GSEXPLORER: INTERACTIVE WORLD GENERATION WITH A PERSISTENT 3D GAUSSIAN WORLD STATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive world generation requires both accurate local control and stable scene consistency over long rollouts. Existing methods improve local generation through camera or geometry conditioning and extend historical context with memory mechanisms. However, long-horizon exploration remains challenging as target-view guidance becomes incomplete and inconsistencies accumulate across repeated scene updates. We present **GSExplorer**, a generation–reconstruction framework built around a persistent 3D Gaussian world state. At each step, the current 3DGS scene provides multimodal rendering conditions for a condition-driven video generator, while camera geometry is injected through a separate Relative Ray Encoding branch. Generated videos are then fed back to a SLAM-based reconstruction backend, where inconsistency filtering prevents conflicting content from corrupting the scene and cross-chunk bundle adjustment improves spatial alignment across generation steps. We further design a data-construction pipeline that reproduces realistic 3DGS degradation during scene extrapolation, yielding 293K training pairs of degraded multimodal renderings and real target videos. GSExplorer achieves the best overall performance among the compared methods on both single-step and three-step WorldScore evaluation, with strong camera control and consistency across both settings, including a Camera Control score of 99.07 in single-step extrapolation. On 481-frame leave-and-return rollouts, it also shows strong revisit consistency, demonstrating the value of maintaining a reliable 3D world state for continued exploration.

1 INTRODUCTION

Starting from a single image, interactive world generation aims to continuously create explorable scenes or videos following user-specified camera trajectories and text prompts (Yu et al., 2024a; 2025a; Duan et al., 2025). This requires two key capabilities: (1) precise camera control with high visual and spatial consistency for each local generation, and (2) stable scene evolution over long rollouts, including revisiting previously explored regions. The former enables reliable scene expansion, while the latter determines whether such capability can persist over continuous exploration.

Existing methods improve local controllability by injecting camera poses, rays, or relative positional encodings into video generators (He et al., 2024; Dai et al., 2026), but these signals provide limited information about the scene structure at the target view. Geometry-guided approaches (Yu et al., 2024b; Ren et al., 2025; Cao et al., 2025; Huang et al., 2025; Zhou et al., 2026; Yang et al., 2026b;a) address this limitation by rendering explicit 3D representations along planned trajectories, providing spatially aligned conditions for generation. Meanwhile, memory-based methods extend temporal context through historical frames, compressed states, or retrieved references to improve long-term consistency (Yu et al., 2025b; Li et al., 2025b; Sun et al., 2025; Team et al., 2026). However, long-horizon generation remains challenging: incomplete 3D renderings provide imperfect guidance (Wu et al., 2025; Zhang et al., 2025), while pose misalignment and conflicting generated content can accumulate as the scene is repeatedly updated (Zhao et al., 2026c; Wang et al., 2026). This motivates a more fundamental question: **how can we maintain a reliable and queryable 3D world state that continuously provides useful target-view conditions throughout exploration?**

To this end, we introduce **GSExplorer**, a generation–reconstruction loop built around a persistent 3D Gaussian world state. Our key insight is to separate long-term spatial memory from short-

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.

horizon video generation: the persistent 3D scene carries accumulated spatial information, while the video model repeatedly solves local generation problems conditioned on the current world state. Generated videos are selectively integrated back into the same scene, allowing the world state to evolve together with exploration.

Specifically, GSExplorer renders degraded RGB, depth, accumulated opacity, and hole masks from the current 3DGS state as multimodal scene conditions, while camera geometry is injected separately through Relative Ray Encoding to provide complementary cross-view guidance. The generated videos are then fed into a SLAM-based 3DGS reconstruction backend. Inconsistency filtering prevents conflicting generated content from corrupting established regions, while cross-chunk bundle adjustment aligns current and historical views for coherent scene updates. To train the generator on realistic incomplete scene observations, we construct 293K degraded-condition–target video pairs from real-world scene datasets.

