PROXY2WORLD: LEARNING TO GENERATE WORLDS FROM LIGHTWEIGHT PROXIES WITHOUT SEEING THEM

Anonymous authors
Paper under double-blind review

ABSTRACT

Lightweight scene proxies let creators control scene layout and motion while leaving room for imagination in appearance, lighting, and visual effects. However, a suitable proxy is not uniquely defined, making paired proxy–video data difficult to construct automatically at scale. We present Proxy2World, a controllable world model that learns these complementary capabilities from ordinary posed RGBD videos, without training on authored proxy–video pairs. The model jointly learns depth-conditioned RGB generation and joint RGBD generation through cross-modal flow matching. Learning both tasks enables proxy–camera hybrid denoising at inference to follow the proxy structure while producing natural, detailed visuals. We further introduce ProxyBench to evaluate this capability across a diverse set of scenes, camera trajectories, and subject motions. Experiments on ProxyBench show that Proxy2World achieves a better balance between structural adherence and visual quality than camera-controlled and geometry-conditioned methods, supported by quantitative metrics, VLM assessments, human evaluations and diverse qualitative results. *Project page:* `https://proxy2world-anonymous-project-page.pages.dev/`

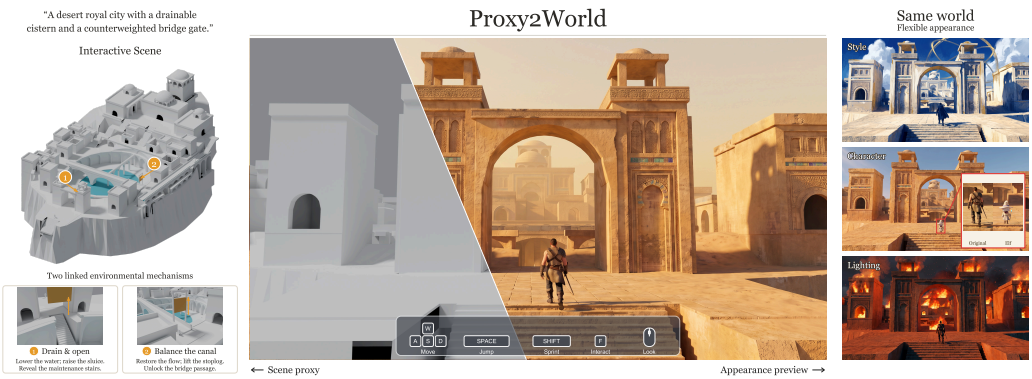


Figure 1: **Build the structure. Generate the world.** Proxy2World transforms lightweight, interactive scene proxies into visually rich, structure-aligned worlds without paired proxy–video training data. Creators specify layouts and interactions through editable proxies (left), while the model generates detailed visuals (center) and variations in style, character, and lighting (right).

1 INTRODUCTION

Creating worlds that users can explore and direct is a central goal in gaming, filmmaking, and embodied AI. Recent models support action-driven interaction and camera-controlled exploration (Bruce et al., 2024; Che et al., 2025; He et al., 2025; Shen et al., 2026), while generative rendering translates explicit scene states into realistic imagery (Gu et al., 2025; Liang et al., 2025; Gómez-Nogales et al., 2026; Chen et al., 2026b). For practical use, visual realism must be accom-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

panied by control over scene layout, camera motion, and object movement. Yet specifying these properties should not require constructing every geometric and visual detail of the final world.

Existing approaches provide different levels of spatial control. Camera-controlled video models, including CameraCtrl, SCoPE, and GEN3C, guide viewpoint changes through pose conditioning, ray-based representations, or reconstructed scene caches (He et al., 2024; Yin et al., 2026; Ren et al., 2025). These methods enable controlled exploration, but a camera trajectory alone does not specify the desired scene layout or object motion. Geometry-aware video models introduce additional spatial structure. VACE and Cosmos-Transfer1 use depth or other spatial signals to guide video synthesis (Jiang et al., 2025; Alhaija et al., 2025), while DAR and UnividX use G-Buffer to connect editable scene states to generated observations (Chen et al., 2026b;a). World-consistent Video Diffusion, FantasyWorld, and Gen3R further couple visual generation with geometry, learning to generate appearance and spatial structure together (Zhang et al., 2025; Dai et al., 2026; Huang et al., 2026a). Video world models draw on learned priors to synthesize rich scenes and dynamics from sparse inputs, while generative rendering offers direct control through explicit scene representations. Bringing these capabilities together would allow users to prescribe scene structure and motion while leaving their visual realization to the generative model.

This raises a fundamental question: *how much of a world must be specified before generation can take over?* We study how lightweight scene proxies can guide world generation without specifying every detail of the final scene. A proxy may use low-poly meshes or simple primitives and may omit parts of the scene; the goal is to follow its intended layout and motion while generating plausible geometry and appearance beyond it.

Learning to generate from such proxies, however, presents several coupled challenges. First, a scene proxy reflects a designer's abstraction of the world—which objects should be represented, how strongly they should be simplified, and which geometry can safely be omitted—so large-scale paired proxy-to-RGB data are difficult to obtain automatically.

Second, structural adherence must coexist with geometric refinement (Yan et al., 2026). A primitive specifies where an object is and how much space it occupies, but its simplified shape should be refined rather than reproduced exactly. Recent work C2R learns control using paired synthetic coarse-to-real examples (Gómez-Nogales et al., 2026); we istead aim to learn from ordinary RGBD videos and transferring that control to proxies at different levels of abstraction. Together, these challenges raise our central question: *can proxy-based control be learned without any paired proxy-to-RGB training data and adapt to different level of abstraction?*

We answer this question with Proxy2World, a controllable world model that learns proxy-based control from ordinary posed RGBD videos without paired proxy–video supervision. As illustrated in Figure 1, it turns editable scene proxies into visually rich worlds while allowing variations in style, character, and lighting. Our key insight is to jointly learn depth-conditioned RGB generation and joint RGBD generation through cross-modal flow matching. Learning both tasks enables proxy–camera hybrid denoising at inference to follow the proxy structure while producing natural, detailed visuals. We retain camera guidance throughout and align camera translations with the metric scale of depth to keep the two control signals consistent. We further introduce ProxyBench to evaluate proxy adherence, generative refinement, and camera control across diverse scenes, proxy abstractions, camera trajectories, and subject motions. Experiments show that Proxy2World achieves a better balance between structural adherence and visual quality than camera-controlled and geometry-conditioned methods, supported by operator-based metrics, VLM assessments, and human evaluations.

Our contributions are threefold:

1. **Proxy-based world generation without paired proxy supervision.** We introduce Proxy2World, a controllable world model that learns structural grounding and generative completion from ordinary posed RGBD videos, without paired proxy–RGB training data.
2. **Cross-Modal hybrid flow matching.** We combine proxy-grounded structure formation with camera-guided joint RGBD refinement, aligning geometry and camera motion in a shared metric scale to preserve the intended layout while refining geometry and appearance.
3. **ProxyBench and comprehensive evaluation.** We introduce **ProxyBench**, an agent-driven benchmark spanning diverse scenes, camera trajectories, subject motions, and proxy ab-

stractions. Evaluations using operator-based metrics, VLM assessments, and human preferences show that Proxy2World achieves a better balance between structural adherence and visual quality than camera-controlled and geometry-conditioned baselines.

2 Related Work

Video foundations such as Stable Video Diffusion, CogVideoX, HunyuanVideo, and Wan provide generative priors for conditional synthesis (Blattmann et al., 2023; Yang et al., 2025; Kong et al., 2024; Wan et al., 2025). We review how subsequent methods expose camera, geometry, and scene-state controls, focusing on their relationship to generation from authored proxies.

2.1 Camera-Controlled Video Generation

Recent video diffusion models have substantially improved explicit camera control.MotionCtrl separates camera and object motion control (Wang et al., 2023b), while CameraCtrl and VD3D inject camera representations into pretrained video models (He et al., 2024; Bahmani et al., 2025). CamCo and CamI2V use epipolar constraints to structure cross-frame interactions (Xu et al., 2024; Zheng et al., 2024). CameraCtrl II extends camera-driven generation to dynamic scene exploration across wider viewpoints (He et al., 2025), and SCoPE incorporates camera sightlines into diffusion-transformer attention (Yin et al., 2026). These methods improve how a generator follows a requested viewpoint sequence. Geometric reprojection provides a complementary control interface. ViewCrafter uses point-based scene clues for novel-view generation (Yu et al., 2024), and GEN3C renders a reconstructed 3D cache along target camera trajectories (Ren et al., 2025). TrajectoryCrafter combines point-cloud renders with a source video to redirect its camera path (Yu et al., 2025). CamTrol obtains training-free camera control by using 3D layout rearrangement to guide noisy latents (Hou & Chen, 2024). CamCtrl3D combines pose, ray, reprojection, and 3D feature conditions (Popov et al., 2025), while RealCam-I2V aligns camera parameters with metric scene depth and applies scene-constrained noise shaping (Li et al., 2025). Proxy2World also couples geometry and camera scale, but accepts an externally authored scene proxy whose layout need not be reconstructed from the reference imagery.

2.2 Geometry-Aware Video World Models

Geometry can guide video synthesis as an input condition or be generated jointly with appearance. DiffusionRenderer learns rendering through G-buffers (Liang et al., 2025), while DaS and DAR synthesize video from mesh-derived controls (Gu et al., 2025; Chen et al., 2026b). VideoComposer, Control-A-Video, VACE, and Cosmos-Transfer1 support depth or other structural conditions (Wang et al., 2023a; Chen et al., 2023; Jiang et al., 2025; Alhaija et al., 2025), complemented by trajectory control in DragNUWA, Motion-I2V, and DragAnything (Yin et al., 2023; Shi et al., 2024; Wu et al., 2024). Geometry Forcing and GeoVideo improve geometric consistency (Wu et al., 2025; Bai et al., 2026). WVD, Aether, Voyager, FantasyWorld, Gen3R, VideoWeave, and DualCamCtrl further couple visual and geometric generation (Zhang et al., 2025; Zhu et al., 2025; Huang et al., 2025b; Dai et al., 2026; Huang et al., 2026a; Xiang et al., 2026; Zhang et al., 2026). We combine depth-conditioned grounding with joint RGBD refinement to follow abstract proxies while allowing their geometry to evolve.

