UNVEIL AS YOU GO: RETHINKING SPATIAL EXPLORATION FOR SINGLE-IMAGE WORLD GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The integration of generative models and reconstruction networks has emerged as a promising paradigm for 3D Gaussian world generation. However, existing methods mostly treat camera exploration as a plug-and-play input rather than scene-adaptive understanding, and the spatial priors encoded in pretrained vision–language models remain underexploited. To address this, we propose UnveilWorld, in which the explorer performs recursive planning and understanding and the generator unveils unobserved views, jointly contributing to the final world reconstruction. To further investigate the synergy between spatial understanding and generation, UnveilWorld is post-trained with hard-state step-wise AWM, pairing stage-wise group-relative credit assignment with failure-biased state propagation. The joint optimization of the full pipeline closes the loop among exploration, generation, and reconstruction end-to-end. Experiments demonstrate that UnveilWorld generates high-quality 3D worlds with strong cross-view consistency in 18.12 seconds on a single NVIDIA A800 GPU.

1 INTRODUCTION

From a single image, humans effortlessly imagine the world beyond the frame and know *where to go* to find out; endowing machines with this ability would benefit embodied simulation, virtual reality, and interactive content creation. Explicit representations such as 3D Gaussian Splatting (Kerbl et al., 2023) have made high-quality, real-time novel-view rendering practical, and a growing line of work couples generative image or video priors with explicit point- or Gaussian-based scene representations (Yu et al., 2024; 2025). Yet the generated worlds remain confined to a narrow, high-quality region around the input view: even when long-range camera trajectories are supplied, the generator tends to fall back on short-range novel views (Li et al., 2026). We argue that this limitation is rooted not in generation, but in how exploration itself is conceived.

In existing pipelines, exploration is treated as an externally prescribed sequence of camera poses rather than a scene-adaptive act of understanding: trajectories are either fixed a priori (Yang et al., 2025; Wang et al., 2026b) or confined to local neighborhoods with highly redundant observations, and they are typically distance-normalized without being aligned to scene geometry, failing to steer the generator toward unexplored regions. Meanwhile, pretrained vision–language models encode rich spatial understanding (Chen et al., 2024a) and ground language-based planning in continuous trajectory generation (Wei et al., 2026), yet these spatial priors remain largely underexploited for world generation.

We propose UnveilWorld, which treats single-image Gaussian world generation as a closed loop of exploration, generation, and reconstruction: the scene itself decides where to look, looking further unveils new content, and what is unveiled becomes an enduring, explorable world. An exploration module reads the input image through the spatial priors of a vision–language model and, grounded by a depth-based spatial anchor, plans a geometry-aligned trajectory together with imagined descriptions of what each step should reveal; a four-step diffusion generator then unveils unobserved views along this trajectory as explicit 3D Gaussians, while rendering-based supervision from the reconstructed world flows back to shape the explorer itself. An automatic, annotation-free data engine supplies the trajectory–description–geometry aligned data that makes this loop learnable.

To optimize this loop, we introduce hard-state step-wise AWM (Xue et al., 2026), an RL-based post-training scheme that assigns credit from group-wise candidate comparisons at every denoising stage

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.

while deliberately propagating the hardest candidate onward, so that later stages learn to recover from failure rather than merely avoid it. Experiments demonstrate that UnveilWorld generates high-quality 3D worlds with strong cross-view consistency in 18.12 seconds on a single NVIDIA A800 GPU. Meanwhile, the recovered videos closely align with the prescribed controls, demonstrating the high trajectory-control fidelity of UnveilWorld.

Our contributions are summarized as follows:

- We propose UnveilWorld, which formulates single-image Gaussian world generation as a mutually reinforcing exploration–generation–reconstruction loop, elevating exploration from externally prescribed camera poses to a scene-adaptive act of understanding.
- We design a learnable exploration module that couples a shared depth-based spatial anchor with vision–language priors to plan scene-adaptive, geometry-aligned exploration trajectories, together with a data engine that supplies the required trajectory–description–geometry aligned training data.
- We introduce hard-state step-wise AWM, which decouples credit assignment from state visitation via group-relative stage-wise advantages and failure-biased state propagation, enabling effective RL-based post-training of the iterative 3D generator.

2 Related Work

3D Scene Reconstruction and Generation. Feed-forward 3D reconstruction has advanced rapidly: DUSt3R and VGGT regress point maps, camera parameters, and depth from uncalibrated images (Wang et al., 2024; 2025), and AnySplat recovers Gaussian primitives from unconstrained image collections, all building on the explicit and efficient 3DGS representation (Kerbl et al., 2023; Jiang et al., 2025). These methods recover geometry supported by the input views, whereas generative world models leverage image or video priors to synthesize plausible content in unobserved regions. Early systems iteratively inpaint novel views into textured meshes or layered surfels (Höllein et al., 2023; Yu et al., 2025); camera-conditioned video diffusion enables controlled world exploration at panoramic and persistent scales (Yang et al., 2025; Shen et al., 2026; Schwarz et al., 2025); and recent feed-forward systems generate 3D Gaussians directly from one or a few views, steadily improving generation speed, cross-view consistency, and scene scale (Li et al., 2026; Ni et al., 2025; Wang et al., 2026b; Team HY-World, 2026). Yet the high-quality explorable region of generated worlds still concentrates near the input views—the gap our work sets out to close.

Camera Control and Scene-Adaptive View Planning. A parallel line of work controls where the camera goes. CameraCtrl conditions video diffusion on prescribed trajectories (He et al., 2025), active-view selection improves reconstruction coverage (Chen et al., 2024b), and recent generative systems incorporate collision- or semantics-aware planning for scene expansion (Duan et al., 2026; Team HY-World, 2026). Meanwhile, vision–language models exhibit strong spatial reasoning (Chen et al., 2024a) and can be coupled with diffusion policies to generate continuous trajectories (Wei et al., 2026). Our planner unifies both roles: from a single reference image, it produces an *explorable trajectory* that decides where to explore and how to observe each location, and its camera poses directly condition iterative 3D Gaussian world generation.

Group-Relative RL Post-Training. Visual generators have been aligned with learned or human preferences through reward feedback, policy optimization, and direct preference objectives (Xu et al., 2023; Black et al., 2024; Wallace et al., 2024), and VLM judgments have been used to curate samples for iterative fine-tuning (Sun et al., 2025). Group-relative methods estimate advantages by comparing samples within a group, as in GRPO (Shao et al., 2024). For diffusion models, AWM retains the pretraining flow-matching objective and simply reweights each sample by its advantage (Xue et al., 2026), while CompGen adds a difficulty-based curriculum to group-relative training (Wang et al., 2026c). Our post-training follows this line but decouples credit assignment from state visitation: group comparisons supply stage-wise advantages, while a hard-state curriculum routes training through the states where the generator fails.

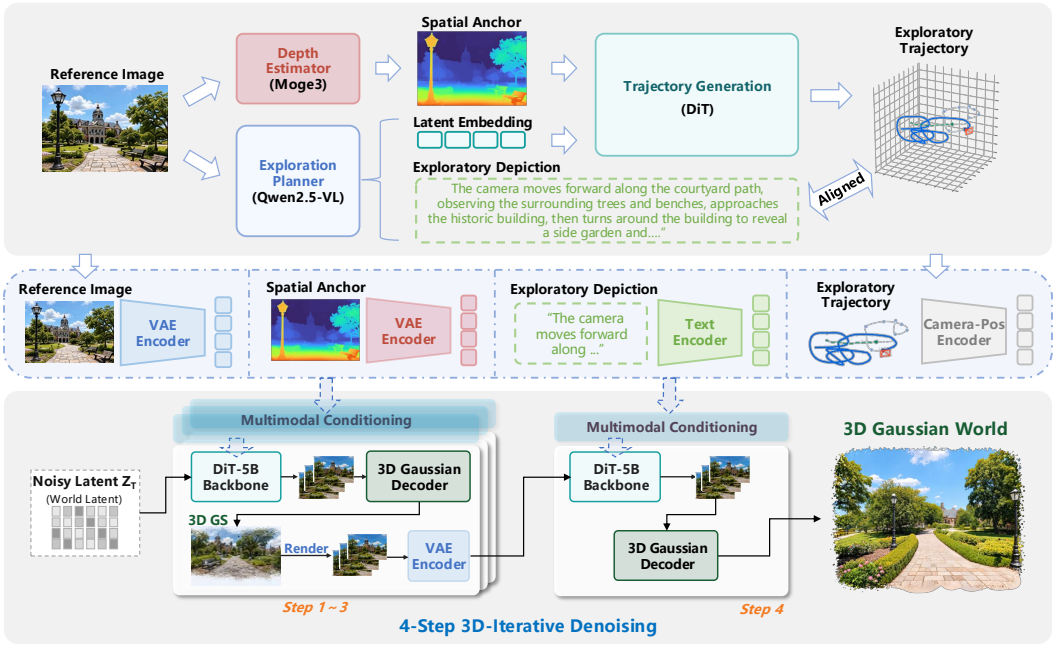

Figure 1: **Overview of UnveilWorld.** Starting from a single image, UnveilWorld formulates world generation as an exploration–generation–reconstruction loop: the scene itself decides where to look, looking further unveils new content, and what is unveiled becomes an enduring, explorable world. A shared spatial anchor keeps the three modules geometrically aligned, so that each reinforces the others far beyond the input view. The entire loop is post-trained with hard-state step-wise AWM.