GSExplorer achieves the best overall performance among the compared methods on both single-step and three-step WorldScore evaluation (Duan et al., 2025). In single-step extrapolation, it achieves a Camera Control score of 99.07 while also attaining leading performance in 3D and Style Consistency. We further evaluate long-horizon revisitation on 50 leave-and-return cases from R2M-Bench (Gu et al., 2026) with 481-frame rollouts, where GSExplorer achieves the best results among the compared methods across PSNR, SSIM, LPIPS, and Match Ratio. These results demonstrate strong controllability and consistency across both local generation and long-horizon exploration.

Our main contributions are summarized as follows:

1. We introduce a generation–reconstruction framework centered on a persistent 3DGS world state, where inconsistency filtering and cross-chunk bundle adjustment enable selective and spatially coherent scene updates across generation steps.
2. We develop a condition-driven video generator that combines multimodal 3DGS renderings with a separate camera-geometry branch, and construct 293K degraded-condition–target video pairs for accurate camera control and scene completion.
3. GSExplorer achieves leading overall performance on single-step and three-step WorldScore evaluation and shows strong long-horizon revisit consistency after 481-frame rollouts.

2 Related Work

2.1 Camera-Controllable World Generation

Camera-controllable video generation commonly injects camera poses (Wang et al., 2024), Plücker rays (He et al., 2024; Dai et al., 2026), or relative positional encodings into DiT backbones to guide viewpoint changes (Li et al., 2026; Zhang et al., 2026). While these signals improve camera control, they provide limited information about the scene structure at the target view. Recent geometry-guided approaches therefore reconstruct explicit 3D representations and render them along planned camera trajectories as spatially aligned conditions (Ren et al., 2025; Cao et al., 2025). However, such conditions can become unreliable under large viewpoint changes.

Recent methods improve geometry conditioning from different perspectives. UniWorld-View (Zhou et al., 2026) addresses visibility and occlusion ambiguity in point-cloud rendering, while NeoVerse (Yang et al., 2026b) combines feed-forward 4D Gaussian reconstruction with online degradation simulation for geometry-conditioned generation. DreamWorld (Yang et al., 2026a) further completes geometry-aware features as an intermediate scaffold before appearance synthesis. GSExplorer combines complementary cues from multimodal 3DGS renderings with a separate camera-geometry encoding, enabling precise camera control and robust scene completion.

2.2 Memory for World Generation

Memory-based video models reuse historical observations to maintain consistency beyond the recent temporal context. Existing methods extend long-term memory by conditioning generation on stored historical frames (Yu et al., 2025b) or retrieving geometry-indexed reference views (Li et al., 2025b). HY-WorldPlay (Sun et al., 2025) and DreamX-World (Team et al., 2026) similarly reconstruct or