2.3 Proxy-to-RGB generation.

C2R learns coarse-simulation control from paired synthetic examples (Gómez-Nogales et al., 2026). Concurrent work CWM and PWM separate programmable world states from visual synthesis, training their renderers on paired proxy–video or structured-control–video data (Chen et al., 2026c; Huang et al., 2026b). Marionette similarly renders predicted articulated states into pose controls for RGB generation (Meng et al., 2026). Proxy2World instead learns from ordinary posed RGBD videos, without paired proxy–RGB training data. Proxy–camera hybrid denoising combines structural grounding with joint geometry–appearance refinement, enabling transfer to low-poly, primitive, and incomplete proxies.

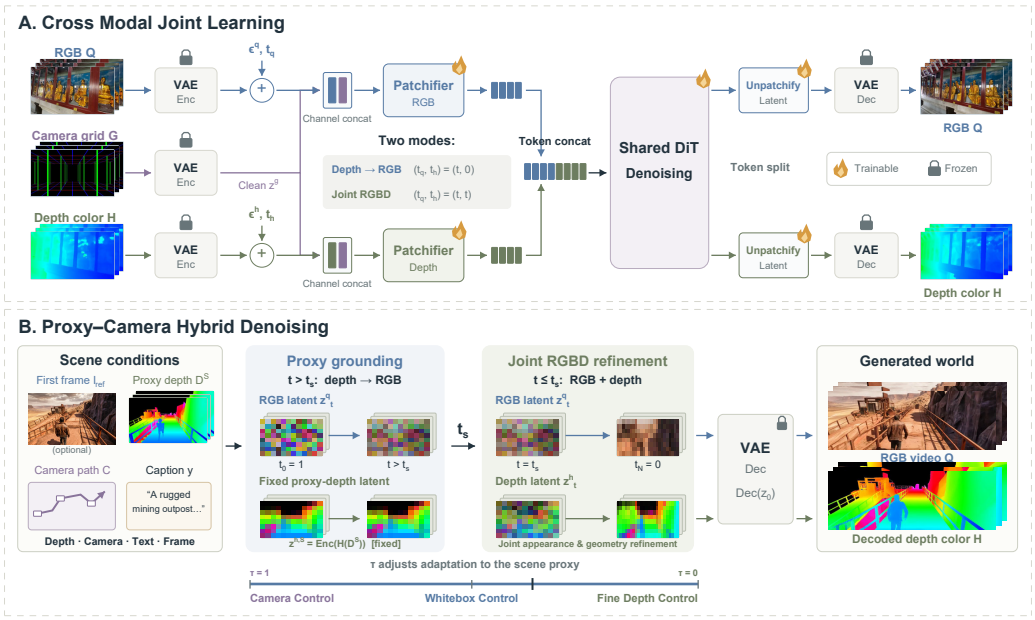


Figure 2: **Overview of the Method. (A) Cross Modal Joint learning.** We use a shared DiT to learn depth-conditioned RGB generation and joint RGBD generation, with camera-grid conditioning for both tasks. We train on ordinary posed RGBD videos without paired proxy–video supervision. **(B) Proxy–camera hybrid denoising.** We combine these learned capabilities during sampling: we first use proxy depth to establish scene structure, then jointly refine RGB and depth to enrich geometry and appearance. We retain camera-grid conditioning throughout. The refinement ratio τ controls the balance between proxy adherence and generative refinement.

3 Method

3.1 Problem Formulation

Given an interactive scene proxy S, a control sequence $\mathcal{A}$, and a camera trajectory $\mathcal{C} = \{(\mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i)\}_{i=1}^{F}$, we render a proxy depth sequence:

$$\mathbf{D}^S = \{D_i^S\}_{i=1}^{F} = \mathcal{R}_{\text{depth}}(S, \mathcal{A}, \mathcal{C}), \tag{1}$$

where F is the number of frames, and $\mathbf{K}_i$, $\mathbf{R}_i$, and $\mathbf{t}_i$ denote camera intrinsics, rotation, and translation, respectively. The controls drive subject motion and scene interactions, which are conveyed to the generative model through rendered depth. Given a text description y and an optional reference image $\mathbf{I}_{\text{ref}}$ spatially aligned with the initial proxy view, our goal is to generate an RGB video:

$$\mathbf{Q} = \{\mathbf{Q}_i\}_{i=1}^{F} \sim p_\theta\big(\mathbf{Q} \mid \mathbf{D}^S, \mathcal{C}, \mathbf{I}_{\text{ref}}, y\big)\,. \tag{2}$$

This task requires balancing structural adherence with generative freedom: strict conditioning can reproduce coarse proxy artifacts, while unconstrained generation can lose the intended layout and behavior. We seek to preserve proxy intent while refining geometry and appearance, learning from ordinary RGBD videos without paired proxy-to-RGB supervision.

3.2 Scene Proxies for Explicit World Control

A scene proxy specifies control-relevant structure and behavior without prescribing the world's final visual realization. Its geometry may be simplified or incomplete, preserving layout, occupancy, and subject motion while leaving visual details to the generative prior. We represent camera motion and scene geometry separately, supporting camera-only generation and additional structural control through proxy depth.

Camera grid. Following OmniDirector (Liu et al., 2026), we represent camera motion by rendering a fixed spatial grid $\mathcal{G}$ along the prescribed trajectory:

$$\mathbf{G} = \mathcal{R}_{\text{grid}}(\mathcal{G}, \mathcal{C}). \tag{3}$$

The grid provides spatial references whose projected motion expresses camera movement without specifying the target scene's surfaces or appearance.

Metric depth. We express video depth and camera translations in consistent metric units. For training sequences whose depth and camera estimates share an arbitrary scale (Yan et al., 2025), we use DA3 (Lin et al., 2025) to estimate first-frame metric depth and align the sequence depth to this reference. The resulting sequence-level scale factor s is applied to both depth and camera translations:

$$D_i^{\text{metric}} = sD_i, \qquad \mathbf{t}_i^{\text{metric}} = s\mathbf{t}_i. \tag{4}$$

This preserves their relative geometry while anchoring both to a common physical scale. Authored proxy depth and camera trajectories are likewise expressed in consistent metric units.

Following Vision Banana (Gabeur et al., 2026), we convert metric depth into a three-channel false-color representation:

$$\mathbf{H}_i = \mathcal{H}(D_i^{\text{metric}}) = h\big(f(D_i^{\text{metric}})\big), \tag{5}$$

where f nonlinearly maps metric distances to a bounded interval and h maps the result along a Hilbert-like path on the RGB cube. The nonlinear transform allocates greater precision to nearby geometry, while the invertible color mapping enables recovery of metric depth. A fixed mapping across frames and scenes preserves scale information and provides a common representation for observed depth and rendered proxy geometry.

3.3 Cross-Modal Joint Learning

We jointly train depth-to-RGB generation and joint RGBD generation within a single flow-matching model. Both tasks are sampled throughout training and share the same parameters, allowing depth to serve as either a fixed geometric condition or a generated variable.

Shared architecture. As shown in Figure 2(A), a frozen video VAE encoder Enc maps RGB, metric-depth color, and camera-grid videos to latents $\mathbf{z}_0^q$, $\mathbf{z}_0^h$, and $\mathbf{z}^g$, respectively. The subscript 0 denotes clean data. Each modality latent is concatenated channel-wise with the clean camera latent and tokenized by a separate patch embedding. RGB and depth tokens are then concatenated along the sequence dimension and processed by a shared DiT, which predicts modality-specific velocities:

$$(\mathbf{v}_\theta^q, \mathbf{v}_\theta^h) = \mathbf{v}_\theta(\mathbf{z}_{t_q}^q, \mathbf{z}_{t_h}^h, t_q, t_h; \mathbf{z}^g, \mathbf{c}), \tag{6}$$

where t_q and t_h are modality-specific noise levels, and $\mathbf{c}$ contains text and optional reference-frame conditions. After sampling, the frozen decoder Dec recovers RGB and depth-color videos. We train the patch embeddings, shared DiT, and output projections.

Cross-modal flow matching. For each modality $m \in \{q, h\}$, we construct a noisy latent by interpolating between the clean latent $\mathbf{z}_0^m$ and independent standard Gaussian noise:

$$\mathbf{z}_{t_m}^m = (1 - t_m)\mathbf{z}_0^m + t_m\boldsymbol{\epsilon}^m, \qquad \boldsymbol{\epsilon}^m \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{7}$$

where $t_m \in [0, 1]$, with 0 denoting clean data and 1 denoting pure noise. The target velocity is $\mathbf{u}^m = \boldsymbol{\epsilon}^m - \mathbf{z}_0^m$. For depth-conditioned RGB generation, we set $(t_q, t_h) = (t, 0)$: depth remains clean and only RGB is supervised. For joint RGBD generation, we set $(t_q, t_h) = (t, t)$ and supervise both modalities. Both tasks share the same model and are optimized through:

$$\mathcal{L} = \mathbb{E}\left[\sum_{m\in\{q,h\}} w_m \left\| \mathbf{v}_\theta^m\left(\mathbf{z}_{t_q}^q, \mathbf{z}_{t_h}^h, t_q, t_h; \mathbf{z}^g, \mathbf{c}\right) - \mathbf{u}^m \right\|_2^2\right]. \tag{8}$$

The expectation covers task sampling, training data, noise levels, and Gaussian noise. We use $(w_q, w_h) = (1, 0)$ for depth-conditioned RGB generation and $(w_q, w_h) = (1, 1)$ for joint RGBD generation. This joint training allows the same model to use depth as a fixed structural condition or generate it together with RGB, providing the two capabilities combined during hybrid denoising.

3.4 PROXY–CAMERA HYBRID DENOISING

We compose the two learned generation modes within a single sampling trajectory. As illustrated in Figure 2(B), early steps use proxy depth to establish scene structure, while later steps jointly refine RGB and depth. Camera-grid guidance is retained throughout both stages. Let $\tau \in [0, 1]$ denote the fraction of sampling steps allocated to joint RGBD refinement. For a sampling schedule $1 = t_0 > \cdots > t_N = 0$, we switch at $k_s = \lfloor (1-\tau)N \rfloor$, with noise level $t_s = t_{k_s}$.

Proxy grounding. We encode the proxy depth-color sequence into $\mathbf{z}^{h,S} = \mathrm{Enc}(\mathcal{H}(\mathbf{D}^S))$, where $\mathcal{H}$ is applied frame-wise, and initialize the RGB latent with Gaussian noise. During the early, high-noise steps, proxy depth remains fixed as a clean condition, and the model updates RGB using its depth-to-RGB mode:

$$\frac{d\mathbf{z}_t^q}{dt} = \mathbf{v}_\theta^q \left(\mathbf{z}_t^q, \mathbf{z}^{h,S}, t, 0; \mathbf{z}^g, \mathbf{c} \right), \qquad t > t_s. \tag{9}$$

Sampling proceeds from $t = 1$ toward $t = 0$. This stage grounds generation in the proxy's layout, occupied regions, and subject motion.