3 METHOD

3.1 OVERVIEW

We address a central challenge in single-image Gaussian world generation: the high-quality explorable region of the generated world remains limited, and even when long-range trajectories are supplied, the generator still tends to default to short-range novel views. To overcome this limitation, we couple exploration with reconstruction in a mutually reinforcing loop. In the forward direction, long-range exploration trajectories explicitly drive the reconstruction of a broader and more explorable world; conversely, reconstructing such a world in turn demands more faithful observation planning along the trajectory. As illustrated in Fig. 1, all three modules of UnveilWorld—exploration, generation, and reconstruction—are learnable, so that this mutual-optimization loop is instantiated end-to-end during training. To align spatial understanding across modules, we employ the depth map of the input image as a shared spatial anchor. This anchor further resolves the ambiguity of distance-normalized trajectories: only geometry-aligned trajectories can reliably steer the generator across arbitrary scenes. For optimization, we adopt a hard-state, step-wise AWM paradigm in which a VLM assessor provides stage-wise feedback.

3.2 SCENE-ADAPTIVE EXPLORATION

Given an arbitrary input image, our goal is to acquire both a visual and a spatial understanding of the scene and to plan a scene-adaptive exploration trajectory. Existing methods either prescribe scene-agnostic trajectories or confine the camera to local neighborhoods with highly redundant observations. Moreover, their trajectories are distance-normalized without being aligned to scene geometry and therefore cannot reliably steer the generator toward unexplored regions; consequently, the practical bottleneck of Gaussian world reconstruction shifts from generation to exploration.

As shown in the upper part of Fig. 1, we introduce a dedicated pipeline to address these issues, comprising three stages. First, the depth estimator MoGe-3 Kong et al. (2026) anchors the spatial

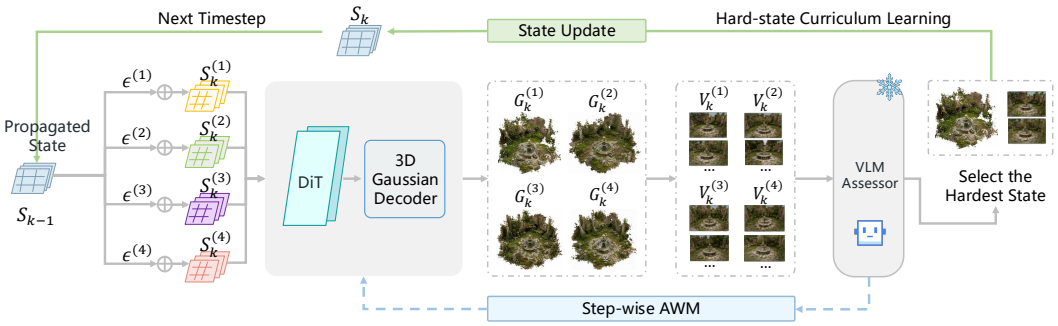

Figure 2: **Overview of Hard-State Step-wise AWM.** Step-wise candidate groups provide group-relative VLM advantages for the latent and render-path AWM losses; the lowest-scoring valid candidate is propagated to the next stage, and completed worlds receive terminal ground-truth feedback. One on-policy update follows each rollout.

geometry of the input image, and the resulting spatial anchor is shared by the subsequent generation and reconstruction stages. Next, we leverage the generalizable visual priors of a vision–language model for scene understanding, yielding an exploration-oriented latent embedding together with imagined step-wise exploration descriptions. Finally, a DiT (Peebles & Xie, 2023) module conditioned on the spatial anchor and the latent embedding plans the scene-adaptive exploration trajectory. The construction of the training data for the exploration module is described in Section 3.5.

3.3 Exploration-conditioned Generator

The exploration trajectory indicates where to go, the reference image and its depth map anchor the appearance and geometry of the scene, and the exploration description imagines the most plausible content; conditioned on these signals, the generator then unveils the world accordingly.

As shown in Fig. 1, these four modalities are encoded into a latent space and injected into the diffusion-based generator as multimodal conditions. The generator performs a 4-step iterative denoising process; all steps share a single DiT backbone, whose outputs are decoded into 3D Gaussians by a shared decoder. Coupling 3D Gaussian reconstruction with the generator enables rendering-based supervision from viewpoints beyond the generated views. This supervision enforces multi-view consistency and, more importantly, drives the exploration-stage models to learn in a world-reconstruction-oriented manner. Among the pretrained components, we employ a VAE to encode 2D visual conditions into latent embeddings, a T5 (Raffel et al., 2020) encoder for textual conditions, and unfold camera poses into a dense ray map. The DiT and the 3DGS decoder are initialized from the pretrained FlashWorld Li et al. (2026) backbone, with only the DiT subsequently post-trained; the DiT additionally injects the spatial anchor through an adapter, aligning trajectory control with spatial understanding.

3.4 Hard-State Step-wise AWM

Post-training an iterative 3D generator presents a dilemma: terminal rewards never distinguish among the intermediate candidates behind the outcome, whereas greedily propagating the best candidate confines later stages to favorable states only. We therefore decouple *credit assignment* from *state visitation*: credit flows from group-wise comparisons at every stage via advantage-weighted matching (AWM) (Xue et al., 2026), while state visitation follows a hard-state curriculum that deliberately routes training through failure.

Setup. Generation proceeds in N stages mirroring the inference-time few-step sampler, with the scene state S_k being a noisy multi-view latent. At stage k, fresh noise $\epsilon^{(i)} \sim \mathcal{N}(0, I)$ turns S_k into M candidate states $S_k^{(i)} = S_k + \epsilon^{(i)}$; the DiT, parameterized by ϕ, predicts the flow-matching velocity v_ϕ and denoises each candidate into a partially denoised estimate $\hat{z}_k^{(i)}$, which a frozen Gaussian decoder Dec converts into Gaussians $G_k^{(i)} = \mathrm{Dec}(\hat{z}_k^{(i)})$. We write $\mathrm{Render}(G, u)$ for

rendering G from camera u and denote by $\mathcal{C}_k$ the stage-specific assessment cameras ($\mathcal{C}_{\mathrm{GT}}$ at the final stage). The generator is kept frozen within each rollout; Fig. 2 illustrates the pipeline with $N{=}4$ and $M{=}4$.

Stage-wise group assessment. For $k < N$, a frozen VLM assessor scores each candidate's rendered views against the reference image and the scene description, yielding $r_k^{(i)\mathrm{VLM}}$ (higher is better), which are converted into continuous, zero-sum group-relative advantages:

$$A_k^{(i)\mathrm{VLM}} = \frac{r_k^{(i)\mathrm{VLM}} - \bar{r}_k^{\mathrm{VLM}}}{\left(\max_j r_k^{(j)\mathrm{VLM}} - \min_j r_k^{(j)\mathrm{VLM}}\right) + \tau_{\mathrm{VLM}}}, \qquad k < N, \tag{1}$$

where $\bar{r}_k^{\mathrm{VLM}}$ is the group mean.

Hard-state propagation. State propagation is deliberately decoupled from advantage assignment: after each of the first $N{-}1$ stages, the hardest candidate—the lowest-scoring valid one—is carried forward as the next state:

$$i_k^- = \arg\min_{i \in \mathcal{I}_k^{\mathrm{valid}}} r_k^{(i)\mathrm{VLM}}, \qquad S_{k+1} = \mathrm{Noisy}\Big(\mathrm{Enc}\big(\mathrm{sg}[G_k^{(i_k^-)}]\big)\Big), \quad k < N, \tag{2}$$

where $\mathcal{I}_k^{\mathrm{valid}}$ excludes candidates without a usable scene state and $\mathrm{Noisy}(\cdot)$ re-noises the encoded state to the next noise level via a frozen encoder Enc. Propagating the hardest candidate trains later stages for error recovery rather than error avoidance; propagation itself confers no reward, as the carried candidate's advantage still follows Eq. equation 1.