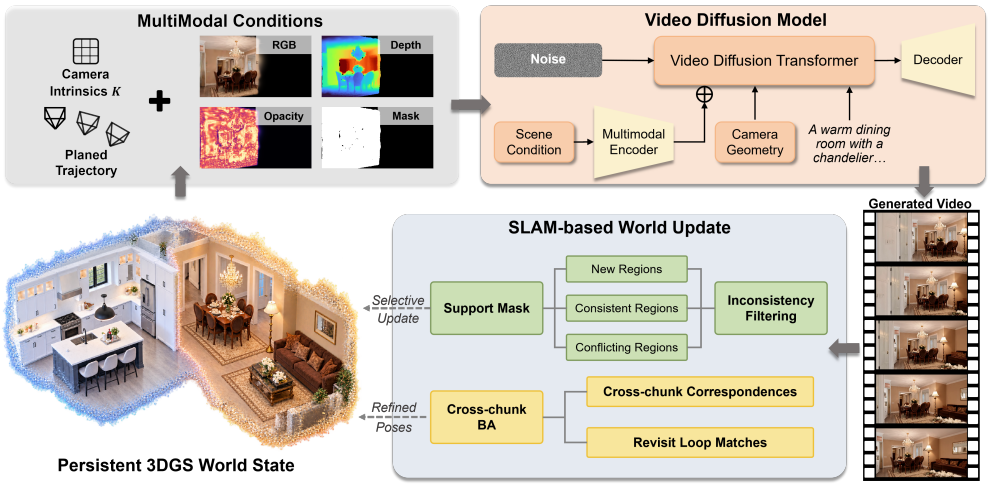

Figure 1: **Overview of GSExplorer.** Starting from a 3DGS scene initialized from a single image, we render multimodal scene conditions along the planned camera trajectory at each step. These conditions, together with camera geometry and text prompts, guide the condition-driven generator to produce a completed video. The generated video is then fed back to the SLAM-based world update, where inconsistency filtering enables selective scene updates and cross-chunk bundle adjustment aligns current and historical views. The updated 3DGS world state then provides reliable conditions for the next step, forming a generation–reconstruction loop.

retrieve historical context during autoregressive generation. These mechanisms extend access to long-term visual information without processing the full history at every step.

Explicit 3D representations provide a more spatially grounded form of memory. Early methods progressively construct and expand explicit 3D scenes (Chung et al., 2023; Yu et al., 2024a; 2025a; 2024a), while later approaches couple generation with scene reconstruction or refinement (Wu et al., 2025; Zhang et al., 2025). Voyager (Huang et al., 2025) and Spatia (Zhao et al., 2026c) maintain updatable point-cloud memories for subsequent generation, with Spatia additionally retrieving historical frames. LSM-World (Wang et al., 2026) stores diffusion features directly in a persistent 3D cache to avoid repeated RGB rendering and re-encoding. In contrast, GSExplorer maintains a persistent 3DGS world state through cross-chunk BA and inconsistency filtering, allowing the updated scene to provide reliable target-view conditions for subsequent short-horizon generation.

3 METHODOLOGY

Figure 1 gives an overview of GSExplorer. We present the persistent 3DGS world update in Sec. 3.1, the condition-driven generation model in Sec. 3.2, and the training-pair construction in Sec. 3.3.

3.1 PERSISTENT 3DGS WORLD STATE

A persistent world state should not only retain past information, but also remain reliable and queryable as new generated videos are incorporated. We maintain a 3DGS scene with a SLAM-based reconstruction backend that carries camera estimates and historical mapping views across generation steps. Instead of blindly integrating generated videos, we selectively update the world state according to their consistency with the existing scene. This enables the scene to expand into unseen regions while preserving previously reconstructed content during exploration and revisit.

Inconsistency Filtering and selective update. Generated videos may contain newly revealed regions, consistent views of existing areas, or conflicting content caused by generation errors. Before updating the Gaussian scene, we evaluate each generated frame against the historical world state and selectively determine which pixels can contribute to reconstruction. Newly visible regions are

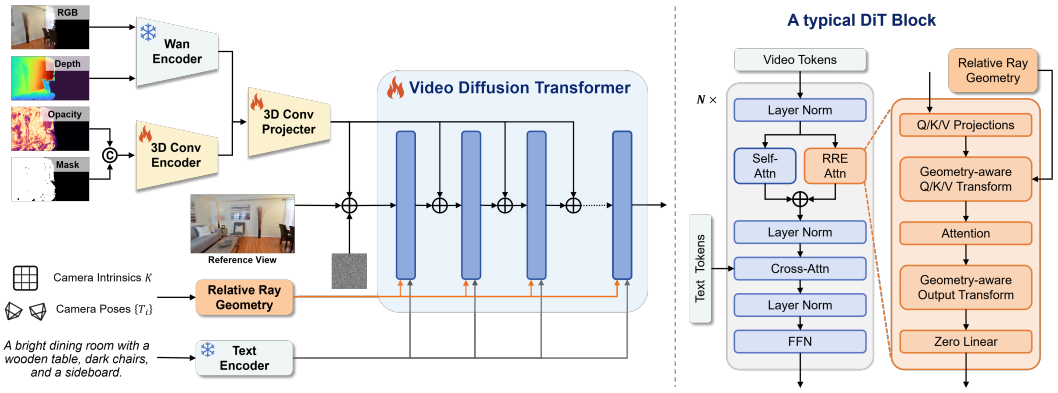

Figure 2: **Architecture of the condition-driven video generator.** Multimodal 3DGS renderings are encoded as scene conditions and injected into the diffusion backbone through one input projection and five residual projections. Camera intrinsics and poses define token-wise ray geometry, which is incorporated through a separate RRE attention branch in every DiT block using geometry-aware Q/K/V and output transforms. Text prompts follow the original Wan cross-attention pathway.

used for scene expansion, consistent regions refine the existing scene, while conflicting regions are excluded from Gaussian initialization and reconstruction optimization.

