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20863–20874. IEEE, 2023. URL https://arxiv.org/abs/2303.13515.

Mingyang Chen, Shengdong Chen, Xiaoxiao Fu, Bosheng Gong, Haoyuan Guo, Bowen Li, Jiawen Li, Kejun Li, Tianpeng Li, Yin Liu, et al. Zing-0.5: Toward Playable Worlds with Real-Time Joint Action and Text Control. *arXiv preprint arXiv:2609.17909*, 2026. URL https://arxiv.org/abs/2609.17909.

Nan Duan, Haoyang Huang, Weiyang Jin, Haoran Li, Yaowei Li, Yuming Li, Yijun Liu, Xin Lu, Xiaoxiao Ma, Yanwen Ma, et al. Long-Horizon Audio-Visual Generation for Persistent Stories and Interactive Worlds. *arXiv preprint arXiv:2608.23383*, 2026. URL https://arxiv.org/abs/2608.23383.

Jean-David Génevaux, Éric Galin, Eric Guérin, Adrien Peytavie, and Bedrich Benes. Terrain Generation Using Procedural Models Based on Hydrology. *ACM Transactions on Graphics (TOG)*, 32(4):1–13, 2013. doi: 10.1145/2461912.2461996. URL https://doi.org/10.1145/2461912.2461996.

Alexander Goslin. InfiniteDiffusion: Bridging Learned Fidelity and Procedural Utility for Open-World Terrain Generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–10, 2026. doi: 10.1145/3799902.3811080. URL https://doi.org/10.1145/3799902.3811080.

Leslie Gu, Junhwa Hur, Charles Herrmann, Fangneng Zhan, Todd Zickler, Deqing Sun, and Hanspeter Pfister. GeCo: Evaluating Geometric Consistency for Video Generation via Motion and Structure. In *The 19th European Conference on Computer Vision (ECCV)*, 2026. URL https://vcg.seas.harvard.edu/publications/geco.

Chunchao Guo, Jinpeng Li, Yang Li, and Zilong Huang. WorldClaw: Agentic 3D Open-World Generation at Scale. *arXiv preprint arXiv:2608.05248*, 2026. URL https://arxiv.org/abs/2608.05248.

Kazuki Higo, Toshiki Kanai, Yuki Endo, and Yoshihiro Kanamori. TerraFusion: Joint Generation of Terrain Geometry and Texture Using Latent Diffusion Models. *Virtual Reality & Intelligent Hardware*, 7(6):560–576, 2025. doi: 10.1016/j.vrih.2025.11.001. URL https://doi.org/10.1016/j.vrih.2025.11.001.

Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2Room: Extracting Textured 3D Meshes from 2D Text-to-Image Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7875–7886. IEEE, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Hollein_Text2Room_Extracting_Textured_3D_Meshes_from_2D_Text-to-Image_Models_ICCV_2023_paper.html.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*, 2021. URL https://arxiv.org/abs/2106.09685.

JiaKui Hu, Jialun Liu, Liying Yang, Xinliang Zhang, Kaiwen Li, Shuang Zeng, Yuanwei Li, Haibin Huang, Chi Zhang, and Yanye Lu. Geometry-as-context: Modulating Explicit 3D in Scene-consistent Video Generation to Geometry Context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4258–4268, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Hu_Geometry-as-context_Modulating_Explicit_3D_in_Scene-consistent_Video_Generation_to_Geometry_CVPR_2026_paper.html.

Ziniu Hu, Ahmet Iscen, Aashi Jain, Thomas Kipf, Yisong Yue, David A Ross, Cordelia Schmid, and Alireza Fathi. SceneCraft: An LLM Agent for Synthesizing 3D Scenes as Blender Code. In *ICML*, pages 19252–19282, 2024. URL https://proceedings.mlr.press/v235/hu24g.html.

Junchao Huang, Guian Fang, Shengju Qian, Xianghao Kong, Zhuoran Zhao, Wei Huang, Yihua Du, Zixin Zhang, Justin Cui, Yuchao Gu, et al. SolarWM: Open Data and Scalable Training for Long-Horizon Video World Models. *arXiv preprint arXiv:2609.02886*, 2026. URL https://arxiv.org/abs/2609.02886.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818. IEEE, 2024. URL https://arxiv.org/abs/2311.17982.

Han-Hung Lee, Qinghong Han, and Angel X Chang. NuiScene: Exploring Efficient Generation of Unbounded Outdoor Scenes. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 26509–26518. IEEE, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Lee_NuiScene_Exploring_Efficient_Generation_of_Unbounded_Outdoor_Scenes_ICCV_2025_paper.html.

Lu Ling, Chen-Hsuan Lin, Tsung-Yi Lin, Yifan Ding, Yu Zeng, Yichen Sheng, Yunhao Ge, Ming-Yu Liu, Aniket Bera, and Max Li. Scenethesis: A Language and Vision Agentic Framework for 3D Scene Generation. In *International Conference on Learning Representations*, volume 2026, pages 136596–136629, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/dd172fa21119391253587a4310a1a426-Abstract-Conference.html.

Yuheng Liu, Xinke Li, Yuning Zhang, Lu Qi, Xin Li, Wenping Wang, Chongshou Li, Xueting Li, and Ming-Hsuan Yang. Controllable 3D Outdoor Scene Generation via Scene Graphs. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 28052–28062. IEEE, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liu_Controllable_3D_Outdoor_Scene_Generation_via_Scene_Graphs_ICCV_2025_paper.html.

Joshua Lochner, James Gain, Simon Perche, Adrien Peytavie, Eric Galin, and Eric Guérin. Interactive Authoring of Terrain using Diffusion Models. In *Computer Graphics Forum*, volume 42, page e14941. Wiley Online Library, 2023. doi: 10.1111/cgf.14941. URL https://doi.org/10.1111/cgf.14941.

Yifan Lu, Xuanchi Ren, Jiawei Yang, Tianchang Shen, Zhangjie Wu, Jun Gao, Yue Wang, Siheng Chen, Mike Chen, Sanja Fidler, et al. InfiniCube: Unbounded and Controllable Dynamic 3D Driving Scene Generation with World-Guided Video Models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 27272–27283. IEEE, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Lu_InfiniCube_Unbounded_and_Controllable_Dynamic_3D_Driving_Scene_Generation_with_ICCV_2025_paper.html.

Quan Meng, Lei Li, Matthias Nießner, and Angela Dai. LT3SD: Latent Trees for 3D Scene Diffusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 650–660. IEEE, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Meng_LT3SD_Latent_Trees_for_3D_Scene_Diffusion_CVPR_2025_paper.html.

Hou In Derek Pun, Hou In Ivan Tam, Austin T Wang, Xiaoliang Huo, Angel X Chang, and Manolis Savva. HSM: Hierarchical Scene Motifs for Multi-Scale Indoor Scene Generation. In *2026 International Conference on 3D Vision (3DV)*, pages 1356–1367. IEEE, 2026. URL https://3dlg-hcvc.github.io/hsm/.

Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, et al. Infinite Photorealistic Worlds Using Procedural Generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12630–12641. IEEE, 2023. URL https://arxiv.org/abs/2306.09310.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, et al. Infinigen Indoors: Photorealistic Indoor Scenes using Procedural Generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21783–21794. IEEE, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Raistrick_Infinigen_Indoors_Photorealistic_Indoor_Scenes_using_Procedural_Generation_CVPR_2024_paper.html.

Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. XCube: Large-Scale 3D Generative Modeling using Sparse Voxel Hierarchies. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4209–4219. IEEE, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ren_XCube_Large-Scale_3D_Generative_Modeling_using_Sparse_Voxel_Hierarchies_CVPR_2024_paper.html.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6121–6132. IEEE, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ren_GEN3C_3D-Informed_World-Consistent_Video_Generation_with_Precise_Camera_Control_CVPR_2025_paper.html.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, et al. Lyra 2.0: Explorable Generative 3D Worlds. *arXiv preprint arXiv:2604.13036*, 2026. URL https://arxiv.org/abs/2604.13036.

Chunyi Sun, Junlin Han, Weijian Deng, Xinlong Wang, Zishan Qin, and Stephen Gould. 3D-GPT: Procedural 3D Modeling with Large Language Models. In *2025 International Conference on 3D Vision (3DV)*, pages 1253–1263. IEEE, 2025. URL https://arxiv.org/abs/2310.12945.