Terminal ground-truth feedback. At stage N, each completed world is scored by its multi-view error against ground truth,

$$e_N^{(i)\mathrm{GT}} = \frac{1}{|\mathcal{C}_{\mathrm{GT}}|} \sum_{u \in \mathcal{C}_{\mathrm{GT}}} d_{\mathrm{GT}}\Big(\mathrm{Render}(G_N^{(i)}, u), I_u^{\mathrm{GT}}\Big), \tag{3}$$

where I_u^{GT} is the ground-truth image at camera u and d_{GT} is a perceptual image distance. Taking the negated error as the reward, the terminal advantage $A_N^{(i)\mathrm{GT}}$ follows the same group normalization as Eq. equation 1 (Appendix A.4).

Advantage-weighted matching. All reward signals enter through AWM applied to flow-matching losses. For a candidate signal $\tilde{x}$ (detached unless otherwise stated), the replay loss $\ell^{\mathrm{FM}}(\tilde{x}; \phi)$ re-noises $\tilde{x}$ to a random flow time and penalizes the velocity-prediction error of v_ϕ. We then form

$$\rho(\tilde{x}) = \exp\big(-\ell^{\mathrm{FM}}(\tilde{x}; \phi) + \mathrm{sg}\big[\ell^{\mathrm{FM}}(\tilde{x}; \phi)\big]\big), \tag{4}$$

the on-policy special case of the likelihood ratio in AWM (Xue et al., 2026). Since ρ evaluates identically to 1 in the forward pass, inserting it as $-A\,\rho(\tilde{x})$ contributes exactly the gradient $A\,\nabla_\phi \ell^{\mathrm{FM}}(\tilde{x}; \phi)$—continuous advantage weighting without backpropagating through the rollout's sampling or assessment.

Training objective. The post-training objective combines four terms:

$$\mathcal{L}_{\mathrm{gen}} = \beta\,\mathcal{L}_{\mathrm{ref}} + \lambda_{\mathrm{lat}}\,\mathcal{L}_{\mathrm{lat}} + \lambda_{\mathrm{nv}}\,\mathcal{L}_{\mathrm{nv}} + \lambda_{\mathrm{term}}\,\mathcal{L}_{\mathrm{term}}. \tag{5}$$

(i) Anchor. $\mathcal{L}_{\mathrm{ref}}$ aligns the velocity predictions of ϕ with those of the frozen initial generator, estimating the KL trust region $D_{\mathrm{KL}}(p_\phi \| p_{\phi_0})$ in velocity space (Xue et al., 2026); we fix $\beta{=}0.5$.

(ii) Latent AWM.

$$\mathcal{L}_{\mathrm{lat}} = -\frac{1}{M(N-1)} \sum_{k=1}^{N-1} \sum_{i=1}^{M} A_k^{(i)\mathrm{VLM}}\, \rho\Big(\hat{z}_k^{(i)}\Big) \tag{6}$$

replays the partially denoised state estimates as targets—error-recovery training on the failure-biased states visited under the hard-state curriculum.

(iii) Render-path AWM.

$$\mathcal{L}_{\mathrm{nv}} = -\frac{1}{M(N-1)} \sum_{k=1}^{N-1} w_k \sum_{i=1}^{M} \frac{A_k^{(i)\mathrm{VLM}}}{|\mathcal{C}_k|} \sum_{u \in \mathcal{C}_k} \rho\Big(\mathrm{Render}(G_k^{(i)}, u)\Big) \tag{7}$$

applies the same advantages where the assessor looks: rendered novel views are re-encoded and matched, with third-stage states recomputed differentiably so that gradients flow through the frozen decoder into ϕ.

(iv) Terminal AWM applies the same replay form at stage N with the GT advantages. Since every advantage is zero-sum within its group, reward signals reallocate mass among candidates without net drift, leaving the anchor as the sole restoring force. Complete loss definitions and hyperparameters are given in Appendix A.4.

3.5 Data Engine

We construct a long-horizon scene-exploration dataset from two public datasets, DL3DV-10K (Ling et al., 2024) and Sekai2 (He et al., 2026), together with a private collection of high-quality videos. Each sequence is conditioned on a reference image and is jointly aligned with a trajectory-level language description and a scene-geometry-anchored camera trajectory. We filter out near-static paths, pose discontinuities, and repetitive object-centric orbits, and rank the remaining trajectories by camera-center coverage and inter-frame visual change. For each selected trajectory, we sample 48 ordered views at approximately uniform intervals along the cumulative camera travel, while explicitly retaining the trajectory endpoints and regions of sharp observation change. This procedure yields approximately 60,000 training tuples in total.

4 Experiments

We evaluate UnveilWorld along three axes: the quality of single-image 3D world generation against existing single-image methods, the impact of the exploration trajectory on world reconstruction, and the fidelity with which generators execute a prescribed trajectory.

4.1 Single-Image 3D World Generation

Setup. We follow the 40-scene protocol of One2Scene (Wang et al., 2026b), sampling ten scenes from each of four WorldScore-Static categories (realistic/stylized, indoor/outdoor) (Duan et al., 2025), and compare against five single-image world generation methods: One2Scene, HY-World 2.0 (Team HY-World, 2026), Matrix-3D (Yang et al., 2025), Marble 0.1-mini (World Labs, 2025), and Moverse (Zhou et al., 2026). All worlds are built from the same reference image and rendered along the One2Scene evaluation trajectories (linear and curvilinear). Camera-motion fidelity is assessed by comparing VGGT-recovered poses (Wang et al., 2025) against the prescribed trajectories.

Results. UnveilWorld achieves the best results on four of the six metrics and the second-best on the remaining two (Table 1); against One2Scene, the strongest baseline on the image-based metrics, it further reduces all three camera-motion errors, so the quality and fidelity gains do not come at the cost of trajectory adherence. Figure 3 corroborates this: across diverse scene styles, UnveilWorld preserves sharper structures where the baselines blur or distort.

Efficiency. Timing covers the full pipeline from the reference image to the final world, including camera planning and preprocessing but excluding one-time initialization, with each baseline run under its standard GPU configuration. UnveilWorld generates a complete 3D world in 18.12 seconds on a single NVIDIA A800 GPU, achieving a wall-clock speedup of approximately 150–212× over One2Scene, HY-World 2.0, and Matrix-3D running on 2–8 A800 GPUs. It also achieves speedups of 2.16× over Marble 0.1-mini and 4.01× over Moverse.

4.2 Exploration Trajectories Benefit World Reconstruction

Setup. To isolate the contribution of the exploration trajectory from that of the generator, we feed the same input image to *the same generator* under three construction trajectories: (i) a *fixed* trajec-

Table 1: **Single-image explorable 3D world generation.** All methods construct a 3D world from the same single reference image and are rendered along identical evaluation trajectories. Best and second-best results are shown in **bold** and underlined, respectively.

Method	Q-Align ↑	CLIP-I ↑	NIQE ↓	TransErr ↓	RotErr ↓	CamMC ↓	GPUs	Time (s) ↓
One2Scene	3.5702	81.3579	4.3023	0.2051	0.1166	0.2770	8× A800	2927.20
HY-World 2.0	2.8736	78.1235	5.5192	0.1559	0.0854	0.2058	8× A800	2726.84
Matrix-3D	2.0886	72.7490	4.7561	**0.0950**	0.0790	**0.1481**	2× A800	3843.57
Marble0.1-mini	1.9335	78.3089	6.2702	0.1543	0.0817	0.1902	-	39.06
Moverse	2.2822	84.5968	6.1316	0.2325	0.0771	0.2357	1× A800	72.68
Ours	**3.6281**	**85.9428**	**3.5289**	0.1519	**0.0674**	0.1645	1× A800	**18.12**