Joint RGBD refinement. At $t = t_s$, we retain the current RGB latent and initialize the depth state by adding noise to the proxy latent at the switching noise level:

$$\mathbf{z}_{t_s}^h = (1 - t_s)\mathbf{z}^{h,S} + t_s \boldsymbol{\epsilon}^h, \qquad \boldsymbol{\epsilon}^h \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{10}$$

Depth then becomes a generated variable, and both modalities evolve under the joint RGBD mode:

$$\frac{d\mathbf{z}_t^m}{dt} = \mathbf{v}_\theta^m \left(\mathbf{z}_t^q, \mathbf{z}_t^h, t, t; \mathbf{z}^g, \mathbf{c} \right), \qquad m \in \{q, h\}, \quad t \le t_s. \tag{11}$$

The proxy is no longer imposed as a fixed depth condition, allowing geometry and appearance to be refined together while camera guidance continues to constrain the viewpoint sequence.

The refinement fraction τ controls the balance between adherence and refinement. A smaller τ retains proxy grounding for more sampling steps, whereas a larger τ allocates more steps to joint refinement. At $\tau = 0$, depth remains fixed throughout sampling; at $\tau = 1$, both modalities start from noise, yielding camera-conditioned generation without proxy grounding. Intermediate settings combine the two capabilities using the same checkpoint, without additional proxy-specific training.

4 EXPERIMENTS

Implementation details. We initialize Proxy2World from Wan2.2-I2V-A14B (Wan et al., 2025) and train on 120,624 RGBD clips from DL3DV (Ling et al., 2024), RealEstate10K (Zhou et al., 2018), and gampeplay dataset(Zhou et al., 2025), each with 81 frames at 480×832 and 16 FPS, without paired proxy–RGB supervision. We freeze the video VAE and text encoder and optimize patch embeddings, DiT blocks, and output projections using AdamW (lr 10^{-5}, weight decay 0.01) on 48 GPUs. High-/low-noise experts are trained for 50,000/35,000 steps, with equal sampling weights for depth-to-RGB and joint RGBD tasks and uniformly sampled discrete noise levels within each expert's shifted-flow schedule. Depth-to-RGB supervises RGB only; joint RGBD assigns unit loss weights to both modalities. The encoded RGB reference and temporal mask are concatenated to both streams, with reference dropout 0.5. At inference, available reference images provide DA3-estimated metric depth for proxy alignment, with the same scale correction applied to camera translations. We use 40-step Euler sampling, flow shift 3.0, and CFG 3.5. The joint-refinement fraction τ determines the switch $k_s = \lfloor (1-\tau)N \rfloor$ for $1 = t_0 > \cdots > t_N = 0$. We set $\tau^\star = 0.925$: three geometry-conditioned steps followed by 37 joint RGBD steps, initializing refinement by noising proxy depth to t_{k_s}.

4.1 EXPERIMENTAL SETUP

ProxyBench. We introduce ProxyBench to evaluate world generation from coarse scene proxies. It comprises 25 scenes—15 medium-scale and 10 large-scale—with 300 five-second video sequences. For evaluation, we select a subset of 78 sequences emphasizing interaction and balanced

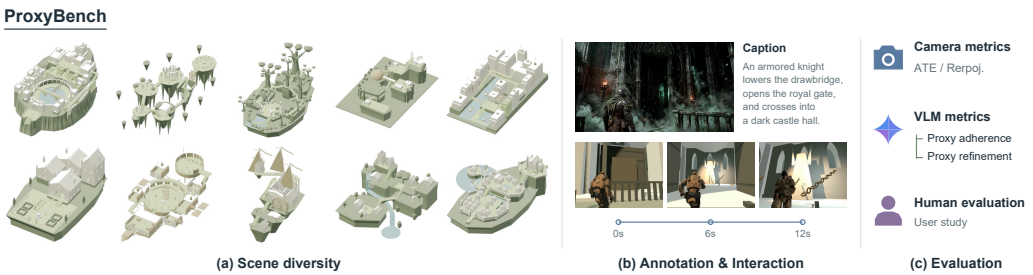


Figure 3: (a) Diverse lightweight scene proxies. (b) Text descriptions and scripted interactions specify the intended scene behavior and subject motion. (c) Evaluation combines camera and spatial consistency metrics, VLM assessments of proxy adherence and refinement, and human preferences.

Table 1: Evaluation on ProxyBench. Input icons indicate text, reference image, camera trajectory, and proxy control; gray icons denote unused inputs. Trajectory errors are computed from VIPE-estimated poses, and reprojection error excludes dynamic objects. Adherence and refinement are evaluated through Gemini pairwise win rates (%). Human ranks reflect overall preference. Bold and underline indicate the best and second-best results within each conditioning group. All our variants share one checkpoint trained without paired proxy–RGB data.

		Camera		Spatial Cons.	Proxy Evaluation				
					Operator-based		VLM-based		Human
Method	Input	$\mathrm{nATE}_t \downarrow$	$\mathrm{nATE}_r \downarrow$	Reproj.↓	Geo. Align↑	Temp. Align↑	Adh. WR↑	Ref. WR↑	Rank↓
Camera-conditioned methods									
Lyra 2.0 (Shen et al., 2026)	T ■ ■	**0.0418**	**0.0413**	2.171	0.4965	**0.7068**	32.21	39.48	10
SCoPE / RAYPE (Yin et al., 2026)	T ■ ■	0.1066	0.1229	4.338	0.4508	0.5780	29.51	47.06	9
FantasyWorld (Dai et al., 2026)	T ■ ■	0.2966	0.3472	2.449	0.4711	0.6158	36.51	50.25	7
Ours (TI2V, $\tau = 1$)	T ■ ■	0.0681	0.0524	**1.956**	**0.5340**	0.7033	**53.55**	**71.44**	**5**
Depth-conditioned methods									
VACE-Depth (Jiang et al., 2025)	T ■	0.0764	0.0584	6.264	0.7576	0.7335	60.14	**45.32**	4
Cosmos-Depth (Alhaija et al., 2025)	T ■	0.0446	0.0362	3.259	0.7713	0.7504	**72.26**	40.28	**3**
Ours (T2V, $\tau = 0$)	T ■	**0.0302**	**0.0270**	**2.280**	**0.7749**	**0.7645**	68.79	31.71	6
Proxy-conditioned methods									
Coarse2Real (Gómez-Nogales et al., 2026)	T ■ ■	0.1637	0.3152	5.312	0.5520	0.6616	40.03	45.21	8
Ours (T2V, $\tau = \tau^\star$)	T ■ ■	0.0863	0.0719	**2.067**	**0.7084**	0.7126	42.81	51.30	2
Ours (TI2V, $\tau = \tau^\star$)	T ■ ■	**0.0659**	**0.0558**	2.524	0.6795	**0.7173**	**64.18**	**77.96**	**1**

coverage of motion types. Each case provides a proxy scene, a target camera trajectory, synchronized proxy renderings, a text prompt, and a reference first frame. Successful generation should preserve the intended layout and behavior while enriching coarse geometry and appearance. Figure 3 provides an overview of ProxyBench, including scene diversity, interaction annotations, and the evaluation protocol.

Baselines. We select baselines covering geometry-conditioned rendering, camera-controlled world generation, and coarse-to-real synthesis. Among geometry-conditioned approaches, we focus on VACE (Jiang et al., 2025) and Cosmos-Transfer (Alhaija et al., 2025) as strong depth-conditioned video generators, adapting them to proxy control using depth rendered from our coarse scenes. For camera-controlled generation, Lyra 2.0 (Shen et al., 2026) represents projection-based approaches, but its static-scene formulation limits dynamic interactions. We therefore also include SCoPE/RayPE (Yin et al., 2026) for ray-based camera control and FantasyWorld (Dai et al., 2026) for joint RGBD world generation. Coarse2Real (Gómez-Nogales et al., 2026) serves as the most direct baseline, explicitly translating coarse simulations into realistic videos. Our camera-only ($\tau = 1$), depth-only ($\tau = 0$), and mixed-control ($\tau = \tau^\star$) variants share one checkpoint. Table 1 specifies the inputs used by each configuration.

Evaluation Protocol. We evaluate camera control using normalized translation and rotation errors (nATE_t and nATE_r) computed from VIPE-estimated poses(Huang et al., 2025a). Spatial consis-

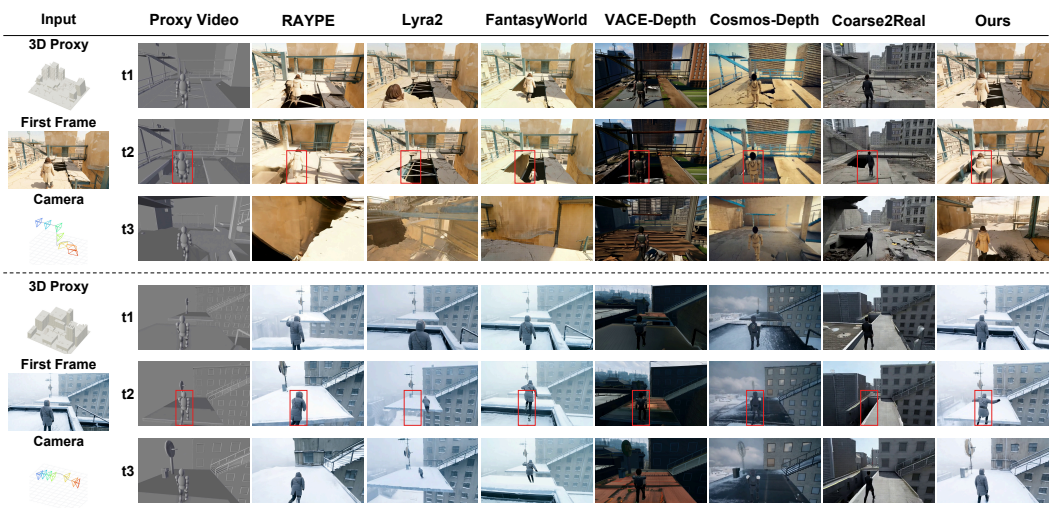


Figure 4: **Qualitative comparison on ProxyBench.** Red boxes highlight the intended subject locations specified by the proxy, highlighting differences in subject placement and shape. Camera-conditioned methods can deviate from the prescribed subject placement, while depth-conditioned methods can retain coarse body shapes and produced distorted characters. Ours preserves the intended placement while generating more natural character shapes and detailed scene appearance.

tency is measured by reprojection error after excluding dynamic objects. Geometric and temporal alignment measure agreement with the proxy. We further use Gemini-v3.7-flash(Team et al., 2023) for pairwise evaluation of *Proxy Adherence* and *Proxy Refinement*. Adherence assesses preservation of the intended layout and subject behavior, while refinement assesses plausible enrichment of geometry, appearance, and motion. We report average win rates across opponents, counting ties as half a win, alongside overall human preference ranks. We evaluate all methods on a common set of 78 sequences using operator-based metrics and VLM assessments, complemented by a human preference study with 10 participants on 10 sequences. Detailed protocols are provided in the appendix.