Dilin Wang, Hyunyoung Jung, Tom Monnier, Kihyuk Sohn, Chuhang Zou, Xiaoyu Xiang, Yu-Ying Yeh, Di Liu, Zixuan Huang, Thu Nguyen-Phuoc, Yuchen Fan, Sergiu Oprea, Ziyan Wang, Roman Shapovalov, Nikolaos Sarafianos, Thibault Groueix, Antoine Toisoul, Prithviraj Dhar, Xiao Chu, Minghao Chen, Geon Yeong Park, Rakesh Ranjan, and Andrea Vedaldi. WorldGen: From Text to Traversable and Interactive 3D Worlds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27124–27135, June 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_WorldGen_From_Text_to_Traversable_and_Interactive_3D_Worlds_CVPR_2026_paper.html.

Pengfei Wang, Liyi Chen, Zhiyuan Ma, Yanjun Guo, Guowen Zhang, and Lei Zhang. One2Scene: Geometric Consistent Explorable 3D Scene Generation from a Single Image. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9aaa0d7db14b9f0c2575a1761b1ab76e-Abstract-Conference.html.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image Technical Report. *arXiv preprint arXiv:2508.02324*, 2025. URL https://arxiv.org/abs/2508.02324.

Zhennan Wu, Yang Li, Han Yan, Taizhang Shang, Weixuan Sun, Senbo Wang, Ruikai Cui, Weizhe Liu, Hiroyuki Sato, Hongdong Li, et al. BlockFusion: Expandable 3D Scene Generation using Latent Tri-plane Extrapolation. *ACM Transactions on Graphics (ToG)*, 43(4):1–17, 2024. doi: 10.1145/3658188. URL https://doi.org/10.1145/3658188.

Chong Xia, Shengjun Zhang, Fangfu Liu, Chang Liu, Khodchaphun Hirunyaratsameewong, and Yueqi Duan. ScenePainter: Semantically Consistent Perpetual 3D Scene Generation with Concept Relation Alignment. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 28808–28817. IEEE, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Xia_ScenePainter_Semantically_Consistent_Perpetual_3D_Scene_Generation_with_Concept_Relation_ICCV_2025_paper.html.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term Consistent World Simulation with Memory. In *Advances in Neural Information Processing Systems*, volume 38, pages 49632–49652, 2025. doi: 10.52202/085713-1659. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/470629a47e2d65ce0606c40055df5d26-Abstract-Conference.html.

Yifei Xue, Yuanchen Fei, Hao Zhang, Chenzhi Nie, Yizhen Lao, et al. Illusion or Integrity? Geometrical Consistency Metric for AIGC Video Quality Evaluation. *arXiv preprint arXiv:2608.09594*, 2026. URL https://arxiv.org/abs/2608.09594.

Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, et al. Holodeck: Language Guided Generation of 3D Embodied AI Environments. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16277–16287. IEEE, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yang_Holodeck_Language_Guided_Generation_of_3D_Embodied_AI_Environments_CVPR_2024_paper.html.

Yuanyang Yin, Gongxuan Wang, Yifan Zhan, Chuanhao Li, Kaipeng Zhang, and Feng Zhao. Alaya-EVOKE: From Linear-Scaling Supervision to Endless World. *arXiv preprint arXiv:2608.13546*, 2026. URL https://arxiv.org/abs/2608.13546.

Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. WonderJourney: Going from Anywhere to Everywhere. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6658–6667. IEEE, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yu_WonderJourney_Going_from_Anywhere_to_Everywhere_CVPR_2024_paper.html.

Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. WonderWorld: Interactive 3D Scene Generation from a Single Image. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5916–5926. IEEE, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yu_WonderWorld_Interactive_3D_Scene_Generation_from_a_Single_Image_CVPR_2025_paper.html.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. ViewCrafter: Taming Video Diffusion Models for High-fidelity Novel View Synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b. doi: 10.1109/TPAMI.2025.3613256. URL https://doi.org/10.1109/TPAMI.2025.3613256.

Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista Martin, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent Video Diffusion with Explicit 3D Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21685–21695, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_World-consistent_Video_Diffusion_with_Explicit_3D_Modeling_CVPR_2025_paper.html.

Yi Zhang, Yunshuang Wang, Zeyu Zhang, and Hao Tang. Code2Worlds: Empowering Coding LLMs for 4D World Generation. *arXiv preprint arXiv:2602.11757*, 2026a. URL https://arxiv.org/abs/2602.11757.

Yisu Zhang, Chenjie Cao, Tengfei Wang, Xuhui Zuo, Junta Wu, Jianke Zhu, and Chunchao Guo. WorldStereo: Bridging Camera-Guided Video Generation and Scene Reconstruction via 3D Geometric Memories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_WorldStereo_Bridging_Camera-Guided_Video_Generation_and_Scene_Reconstruction_via_3D_CVPR_2026_paper.html.

Mengqi Zhou, Yuxi Wang, Jun Hou, Shougao Zhang, Yiwei Li, Chuanchen Luo, Junran Peng, and Zhaoxiang Zhang. SceneX: Procedural Controllable Large-scale Scene Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10806–10814, 2025. URL https://arxiv.org/abs/2403.15698.

Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. SANA-WM: Efficient Minute-Scale World Modeling with Hybrid Linear Diffusion Transformer. *arXiv preprint arXiv:2605.15178*, 2026. URL https://arxiv.org/abs/2605.15178.

A Supplementary Ablation Studies

A.1 Ablation Protocol

The codec test compares five encodings on 64 matched tasks through one frozen VAE. Terrain input evaluation uses four neighbor configurations, 64 targets each and two protocols, matching encoding, inference steps and seeds. Canvas uses pooled-context training and multi-image models use context subsets; targets are in-domain authored RDR2 terrain. The default planning model is GPT-5.6-Sol (`gpt-5.6-sol`, low reasoning effort), compared with Qwen3.8-Max. For planning-model substitution, terrain, intent, asset catalog, compiler, revision budget and cameras are fixed. We compare relation realization, geometric validity and revision cost alongside asset layouts.

A.2 Visual Comparison of Terrain Conditioning

Figure 6 complements Table 2(b) with matched terrain completions. Spatial canvas conditioning maintains more coherent relief across the marked interfaces in these examples, supporting the boundary-continuity gains in the quantitative comparison.

A.3 Planning-model Comparison

Figure 7 compares full layouts with identical terrain and an unchanged existing region on the left. The alternative planner favors wooded open space and dispersed settlements; the default planner produces a denser settlement and road network. Both retain the inherited terrain interface and organized connections.

A.4 Random-seed Stability

A.4.1 Controlled Comparison

We evaluate five fixed scene–trajectory pairs with seeds 0, 1, 2, 42 and 100. The world version, first frame, prompt, depth and non-seed generation settings are held fixed. Smoothness and MC-GeCo use the main comparison's frozen task-wise motion reference; MC-GeCo divides each video's co-visible residual by its relative motion before aggregation. Table 4 reports all 25 controlled outputs together with the five main-experiment references. Seed dispersion is computed from the 25 controlled outputs.

A.4.2 Metric Variation and Visual Outcomes

Table 5 first averages the same five tasks for each seed. Its standard deviation and coefficient of variation are computed across those five seed-wise means, using the population standard deviation. This separates aggregate seed sensitivity from differences between scenes.

Computational measures. The aggregate IQ, AQ and MN-MS vary narrowly across seeds, with CVs of 0.39%, 0.40% and 0.05%. The geometric scores have larger but still comparatively moderate CVs: 5.29% for MC-GeCo, 3.11% for MC-MEt3R and 6.33% for MC-GeoCon. These results support stability of the aggregate appearance and geometry measures under the tested noise changes. Per-task results remain important: averaging can reduce dispersion and does not imply that every individual video is equally stable.

Occasional sky corruption. Visual inspection identifies a black-sky interval in one of the 25 controlled outputs, the mountain scene with seed 100. Its IQ is 0.7667, compared with 0.8021–0.8044 for the other four seeds, and MC-MEt3R rises to 0.1576 from 0.1093–0.1180. The outcome is consistent with the base video generator occasionally transferring dark depth-conditioning regions into sky appearance, although the seed comparison alone does not isolate that cause. The affected video remains in all statistics. This observation identifies an appearance-synthesis failure under unchanged geometry; one occurrence in this small study does not establish its deployment frequency.