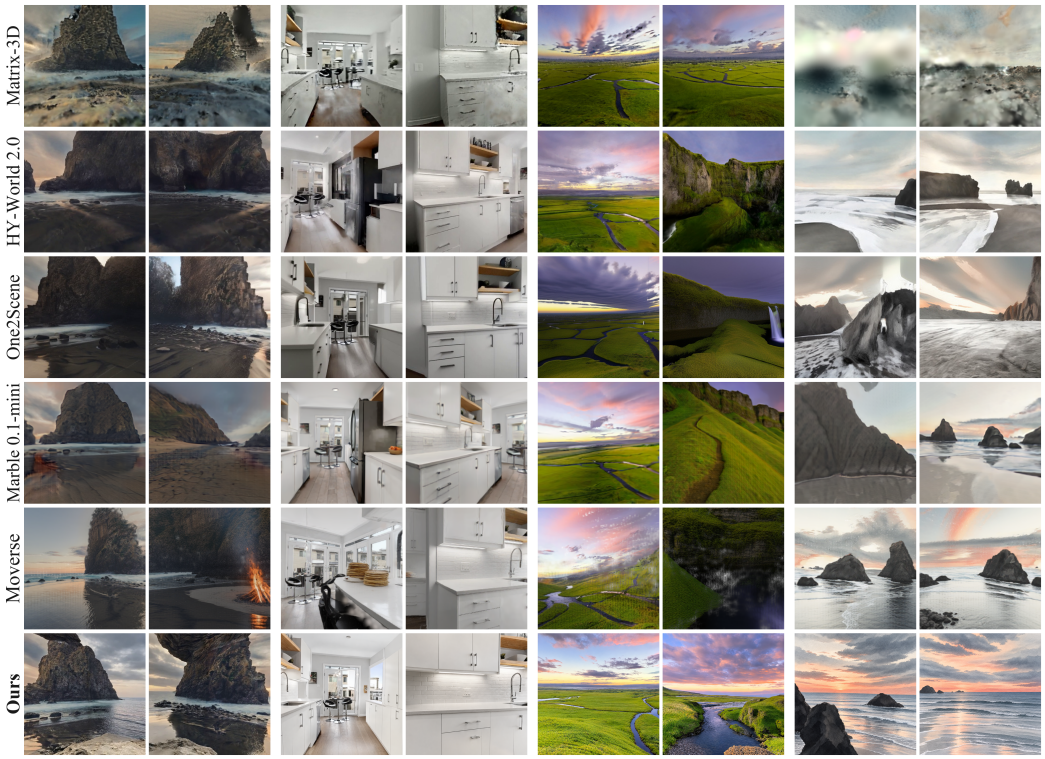

Figure 3: **Qualitative comparison of single-image world generation.** Across diverse scene styles, UnveilWorld consistently produces coherent worlds with plausible continuations beyond the input view, even under large viewpoint changes.

tory that observes in place, agnostic to scene content; (ii) a *common VLA* trajectory, scene-aware but not reconstruction-oriented; and (iii) our *scene-adaptive exploration* trajectory. The three resulting Gaussian worlds are compared on 100 held-out DL3DV scenes with 48 construction views and 20 GT viewpoints per scene (five in each of four camera-distance bins); FID pools 2,000 rendered and 2,000 GT images, and KID-20 averages per-scene estimates (Appendix B.3).

Results. Scene-adaptive exploration achieves the best results on all six metrics (Table 2): relative to the fixed trajectory, PSNR and SSIM increase by 2.533 dB and 0.0595, while the remaining metrics decrease by 21.9–73.7%. That the scene-aware common VLA trajectory still falls short everywhere confirms that what matters is not motion per se but reconstruction-oriented exploration. Figure 4 shows why: the fixed trajectory repeatedly observes the initial forward-facing region, whereas scene-adaptive exploration unveils exactly the regions that held-out viewpoints later query.

Table 2: **Ablation on the exploration trajectory.** All variants drive the same generator with the same input image; only the construction trajectory differs. Scene-adaptive exploration achieves the best world reconstruction on all metrics.

Construction trajectory	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	FID ↓	KID-20 ↓
Fixed	11.7931	0.3002	0.6156	0.2799	294.3891	0.25774
Common VLA	13.4782	0.3139	0.5376	0.2981	206.6486	0.19362
Scene-adaptive (Ours)	**14.3260**	**0.3597**	**0.4768**	**0.2186**	**174.1745**	**0.06789**

Table 3: **Ablation of Hard-State Step-wise AWM and joint optimization.** Step-wise AWM improves all metrics over the single-sample baseline, and joint training brings further gains on fidelity and perceptual quality.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	Subjective Quality ↑	FID ↓
Single-Sample Baseline	12.2682	0.4079	0.5908	0.2818	14.1232	94.5636
Hard-State AWM	12.6773	0.4235	**0.5833**	0.2727	19.0516	89.7532
Hard-State AWM + Joint Training	**13.5126**	**0.4242**	0.5867	**0.2442**	**21.4933**	**87.8269**

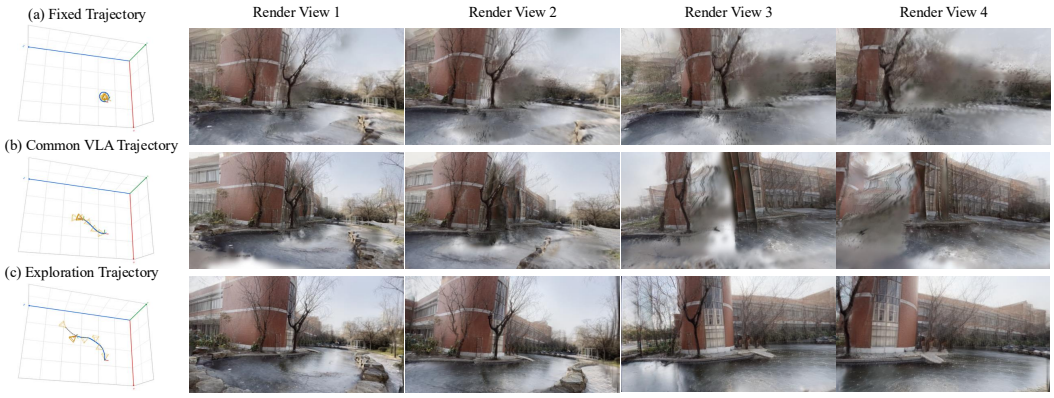

Figure 4: **Worlds reconstructed from three trajectory types.** Each Gaussian world is rendered from four held-out viewpoints. At nearby viewpoints (render views 1–2) all trajectories yield comparable quality; at distant viewpoints that demand the occluded backsides of buildings (render views 3–4), only the world built from our scene-adaptive exploration trajectory remains faithful.

4.3 Ablation of Hard-State AWM and Joint Optimization

We compare three configurations, each updating the generation DiT with the Gaussian decoder frozen: a single-sample baseline trained with the native GT-supervised objective for 300 updates, Hard-State AWM with four candidates per stage for 300 updates, and a subsequent joint-training stage of 300 updates that additionally optimizes the Exploration Planner and the trajectory-generation DiT. As shown in Table 3, Hard-State AWM improves all six metrics over the baseline, and joint training brings further gains—PSNR increases by 0.8353 dB while DISTS and pooled FID decrease by 10.45% and 2.1%—confirming the effectiveness of both stage-wise advantage-weighted post-training and end-to-end joint optimization.

4.4 Alignment between Generators and Trajectory Control

Setup. In this section, we validate the effectiveness of prescribed trajectory. The same single image, together with the exploration trajectory planned by our exploration module, is fed to a range of generators—LucidDreamer (Chung et al., 2025), WonderWorld (Yu et al., 2025), FlashWorld (Li et al., 2026), GEN3C (Ren et al., 2025), Lyra 2.0 (Shen et al., 2026), WorldStereo 2.0 (Zhang et al., 2026; Team HY-World, 2026), and LingBot-World 2.0 (Gao et al., 2026)—and VGGT-Ω (Wang et al., 2026a) recovers the trajectory *actually executed* in each output video, with camera centers rel-

Table 4: **World quality under an identical input trajectory.** All generators are driven by the same single image and the same exploration trajectory. Best and second-best results are shown in **bold** and underlined, respectively.