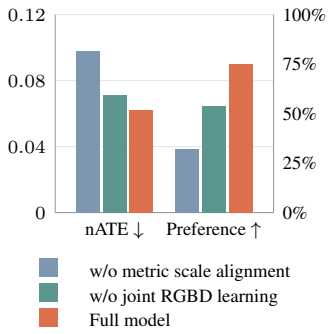
4.2 Evaluation on ProxyBench

Quantitative comparison. Table 1 shows that Proxy2World achieves a strong balance between proxy adherence and generative refinement. Compared with camera-controlled methods, our mixed TI2V model achieves higher geometric alignment and VLM adherence while also attaining the highest refinement win rate. The comparison with our own camera-only TI2V variant isolates the benefit of proxy guidance: geometric alignment improves by 27.2%, while adherence and refinement win rates increase by 10.63 and 6.52 percentage points, respectively. Thus, explicit proxy control improves structural adherence without sacrificing the model's generative flexibility. Compared with depth-conditioned methods, mixed generation allows greater refinement of the supplied coarse geometry. Cosmos-Depth achieves stronger VLM adherence (72.26% vs. 64.18%), but our mixed TI2V model achieves a substantially higher refinement win rate (77.96% vs. 40.28%). Together with its first-place ranking in human preference, these results support the benefit of balancing structural constraints with the freedom to refine coarse inputs. Compared with the proxy-conditioned baseline Coarse2Real as the same proxy method, our mixed T2V model reduces translation and rotation errors by 47.3% and 77.2%, respectively, and improves geometric alignment by 28.3%. These gains are obtained without authored proxy–video training pairs, demonstrating that ordinary RGBD supervision can support effective structural control on authored proxies.

Qualitative comparison. Figure 4 highlights two common failure modes. Depth-conditioned methods follow the supplied structure but can reproduce the coarse proxy's simplified body shapes, resulting in distorted characters and unnatural proportions. Camera-conditioned methods generate more natural-looking content, but can deviate from the proxy's scene layout, subject placement, and motion. Proxy2World combines structural adherence with geometric refinement: it preserves the



(a) Effect of the refinement ratio τ



(b) Component ablation

Figure 5: Effect of the refinement ratio and model components. (a) Varying τ with all other inputs and the random seed fixed illustrates the trade-off between proxy adherence and refinement. (b) Ablation of metric scale alignment and joint RGBD learning, evaluated through camera accuracy and human preference.

prescribed spatial relationships and subject motion while producing more natural character shapes and detailed scene appearance. The highlighted regions illustrate this difference, showing how our method refines coarse subjects while retaining their intended placement within the scene.

4.3 Ablation Studies

Hybrid denoising. Figure 5(a) examines how the two generation modes contribute to proxy-based control. Proxy-only generation preserves the prescribed layout and subject placement, but tends to carry simplified proxy shapes into the output, particularly for characters. Camera-only generation produces more natural shapes and richer appearance, but can alter subject placement and scene structure. Mixed generation preserves the major spatial relationships while refining coarse characters and adding scene details. These results support our use of early proxy grounding to establish structure and subsequent joint RGBD refinement to improve its visual realization, combining the strengths of both learned modes without additional proxy-specific training.

Metric scale alignment. We align depth and camera translations to a common metric scale across training scenes. Normalizing both consistently within each scene preserves their relative geometry, but does not establish a shared physical scale across scenes. Metric alignment therefore provides a consistent scale convention for learning from depth and camera-grid conditioning. Figure 5(b) shows that removing this alignment during training yields the largest camera error among the tested variants and substantially lowers human preference. These results highlight the importance of a shared metric scale across training scenes for accurate camera control and preferred visual results.

Joint RGBD learning. To test whether refinement requires joint geometry–appearance modeling, we replace joint RGBD generation with camera-conditioned RGB-only generation while retaining the initial depth-conditioned grounding stage. This variant still releases the fixed proxy-depth constraint during later sampling steps, but no longer generates depth alongside RGB. As shown in Figure 5(b), it yields higher camera error and lower human preference than the full model. The improvement therefore cannot be attributed solely to removing the depth constraint: explicitly generating geometry together with appearance contributes to both camera accuracy and visual quality. This supports joint RGBD learning as a key component of our refinement stage.

5 Conclusion

We presented Proxy2World, a unified RGBD world model that generates camera-controllable videos from lightweight proxies without paired proxy–video training data. Proxy–camera hybrid denoising combines structural grounding with joint RGBD refinement, balancing proxy adherence and visual quality as demonstrated on ProxyBench. Our approach lets creators specify coarse structure while leaving visual details to generation. Long-horizon interactive generation remains future work.

AI USE STATEMENT

We used generative AI tools to assist with drafting and polishing the manuscript, discovering related literature, and constructing scene proxies for ProxyBench. We also used a vision-language model for video evaluation, as described in our evaluation protocol. The authors take responsibility for the final manuscript, references, benchmark artifacts, and reported results.

ETHICS STATEMENT

Proxy2World generates synthetic videos from editable scene proxies. Like other video generation methods, it could be misused to produce misleading content or reproduce biases inherited from pretrained models and training data. Generated videos should not be presented as recordings of real events. Our human study compares synthetic videos to assess generation quality and controllability. VLM-based assessments may also reflect evaluator biases; we therefore complement them with operator-based metrics and human judgments.

REPRODUCIBILITY STATEMENT

Section 3 describes the control representations, model architecture, training objective, and hybrid sampling procedure. The experimental setup reports training data, optimization settings, and inference hyperparameters. Our ablation studies specify the model variants used to examine hybrid denoising, metric scale alignment, and joint RGBD learning.

REFERENCES

Nvidia Hassan Abu Alhaija, Jose M. Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, Dieter Fox, Yunhao Ge, Jinwei Gu, Ali Hassani, Michael Isaev, Pooya Jannaty, Shiyi Lan, Tobias Lasser, Huan Ling, Ming-Yu Liu, Xian Liu, Yifan Lu, Alice Luo, Qianli Ma, Hanzi Mao, Fabio Ramos, Xuanchi Ren, Tianchang Shen, Shitao Tang, Tingjun Wang, Jay Zhangjie Wu, Jiashu Xu, Stella Xu, Kevin Xie, Yunyao Ye, Xiaodong Yang, Xiaohui Zeng, and Yuan Zeng. Cosmos-transfer1: Conditional world generation with adaptive multimodal control. *ArXiv*, abs/2503.14492, 2025. URL `https://arxiv.org/abs/2503.14492`.

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. In *International Conference on Learning Representations*, volume 2025, pp. 66712–66737, 2025.

Yunpeng Bai, Shaoheng Fang, Chaohui Yu, Fan Wang, and Qixing Huang. Geovideo: Introducing geometric regularization into video generation model. *Advances in Neural Information Processing Systems*, 38:57602–57622, 2026.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*, 2024.

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. In *International Conference on Learning Representations*, volume 2025, pp. 37546–37593, 2025.

Houyuan Chen, Hong Li, Xianghao Kong, Tianrui Zhu, Shaocong Xu, Weiqing Xiao, Yuwei Guo, Chongjie Ye, Lvmin Zhang, Hao Zhao, et al. Unividx: A unified multimodal framework for versatile video generation via diffusion priors. *arXiv preprint arXiv:2605.00658*, 2026a.

Junhao Chen, Mingjin Chen, Henghaofan Zhang, Minglin Chen, Liaoyuan Fan, Boran Zhang, Saining Zhang, Mingze Sun, Hao Zhao, Ruqi Huang, Zhihao Li, and Yufei Wang. Video models as native 4d renderers: World-grounded conditioning from animated mesh. *arXiv preprint arXiv:2608.00094*, 2026b.

Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning. *arXiv preprint arXiv:2305.13840*, 2023.

Yiwen Chen, Guosheng Lin, and Chi Zhang. Code world model: Coding agent as world brain. *arXiv preprint arXiv:2608.25927*, 2026c.

Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=3q9vHEqsNx.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T Barron, et al. Image generators are generalist vision learners. *arXiv preprint arXiv:2604.20329*, 2026.

Gonzalo Gómez-Nogales, Yicong Hong, Chongjian Ge, Marc Comino-Trinidad, Dan Casas, and Yi Zhou. Coarse-to-real: Generative rendering for populated dynamic scenes. *ArXiv*, abs/2601.22301, 2026. URL https://arxiv.org/abs/2601.22301.

Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2025.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13416–13426. IEEE, 2025.

Chen Hou and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025a.

Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3r: 3d scene generation meets feed-forward reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25358–25369, June 2026a.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics (TOG)*, 44 (6):1–15, 2025b.

Zheng-Hui Huang, Guixu Lin, Jiacheng Lin, Yi-Chuan Huang, Ruihan Yu, Muyao Niu, Siqi Yang, Yu-Lun Liu, Yung-Yu Chuang, Kaipeng Zhang, et al. Programmable world model. *arXiv preprint arXiv:2609.10540*, 2026b.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 17191–17202, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yepan Xiong, Min Chen, Lin Cheng, and Xi Li. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28785–28796, October 2025.

Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, et al. Diffusionrenderer: Neural inverse and forward rendering with video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26069–26080. IEEE, 2025.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views,(2025). *arXiv preprint arXiv:2511.10647*, 5, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Jiwen Liu, Shujuan Li, Zhixue Fang, Xiaohan Li, Yan Zhou, Zijie Meng, Zhimin Zhang, Yawen Luo, Guoxin Zhang, Yu-Shen Liu, et al. Omnidirector: General multi-shot camera cloning without cross-paired data. *arXiv preprint arXiv:2606.13432*, 2026.

Zian Meng, Zhen Li, Chuanhao Li, Qiang Li, and Kaipeng Zhang. Marionette: Predicting world states, rendering geometry, painting appearance. *arXiv preprint arXiv:2608.14530*, 2026.