Event-based evaluation. SMC and Cam have CVs of 27.61% and 15.60%, respectively, and vary more than the computational measures. Their discrete disappearance, return and motion-clause decisions can amplify small changes in generated evidence; the visual judgments additionally involve a large-model evaluator. These scores therefore reflect event realization and evaluation sensitivity as well as visual consistency. The present experiment varies generation seeds, not repeated judgments of an identical video, so it does not attribute all dispersion to evaluator randomness. We report their full ranges alongside the more stable computational measures.

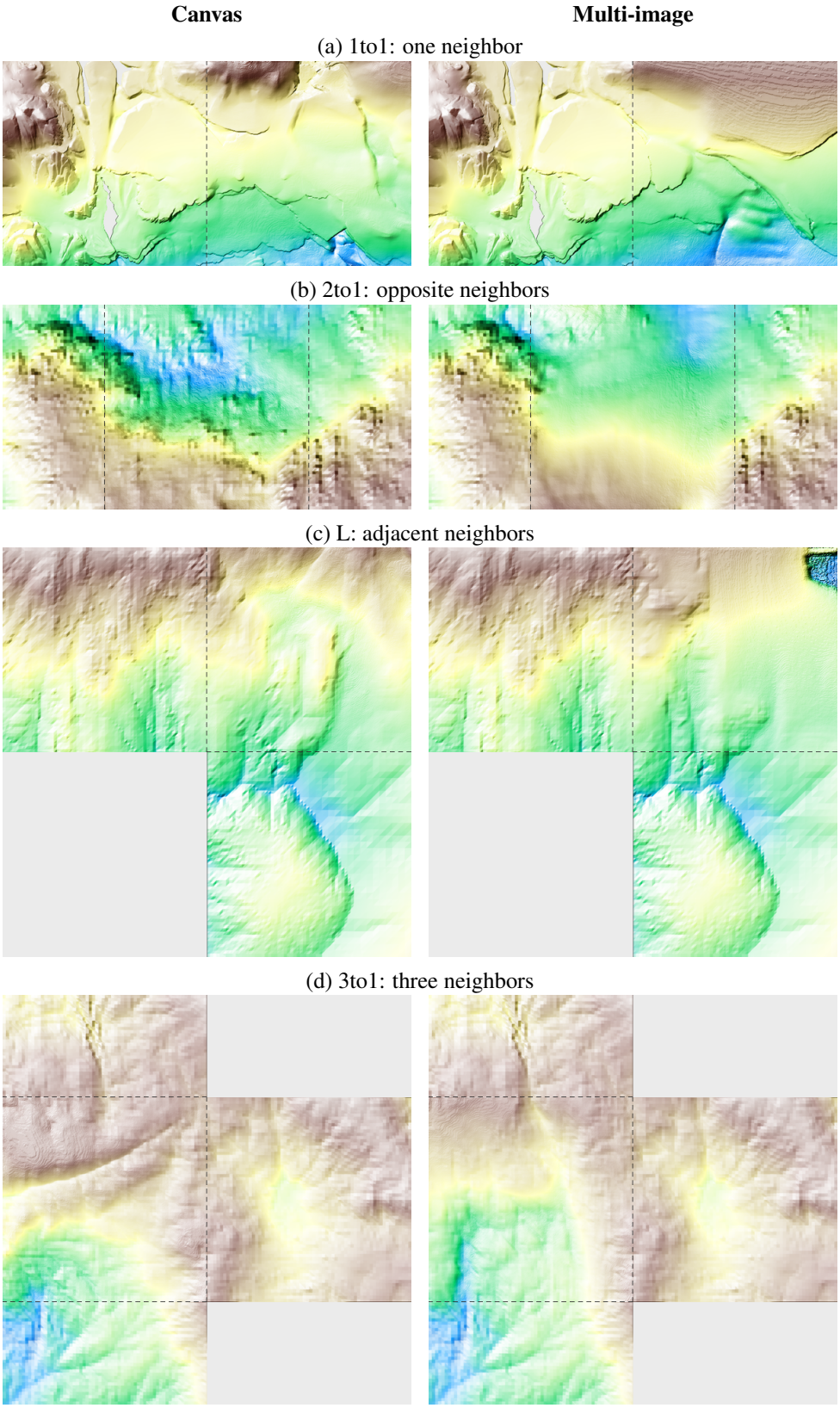

Figure 6: Terrain completion with spatial canvas (left) and separate images (right). Each pair shares known terrain, elevation colors and illumination. Dashed lines mark context–target interfaces; gray regions are unavailable.

(a) Alternative planner (b) Default planner

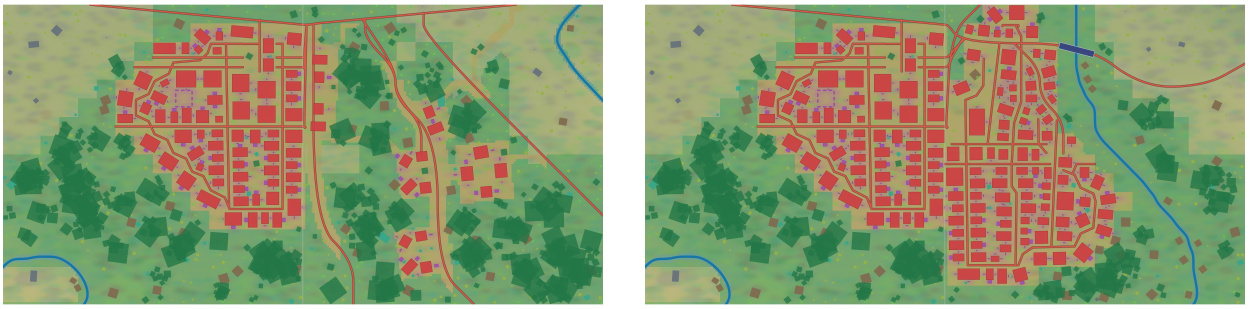

Figure 7: Full planning-model comparison: (a) alternative and (b) default planner. Within each map, the shared existing region is on the left and the new region on the right. Terrain and drawing conventions are identical.

Table 4: Per-task seed comparison. Each scene contains one main-experiment reference and five controlled seed variants. All values are retained; † marks the observed black-sky case.

Version	Seed	IQ↑	AQ↑	MN-MS↑	MC-GeCo↓	SMC↑	Cam↑	MC-MEt3R↓	MC-GeoCon↓
Forest									
Reference	–	0.7927	0.5921	0.9853	0.0468	1.0000	97.62	0.1080	0.0768
Seed	0	0.7943	0.5928	0.9834	0.0463	1.0000	97.62	0.1075	0.1002
Seed	1	0.7917	0.5929	0.9825	0.0448	1.0000	97.62	0.1053	0.0957
Seed	2	0.7916	0.5993	0.9828	0.0431	1.0000	97.62	0.1095	0.0902
Seed	42	0.7896	0.5929	0.9847	0.0423	1.0000	97.62	0.1078	0.0860
Seed	100	0.7930	0.6025	0.9862	0.0374	1.0000	95.24	0.1085	0.0655
Rocky grassland									
Reference	–	0.7930	0.6765	0.9842	0.0687	1.0000	97.62	0.1100	0.0660
Seed	0	0.7859	0.6473	0.9850	0.0783	1.7500	97.62	0.1128	0.0871
Seed	1	0.7874	0.6554	0.9840	0.0881	0.0000	40.48	0.1110	0.0806
Seed	2	0.7873	0.6443	0.9862	0.0743	1.7500	97.62	0.1110	0.0814
Seed	42	0.7918	0.6631	0.9853	0.0628	0.0000	40.48	0.1132	0.0779
Seed	100	0.7854	0.6407	0.9859	0.0652	0.0000	32.14	0.1120	0.0805
Town									
Reference	–	0.7958	0.7683	0.9789	0.0513	2.0000	97.96	0.1357	0.1563
Seed	0	0.7960	0.7641	0.9789	0.0496	1.5000	97.96	0.1378	0.1880
Seed	1	0.7966	0.7795	0.9806	0.0526	0.0000	41.84	0.1335	0.1490
Seed	2	0.7963	0.7716	0.9795	0.0524	1.0000	97.96	0.1355	0.1790
Seed	42	0.7974	0.7683	0.9805	0.0530	1.5000	97.96	0.1348	0.1988
Seed	100	0.7964	0.7714	0.9784	0.0502	1.5000	97.96	0.1363	0.1809
Mountain									
Reference	–	0.8030	0.7135	0.9809	0.0490	3.5000	97.96	0.1166	0.0985
Seed	0	0.8044	0.7259	0.9810	0.0460	3.5000	97.96	0.1180	0.1274
Seed	1	0.8027	0.7207	0.9811	0.0436	2.5000	97.96	0.1107	0.1027
Seed	2	0.8021	0.7190	0.9807	0.0475	3.5000	97.96	0.1118	0.1300
Seed	42	0.8028	0.7297	0.9819	0.0462	3.3750	97.96	0.1093	0.0891
Seed	100†	0.7667	0.7153	0.9818	0.0431	3.5000	97.96	0.1576	0.0949
Winter									
Reference	–	0.7921	0.7029	0.9751	0.0525	0.0000	34.69	0.1764	0.1709
Seed	0	0.7922	0.6954	0.9772	0.0403	1.0000	97.96	0.1669	0.1817
Seed	1	0.7899	0.6930	0.9759	0.0451	0.0000	34.69	0.1634	0.1607
Seed	2	0.7886	0.6785	0.9777	0.0384	0.0000	34.69	0.1638	0.1661
Seed	42	0.7943	0.6975	0.9777	0.0388	0.0000	34.69	0.1612	0.1583
Seed	100	0.7914	0.6953	0.9783	0.0399	0.0000	34.69	0.1638	0.1555