Method	3D Cons. ↑	Photo. Cons. ↑	Object Ctrl. ↑	Content Align. ↑	Style Cons. ↑	Subj. Quality ↑	Average ↑
LucidDreamer	89.36	73.61	36.25	45.86	57.05	47.91	58.34
WonderWorld	82.64	79.70	53.17	55.64	62.25	36.99	61.73
FlashWorld	87.66	88.66	48.23	53.57	78.68	49.08	67.65
GEN3C	83.10	85.27	54.25	51.10	73.99	47.65	65.89
WorldStereo 2.0	88.63	78.95	55.63	51.63	79.67	46.57	66.85
Lyra 2.0	83.95	87.59	55.88	54.51	79.37	45.85	67.86
LingBot-World 2.0	78.50	83.79	54.13	53.95	**82.05**	**61.44**	68.98
Ours	**92.23**	**93.40**	**58.09**	**58.23**	80.38	52.35	**72.45**

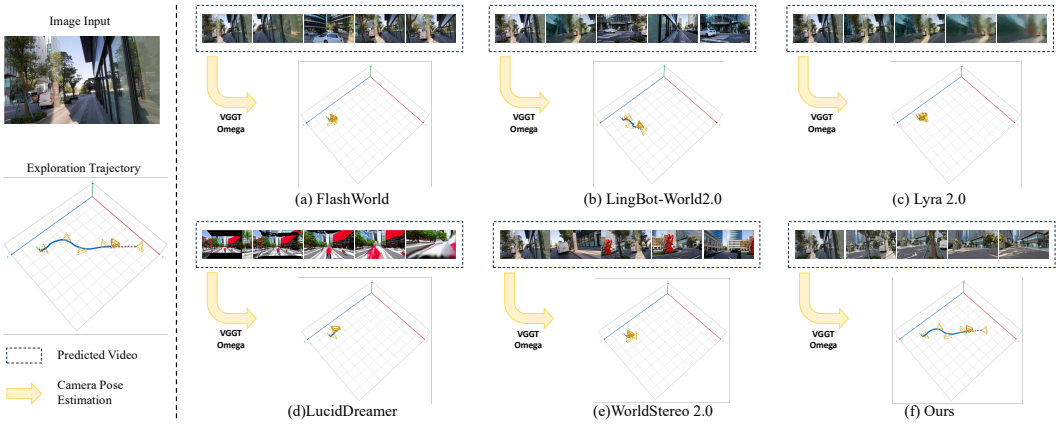

Figure 5: **Trajectory control fidelity.** Given the same input image and control trajectory, VGGT-Ω recovers the executed trajectory from each generator's output video. Existing methods deviate substantially whereas the trajectory executed by UnveilWorld closely overlays the input one.

ative to the first view and path lengths normalized by scene extent. World quality is assessed with the WorldScore-Static metrics (Duan et al., 2025) under the interval-sampled protocol of FlashWorld.

Results. As shown in Table 4, under the identical input trajectory UnveilWorld ranks first on four of the six metrics, exceeding the strongest baseline's average score by 3.47 points. Figure 5 reveals why: the VGGT-Ω trajectories recovered from the baseline outputs deviate markedly from the prescribed control trajectory—exhibiting reduced traversal or drift—whereas the trajectory executed by UnveilWorld closely matches the input, indicating that its gains stem from genuinely more faithful execution of camera control rather than from any advantage in the prescribed motion itself.

5 Conclusion

We have presented UnveilWorld, which couples scene-adaptive exploration, generation, and reconstruction for single-image 3D Gaussian world generation. A shared depth anchor and vision–language priors guide geometry-aligned trajectories and accompanying descriptions, conditioning a four-step generator to unveil unobserved content. Hard-State Step-wise AWM combines stage-wise group-relative learning with hard-state propagation, while joint optimization allows reconstruction feedback to shape exploration. UnveilWorld achieves the highest average score among the compared methods under the revised six-metric WorldScore-Static protocol, generating a complete world in 18.12 seconds on a single NVIDIA A800 GPU. These results establish exploration as a learnable interface between spatial understanding and world generation.

AI USE STATEMENT

Generative AI tools were used only to assist with manuscript editing and literature research, including language polishing, improving clarity and conciseness, and organizing relevant research materials. They were not used for method design, experimental design, experiment execution, result analysis, or drawing conclusions. All research decisions, experiments, analyses, and conclusions were independently conducted and verified by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

All key technical details and experimental settings are provided in the main paper and Appendix. We provide detailed descriptions of the training procedure, camera configurations, and evaluation protocols to facilitate reproduction of our results. We will release the source code, model checkpoints, and evaluation scripts upon acceptance of the paper.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. doi: 10.48550/arXiv.2511.21631. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, pp. 4965–4987, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/14f75513f0f1ca01de1e826b52e6b840-Abstract-Conference.html.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, 2024a. doi: 10.1109/CVPR52733.2024.01370. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_SpatialVLM_Endowing_Vision-Language_Models_with_Spatial_Reasoning_Capabilities_CVPR_2024_paper.html.

Chaofeng Chen and Jiadi Mo. IQA-PyTorch: PyTorch toolbox for image quality assessment. GitHub repository, 2022. URL https://github.com/chaofengc/IQA-PyTorch. Accessed: 2026-09-26.

Xiao Chen, Quanyi Li, Tai Wang, Tianfan Xue, and Jiangmiao Pang. GenNBV: Generalizable next-best-view policy for active 3D reconstruction. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16436–16445, 2024b. doi: 10.1109/CVPR52733.2024.01555. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_GenNBV_Generalizable_Next-Best-View_Policy_for_Active_3D_Reconstruction_CVPR_2024_paper.html.

Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. LucidDreamer: Domain-free generation of 3D Gaussian Splatting scenes. *IEEE Transactions on Visualization and Computer Graphics*, 31(12):10640–10651, 2025. doi: 10.1109/TVCG.2025.3611489. URL https://doi.org/10.1109/TVCG.2025.3611489.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810. URL https://doi.org/10.1109/TPAMI.2020.3045810.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. WorldScore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27713–27724, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Duan_WorldScore_A_Unified_Evaluation_Benchmark_for_World_Generation_ICCV_2025_paper.html.

Hongbo Duan, Peiyu Zhuang, Yi Liu, Zhengyang Zhang, Yuxin Zhang, Pengting Luo, Fangming Liu, and Xueqian Wang. NavCrafter: Exploring 3D scenes from a single image. In *IEEE International Conference on Robotics and Automation*, 2026. URL https://arxiv.org/abs/2604.02828.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026. doi: 10.48550/arXiv.2607.07534. URL https://arxiv.org/abs/2607.07534.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations*, pp. 100433–100464, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f98fd73d59d8494489ea970747b91fe4-Abstract-Conference.html.

Kang He, Wenshuo Peng, Zihui Gao, Jiaming Tan, Kaipeng Zhang, and Yongtao Ge. Sekai2: From world exploration to interactive world modeling. *arXiv preprint arXiv:2608.09449*, 2026. doi: 10.48550/arXiv.2608.09449. URL https://arxiv.org/abs/2608.09449.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2Room: Extracting textured 3D meshes from 2D text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7909–7920, 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.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, Dahua Lin, and Bo Dai. AnySplat: Feed-forward 3D Gaussian Splatting from unconstrained views. *ACM Transactions on Graphics*, 44(6):257:1–257:16, 2025. doi: 10.1145/3763326. URL https://doi.org/10.1145/3763326.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Lingyu Kong, Ruicheng Li, Ruicheng Wang, Sicheng Xu, Chengtang Yao, Jianfeng Xiang, and Jiaolong Yang. MoGe-3: Fine-detail monocular geometry estimation with self-guided sparse volumetric refinement. *arXiv preprint arXiv:2607.17967*, 2026. doi: 10.48550/arXiv.2607.17967. URL https://arxiv.org/abs/2607.17967.

Xinyang Li, Tengfei Wang, Zixiao Gu, Shengchuan Zhang, Chunchao Guo, and Liujuan Cao. FlashWorld: High-quality 3D scene generation within seconds. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2IftRjRB07.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. 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 (CVPR)*, pp. 22160–22169, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ling_DL3DV-10K_A_Large-Scale_Scene_Dataset_for_Deep_Learning-based_3D_Vision_CVPR_2024_paper.html.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726. URL https://doi.org/10.1109/LSP.2012.2227726.

Chaojun Ni, Jie Li, Haoyun Li, Hengyu Liu, Xiaofeng Wang, Zheng Zhu, Guosheng Zhao, Boyuan Wang, Chenxin Li, Guan Huang, and Wenjun Mei. WonderFree: Enhancing novel view quality and cross-view consistency for 3D scene exploration. *arXiv preprint arXiv:2506.20590*, 2025. doi: 10.48550/arXiv.2506.20590. URL https://arxiv.org/abs/2506.20590.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11410–11420, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Parmar_On_Aliased_Resizing_and_Surprising_Subtleties_in_GAN_Evaluation_CVPR_2022_paper.html.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132, 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.

Christoph Schuhmann. CLIP+MLP aesthetic score predictor. Software repository, 2022. URL https://github.com/christophschuhmann/improved-aesthetic-predictor. Model weights: sac+logos+ava1-l14-linearMSE.pth. Accessed 2026-09-26.