For each frame I_i in a generated video, we render the historical Gaussian subset at its estimated camera pose, obtaining the historical rendering $\hat{I}_i^{\text{old}}$ and accumulated opacity A_i^{old}. We measure the appearance discrepancy between the generated frame and historical rendering as:

$$e_i(u) = \frac{1}{3}\left\| I_i(u) - \hat{I}_i^{\text{old}}(u) \right\|_1 . \tag{1}$$

Only historically covered pixels are evaluated for consistency, since uncovered regions provide no evidence of conflict. We therefore define a pixel-wise consistency score:

$$c_i(u) = \begin{cases} \exp(-\lambda e_i(u)), & A_i^{\text{old}}(u) \geq \eta, \\ 1, & \text{otherwise}, \end{cases} \tag{2}$$

where η is the historical coverage threshold and λ controls the discrepancy scaling. Combining appearance consistency with geometric validity gives the reconstruction support mask:

$$S_i(u) = V_i(u)\,\mathbf{1}[c_i(u) \geq \tau_c], \tag{3}$$

where $V_i(u) \in \{0, 1\}$ indicates valid geometry, e.g., finite positive depth, and τ_c is the consistency threshold. The support mask guides Gaussian initialization and reconstruction optimization, allowing new content to be incorporated while preventing inconsistent generated content from overwriting the existing scene. Detailed validity checks and threshold settings are provided in the Appendix.

Cross-chunk bundle adjustment. To maintain scene consistency across generation chunks, we further refine camera alignment using the SLAM backend. Since generated videos may slightly deviate from the planned trajectory, directly fusing their frames with the existing Gaussian scene can introduce spatial misalignment. Because these poses determine how generated frames are integrated into the shared Gaussian map, better cross-chunk camera alignment directly improves scene continuity. Cross-chunk correspondences and reliable revisit matches provide constraints between current and historical keyframes, allowing the camera graph to be jointly optimized across chunks.

Specifically, we optimize camera poses by minimizing correspondence residuals over accepted cross-chunk constraints:

$$\min_{\{T_i\}} \sum_{(i,j)\in\mathcal{E}} \rho\Big(\|r_{ij}(T_i, T_j)\|_2^2 \Big) , \tag{4}$$

where $\mathcal{E}$ denotes accepted cross-chunk and revisit constraints, r_{ij} represents correspondence residuals, and ρ is a robust loss function. The refined camera states are then propagated to reconstruction,

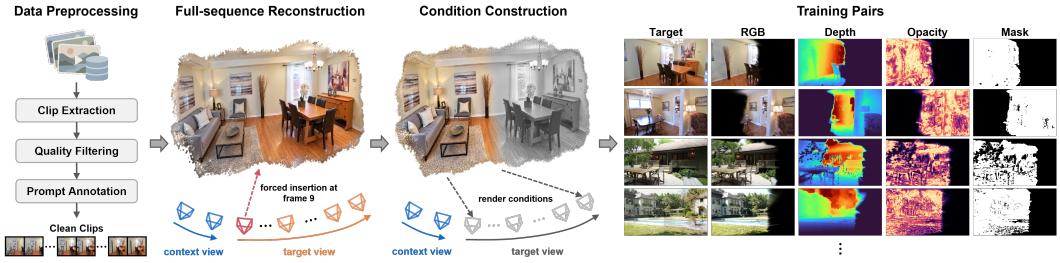

Figure 3: **Training-pair construction pipeline.** First, long sequences are filtered and sampled into 57-frame clips containing 8 context frames and 49 target frames. Next, we reconstruct the full sequence with 3DGS-SLAM while forcing Gaussian insertion at the first target frame. Together, these steps improve target-view alignment while preserving the desired context-limited degradation. We then retain only Gaussians born within the context frames, recovering the desired context-limited degradation. Finally, the retained 3DGS is rendered at the target poses to produce multimodal degraded conditions paired with the real target frames.