Stefan Popov, Amit Raj, Michael Krainin, Yuanzhen Li, William T Freeman, and Michael Rubinstein. Camctrl3d: Single-image scene exploration with precise 3d camera control. In *2025 International Conference on 3D Vision (3DV)*, pp. 649–658. IEEE, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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.

Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36:7594–7611, 2023a.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. 2023b.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025.

Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, pp. 331–348. Springer, 2024.

Xunzhi Xiang, Zixuan Duan, Yabo Chen, Zhengxuan Wei, Guiyu Zhang, Zixiao Gu, Zhe Gao, Haibin Huang, Chi Zhang, Qi Fan, et al. Videoweave: Unlocking geometric consistency in video generation via joint geometry-video modeling. *arXiv preprint arXiv:2606.14162*, 2026.

Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024.

Weilong Yan, Ming Li, Haipeng Li, Shuwei Shao, and Robby T. Tan. Synthetic-to-real self-supervised robust depth estimation via learning with motion and structure priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 21880–21890, June 2025.

Weilong Yan, Haipeng Li, Hao Xu, Nianjin Ye, Yihao Ai, Shuaicheng Liu, and Jingyu Hu. LaS-Comp: Zero-shot 3D completion with latent-spatial consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7588–7599, June 2026.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Minghao Yin, Jiahao Lu, Wenbo Hu, Wang Zhao, Shan Ying, and Kai Han. Scope: Sightline-coordinate positional encoding for video diffusion transformers. *ArXiv*, abs/2606.27345, 2026. URL `https://arxiv.org/abs/2606.27345`.

Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 100–111. IEEE, 2025.

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. *arXiv preprint arXiv:2409.02048*, 2024.

Hongfei Zhang, Kanghao Chen, Zixin Zhang, Harold H Chen, Yuanhuiyi Lyu, Kun Zhou, Yuqi Zhang, Shuai Yang, and Ying-cong Chen. Dualcamctrl: Dual-branch diffusion model for geometry-aware camera-controlled video generation. In *European Conference on Computer Vision*, pp. 655–678. Springer, 2026.

Qihang Zhang, Shuangfei Zhai, Miguel Martin, Kevin Miao, Alexander T. Toshev, Joshua M. Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21685–21695, 2025.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *TOG*, 37(4), July 2018. ISSN 0730-0301. doi: 10.1145/3197517.3201323. URL https://doi.org/10.1145/3197517.3201323.

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, et al. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. *arXiv preprint arXiv:2509.12201*, 2025.

Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8535–8546. IEEE, 2025.

A Appendix

A.1 Implementation Details

A.1.1 Training Data and Preprocessing

Our training formulation uses RGB videos with aligned depth, camera poses, and text descriptions rather than paired authored-proxy and realistic-video data. We train on 120,624 clips from DL3DV, RealEstate10K, and the gameplay data described in the main paper. RGB, depth, and rendered camera-grid signals are sampled on a common temporal grid and encoded into modality-specific latents. The five-second evaluation setting uses 81 frames at 16 fps: frame timestamps span 0–5 seconds, whereas the encoded 81-frame video lasts 5.0625 seconds. Cropping and resizing are applied consistently to the image-aligned signals, and camera intrinsics are adjusted for the crop and output resolution.

A.1.2 Metric Scale Alignment and Control Representations

Training-time metric alignment. We express training depth and camera translations in a common metric scale across scenes. For sequences with an arbitrary reconstruction scale, first-frame DA3 metric depth provides a scale anchor. The resulting sequence-level factor is applied to both depth and camera translations, preserving their relative geometry. Consistent within-scene normalization alone does not establish a shared physical scale across training scenes.

Reference-based alignment at inference. For ProxyBench cases with a reference image, DA3METRIC-LARGE predicts depth D_0^{DA3} from the selected first-frame image after the same center crop used for conditioning. Let D_0^P be the rendered proxy depth at that view. On finite, strictly positive overlapping pixels Ω, define

$$r_p = \log D_0^{\mathrm{DA3}}(p) - \log D_0^P(p), \tag{12}$$

$$m = \mathrm{median}_{p\in\Omega}\, r_p, \qquad a = \mathrm{median}_{p\in\Omega}\, |r_p - m|, \tag{13}$$

$$\Omega' = \{p \in \Omega : |r_p - m| \le \max(4.5 \cdot 1.4826a, 0.15)\}, \tag{14}$$

$$s = \exp(\mathrm{median}_{p\in\Omega'}\, r_p)\,. \tag{15}$$

DA3 remains the scale anchor: proxy depths are multiplied by s, and camera centers are transformed as $\mathbf{c}'_t = \mathbf{c}_0 + s(\mathbf{c}_t - \mathbf{c}_0)$. Rotations are unchanged; intrinsics change only through image preprocessing.

Why the common scale matters. Jointly scaling depth and translation preserves projective geometry, whereas scaling only one changes the displacement-to-depth ratio and the expected parallax. Moreover, our encodings are not independently scale invariant: depth colors use a fixed metric transfer function and camera grids use fixed metric spacing. A shared metric convention therefore gives depth and camera-grid conditioning a consistent physical interpretation across training scenes and at inference. DA3 supplies an estimated metric anchor; its estimation error can affect the resulting scale.

Table 2: Training and inference settings for Proxy2World.

Setting	Value
Base	Wan2.2 A14B, high/low experts
Expert boundary / noise shift	0.9 / 3.0
Evaluation resolution	480×832 (height $\times$ width)
Evaluation temporal sampling	81 frames, 16 fps
Staged sampler	Flow-matching Euler, 40 steps
Text CFG	3.5
AdamW learning rate	10^{-5}
Reference-RGB dropout	0.5
Depth-to-RGB / joint RGBD task probabilities	0.5 / 0.5

Depth-color mapping. For nonnegative depth d in meters, we use

$$u(d) = 1 - (1 + d/10)^{-2}, \quad z = 7u, \quad j = \min(\lfloor z \rfloor, 6), \quad H(d) = (1 - z + j)\mathbf{h}_j + (z - j)\mathbf{h}_{j+1}, \tag{16}$$

where the RGB cube vertices are $(000, 001, 011, 010, 110, 100, 101, 111)$ in this order. Invalid values map to black; valid zero depth also maps to black. This is a fixed continuous seven-segment color encoding, not a per-frame min–max normalization. The default mapping has no finite maximum-depth cutoff. Metric depth can be recovered from continuous decoded colors by projection onto the color polyline followed by inversion, $d = 10((1-u)^{-1/2} - 1)$. Decoding errors and color saturation limit the precision of this recovery.

Camera grids. The camera-grid signal is a video rendered from camera poses and intrinsics in the same spatial convention. The camera-grid rendering configuration specifies 2 m lattice spacing, 6 m grid height, 0.08 m line width, 0.045 m pillar radius, pillars every two cells, a 0.05–16 m ray range, 72 ray-marching steps, and a 1 m room-transition setting.

A.1.3 Training and Sampling Configuration

Our model extends Wan2.2 A14B with high- and low-noise experts and modality-dependent timesteps. RGB and depth are separate token sequences, allowing depth to remain clean while RGB is noisy. Grid latents are concatenated in channels within each modality stream. Each modality input has 52 channels: 16 state channels, 16 camera-grid channels, four reference-mask channels, and 16 reference-RGB channels. First-frame depth is not concatenated as an additional input.

We sample depth-conditioned RGB generation and joint RGBD generation with equal probability. In depth-conditioned RGB generation, depth remains clean and only RGB is supervised. In joint RGBD generation, both modalities share the noise level and expert route, with independent Gaussian noise and unit loss weights. The corruption and velocity targets follow the formulation in the main paper.

Hybrid sampling. For the descending schedule $1 = t_0 > \cdots > t_N = 0$, we use $k_s = \lfloor (1-\tau)N \rfloor$ proxy-grounding steps and $N - k_s$ joint-refinement steps. With $N = 40$ and $\tau^\star = 0.925$, these are three and 37 steps, respectively. At the switch, we retain the RGB state and initialize depth as $\mathbf{z}^h_{t_{k_s}} = (1 - t_{k_s})\mathbf{z}^{h,S} + t_{k_s}\boldsymbol{\epsilon}^h$. Each active modality is updated by Euler integration, $\mathbf{z}^m_{t_{k+1}} = \mathbf{z}^m_{t_k} + (t_{k+1} - t_k)\mathbf{v}^m_\theta$. The depth state remains fixed before the switch and is updated jointly with RGB afterward. The refinement fraction τ specifies the step allocation; t_{k_s} specifies the noise level used for depth initialization.

A.2 ProxyBench Construction

A.2.1 Agent-Assisted Proxy Construction

Our authoring workflow separates spatial specification from appearance. A scene description guides an agent to construct editable geometry and behavior scripts; a renderer then records synchronized RGB, depth, normals, and camera parameters. Appearance prompts and first-frame synthesis provide the visual realization separately from the geometric recording. The resulting clip package

retains the proxy streams, camera trajectory, first-frame image, first-frame prompt, and video caption.

A.2.2 Scene, Camera, and Interaction Design

Proxy geometry specifies layout, occupied regions, openings, subject paths, and interaction timing while leaving material, texture, and fine anatomy open. Camera poses and intrinsics are saved per frame, and all recorded streams are segmented using identical temporal boundaries. Incomplete terminal segments are excluded.

Evaluation subset. ProxyBench contains 25 scenes and 300 five-second clips. Quantitative operator-based and VLM evaluations use 78 matched clips from five worlds: Mountain Harbor, Neon Lockdown, Redrock Mining, Sunken Royal City, and Collapsed City Rescue. We select these five worlds for their rich interaction designs, which allow us to evaluate control over subject behavior and scene interactions in addition to camera motion. Human evaluation uses ten clips, as detailed below.

A.3 Evaluation Protocols

A.3.1 Baseline Configurations

Each baseline uses its supported input modalities. Camera-conditioned baselines receive camera/appearance inputs; geometry-conditioned methods receive their appropriate rendered controls. FantasyWorld uses official scene normalization and 50 sampling steps; SCoPE/RAYPE uses first-depth 25th-percentile normalization and 40 steps; both use seed 1024, 81 frames, 480×832, and 16 fps. Lyra2.0 DMD uses the first 81 source frames, raw DA3 first-frame depth, and a pose-scale multiplier of 1.0.

A.3.2 Camera-Control Evaluation via Video Reconstruction

Evaluation Setup We evaluate camera adherence on 78 matched clips from five authored worlds using ten model configurations, yielding 780 generated videos. Each video contains 81 frames at 16 FPS with a resolution of 832×480. We reconstruct each complete sequence using ViPE and compare the estimated camera trajectory with the prescribed input trajectory. Only generated RGB videos are used for reconstruction; reference poses and generated depth predictions are not provided to ViPE. All methods use the same reconstruction configuration: estimated pinhole intrinsics, the DA3 depth pipeline, a multi-view window of 81 frames, an overlap setting of 20 frames, and a processing resolution of 504.