Katja Schwarz, Norman Müller, and Peter Kontschieder. Generative Gaussian splatting: Generating 3D scenes with video diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27510–27520, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Schwarz_Generative_Gaussian_Splatting_Generating_3D_Scenes_with_Video_Diffusion_Priors_ICCV_2025_paper.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, Sanja Fidler, Jiahui Huang, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra 2.0: Explorable generative 3D worlds. *arXiv preprint arXiv:2604.13036*, 2026. doi: 10.48550/arXiv.2604.13036. URL https://arxiv.org/abs/2604.13036. Accepted to SIGGRAPH Asia 2026.

Jiao Sun, Deqing Fu, Yushi Hu, Su Wang, Royi Rassin, Da-Cheng Juan, Dana Alon, Charles Herrmann, Sjoerd van Steenkiste, Ranjay Krishna, and Cyrus Rashtchian. DreamSync: Aligning text-to-image generation with image understanding feedback. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5920–5945. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.304. URL https://aclanthology.org/2025.naacl-long.304/.

Team HY-World. HY-World 2.0: A multi-modal world model for reconstructing, generating, and simulating 3D worlds. *arXiv preprint arXiv:2604.14268*, 2026. doi: 10.48550/arXiv.2604.14268. URL https://arxiv.org/abs/2604.14268.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238, 2024. doi: 10.1109/CVPR52733.2024.00786. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wallace_Diffusion_Model_Alignment_Using_Direct_Preference_Optimization_CVPR_2024_paper.html.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2555–2563, 2023. doi: 10.1609/aaai.v37i2.25353. URL https://ojs.aaai.org/index.php/AAAI/article/view/25353.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VGGT_Visual_Geometry_Grounded_Transformer_CVPR_2025_paper.html.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω, 2026a. URL https://arxiv.org/abs/2605.15195.

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*, pp. 95755–95770, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9aaa0d7db14b9f0c2575a1761b1ab76e-Abstract-Conference.html.

Shijian Wang, Runhao Fu, Siyi Zhao, Qingqin Zhan, Xingjian Wang, Jiarui Jin, Yuan Lu, Hanqian Wu, and Cunjian Chen. Synthetic curriculum reinforces compositional text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21953–21963, 2026c. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_Synthetic_Curriculum_Reinforces_Compositional_Text-to-Image_Generation_CVPR_2026_paper.html.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024. doi: 10.1109/CVPR52733.2024.01956. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_DUSt3R_Geometric_3D_Vision_Made_Easy_CVPR_2024_paper.html.

Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://doi.org/10.1109/TIP.2003.819861.

Meng Wei, Chenyang Wan, Jiaqi Peng, Xiqian Yu, Yuqiang Yang, Delin Feng, Wenzhe Cai, Chenming Zhu, Tai Wang, Jiangmiao Pang, and Xihui Liu. Ground slow, move fast: A dual-system foundation model for generalizable vision-language navigation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GK4rznYwhn.

World Labs. Marble: A multimodal world model. Official product announcement, 2025. URL https://www.worldlabs.ai/blog/marble-world-model. Model variant: Marble 0.1-mini; accessed 2026-09-26.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54015–54029. PMLR, 2024. URL https://proceedings.mlr.press/v235/wu24ah.html.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15903–15935. Curran Associates, Inc., 2023. doi: 10.52202/075280-0700. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/33646ef0ed554145eab65f6250fab0c9-Abstract-Conference.html.

Shuchen Xue, Chongjian Ge, Shilong Zhang, Yichen Li, and Zhi-Ming Ma. Advantage Weighted Matching: Aligning RL with pretraining in diffusion models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2509.25050.

Zhongqi Yang, Wenhang Ge, Yuqi Li, Jiaqi Chen, Haoyuan Li, Mengyin An, Fei Kang, Hua Xue, Baixin Xu, Yuyang Yin, Eric Li, Yang Liu, Yikai Wang, Hao-Xiang Guo, and Yahui Zhou. Matrix-3D: Omnidirectional explorable 3D world generation. *arXiv preprint arXiv:2508.08086*, 2025. doi: 10.48550/arXiv.2508.08086. URL https://arxiv.org/abs/2508.08086.

Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T. Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, and Charles Herrmann. WonderJourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6658–6667, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5916–5926,

2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yu_WonderWorld_Interactive_3D_Scene_Generation_from_a_Single_Image_CVPR_2025_paper.html.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

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*, pp. 40327–40339, 2026. URL https://arxiv.org/abs/2603.02049.

Yang Zhou, Ziheng Wang, Yuqin Lu, Haofeng Liu, Jun Liang, Shengfeng He, and Jing Li. MoVerse: Real-time video world modeling with panoramic gaussian scaffold. *arXiv preprint arXiv:2606.13376*, 2026. doi: 10.48550/arXiv.2606.13376. URL https://arxiv.org/abs/2606.13376.

Appendix organization. Appendix A provides implementation and optimization details, including model initialization, training-data construction, depth-anchored normalization, Hard-State Stepwise AWM, and inference. Appendix B specifies the evaluation data, construction and evaluation views, metric computation, baseline settings, and runtime protocol. Appendix C presents additional generator ablations and qualitative results.

A Implementation Details

A.1 Data engine

We render each sequence as an ordered trajectory video and annotate it with Qwen3-VL-8B-Instruct Bai et al. (2025a) to obtain a single trajectory-aligned textual description. We then estimate the first-frame depth map with MoGe-3 and derive a robust scene scale from its valid depth values. Camera translations are recentered to the first camera position and normalized by this scale, while rotations and intrinsics are preserved. Each tuple comprises the first view as the reference image, 48 ordered frames, the original and depth-normalized camera trajectories, camera intrinsics, and the corresponding textual description.

A.2 Model Initialization and Trainable Components

The Exploration Planner and trajectory-generation DiT are initialized from the System 2 and DiT-based System 1 components of InternVLA-N1-DualVLN (Wei et al., 2026), respectively. System 2 employs a navigation-tuned Qwen2.5-VL-7B (Bai et al., 2025b) backbone. MoGe-3 estimates the reference depth, while the 5B generation DiT and 3D Gaussian decoder are initialized from FlashWorld. The pretrained VAE and T5 encoder encode the reference image and exploratory descriptions, respectively. A depth-conditioning adapter maps the normalized reference depth and validity mask into the generator's latent conditioning space, while dense ray maps represent camera poses and intrinsics. A frozen Qwen3-VL-8B-Instruct model provides training-time assessment.

AWM post-training updates the generation DiT while keeping the Gaussian decoder and remaining modules frozen. A copy of the generation DiT at the start of this stage serves as the fixed reference model. The reference term regularizes the DiT parameters, while render-path gradients pass through the frozen decoder, rendering, and frozen VAE encoder to update the DiT.

Joint training initializes the generation DiT from its AWM-trained checkpoint and continues with the same four-term objective in Eq. (5). It jointly updates the generation DiT, Exploration Planner, and trajectory-generation DiT, while the remaining modules stay frozen.

A.3 Training Data and Geometric Normalization

Training Data and View Sampling. The training set contains approximately 60,000 reference-conditioned sequences, collected from two public datasets, DL3DV-10K (Ling et al., 2024) and Sekai2 (He et al., 2026), together with a private collection of high-quality videos. Each sequence contains a reference image, 48 ordered observations with camera poses and intrinsics, and a trajectory-aligned textual description. We use all 48 views at 704×480 pixels (width $\times$ height).

Views are sampled approximately uniformly along cumulative camera travel, while retaining trajectory endpoints and regions of sharp observation change. We exclude near-static scans, pose-discontinuous sequences, and repetitive single-object orbits. Images and camera parameters remain frame-wise aligned, and camera poses are expressed relative to the reference camera.

Depth-Anchored Normalization. We estimate a shared scene scale from the reference depth predicted by MoGe-3. Valid pixels with finite, positive depths are back-projected into the reference-camera coordinate system. After filtering depths outside the 1st–99th percentile range, we compute the coordinate-wise 1st and 99th percentiles of the reference point cloud, denoted by $\mathbf{q}_{0.01}$ and $\mathbf{q}_{0.99}$. The shared isotropic scale is

$$s = \max_{a \in \{x,y,z\}} \left(q_{0.99,a} - q_{0.01,a}\right).$$

During training, source-camera translations are first aligned to the reference-depth scale and expressed relative to the reference camera. Depth and translations are then normalized using the same scale:

$$\widehat{\mathbf{t}}_i = \mathbf{t}_i / s, \qquad \widehat{R}_i = R_i, \qquad \widehat{D}(u,v) = \begin{cases} D(u,v)/s, & \mathcal{M}(u,v) = 1, \\ 0, & \mathcal{M}(u,v) = 0. \end{cases}$$

Here D is the reference depth and $\mathcal{M}$ is its validity mask. Percentile filtering is used to estimate the scale; all valid depth values are retained in the final condition. We apply the same isotropic scaling to depth and camera translations, without independently normalizing the trajectory.