Table 5: Seed-wise means over five tasks and their across-seed dispersion. The five main-experiment references are excluded. SD is the population standard deviation and CV is SD divided by the mean, expressed as a percentage.

Seed	IQ	AQ	MN-MS	MC-GeCo	SMC	Cam	MC-MEt3R	MC-GeoCon
0	0.7945	0.6851	0.9811	0.0521	1.7500	97.82	0.1286	0.1369
1	0.7937	0.6883	0.9808	0.0548	0.7000	62.52	0.1248	0.1177
2	0.7932	0.6825	0.9814	0.0512	1.4500	85.17	0.1263	0.1293
42	0.7952	0.6903	0.9820	0.0486	1.1750	73.74	0.1253	0.1220
100	0.7866	0.6850	0.9821	0.0472	1.2000	71.60	0.1356	0.1155
Mean	0.7926	0.6863	0.9815	0.0508	1.2550	78.17	0.1281	0.1243
SD	0.0031	0.0027	0.0005	0.0027	0.3466	12.19	0.0040	0.0079
CV (%)	0.39	0.40	0.05	5.29	27.61	15.60	3.11	6.33

B Evaluation Metrics

B.1 Metric Definitions

Visual quality. IQ averages MUSIQ frame-quality estimates, AQ averages CLIP-based aesthetic predictions, and MS measures agreement with AMT-interpolated frames, following VBench (Huang et al., 2024). All measures evaluate generated RGB under the same public task description.

Structural Memory Consistency (SMC). SMC requires dedicated camera trajectories that induce a disappearance–return event for a designated target. To enable this evaluation, we randomly select 40% of all samples and adapt their sampling trajectories and camera motions accordingly; SMC is computed on this subset. A method-blind Qwen3.5-27B evaluator samples each video at 1 Hz, identifies the named target and its neighbors, and selects the earliest usable reference. It verifies complete target absence, then compares the reference with the earliest and latest usable return views. Each pair receives anchored 1–5 grades for identity, structure, neighbor relations and texture: 5 denotes preservation, 3 a clear local change, and 1 replacement or unrecognizable structure. Legitimate viewpoint and lighting changes are allowed. With grades g_{jd} for return view j and criterion d, the score is

$$s_j = \mathrm{Cap}\left(\frac{\sum_{d\in D_j} w_d g_{jd}}{\sum_{d\in D_j} w_d}\right), \qquad \mathrm{SMC} = \frac{1}{|\mathcal{R}|}\sum_{j\in\mathcal{R}} s_j, \qquad w = (0.30, 0.30, 0.25, 0.15). \tag{9}$$

Here D_j contains graded criteria and $\mathcal{R}$ the return views. Identity, structure and neighbor evidence are required; an unobservable texture grade is omitted. Identity grade 1 fixes the pair score at 1; grade 2 caps it at 2. Structure at most 2 caps it at 2.5, and neighbor relations at most 2 cap it at 3. A designated memory task scores zero when its required disappearance–return event is absent.

Camera Compliance (Cam). The public prompt is decomposed into motion and visibility clauses. Translation and turning are evaluated from estimated camera poses at 1 Hz; target framing, absence and return use visual evidence. An event clause receives 1 when observed and 0 otherwise; a sustained clause receives the fraction of intervals satisfying it. Clause degrees d_c are combined as

$$\mathrm{Cam} = \frac{100\, o}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} d_c, \qquad o = 1 \text{ for correct event order}, \quad o = \tfrac{1}{2} \text{ otherwise}. \tag{10}$$

Thus Cam rewards executing the requested observation sequence, while SMC evaluates what is preserved when the target returns.

Geometric consistency. GeCo combines motion and depth residuals (Gu et al., 2026); MEt3R compares reprojected image features (Asim et al., 2025); GeoCon-Bench measures correspondence inlier ratio and geometric fitting error (Xue et al., 2026). We use a correspondence-based local reproduction for GeoCon and apply the corrections below.

B.2 Motion-Calibrated Consistency and Smoothness

Realized motion. Motion amplitude is the evaluator's mean optical-flow norm, reflecting scene depth, camera rotation and translation, and scene motion. Across 120 paired tasks, WorldWeave exceeds the six baselines listed in the main text on 74–93 tasks each.

The task requires both scene preservation and camera movement. Absolute disagreement can decrease when a model barely changes its view, even though it misses the requested exploration. Figure 8 shows this mismatch: favorable raw scores coexist with distorted texture or nearly stationary views. We therefore measure geometric disagreement relative to the motion actually realized and evaluate camera compliance separately.

Relative residuals on common visibility. Let u be observed angular flow, u_r the flow induced by estimated camera motion and depth, and z_w, z_t warped and target depths. On pixels that project inside the target and pass depth agreement, the co-visible GeCo residual uses

$$e_m = \frac{\|u - u_r\|_2}{\|u\|_2 + \|u_r\|_2 + \epsilon}, \qquad e_z = \frac{|z_w - z_t|}{|z_w| + |z_t| + \epsilon}. \tag{11}$$

The flow denominator expresses deformation relative to observed and predicted displacement; the visibility mask prevents newly exposed content from being treated as a correspondence failure. Valid motion and depth cues are averaged equally, retaining GeCo's single-cue fallback.

Task-relative motion normalization. For GeCo, e is the co-visible relative residual above; for MEt3R and GeoCon, it is the original feature error or combined geometric error. Each video v is normalized by realized motion $m(v)$ relative

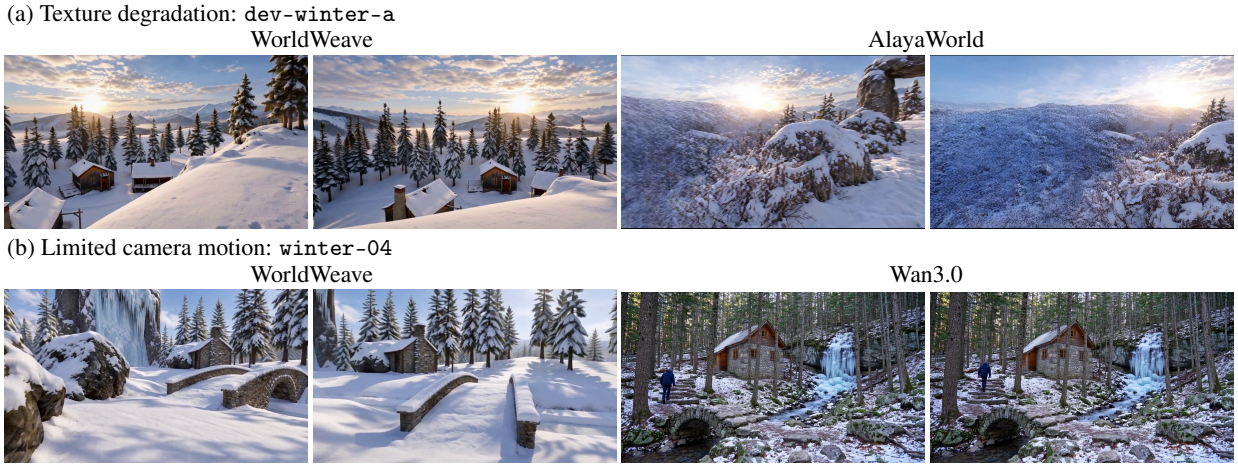

Figure 8: Motion and consistency are complementary. AlayaWorld's low raw MEt3R error coexists with distorted snow texture; Wan3.0's low raw GeCo error accompanies little camera movement. Times are shown; scores refer to whole videos.