Foreground Exclusion Moving foreground objects can bias both camera estimation and reprojection-based evaluation. We therefore enable ViPE's instance segmentation and tracking with the prompts `person`, `vehicle`, `car`, `truck`, `bus`, `motorcycle`, and `bicycle`. Detection is refreshed every second, with masks tracked across all frames. Sky segmentation is disabled. All detected instances of these categories are excluded, including stationary vehicles. A 5×5 dilation is applied to the foreground masks to reduce boundary contamination. The corresponding regions are excluded from ViPE's pose optimization. For reprojection evaluation, we additionally reject any correspondence whose observed source or target location lies inside the dilated masks. We retain per-frame mask coverage and overlay videos for inspection.

Normalized Trajectory Errors We report normalized absolute trajectory errors for translation and rotation, denoted by nATE_t and nATE_r. Our protocol uses the actual prescribed camera trajectory as the reference.

Let $\widehat{T}_i = [\widehat{R}_i, \widehat{\mathbf{p}}_i]$ and $T_i = [R_i, \mathbf{p}_i]$ denote the reconstructed and reference camera-to-world poses. Each trajectory is first expressed in its own first-camera coordinate system:

$$\widehat{T}_i^0 = \widehat{T}_0^{-1}\widehat{T}_i, \qquad T_i^0 = T_0^{-1}T_i. \tag{17}$$

Below, we omit the superscript for readability.

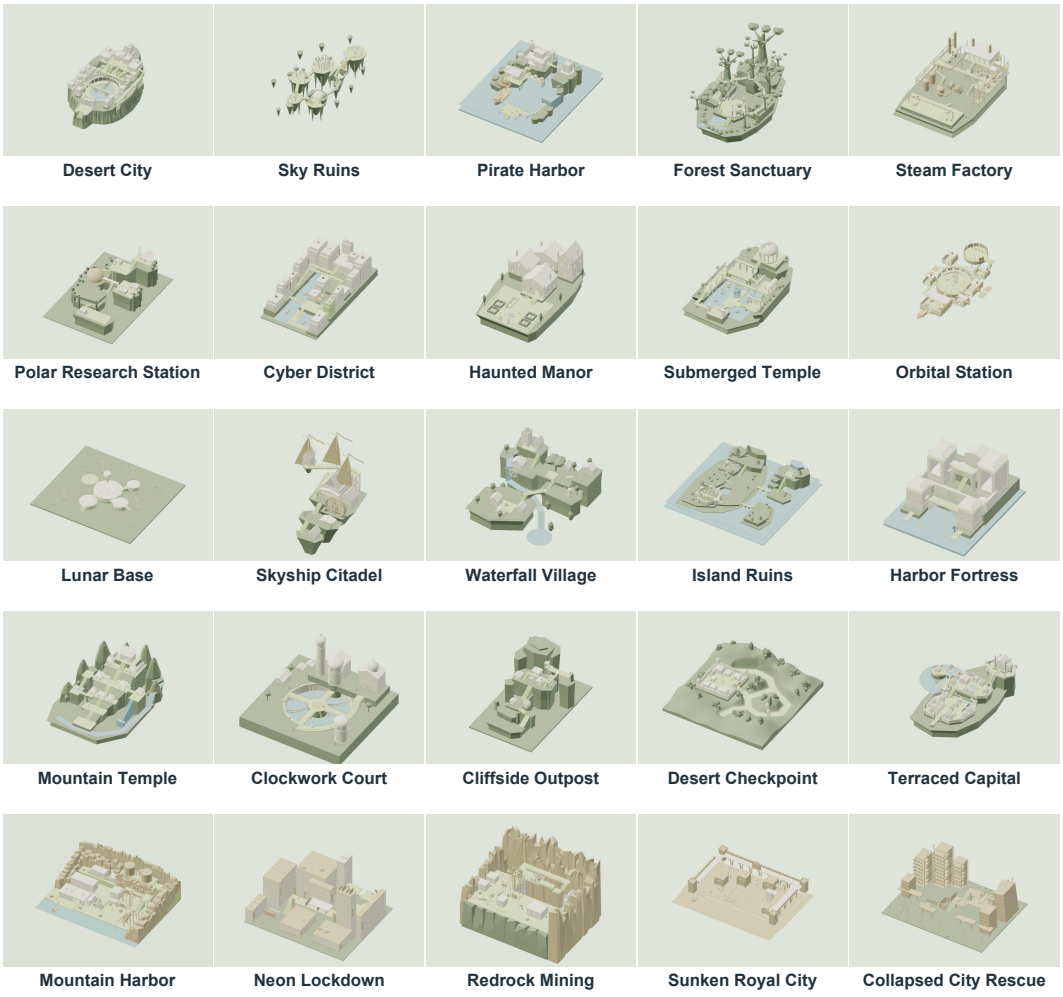


Figure 6: Overview of 25 authored proxy maps. The overview illustrates layouts and scene types.

To remove global translation-scale ambiguity, we match the total path length of the reconstructed trajectory to that of the reference trajectory:

$$\widehat{L} = \sum_{i=1}^{F-1} \|\widehat{\mathbf{p}}_i - \widehat{\mathbf{p}}_{i-1}\|_2, \qquad L = \sum_{i=1}^{F-1} \|\mathbf{p}_i - \mathbf{p}_{i-1}\|_2, \qquad s = \frac{L}{\widehat{L}}, \tag{18}$$

where $F = 81$. Only a positive scalar is applied to translation; we do not fit an additional spatial rotation or reflection.

Both trajectories are independently resampled to $N = 60$ uniformly spaced points along their respective translation arc lengths. Positions are linearly interpolated, and orientations are interpolated using spherical linear interpolation (SLERP). Let the resulting samples be $(\widehat{\mathbf{p}}^{(k)}, \widehat{R}^{(k)})$ and $(\mathbf{p}^{(k)}, R^{(k)})$. We compute

$$\mathrm{ATE}_t = \frac{1}{N} \sum_{k=1}^{N} \|s\widehat{\mathbf{p}}^{(k)} - \mathbf{p}^{(k)}\|_2, \tag{19}$$

$$\mathrm{ATE}_r = \frac{1}{N} \sum_{k=1}^{N} \theta\Big((\widehat{R}^{(k)})^\top R^{(k)} \Big), \tag{20}$$

where $\theta(\cdot)$ is the rotation geodesic angle in degrees. These are mean errors, not root-mean-square errors.

The normalized metrics are

$$\mathrm{nATE}_t = \mathrm{clip}_{[0,1]} \left(\frac{\mathrm{ATE}_t}{\max(s\widehat{L},\, 0.5)} \right), \tag{21}$$

$$\mathrm{nATE}_r = \mathrm{clip}_{[0,1]} \left(\frac{\mathrm{ATE}_r}{\max(\widehat{\Phi},\, 10^\circ)} \right), \tag{22}$$

where

$$\widehat{\Phi} = \sum_{i=1}^{F-1} \theta(\widehat{R}_{i-1}^\top \widehat{R}_i) \tag{23}$$

is the accumulated rotation of the original reconstructed trajectory.

Background Reprojection Error We independently evaluate whether the reconstructed geometry is consistent with observed background motion. For each selected source frame, we detect up to 1,000 Shi–Tomasi corners outside the foreground mask and track them using pyramidal Lucas–Kanade optical flow with a 21×21 window and `maxLevel=4`. We evaluate frame gaps of 1 and 8, with source frames sampled at a stride of 4. Correspondences are retained only when their forward–backward tracking error is below one pixel, both observed endpoints lie outside the dilated foreground masks, and the depth and projection are valid. No reprojection-residual threshold is used to remove large errors. For a retained correspondence $(\mathbf{u}_i, \mathbf{u}_j)$, the reconstructed depth $\widehat{D}_i$ and intrinsics $\widehat{K}_i$ lift the source point into 3D:

$$\mathbf{X}_i = \widehat{D}_i(\mathbf{u}_i) \widehat{K}_i^{-1} \widetilde{\mathbf{u}}_i. \tag{24}$$

It is transformed and projected into the target frame using

$$\widehat{\mathbf{u}}_j = \pi\left(\widehat{K}_j \left[\widehat{T}_j^{-1} \widehat{T}_i \begin{pmatrix} \mathbf{X}_i \\ 1 \end{pmatrix} \right]_{1:3} \right). \tag{25}$$

The reprojection residual is $\|\widehat{\mathbf{u}}_j - \mathbf{u}_j\|_2$ in pixels.

We first average residuals within each video and then average equally across videos. Thus, videos with more valid matches do not receive greater weight. This metric measures reconstructed geometric consistency, not adherence to the prescribed camera: a video can have a low reprojection error while following an incorrect trajectory.

Method	$\mathrm{nATE}_t \downarrow$	$\mathrm{nATE}_r \downarrow$	Reproj. (px)↓
RAYPE / SCoPE	0.1066	0.1229	4.338
Lyra2	0.0418	0.0413	2.171
FantasyWorld	0.2966	0.3472	2.449
Ours, TI2V $P = 1$	0.0681	0.0524	**1.956**
Cosmos-Depth	0.0446	0.0362	3.259
VACE-Depth	0.0764	0.0584	6.264
Ours, T2V $P = 0$	**0.0302**	**0.0270**	**2.280**
C2R	0.1637	0.3152	5.312
Ours, T2V $P = P^*$	0.0863	0.0719	**2.067**
Ours, TI2V $P = P^*$	**0.0659**	**0.0558**	2.524

Table 3: Full camera-control evaluation on 78 matched clips. Reprojection errors are measured outside detected person/vehicle masks. All values are per-video macro averages.

Results The common evaluation set requires valid nondegenerate translation paths and at least 100 valid background correspondences for every configuration. All 78 clips satisfy these criteria. Table 3 reports the full 78-clip evaluation. Lower values are better for all metrics. Our T2V $P = 0$ configuration achieves the lowest normalized translation and rotation errors, while our TI2V $P = 1$ configuration obtains the lowest background reprojection error.

A.3.3 Operator-based Evaluation

The main table reports two operator-based, condition-aware scores for each matched generated video and proxy-depth sequence. Both are in $[0, 1]$ (higher is better), and we average clip scores over the 78 evaluation sequences. These scores are separate from VBench and from our VLM evaluation.