enabling generated frames to be integrated into the persistent 3DGS world with improved spatial alignment. Together with inconsistency filtering, this scene update maintains a coherent world state that provides reliable conditions for future generation and revisit.

3.2 Condition-Driven Video Generation

We build our generator on Wan2.2-TI2V-5B (Wan et al., 2025) and condition it on multimodal 3DGS renderings, camera geometry, and text prompts. As shown in Fig. 2, rendered observations provide the main scene guidance, while camera geometry is injected through a separate branch as complementary guidance. This keeps camera geometry separate from rendered content while allowing the two condition streams to cooperate during generation.

Multimodal conditioning. Given a planned camera trajectory, we render degraded RGB, depth, and accumulated opacity from the current 3DGS state, and derive a binary hole mask from the opacity. RGB provides observed appearance, depth provides explicit geometry, accumulated opacity describes continuous rendering coverage, and the mask indicates hole regions to be completed.

RGB and normalized depth are encoded separately with the same pretrained Wan VAE, while opacity and the hole mask are mapped to the same latent grid using a lightweight 3D convolutional encoder. Their features are concatenated and projected to the DiT hidden space. The fused condition is injected through a zero-initialized input projection and five independent zero-initialized residual projections after selected DiT blocks, yielding six content-conditioning sites in total.

Unlike prior designs that inject camera cues together with rendered modalities (Yang et al., 2026b), we encode camera geometry through a separate Relative Ray Encoding (RRE) branch from UCPE (Zhang et al., 2026). In this branch, camera intrinsics and poses define token-wise ray geometry, which is used to transform the queries, keys, and values of a lightweight attention module, followed by a geometry-aware output transform before projection back to the backbone features. The resulting RRE attention output is added in parallel to the original self-attention in every DiT block. This design allows rendered conditions to dominate scene completion, while camera geometry provides complementary cross-view guidance. For geometric consistency, depth and camera translation are normalized using the same clip-level scale, while text prompts follow the original Wan cross-attention pathway.

Training scheme. We train the generator on the paired videos constructed in Sec. 3.3 using a weighted flow-matching objective. Given the target video latent z_0 and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, we form $z_\sigma = (1-\sigma)z_0 + \sigma\epsilon$ and predict the velocity target $\epsilon - z_0$. Following the TI2V formulation, the first temporal latent slice is replaced with the clean encoding of the first target frame and excluded from the training loss. Denoting the resulting input by $\bar{z}_\sigma$, we optimize:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{z_0,\epsilon,\sigma}\left[w(\sigma)\ \mathrm{MSE}_{\neg\mathrm{ref}}\left(v_\theta(\bar{z}_\sigma, \sigma; \mathcal{C}, y), \epsilon - z_0\right)\right], \tag{5}$$

Table 1: **Results on WorldScore for single-step extrapolation.** GSExplorer achieves the best overall performance with strong controllability and consistency. The first, second, and third best results are highlighted accordingly.