to the fixed task-wise reference m_{ref}, then scores are averaged:

$$\begin{aligned} r_m(v) &= \frac{m(v)}{m_{\text{ref}}}, & e_{\text{MC}}(v) &= \frac{e(v)}{r_m(v)}, \\ \overline{e}_{\text{MC}} &= \frac{1}{N}\sum_{i=1}^{N} e_{\text{MC}}(v_i), & e_{\text{GeoCon}} &= \sqrt{\max(1-\text{IR},\epsilon)\max(\text{GE},\epsilon)}. \end{aligned} \tag{12}$$

This reports error per relative amount of exploration: small raw error receives less credit when accompanied by little movement. The reference is the task-wise median from the frozen comparison set, shared across methods and held fixed when adding the Base or seed variants. Aggregation uses the mean of video-wise ratios, not the ratio of aggregate errors and motion. Non-positive motion and invalid correspondences are excluded. GeoCon remains supplementary because the local reproduction is insensitive to the tested local warp. The geometric scores and Cam jointly assess preservation under completed camera motion.

Motion-normalized smoothness. We apply the same task-relative motion calibration to VBench's smoothness error:

$$\text{MN-MS}(v) = 1 - \frac{1-\text{MS}(v)}{r_m(v)}. \tag{13}$$

Higher is better. Scores are computed per video before aggregation and interpreted jointly with geometric consistency and camera compliance.

C User-study Protocol

Cases and presentation. The study contains 20 matched scene–trajectory cases with seven methods per case. Cases were selected using automatic metrics and visual screening before human rating; the reported means characterize this curated evaluation set. Participants view anonymized videos under shared task instructions; method names and automatic scores are hidden. Each account is assigned ten distinct videos, with presentation order shuffled and allocation prioritizing rating coverage. Participation is voluntary and requires informed consent.

Rating criteria. Participants assign integer scores from 1 to 5, with higher scores indicating better consistency. Camera consistency evaluates the requested movement, turns and observation order. Object consistency evaluates preservation of object identity, shape and number. Structure/texture consistency evaluates stable scene layout and coherent surface details. The same three questions accompany every video.

Aggregation. All 140 videos meet the minimum of three ratings: 124 have three and 16 have four, yielding 436 video evaluations and 1,308 criterion scores. We first average ratings per video and criterion, then average the 20 cases with equal weight. Let s_{cmdr} be rating r for case c, method m and criterion d, with n_{cm} ratings per video. The method-level

mean is

$$\bar{s}_{cmd} = \frac{1}{n_{cm}} \sum_{r=1}^{n_{cm}} s_{cmdr}, \qquad \mu_d(m) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \bar{s}_{cmd}. \tag{14}$$

Here $\mathcal{C}$ contains all 20 cases. Each case has equal weight, so videos with four ratings do not contribute more than those with three. Scores remain on the original 1–5 scale; higher values indicate stronger perceived consistency. WorldWeave attains the unique highest object mean in 19 of 20 cases, and attains or shares the highest camera and structure/texture means in 16 and 18 cases.

D Metric Terrain Implementation

D.1 Context Reference and Coordinate Convention

Let Δ be the metric sample spacing and o_q the target origin. Sample centers are $x_{ij} = o_q + \Delta(i + \frac{1}{2}, j + \frac{1}{2})$. Rotation into a canonical neighbor arrangement acts jointly on elevation, validity and side labels; outputs are mapped back before joining. Let $\mathcal{B}_{\mathrm{ctx}}$ collect samples from inward-facing context bands and $X_a = (1, x_a, y_a)$. The reference plane is initialized by least squares and refined by reweighted fits:

$$\begin{aligned} e_a^{(l)} &= X_a \theta^{(l)} - h_a, \quad \sigma_l = \max\{\sigma_{\min}, 1.4826\ \mathrm{median}_a\, |e_a^{(l)}|\}, \\ \omega_a^{(l)} &= \min\{1, 1.5\sigma_l / \max(|e_a^{(l)}|, \epsilon)\}, \\ \theta^{(l+1)} &= \arg\min_{\theta} \sum_{a \in \mathcal{B}_{\mathrm{ctx}}} (\omega_a^{(l)})^2 (X_a \theta - h_a)^2. \end{aligned} \tag{15}$$

The squared weights reflect weighting both the design matrix and observations. Multiple neighbors share one fit. With a single neighbor, the normal trend is tapered away from the interface to prevent indefinite linear extrapolation. Writing the plane coefficients as (a, b, c) in normal–tangential coordinates (d, y) gives

$$b_q(d, y) = a + cy + b\,\psi_L(d), \qquad \psi_L(d) = \begin{cases} d - d^2/(2L), & 0 \le d < L, \\ L/2, & d \ge L. \end{cases} \tag{16}$$

The normal derivative therefore decreases continuously to zero. Multi-neighbor configurations use the joint plane without this single-edge taper. Context bands are 64 samples wide; the fit uses four reweighting passes, $\sigma_{\min} = 1\,\mathrm{m}$ and $L = 64\,\mathrm{m}$. Missing elevations are filled from the nearest valid sample for numerical processing, while their original validity remains excluded from supervision. Target elevations never enter the context fit.

D.2 Redundant Encoding and Robust Metric Decoding

The channel map $f_j(r) = \frac{1}{2} + \frac{1}{2}\tanh(r/s_j)$ defines a one-dimensional curve in RGB space. Spatial repetition adds image-space redundancy, and block averaging restores the metric grid after VAE decoding:

$$(U_\rho c)_{\rho i + a, \rho j + b} = c_{ij}, \qquad (A_\rho \widetilde{c})_{ij} = \rho^{-2} \sum_{a,b=0}^{\rho-1} \widetilde{c}_{\rho i + a, \rho j + b}, \qquad \rho = 2. \tag{17}$$

For an averaged decoded sample c_j, clip channel endpoints before inversion. Define $y_j = 2c_j - 1$ and use inverse-sensitivity weighting to initialize the residual:

$$r_j = s_j\,\mathrm{atanh}(y_j), \quad G_j = \frac{2s_j}{\max(1 - y_j^2, \epsilon)}, \quad r^{(0)} = \frac{\sum_j G_j^{-2} r_j}{\sum_j G_j^{-2}}. \tag{18}$$

Saturated channels have larger inverse gain and receive less weight. The channels need not agree after image decoding, so a fixed robust projection brings the estimate back toward their common encoding curve. With $J_j(r) = f_j'(r)$ and $\eta_j^{(l)} = \min\{1, \kappa / \max(|f_j(r^{(l)}) - c_j|, \epsilon)\}$, the scalar update is

$$r^{(l+1)} = r^{(l)} - \frac{\sum_j \eta_j^{(l)} J_j(r^{(l)})[f_j(r^{(l)}) - c_j]}{\max\{\sum_j \eta_j^{(l)} J_j(r^{(l)})^2, \epsilon\}}, \qquad \widehat{h}_q = b_q + r^{(L_d)}. \tag{19}$$

We use $L_d = 5$ and $\kappa = 2/255$, clipping channel values to $[10^{-5}, 1 - 10^{-5}]$. The projection is determined by decoded channels and the analytic codec.