Geo Align. We sample 9 frames from each 81-frame clip, decode the proxy depth, and estimate the RGB video's depth with Depth Anything V2 (Yang et al., 2024). Each depth map is independently normalized by its 2nd and 98th percentiles. At each frame, r is the absolute Spearman depth-rank correlation, a is 1 minus the standard-deviation-normalized RMSE after the best affine fit (clipped to $[0, 1]$), and e is depth-edge F1 with a two-pixel tolerance. For any framewise score x, let $L(x)$ combine its mean and its lowest 20%:

$$L(x) = 0.7\,\overline{x} + 0.3\,\overline{x}_{\text{worst }20\%}, \qquad S_{\text{Geo}} = \frac{L(r) + L(a) + L(e)}{3}. \tag{26}$$

This tests coarse geometric layout without requiring the generated image to copy the proxy's appearance.

Temp Align. We compare Farneback optical flow between matched adjacent frames (sampling every other transition). Let s_{flow} denote flow agreement based on endpoint error normalized by the two flow magnitudes; s_{dir} is flow-direction cosine on moving proxy pixels, mapped to $[0, 1]$; and s_{prof} is the similarly mapped Spearman correlation between the two temporal motion-magnitude profiles. To capture abrupt failures, we also compare frame-to-frame edge changes using median/MAD-normalized scores: a spike is *unmatched* when the output score exceeds 6 but the proxy score is below 3. If u is the fraction of unmatched transitions, set $s_{\text{spike}} = \exp(-10u)$. The final score is

$$\begin{aligned} S_{\text{Temp}} = {} & 0.35\,\overline{s}_{\text{flow}} + 0.20\,\overline{s}_{\text{flow,worst }20\%} \\ & + 0.15\,s_{\text{dir}} + 0.15\,s_{\text{prof}} + 0.15\,s_{\text{spike}}. \end{aligned} \tag{27}$$

Undefined components are omitted and the remaining weights renormalized.

A.3.4 VLM Evaluation

We use `gemini-3.7-flash` to evaluate proxy adherence and generative refinement on all 78 matched clips. For each pair of generated videos, the model receives the proxy reference and is asked the same criterion-specific question used in the human evaluation. Adherence concerns preservation of the intended layout, subject placement, and behavior. Refinement concerns plausible enrichment of coarse geometry, appearance, and motion. The comparison allows either candidate to be preferred or the pair to be judged approximately equal.

Table 4: Human preference rates (%) over valid responses, including partial sessions. Parentheses give comparison counts.

Method	Adherence	Refinement	Overall
RAYPE	24.0% (52)	30.4% (51)	20.2% (52)
Lyra2	12.3% (53)	18.3% (52)	10.4% (53)
FantasyWorld	18.4% (57)	23.5% (51)	32.0% (50)
Ours CAM-TI2V	42.3% (65)	71.4% (56)	57.9% (38)
Cosmos-Depth	72.2% (54)	61.1% (54)	71.4% (49)
VACE-Depth	81.8% (44)	44.0% (50)	65.9% (63)
C2R	43.1% (51)	33.0% (56)	27.0% (50)
Ours T2V $P = P^*$	75.0% (58)	88.2% (51)	82.3% (48)
Ours TI2V $P = P^*$	69.0% (58)	90.4% (47)	84.9% (53)
Ours T2V $P = 0$	67.7% (48)	38.9% (36)	47.9% (24)

For methods a and b, we compute the criterion-specific pairwise win rate

$$\mathrm{WR}_{a,b} = \frac{W_{a,b} + \frac{1}{2}T_{a,b}}{W_{a,b} + L_{a,b} + T_{a,b}}, \tag{28}$$

where $W_{a,b}$, $L_{a,b}$, and $T_{a,b}$ count wins, losses, and ties across evaluated clips. We report the mean pairwise win rate across the other methods, expressed as a percentage. VLM and human results use the same evaluation criteria but different clip coverage and are reported separately.

A.3.5 Human Evaluation

Each trial displays a whitebox reference and two anonymized generated videos with synchronized five-second playback. A criterion-specific question asks about adherence, refinement, or overall preference. Participants select left, right, or "About equal."

We evaluate ten method configurations on ten clips using 762 valid pairwise judgments: 720 from ten completed questionnaires and 42 from partially completed questionnaires. Each completed questionnaire contains 72 comparisons, with 24 questions each on adherence, refinement, and overall preference. The final analysis includes 270 adherence judgments, 252 refinement judgments, and 240 overall-preference judgments. Comparison counts are reported for each method and each method pair.

For each criterion, empirical preference is $\mathrm{PR} = (W + T/2)/(W + L + T)$. We report counts, retain valid partial-session responses, and exclude TEST/practice, undone and skipped responses. Each cell of the pairwise tables is the row method preferred over the column method. The aggregate rates pool the valid comparisons involving each method. Because comparison counts and opponent coverage vary, these are descriptive preference estimates. The pairwise matrices provide the corresponding comparison counts.

B Additional Qualitative Results

We present the first matched example from each of the five evaluation worlds to illustrate structural control and generative refinement. Static comparisons use 1 s and 3 s; temporal examples use 0–4 s. Conditions differ across methods according to their supported control modalities.

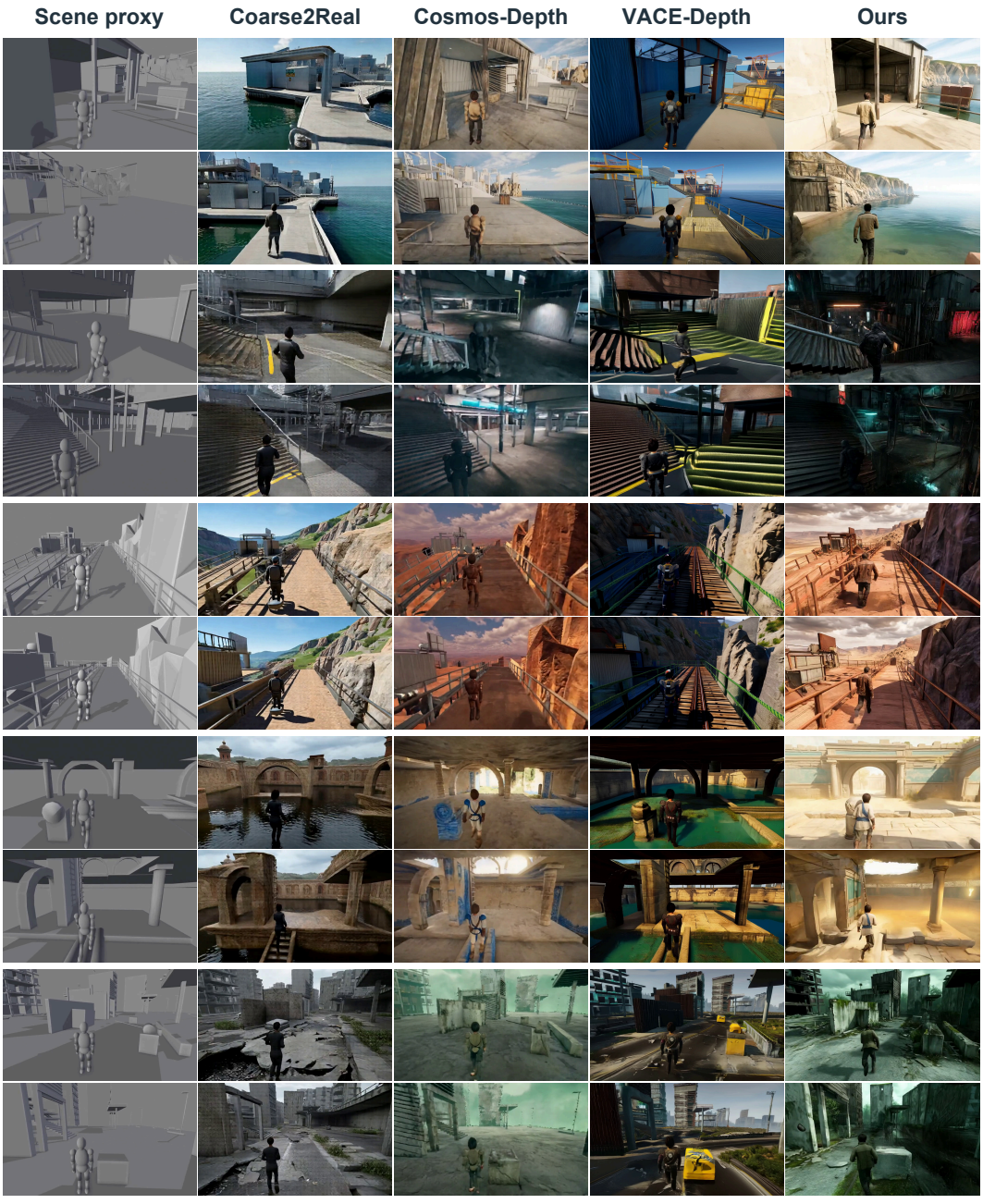


Figure 7: Geometry-conditioned comparison. Each pair of rows shows 1 s and 3 s from one clip, ordered Mountain Harbor, Neon Lockdown, Redrock Mining, Sunken Royal City, and Collapsed City Rescue. Columns show the proxy, Coarse2Real, Cosmos-Depth, VACE-Depth, and our staged proxy-conditioned TI2V result. The examples illustrate the trade-off between reproducing coarse proxy geometry and generating natural subject shapes. Our mixed TI2V results retain major spatial relationships while enriching appearance. Each system uses its stated input conditions.



Figure 8: Camera-controlled comparison at the same two times and in the same world order as Figure 7. The proxy is a visual reference and is not supplied to all baselines. Viewpoint control alone does not impose the proxy terrain or subject motion. The Ours column uses staged proxy-conditioned TI2V, not the camera-only configuration in the quantitative table.