Method	3D Consist.	Photo Consist.	Style Consist.	Camera Ctrl.	Object Ctrl.	Content Align.	Subjective Qual.	Avg.
WonderWorld	82.85	67.86	55.79	92.32	47.63	79.09	69.03	**70.65**
AETHER	79.84	58.68	72.09	57.44	52.26	28.06	41.11	**55.64**
Uni3C	78.59	85.48	88.32	62.94	45.83	47.40	57.00	**66.51**
Voyager	56.00	80.68	72.89	45.92	57.69	48.36	44.74	**58.04**
FantasyWorld	83.31	86.11	94.22	57.05	34.46	38.45	57.40	**64.43**
Wan2.2-5B-Camera	78.19	83.07	84.16	75.83	62.59	63.12	47.46	**70.63**
NeoVerse	83.67	78.65	89.33	90.81	31.63	69.02	63.03	**72.31**
Spatia	86.81	86.63	85.33	73.98	58.84	51.36	49.74	**70.38**
DreamWorld	84.96	92.70	92.98	81.97	63.11	50.37	59.17	**75.04**
DreamX-World	77.93	64.81	86.21	73.30	63.56	68.00	50.84	**69.24**
UniWorld-View	89.53	89.41	84.38	94.57	46.09	68.51	63.76	**76.61**
GSExplorer (Ours)	90.83	88.47	94.40	99.07	65.91	74.70	66.59	**82.85**

where $\mathcal{C}$ denotes the multimodal conditions, y is the text prompt, $w(\sigma)$ is a timestep-dependent weight, and $\mathrm{MSE}_{\neg\mathrm{ref}}$ averages over all latent elements except the first temporal slice.

Training follows two stages. We first keep the DiT backbone frozen and optimize only the multimodal adapter and pose branch to establish basic condition following. We then jointly fine-tune the DiT and both conditioning branches. The VAE and text encoder remain frozen throughout.

3.3 Training Pair Construction

To train our model to complete videos from multimodal conditions and text prompts, we construct degraded conditions paired with target videos from diverse real-world scene datasets, including RealEstate10K (Zhou et al., 2018), DL3DV (Ling et al., 2024), ACID (Liu et al., 2021), and ScanNet (Dai et al., 2017). As shown in Fig. 3, the key is to reproduce realistic 3DGS degradation while maintaining pixel alignment with the real targets.

Data preprocessing. We segment long sequences into 57-frame clips after quality filtering and use Qwen3.5 (Qwen Team, 2026) to annotate a scene description for each clip. The first 8 frames serve as context and the remaining 49 as targets. More details are provided in the Appendix.

Degraded condition construction. Reconstructing only the 8 context frames provides insufficient geometric constraints, causing rendered conditions to become severely misaligned with real targets at farther extrapolated views. We therefore run the same 3DGS-SLAM system in our extrapolation pipeline over the full 57-frame sequence and record the birth frame b_g of each Gaussian primitive.

Full-sequence optimization, however, may cause context Gaussians to adapt to future observations and weaken the degradation we aim to simulate. We therefore force the first target frame (frame 9) to be a keyframe and explicitly insert new Gaussians, providing dedicated capacity for newly revealed content. After reconstruction, we retain only Gaussians born within the context:

$$\mathcal{G}^{\mathrm{ctx}} = \left\{ g \in \mathcal{G}^{\mathrm{full}} \mid b_g < 8 \right\}. \tag{6}$$

The resulting $\mathcal{G}^{\mathrm{ctx}}$ retains context-limited scene coverage while benefiting from the improved alignment of full-sequence SLAM.

Condition rendering and alignment filtering. We render $\mathcal{G}^{\mathrm{ctx}}$ at the 49 SLAM-estimated target poses to obtain degraded RGB, depth, and accumulated opacity, derive the corresponding hole masks, and retain the camera poses as conditions. To filter remaining misaligned samples, we apply ECC registration (Evangelidis & Psarakis, 2008) between the valid region of the last rendered frame and its target and compute the mean pixel displacement d_{ECC}. Clips with $d_{\mathrm{ECC}} \geq \tau$ are discarded. The final dataset contains 293K training video pairs.

Table 2: **Results on WorldScore for three-step extrapolation.** GSExplorer maintains the best overall performance under multi-step generation, with strong camera control and scene consistency.