D.3 Conditioning, Supervision and Boundary Acceptance

Let K_q, T_q, V_q denote known-context, target and valid masks on the spatial canvas. The condition retains encoded context and replaces target content with a neutral value. A separate region image distinguishes target, validity and unavailable locations:

$$C_{\mathrm{cond}}(x) = \begin{cases} f(H(x) - B(x)), & x \in K_q, \\ (1/2, 1/2, 1/2), & x \in T_q, \\ (0, 0, 1), & x \notin K_q \cup T_q, \end{cases} \qquad R_q = (T_q V_q, V_q, 1 - K_q - T_q). \tag{20}$$

Canvas slots preserve neighbor direction; half-neighbor bands are used where the canonical layout places opposite neighbors around the full target. Only target-valid locations contribute to the normalized loss in the main text. Distances for edge weights are evaluated in target-local metric coordinates, so image repetition and the loss-grid resolution do not change the physical correction-band width.

Joining is evaluated on the shared interface, not by equating distinct sample centers on opposite sides. Let γ_e be interface height and derivative extraction with a common normal. Its residual is

$$E_{\partial q}(h) = \{\gamma_e(h) - (g_e, v_e)\}_{e \in \mathcal{E}_q}, \quad \gamma_e(h) = (h|_{\Gamma_e}, \partial_{n_e} h|_{\Gamma_e}). \tag{21}$$

The connector inherits height and derivatives at the boundary of each preserved old surface. Corrections vanish with their normal derivatives at their inner support boundaries, while corner reconciliation acts on new-owned nodes. Continuity and modification magnitude are measured separately; all targets remain in the joining comparison, including those requiring large corrections.

D.4 Ownership-aware Continuous Joining

Committed geometry is defined over the cells between existing sample centers. The gap between neighboring center grids and uncovered corner cells belong to the new connector surface. Let $J = (h, h_x, h_y, h_{xy})$ denote nodal height and derivatives. Bicubic Hermite connector cells inherit J on their old-facing edges; new-owned nodes supply the other constraints. Adjacent cells therefore share height and first derivatives without forcing spatially distinct old samples to coincide.

Missing context is completed before connection, preserving all original valid values and derivatives. Missing regions use nearest-valid completion, with a finite constant fallback only for entirely missing tiles. Generated targets retain their actual finite mask. Height and normal-slope corrections are applied on the new side, with normalized corner weights to avoid summing multiple full-strength side corrections. The fixed variant uses a 16 m band for both terms; slope taper keeps the height band fixed and shortens the derivative term's support, targeting at most 0.5 m additional excursion with a minimum support of one sample spacing.

The joining experiment retains all 256 targets (64 terrain targets under four neighbor configurations). Large corrections receive quality diagnostics rather than being rejected from the comparison. The reported modification is

$$\Delta h_q = \left(\frac{1}{N_q} \sum_{i=1}^{N_q} [h_q^+(x_i) - \widehat{h}_q(x_i)]^2 \right)^{1/2}. \tag{22}$$

Here N_q counts the fixed full-target samples. H_c and S_c measure height and slope disagreement through continuous queries at the preserved geometry boundary; they are distinct from the raw midpoint-extrapolated H/S in Table 2(b). Zeros in Table 2(c) denote residuals below 10^{-9}. Their numerical closure follows from derivative inheritance, while Δh_q quantifies the cost of modifying the candidate. Relative to fixed C^1 joining, slope taper reduces mean modification by 0.1263 m (target-clustered 95% bootstrap interval: 0.0829–0.1806 m). The largest pointwise change remains 35.37 m; exact boundary closure does not imply uniformly small modifications or gentle slopes throughout the connector.

E Hierarchical Scene Compilation

E.1 Terrain Evidence and Spatial Responsibilities

Evidence is computed before semantic allocation. For metric elevation h, the local gradient determines slope and the Laplacian summarizes curvature. After depression filling, drainage uses the steepest positive descent among adjacent cells:

$$\begin{aligned} s(x) &= \|\nabla h(x)\|_2, \qquad k(x) = \nabla^2 h(x), \\ d(x) &= \arg\max_{y \in \mathcal{N}_8(x)} \frac{\widetilde{h}(x) - \widetilde{h}(y)}{\|x - y\|_2}, \qquad A(x) = 1 + \sum_{y : d(y) = x} A(y). \end{aligned} \tag{23}$$

Here $\widetilde{h}$ is the depression-filled field; cells without positive descent have no receiver. Accumulation A counts upstream contributing cells. The agent uses these fields together with inherited interfaces and asset capabilities to assign regional roles; evidence does not itself impose a semantic label.

Regions assign spatial responsibility, while districts refine it. Let R_a be a region and D_{ab} its districts on the target domain Ω_q. Coverage and containment require

$$\bigcup_a R_a = \Omega_q, \qquad D_{ab} \subseteq R_a, \qquad \bigcup_b D_{ab} = R_a, \qquad \operatorname{int}(D_{ab}) \cap \operatorname{int}(D_{ac}) = \emptyset \ (b \neq c). \tag{24}$$

These conditions apply to ownership layers; roads, water corridors and vegetation overlays carry separate typed responsibilities. Boundary ports additionally constrain the position, direction and type of continuations. Region feasibility is checked before district and instance refinement, so later placement operates on accepted spatial responsibilities.

E.2 Typed Relations and Asset Transforms

An asset plan is a typed relation graph $\mathcal{G}_q = (\mathcal{V}_q^{\text{asset}}, \mathcal{E}_q^{\text{rel}})$. Nodes carry catalog choice, role and district ownership; edges describe shared space, orientation, support or access. The catalog provides local geometry, support anchors, entrance axes and connector endpoints. For local point p of instance i, the compiled transform is

$$p^{\,w} = A_i p + t_i, \qquad T_i = \begin{bmatrix} A_i & t_i \\ 0 & 1 \end{bmatrix}, \qquad \mathcal{I}_q = \{(\mathrm{id}_i, \mathrm{asset}_i, T_i, \mathrm{role}_i, \mathrm{district}_i)\}_i. \tag{25}$$

Allowed scaling is determined by asset type. For a two-ended connector module, let p_-, p_+ be local endpoints and q_-, q_+ the target endpoints. With $F(v)$ an orthonormal frame whose first axis follows v, alignment is

$$\begin{aligned} \alpha &= \frac{\|q_+ - q_-\|_2}{\|p_+ - p_-\|_2}, \\ A_i &= F(q_+ - q_-)\,\mathrm{diag}(\alpha, 1, 1) F(p_+ - p_-)^\top, \\ t_i &= (q_- + q_+)/2 - A_i (p_- + p_+)/2. \end{aligned} \tag{26}$$

Only the longitudinal axis is scaled, and forbidden span changes are rejected. This transform aligns connector endpoints with their prescribed world positions. Other asset families use their own support and orientation contracts under the same typed interface.

E.3 Geometric Validation and Repair Attribution

The compiler evaluates the predicates required by each asset contract. With world support anchors S_i, admissible support surface $\mathcal{S}_i$, collision envelopes B_i, and access graph $\mathcal{H}$, representative constraints are

$$\begin{aligned} &\max_{p \in S_i} \operatorname{dist}(p, \mathcal{S}_i) \leq \tau_i^{\text{sup}}, \\ &\operatorname{dist}(B_i, B_j) \geq c_{ij} \quad \text{for pairs requiring clearance}, \\ &\mathrm{Reach}_{\mathcal{H}}(\mathrm{entry}_i, \mathrm{road}_i) = 1 \quad \text{for assets requiring road access}. \end{aligned} \tag{27}$$

Support relations permit intended contact; clearance applies to incompatible object pairs. The tolerances and required predicates belong to the asset contract. Each failed predicate retains its instance, relation and district identifiers, which

localize the plan entries to revise. Relation graphs therefore serve both construction and failure attribution, while deterministic geometry tools remain responsible for metric realization.