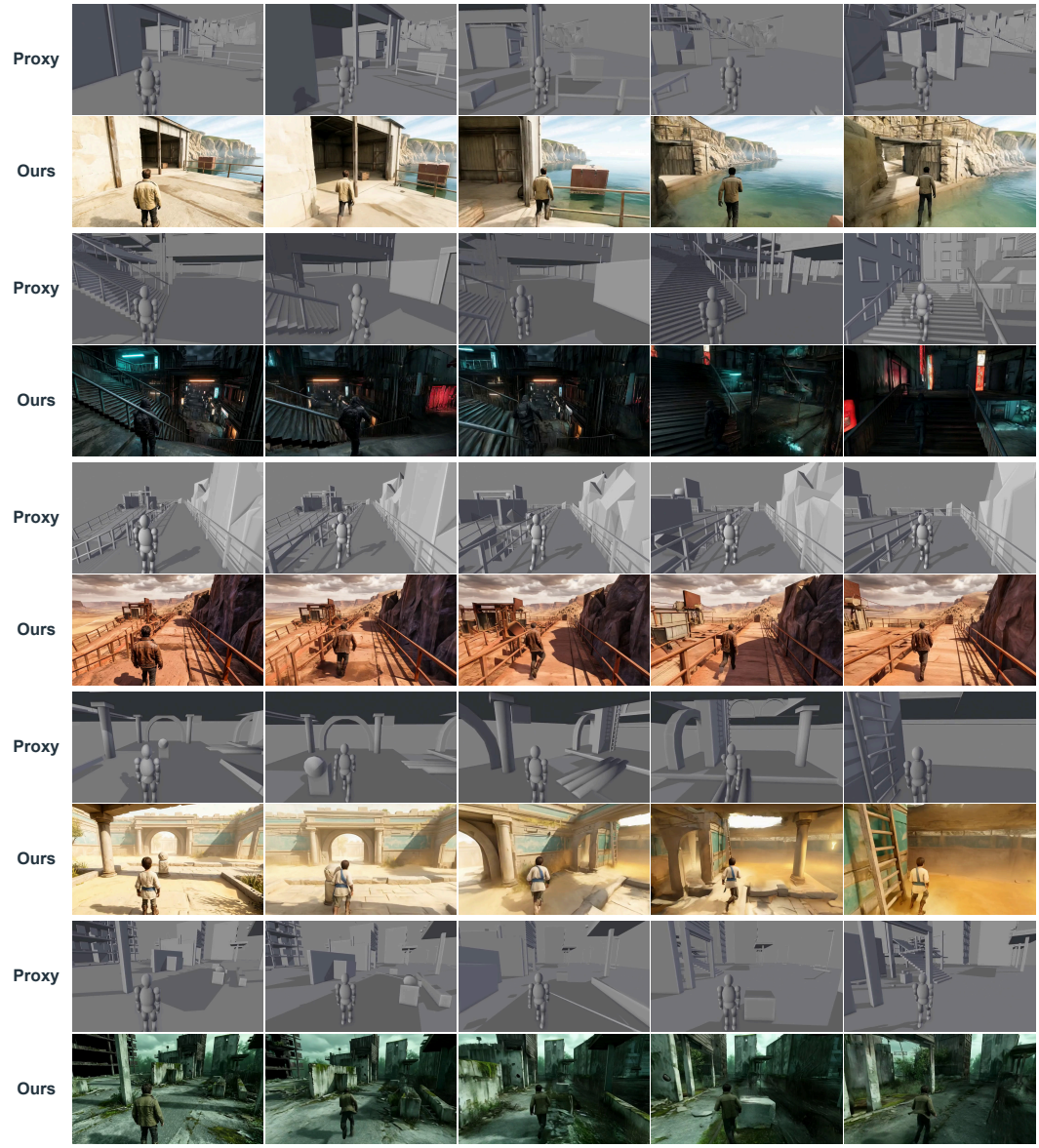


Figure 9: Proxy-to-world sequences in five authored worlds. In each two-row block the proxy appears above the generated RGB; columns sample 0, 1, 2, 3, and 4 s. The sequences illustrate how proxy layout and subject motion guide generation while materials, lighting, and fine geometry are supplied by the model.

Table 5: Whitebox adherence: empirical pairwise preference rates. Entries report the percentage preferring the row method over the column method, with ties counted as half a win; comparison counts are in parentheses. This interim snapshot includes valid responses from incomplete sessions.

Row / Column	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	–	77.8% (9)	58.3% (6)	8.3% (6)	8.3% (6)	0.0% (4)	25.0% (2)	0.0% (8)	0.0% (6)	10.0% (5)
M2	22.2% (9)	–	41.7% (6)	0.0% (5)	0.0% (6)	0.0% (6)	16.7% (9)	0.0% (6)	0.0% (1)	10.0% (5)
M3	41.7% (6)	58.3% (6)	–	12.5% (8)	0.0% (7)	0.0% (3)	25.0% (8)	14.3% (7)	8.3% (6)	0.0% (6)
M4	91.7% (6)	100.0% (5)	87.5% (8)	–	7.1% (7)	25.0% (4)	43.8% (8)	12.5% (8)	17.9% (14)	30.0% (5)
M5	91.7% (6)	100.0% (6)	100.0% (7)	92.9% (7)	–	41.7% (6)	66.7% (3)	50.0% (8)	58.3% (6)	40.0% (5)
M6	100.0% (4)	100.0% (6)	100.0% (3)	75.0% (4)	58.3% (6)	–	83.3% (6)	66.7% (3)	100.0% (6)	58.3% (6)
M7	75.0% (2)	83.3% (9)	75.0% (8)	56.3% (8)	33.3% (3)	16.7% (6)	–	0.0% (5)	0.0% (5)	10.0% (5)
M8	100.0% (8)	100.0% (6)	85.7% (7)	87.5% (8)	50.0% (8)	33.3% (3)	100.0% (5)	–	50.0% (8)	50.0% (5)
M9	100.0% (6)	100.0% (1)	91.7% (6)	82.1% (14)	41.7% (6)	0.0% (6)	100.0% (5)	50.0% (8)	–	75.0% (6)
M10	90.0% (5)	90.0% (5)	100.0% (6)	70.0% (5)	60.0% (5)	41.7% (6)	90.0% (5)	50.0% (5)	25.0% (6)	–

M1: RAYPE; M2: Lyra2; M3: FantasyWorld; M4: Ours CAM-TI2V; M5: Cosmos-Depth; M6: VACE-Depth; M7: C2R; M8: Ours T2V $P = P^*$; M9: Ours TI2V $P = P^*$; M10: Ours T2V $P = 0$ (depth40).

Table 6: Refinement: empirical pairwise preference rates. Entries report the percentage preferring the row method over the column method, with ties counted as half a win; comparison counts are in parentheses. This interim snapshot includes valid responses from incomplete sessions.

Row / Column	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	–	83.3% (6)	25.0% (4)	16.7% (9)	16.7% (6)	37.5% (8)	28.6% (7)	0.0% (4)	0.0% (4)	66.7% (3)
M2	16.7% (6)	–	50.0% (6)	11.1% (9)	0.0% (6)	0.0% (2)	62.5% (4)	0.0% (6)	0.0% (9)	50.0% (4)
M3	75.0% (4)	50.0% (6)	–	25.0% (6)	0.0% (8)	14.3% (7)	42.9% (7)	0.0% (3)	0.0% (7)	16.7% (3)
M4	83.3% (9)	88.9% (9)	75.0% (6)	–	75.0% (8)	83.3% (6)	70.0% (5)	40.0% (5)	0.0% (2)	58.3% (6)
M5	83.3% (6)	100.0% (6)	100.0% (8)	25.0% (8)	–	42.9% (7)	91.7% (6)	8.3% (6)	25.0% (4)	66.7% (3)
M6	62.5% (8)	100.0% (2)	85.7% (7)	16.7% (6)	57.1% (7)	–	20.0% (5)	0.0% (7)	20.0% (5)	66.7% (3)
M7	71.4% (7)	37.5% (4)	57.1% (7)	30.0% (5)	8.3% (6)	80.0% (5)	–	0.0% (10)	7.1% (7)	30.0% (5)
M8	100.0% (4)	100.0% (6)	100.0% (3)	60.0% (5)	91.7% (6)	100.0% (7)	100.0% (10)	–	40.0% (5)	90.0% (5)
M9	100.0% (4)	100.0% (9)	100.0% (7)	100.0% (2)	75.0% (4)	80.0% (5)	92.9% (7)	60.0% (5)	–	100.0% (4)
M10	33.3% (3)	50.0% (4)	83.3% (3)	41.7% (6)	33.3% (3)	33.3% (3)	70.0% (5)	10.0% (5)	0.0% (4)	–

M1: RAYPE; M2: Lyra2; M3: FantasyWorld; M4: Ours CAM-TI2V; M5: Cosmos-Depth; M6: VACE-Depth; M7: C2R; M8: Ours T2V $P = P^*$; M9: Ours TI2V $P = P^*$; M10: Ours T2V $P = 0$ (depth40).

Table 7: Overall preference: empirical pairwise preference rates. Entries report the percentage preferring the row method over the column method, with ties counted as half a win; comparison counts are in parentheses. This interim snapshot includes valid responses from incomplete sessions.

Row / Column	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	–	50.0% (4)	31.3% (8)	0.0% (3)	0.0% (6)	16.7% (6)	55.6% (9)	0.0% (6)	0.0% (8)	0.0% (2)
M2	50.0% (4)	–	42.9% (7)	0.0% (4)	0.0% (6)	5.0% (10)	0.0% (5)	0.0% (6)	0.0% (8)	0.0% (3)
M3	68.8% (8)	57.1% (7)	–	12.5% (4)	0.0% (4)	25.0% (8)	66.7% (3)	12.5% (8)	20.0% (5)	0.0% (3)
M4	100.0% (3)	100.0% (4)	87.5% (4)	–	25.0% (4)	37.5% (8)	83.3% (6)	10.0% (5)	0.0% (2)	100.0% (2)
M5	100.0% (6)	100.0% (6)	100.0% (4)	75.0% (4)	–	40.0% (5)	77.8% (9)	37.5% (4)	31.3% (8)	100.0% (3)
M6	83.3% (6)	95.0% (10)	75.0% (8)	62.5% (8)	60.0% (5)	–	100.0% (7)	37.5% (8)	25.0% (8)	33.3% (3)
M7	44.4% (9)	100.0% (5)	33.3% (3)	16.7% (6)	22.2% (9)	0.0% (7)	–	0.0% (3)	0.0% (6)	25.0% (2)
M8	100.0% (6)	100.0% (6)	87.5% (8)	90.0% (5)	62.5% (4)	62.5% (8)	100.0% (3)	–	50.0% (5)	100.0% (3)
M9	100.0% (8)	100.0% (8)	80.0% (5)	100.0% (2)	68.8% (8)	75.0% (8)	100.0% (6)	50.0% (5)	–	100.0% (3)
M10	100.0% (2)	100.0% (3)	100.0% (3)	0.0% (2)	0.0% (3)	66.7% (3)	75.0% (2)	0.0% (3)	0.0% (3)	–

M1: RAYPE; M2: Lyra2; M3: FantasyWorld; M4: Ours CAM-TI2V; M5: Cosmos-Depth; M6: VACE-Depth; M7: C2R; M8: Ours T2V $P = P^*$; M9: Ours TI2V $P = P^*$; M10: Ours T2V $P = 0$ (depth40).