RGB images, depth maps, and validity masks undergo aligned resizing and padding, with corresponding updates to camera intrinsics. Depth maps and masks use nearest-neighbor interpolation. The source-to-depth calibration and shared scale are stored in the sample metadata.

A.4 Loss and Implementation Details of Hard-State Step-wise AWM

Full loss definitions. The terminal advantage mirrors Eq. equation 1 with reward $r_N^{(i)\mathrm{GT}} = -e_N^{(i)\mathrm{GT}}$ and stabilizer τ_{GT}:

$$A_N^{(i)\mathrm{GT}} = \frac{r_N^{(i)\mathrm{GT}} - \bar{r}_N^{\mathrm{GT}}}{\left(\max_j r_N^{(j)\mathrm{GT}} - \min_j r_N^{(j)\mathrm{GT}}\right) + \tau_{\mathrm{GT}}}. \tag{8}$$

The anchor term is a velocity MSE to the frozen initial generator,

$$\mathcal{L}_{\mathrm{ref}} = \mathbb{E}_{z,t,c}\left[\left\| v_\phi(z_t, t, c) - v_{\phi_0}(z_t, t, c) \right\|^2\right], \tag{9}$$

where z, t, and c are a data latent, a flow time, and the multimodal conditions from a training sequence paired with the rollout, and both velocities are evaluated on the same noised latent z_t; sharing the time and noise between the two evaluations reduces the variance of the estimate. The terminal term applies the same replay form as Eq. equation 6 at stage N:

$$\mathcal{L}_{\mathrm{term}} = -\frac{1}{M} \sum_{i=1}^{M} A_N^{(i)\mathrm{GT}} \, \rho\left(\hat{z}_N^{(i)}\right). \tag{10}$$

Since ρ equals 1 in the forward pass and every advantage is zero-sum within its group, all AWM terms vanish in the forward pass and carry signal only through gradients. For render-path replay, rendered views are encoded into the DiT's input space by the frozen VAE encoder before re-noising.

Remarks on the main-text formulation. The VLM and GT signals are complementary: the VLM provides relative perceptual judgments at every stage, while the GT error offers an absolute fidelity signal anchored to measured ground truth, thereby helping guard against reward hacking. The weighting function ρ in Eq. equation 4 is the on-policy special case of the ELBO-based likelihood ratio in AWM (Xue et al., 2026), with the stop-gradient standing in for the old-policy evaluation. The anchor loss is computed on a paired training sequence, and the per-stage weight w_k in Eq. equation 7 increases with k, since residual noise in early renders degrades the signal-to-noise ratio; the third-stage states are recomputed differentiably, shaping latents whose Gaussian decodings render well.

Implementation details. Each rollout uses $N = 4$ denoising stages with $M = 4$ candidates per stage. Candidates share the reference image, spatial anchor, camera conditions, and text. Candidate generation and scoring do not retain a gradient graph.

Candidates within each stage share the assessment-camera set $\mathcal{C}_k$, and terminal candidates share the training ground-truth camera set $\mathcal{C}_{\mathrm{GT}}$. The objective follows Eq. (5), with $\beta = 0.5$. The latent and render-path terms retain the stage, candidate, and view averaging defined in Eqs. (6) and (7).

Advantages are detached and used without clipping. Invalid candidates are excluded from group statistics, advantage-weighted losses, and propagation. For groups containing invalid candidates, the candidate sums, means, and extrema in the AWM definitions are restricted to the valid candidate set, and candidate averages use its size in place of M. The stage normalization $1/(N-1)$ remains unchanged. Groups with fewer than two valid candidates contribute no advantage-weighted term. A rollout is discarded if no valid continuation is available.

FM replay follows the base generator's noise schedule and prediction target, with replay time and noise shared within each candidate group. Replay targets are detached for the latent and terminal terms. For the render-path term, third-stage states are recomputed differentiably from cached rollout inputs. Gradients propagate through the frozen Gaussian decoder, rendering, and frozen VAE encoder to update the generation DiT.

Optimization. The single-sample baseline and Hard-State Step-wise AWM each receive 300 optimizer updates. The baseline uses the native GT-supervised generator objective, while AWM updates the generation DiT with the four-term objective in Eq. (5). Joint training starts from the AWM-trained generation DiT and continues for 300 additional updates with the same objective, jointly optimizing the generation DiT, Exploration Planner, and trajectory-generation DiT. The generation DiT receives 600 updates in total; the Gaussian decoder remains frozen during AWM and joint training.

A.5 Inference and Runtime

Inference. Given a reference image, MoGe-3 estimates its depth, from which we compute the shared scene scale described in Section A.3. The normalized depth establishes the spatial anchor. The Exploration Planner produces an exploration embedding and textual descriptions, and the trajectory-generation DiT predicts an ordered trajectory of 48 construction views in the same normalized, reference-camera coordinate system. The predicted translations are used directly, without additional trajectory-dependent rescaling. The reference image, spatial anchor, descriptions, and construction cameras jointly condition the world generator.

The generation DiT performs a single four-stage rollout. After each of the first three stages, the intermediate representation is decoded into 3D Gaussians, rendered, and re-encoded as feedback for the next stage. The final stage produces the explicit 3D Gaussian world for novel-view rendering. Inference uses one predicted trajectory without candidate grouping or ranking. VLM assessment, candidate selection, GT rewards, and FM replay are used only during training.

Runtime. End-to-end generation requires 18.12 seconds on a single NVIDIA A800 GPU. The measurement spans the complete pipeline from the input reference image to the final 3D Gaussian world, including depth estimation, trajectory planning, condition encoding, four-stage denoising, intermediate rendering and re-encoding, and Gaussian decoding. Model loading and process initialization are excluded.

B EVALUATION PROTOCOLS

We evaluate UnveilWorld under two main settings: complete-pipeline explorable-scene generation and generator evaluation under shared exploration trajectories on WorldScore-Static. Additional GT-based evaluations examine construction trajectories and generator optimization.

B.1 EVALUATION DATA AND SPLITS

Explorable 3D Scene Generation. For Table 1, we follow the 40-scene protocol introduced by One2Scene (Wang et al., 2026b). The evaluation set contains ten scenes from each of four WorldScore-Static categories: realistic indoor, stylized indoor, realistic outdoor, and stylized outdoor. We compare UnveilWorld with One2Scene, HY-World 2.0, Matrix-3D, Marble 0.1-mini, and MoVerse, with all methods receiving the same reference image for each scene.

WorldScore-Static. For Table 4, we evaluate all 2,000 test cases in the WorldScore-Static split. Our exploration module plans one trajectory per reference image, shared by all generators. We compare with LucidDreamer, WonderWorld, FlashWorld, GEN3C, Lyra 2.0, WorldStereo 2.0 and LingBot-World 2.0 on the same cases. The 40-scene set is a subset of this benchmark; the two tables use distinct evaluation protocols.

GT-Based Ablation Data. Tables 2 and 3 each use 100 held-out scenes from DL3DV-10K. For both evaluations, the original source scenes are excluded from those used to construct the training set. The split is performed at the original-scene level. GT images are the corresponding captured RGB frames from DL3DV, paired with their camera poses and intrinsics.

B.2 CONSTRUCTION AND EVALUATION VIEWS

Complete-Pipeline Evaluation. Construction cameras condition world generation, whereas evaluation cameras define the viewpoints used for scoring. In the 40-scene setting, UnveilWorld executes the complete pipeline described in Section A.5. The Exploration Planner and trajectory-generation DiT predict one ordered trajectory of 48 construction views, which conditions the four-stage world generator. The resulting worlds are evaluated along the shared One2Scene trajectories, which include both linear and curvilinear camera motion. The 48-view count defines the construction budget and is independent of the number of rendered frames used for scoring.

Shared-Trajectory Evaluation. On WorldScore-Static, our Exploration Planner and trajectory-generation DiT predict one trajectory per reference image. All generators receive the same image and this planned trajectory. For UnveilWorld, frozen MoGe-3 estimates reference depth for the depth-conditioning adapter. The conditioning encoders, Gaussian decoder, and four-stage denoising procedure remain unchanged. Neither GT depth nor additional observed views are generation inputs. All methods are evaluated along the shared exploration trajectories using the frame-sampling protocol in Section B.4.