F Revision, Publication and Read-only Observation

F.1 Bounded Candidate Revision

Program checks and visual review consume the same candidate revision. Let $g_j(\Delta M_q)$ be its required geometric predicates and $a_{\mathrm{reg}}, a_{\mathrm{vis}}$ the region and compiled-scene review decisions. Publication requires every geometric predicate and both review decisions to pass:

$$\mathrm{Accept}(\Delta M_q) = a_{\mathrm{reg}} \wedge a_{\mathrm{vis}} \wedge \bigwedge_{j \in \mathcal{J}_{\mathrm{req}}} g_j(\Delta M_q). \tag{28}$$

A revision delta identifies a set of editable candidate records $\mathcal{U}_t$. Unaffected semantic decisions remain fixed, while the compiler recomputes geometry that depends on edited records. The plan-level restriction and old-world protection are

$$\pi_q^{(t+1)}|_{\mathcal{U}_t^c} = \pi_q^{(t)}|_{\mathcal{U}_t^c}, \qquad \mathrm{WriteSet}(\Delta M_q^{(t)}) \cap \Omega_k = \emptyset. \tag{29}$$

A failed revision does not relax acceptance predicates. When the budget is exhausted, the candidate is rejected and the previous version remains available. Each revision regenerates reports and matched previews for its own candidate.

F.2 Versioned Publication

Publication binds terrain, instance transforms, plans and interface records into one version. Let v be the visible version identifier and $\widehat{M}$ the validated candidate world. Readers select a version before querying:

$$\widehat{M} = M_k \oplus \Delta M_q, \qquad v_{\mathrm{after}} = \begin{cases} k+1, & \mathrm{Accept}(\Delta M_q) = 1 \text{ and publication succeeds,} \\ k, & \text{otherwise.} \end{cases} \tag{30}$$

New records expose interfaces for subsequent additions, while existing instance identities and transforms persist. A query retains its selected version throughout a trajectory. Adding geometry may change visibility in a later version, but does not rewrite the earlier version's structural records.

F.3 Camera Geometry and Video Conditions

A viewing request provides a trajectory or lets the agent sample one from the requested motion and targets. The resulting camera schedule is $C_t = (K_t, R_t, c_t, \tau_t)$, where R_t maps camera axes to world axes, c_t is the camera center and τ_t is the timestamp. Clearance and requested visibility events are evaluated against the selected world. For homogeneous pixel $\bar{u} = (u, v, 1)^\top$, the normalized world ray is

$$d_t(u) = \frac{R_t K_t^{-1} \bar{u}}{\|R_t K_t^{-1} \bar{u}\|_2}, \qquad \lambda_t^*(u) = \min\{\lambda > 0 : c_t + \lambda d_t(u) \in \mathcal{G}(M_v)\}. \tag{31}$$

The geometry set $\mathcal{G}(M_v)$ includes terrain and transformed assets. The nearest valid intersection determines visibility, hit identity and depth. Ray distance and camera-axis depth are distinct quantities:

$$X_t(u) = c_t + \lambda_t^*(u) d_t(u), \qquad D_t^{\mathrm{ray}}(u) = \lambda_t^*(u), \qquad D_t^z(u) = e_3^\top R_t^\top [X_t(u) - c_t]. \tag{32}$$

Here $e_3 = (0, 0, 1)^\top$ is the camera's optical-axis unit vector. Pixels without a valid hit carry an explicit validity mask. The adapter preserves the depth convention and calibration while converting resolution, temporal sampling and encoding to the video model's interface. Depth conditions are paired with optional text or image rendering-style references supported by the selected generator. All queries and RGB synthesis operate on the selected world without a write-back path.

F.4 Relation to Joint-Embedding Predictive Architectures

I-JEPA learns semantic image representations by predicting target-region embeddings from context (Assran et al., 2023). V-JEPA extends feature prediction to video, learning representations through a latent prediction objective without pixel reconstruction (Bardes et al., 2024). V-JEPA 2 further uses an action-conditioned predictor over learned representations for planning (Assran et al., 2025). These approaches support a broader distinction between internal world modeling and the synthesis of visual observations.

WorldWeave shares this motivation but implements the separation at the level of explicit world-state maintenance. Its persistent state comprises metric terrain, asset instances, semantic organization and cross-region interfaces. Hierarchical planning, geometric compilation and validated commits extend that state; a video generator receives depth queried from a selected version and optional appearance references. JEPA concerns learning predictive representations, whereas WorldWeave concerns constructing and preserving a geometrically addressable world for subsequent observation. The connection is conceptual rather than an adoption of the JEPA training objective: our contribution is the concrete state-construction, extension and read-only query mechanism.

G Additional Visual Comparisons

Figures 9, 10 and 11 compare target persistence during occlusion and off-screen return in farmstead, town and woodland scenes. The sequences allow comparison of object identity, neighboring layout and completion of the requested camera motion. Each comparison includes all eleven methods, including MiniMax-H3 (Base).

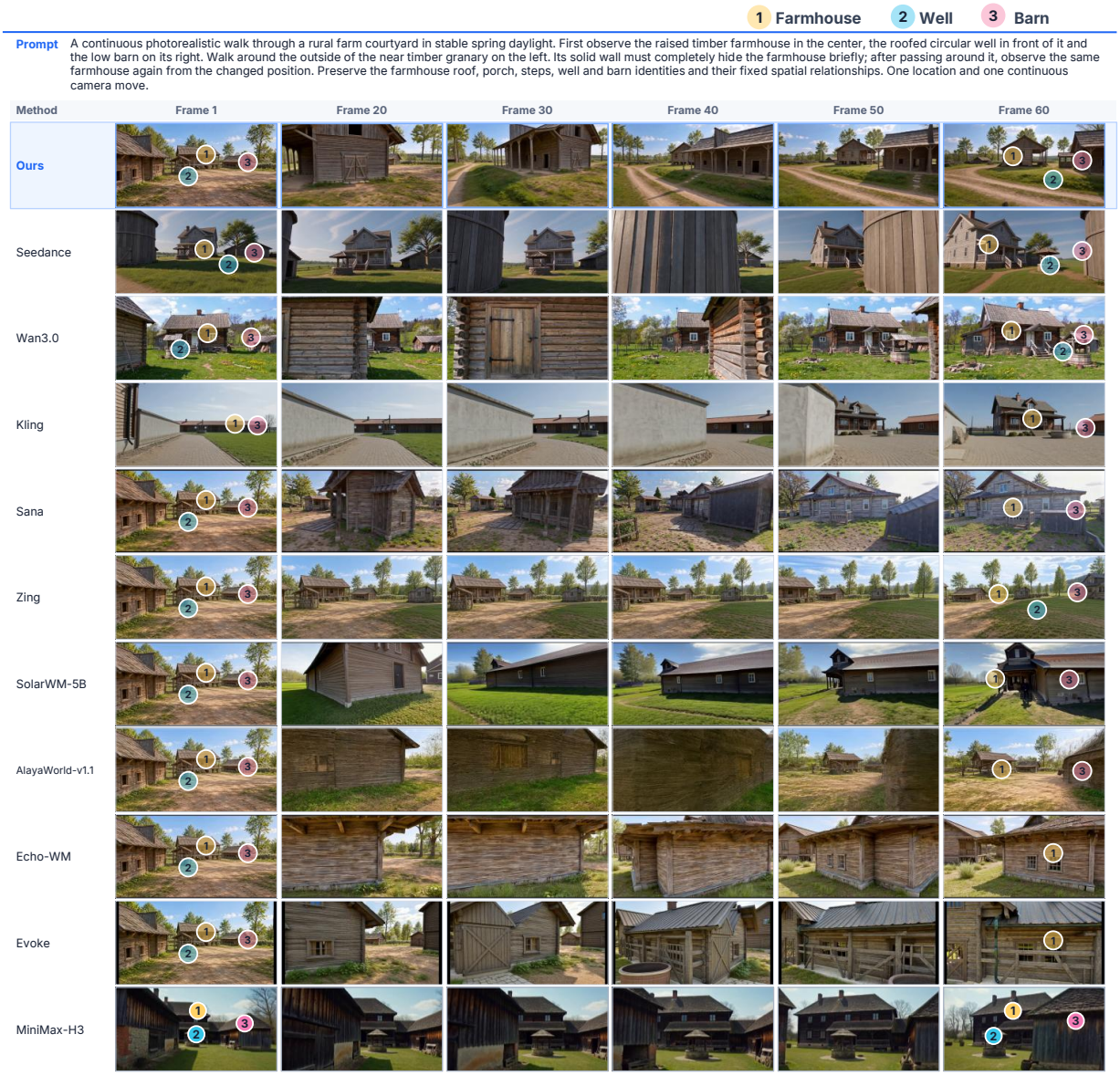

Figure 9: Occlusion-and-return comparison in a farmstead. Rows show different methods and columns follow the video sequence. Numbered markers identify the farmhouse, well and barn for comparison across disappearance and return.