Method	3D Consist.	Photo Consist.	Style Consist.	Camera Ctrl.	Object Ctrl.	Content Align.	Subjective Qual.	Avg.
WonderWorld	63.70	3.22	35.95	96.28	38.61	97.10	72.46	**58.19**
AETHER	63.97	33.11	61.99	4.43	34.78	33.69	35.09	**38.15**
Uni3C	79.95	46.78	71.43	8.69	34.28	77.88	51.12	**52.88**
Voyager	13.82	9.52	61.34	0.00	49.23	64.10	39.21	**33.89**
FantasyWorld	74.83	60.61	82.02	11.24	31.96	77.20	50.46	**55.47**
Wan2.2-5B-Camera	64.73	80.05	71.07	49.38	51.62	39.57	34.97	**55.91**
NeoVerse	66.48	75.68	83.99	65.09	26.50	41.74	61.30	**60.11**
Spatia	65.95	77.33	75.17	29.85	44.88	10.58	37.95	**48.82**
DreamX-World	56.69	37.59	80.29	33.28	52.63	46.27	38.80	**49.36**
UniWorld-View	75.02	88.16	70.45	71.21	42.13	35.12	66.49	**64.08**
GSExplorer (Ours)	77.50	84.51	84.36	85.17	56.25	52.82	60.34	**71.56**



Figure 4: **Qualitative WorldScore results across one-, two-, and three-step extrapolation.** P and C denote the text prompt and camera instruction. GSExplorer accurately follows both controls while maintaining high visual quality and coherent spatial structure over multiple generation steps.

4 Experiments

4.1 Experimental Setup

Benchmarks and metrics. We evaluate both single-step and three-step world extrapolation on the photorealistic subset of the WorldScore static benchmark (Duan et al., 2025). Following its protocol, we report seven metrics covering controllability and generation quality: Camera Control, Object Control, Content Alignment, 3D Consistency, Photometric Consistency, Style Consistency, and Subjective Quality. For long-horizon revisit evaluation, we select 50 cases from R2M-Bench (Gu et al., 2026) and use its leave-and-return trajectories. Each method generates a 10-step, 481-frame rollout that returns to the initial viewpoint. We compare the final frame with the input image using PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and Match Ratio. Following R2M-Bench, Match Ratio is computed using SuperPoint (DeTone et al., 2018) and LightGlue (Lindenberger et al., 2023), normalized by the smaller number of detected keypoints in the image pair.

Baselines. For WorldScore, we compare GSExplorer with WonderWorld (Yu et al., 2025a), AETHER (Zhu et al., 2025), Uni3C (Cao et al., 2025), Voyager (Huang et al., 2025), FantasyWorld (Dai et al., 2026), NeoVerse (Yang et al., 2026b), Spatia (Zhao et al., 2026c), DreamWorld (Yang et al., 2026a), DreamX-World (Team et al., 2026), UniWorld-View (Zhou et al., 2026), and Wan2.2-Fun-5B-Control-Camera baseline (Wan et al., 2025), covering explicit 3D scene generation, geometry-guided video generation, and recent interactive world models. For long-horizon revisit evaluation, we compare against UniWorld-View, DreamX-World, WonderWorld, and Spatia.

Table 3: **Long-horizon revisit evaluation.** GSExplorer achieves the best revisit consistency.

Metric	UniWorld-View	DreamX-World	WonderWorld	Spatia	GSExplorer (Ours)
PSNR↑	11.73	12.02	15.09	17.86	**19.85**
SSIM↑	0.375	0.417	0.592	0.565	**0.623**
LPIPS↓	0.544	0.649	0.427	0.415	**0.369**
Match Ratio↑	0.210	0.131	0.431	0.423	**0.481**

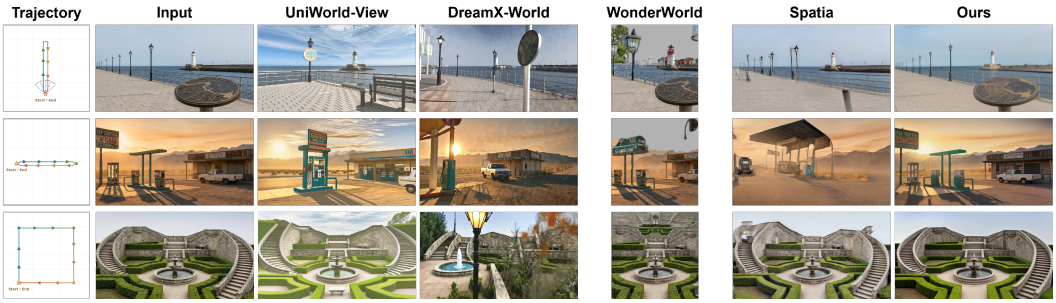

Figure 5: **Qualitative comparison of long-horizon revisit results.** GSExplorer better preserves the content and details of the starting view after long-range exploration.