B.3 GT-BASED ABLATION PROTOCOLS

Construction-Trajectory Ablation. Table 2 compares Fixed, Common VLA, and Scene-adaptive VLA on 100 scenes. All variants use the same generator and 48 construction views per scene. Within each scene, all variants are evaluated at the same 20 GT camera poses, yielding 2,000 generated/GT image pairs per method.

Camera distance is the Euclidean distance from the reference-camera center in the normalized coordinate system of Section A.3. With the reference center at the origin, $d_i = \|\widehat{\mathbf{t}}_i\|_2$. We use four distance intervals: $d \leq 0.4$, $0.4 < d \leq 0.6$, $0.6 < d \leq 0.8$, and $d > 0.8$, selecting five GT evaluation views per interval. The same thresholds are used across scenes, and the selected evaluation views are shared by all compared methods. Construction cameras specify the generation trajectory, while GT evaluation cameras specify where the resulting world is rendered for scoring.

Generator-Optimization Ablation. Table 3 evaluates each training configuration on 100 scenes with 48 GT evaluation views per scene, yielding 4,800 generated/GT image pairs per method. All

configurations use a 48-view construction budget and share the same reference images and GT evaluation camera sets. The GT evaluation-view count is distinct from the construction budget and the training-time reward-view counts.

Paired Image Metrics. PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and DISTS (Ding et al., 2022) are computed on corresponding rendered and GT images. Each metric is averaged over evaluation views within each scene and then equally over the 100 scenes. For Table 2, each of the four five-view distance intervals therefore receives equal weight.

Pooled FID. FID (Heusel et al., 2017) is computed once per method from the means and covariances of Inception-v3 features pooled across all evaluation scenes. Table 2 uses 2,000 generated images and 2,000 GT images; Table 3 uses 4,800 generated images and 4,800 GT images. We use the clean-mode implementation (Parmar et al., 2022) in `pyiqa` (Chen & Mo, 2022) with identical preprocessing for generated and GT images. FID comparisons are made within each table under its fixed evaluation set and sample count.

Per-Scene KID-20. For Table 2, KID (Bińkowski et al., 2018) is computed separately for each scene using its 20 generated images and 20 corresponding GT images, and the 100 scene scores are averaged equally. Each scene score is the unbiased squared-MMD estimate in Inception feature space with the polynomial kernel $k(x, y) = (x^{\mathsf{T}}y/D + 1)^3$, where D is the feature dimension. Within-set diagonal terms are excluded. KID-20 is reported without an additional scaling factor.

Subjective Quality in the Generator Ablation. Subjective Quality in Table 3 combines automatic CLIP-IQA+ (Wang et al., 2023) and CLIP aesthetic predictions (Schuhmann, 2022). RGB images are resized to 512×512 using the PIL pipeline and converted to tensors in $[0, 1]$. Each predictor's raw scores are averaged over the 48 evaluation frames within each scene, giving q_s and a_s. We use the fixed normalization

$$n(x; \mu, \sigma, l, u) = \mathrm{clip}_{[0,1]}\left(0.25 + 0.5\frac{(x-\mu)/\sigma - l}{u - l}\right), \tag{11}$$

with constants

$$(\mu, \sigma, l, u)_{\mathrm{IQA}} = (0.5842, 0.0441, -1.8342, 1.8703),$$
$$(\mu, \sigma, l, u)_{\mathrm{aesthetic}} = (5.5952, 0.2561, -2.9342, 1.9741).$$

The reported score is

$$\text{Subjective Quality} = \frac{1}{100}\sum_{s=1}^{100} 50\left[n_{\mathrm{IQA}}(q_s) + n_{\mathrm{aesthetic}}(a_s)\right]. \tag{12}$$

Scores lie in $[0, 100]$, with higher values indicating better predicted quality.

B.4 Metrics and Baseline Settings

Explorable-Scene Metrics. Table 1 reports six metrics. Q-Align (Wu et al., 2024) and NIQE (Mittal et al., 2013) measure no-reference visual quality, while CLIP-I measures similarity between generated views and the reference image in CLIP (Radford et al., 2021) image embedding space. VGGT (Wang et al., 2025) estimates camera poses from the rendered views. TransErr, RotErr, and CamMC compare these estimates with the prescribed evaluation trajectory (Wang et al., 2026b). Higher is better for Q-Align and CLIP-I; lower is better for NIQE, TransErr, RotErr, and CamMC.

WorldScore Metrics and Frame Sampling. We follow the revised evaluation protocol for 3D methods adopted by FlashWorld (Li et al., 2026). Object Control, Content Alignment, Style Consistency, and Subjective Quality are computed from frames randomly sampled within predefined intervals, using identical frame indices across methods. This evaluates views throughout each interval. 3D Consistency and Photometric Consistency follow the official WorldScore computation, while Camera Control is omitted under the revised protocol. Higher is better for all six reported metrics. Let S_j be the dataset-level score for metric j. The aggregate is the unweighted arithmetic mean:

$$\text{Average} = \frac{1}{6}\sum_{j=1}^{6} S_j. \tag{13}$$

Table 5: Ablation of the textual condition used for world generation. Trajectory-aligned exploration descriptions improve both visual quality and camera-trajectory consistency.

Method	Q-Align ↑	CLIP-I ↑	NIQE ↓	TransErr ↓	RotErr ↓	CamMC ↓
Generic Image Description	3.5369	83.9368	3.5876	0.1753	0.0998	0.2499
Trajectory-Aligned Exploration Description	**3.6281**	**85.9428**	**3.5289**	**0.1519**	**0.0674**	**0.1645**

Baseline Execution and Runtime. Each baseline retains its method-specific generation pipeline while receiving the shared external inputs for the corresponding setting. In Table 1, UnveilWorld and Moverse each use one NVIDIA A800 GPU; One2Scene and HY-World 2.0 each use eight A800 GPUs; and Matrix-3D uses two A800 GPUs. End-to-end timing follows the boundary in Section A.5, including method-specific preprocessing and planning and excluding one-time model loading and process initialization. GPU allocations are reported alongside wall-clock runtimes.

C Additional Experiments and Qualitative Results

This section provides additional evidence for the proposed scene-adaptive exploration and generator optimization framework. We first examine the effects of trajectory-aligned textual conditioning, Hard-State Step-wise AWM, and joint optimization. We then provide additional qualitative results that visualize scene-adaptive exploration and the resulting 3D worlds across diverse settings.

C.1 Additional Generator Ablations

We further examine two components that directly affect world generation: the trajectory-aligned exploration description used for conditioning and the post-training strategy used to optimize the generator. The former studies whether the textual condition should describe only the observed image or explicitly reflect the intended exploration trajectory, while the latter evaluates the contribution of Hard-State Step-wise AWM and subsequent joint optimization.

C.1.1 Trajectory-Aligned Exploration Descriptions

A generic image caption describes the visible content of the reference view but does not account for how the scene will be explored. In contrast, our trajectory-aligned exploration description conditions generation on both the observed scene and the intended camera movement. To isolate the effect of this conditioning design, we replace the trajectory-aligned description with a generic image description while keeping the remaining generation and evaluation settings unchanged.

As shown in Table 5, trajectory-aligned descriptions improve both appearance-related and trajectory-related metrics. Compared with a generic image description, they increase Q-Align and CLIP-I by 2.58% and 2.39%, respectively, while reducing NIQE by 1.64%. The improvement is more pronounced in camera consistency: TransErr, RotErr, and CamMC decrease by 13.35%, 32.46%, and 34.17%, respectively. These results show that aligning the textual condition with the intended exploration provides guidance beyond reference-view semantics, helping the generator preserve visual quality while following the specified camera motion more accurately.

The single-sample baseline uses one continuation per stage and is trained with the native GT-supervised generator objective for 300 optimizer updates. Hard-State Step-wise AWM uses four candidates per stage and updates the generation DiT with Eq. (5) for 300 updates. The full model starts from the AWM-trained generation DiT and continues for 300 additional updates with the same four-term objective, jointly optimizing the generation DiT, Exploration Planner, and trajectory-generation DiT. The generation DiT receives 600 updates in total; the Gaussian decoder remains frozen during AWM and joint training.

C.2 Additional Qualitative Results

Figures 6 and 7 present additional qualitative results across diverse indoor and outdoor environments. The results demonstrate coherent novel-view synthesis over large camera rotations while preserving the appearance and structural characteristics of the input scene.



Figure 6: Additional qualitative results on diverse outdoor scenes.



Figure 7: Additional qualitative results on diverse indoor scenes.