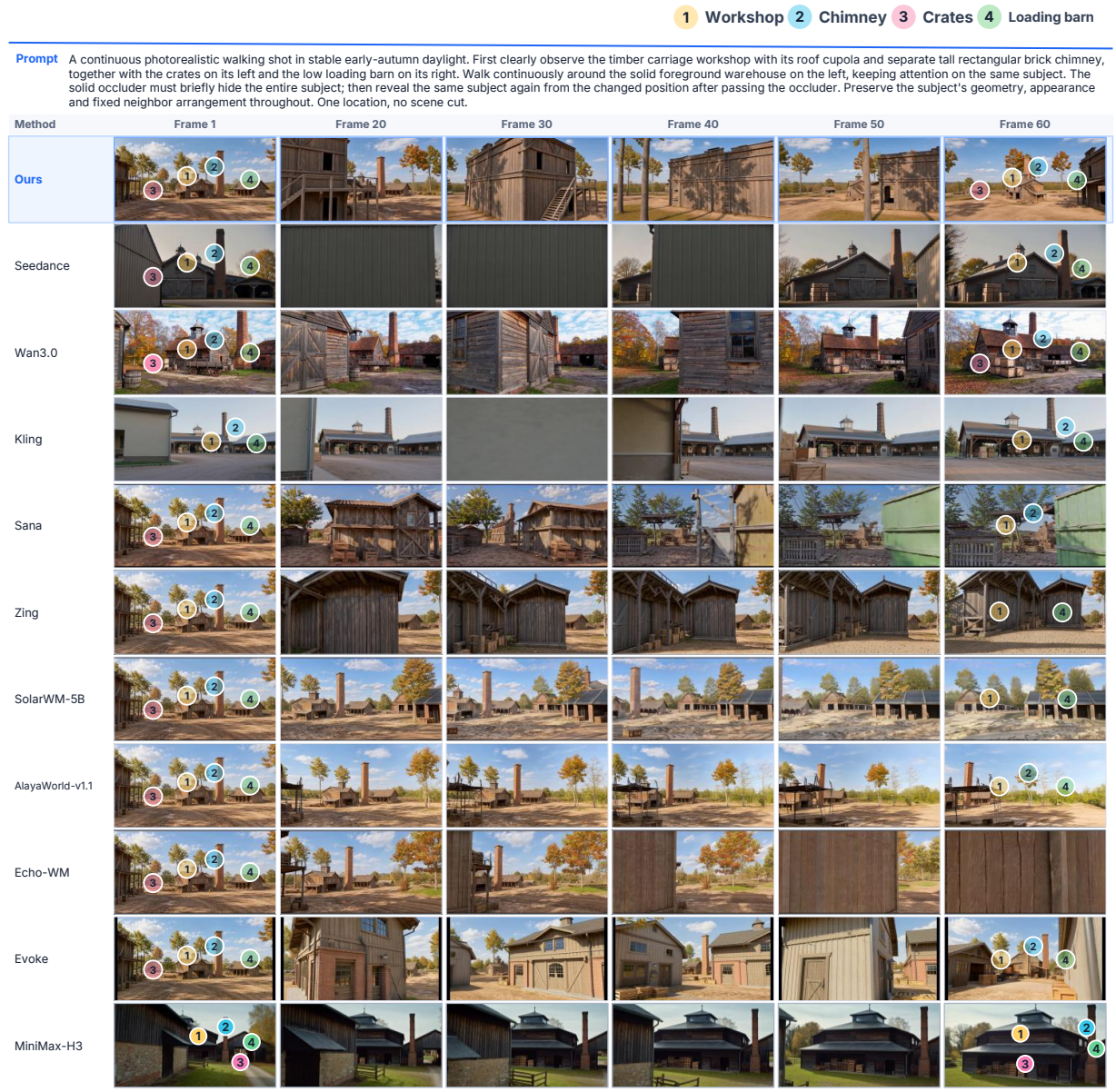

Figure 10: Occlusion-and-return comparison in a town. Rows show different methods and columns follow the video sequence. Markers track the workshop, chimney, crates and loading barn as the camera passes behind the warehouse.

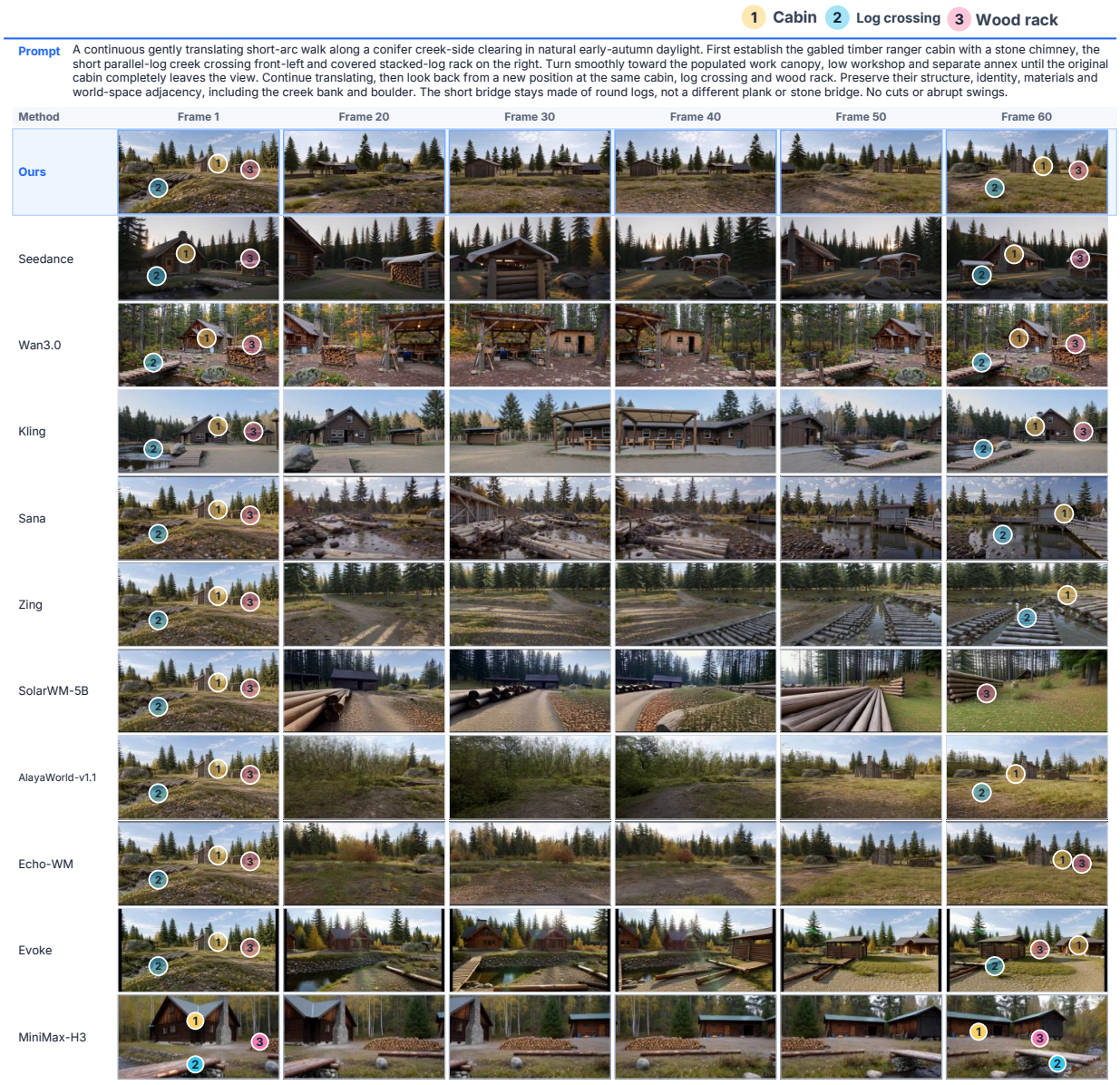

Figure 11: Off-screen-return comparison in a woodland scene. Rows show different methods and columns follow the video sequence. Markers track the ranger cabin, log crossing and wood rack during the turn-away and return sequence.