Implementation details. We initialize the generator from Wan2.2-TI2V-5B and train it for 30K iterations on 32 NVIDIA H20 GPUs with a global batch size of 32. Training clips contain 49 frames at 640×352 resolution, while ScanNet clips use 640×480. For the first 10K iterations, we freeze the DiT backbone and train the multimodal adapter and pose branch with a learning rate of 10^{-4}. We then jointly fine-tune the DiT, multimodal adapter, and pose branch for 20K iterations with learning rates of 1×10^{-7}, 1×10^{-5}, and 1×10^{-5}, respectively. We apply clean-reference and text-prompt dropout with probabilities of 0.5 and 0.1. Additional training details are provided in the Appendix.

4.2 Main Results

World generation on WorldScore. Tables 1 and 2 report the single-step and three-step extrapolation results on WorldScore, respectively. In the single-step setting, GSExplorer achieves the highest average score of **82.85**, outperforming the strongest baseline by 6.24 points. It also reaches **99.07** Camera Control and ranks first in 3D Consistency, Style Consistency, and Object Control, demonstrating precise local control together with scene consistency. The improvement is not limited to one metric category, but spans camera following, geometric consistency, and generation quality, suggesting that the multimodal scene conditions and explicit camera guidance are complementary.

The advantage remains under three-step extrapolation, where GSExplorer achieves the best average score of **71.56**, exceeding the strongest baseline by 7.48 points. It ranks first in Style Consistency and Object Control and second in 3D Consistency, Photometric Consistency, and Camera Control. The margin over the strongest baseline increases from 6.24 points in the single-step setting to 7.48 points after three steps, indicating that the advantage is maintained as errors accumulate over successive generation chunks. Together with the qualitative results in Fig. 4, these results show that GSExplorer preserves strong controllability and generation quality throughout multi-step scene extrapolation.

Long-horizon revisit consistency. As shown in Table 3, GSExplorer achieves the best performance on all revisit metrics, with **19.85** PSNR, **0.623** SSIM, **0.369** LPIPS, and **0.481** Match Ratio. Compared with the strongest competing results, this corresponds to a 3.99 dB improvement in PSNR and a 0.050 gain in Match Ratio. The qualitative results in Fig. 5 further show better recovery of the starting-view content and fine details after long-range exploration. These results indicate that GSExplorer maintains a consistent persistent 3D world state over long rollouts, providing reliable conditions when revisiting previously explored regions.

Table 4: **Ablation of conditioning modalities and pose injection.** All conditioning modalities contribute to the final performance, while separate pose encoding through RRE achieves the best overall result. "w/o RRE-Attn" injects pose together with the rendered conditions instead of using the separate RRE attention branch.

Method	3D Consist.	Photo Consist.	Style Consist.	Camera Ctrl.	Object Ctrl.	Content Align.	Subjective Qual.	Avg.
w/o RRE-Attn	88.14	84.51	93.80	96.03	65.84	73.70	65.56	81.08
w/o pose	88.64	85.74	93.68	95.58	65.72	73.41	65.75	81.22
w/o depth	87.13	88.02	93.84	97.65	65.53	74.30	66.20	81.81
w/o opacity	90.06	88.91	94.06	98.82	64.34	73.06	65.87	82.16
w/o pose & opacity	87.66	85.37	93.80	95.02	64.18	72.59	65.30	80.56
Ours	**90.83**	**88.47**	**94.40**	**99.07**	**65.91**	**74.70**	**66.59**	**82.85**

Table 5: **Ablation of persistent world-state update.** Both observation filtering and cross-chunk bundle adjustment improve multi-step extrapolation and revisit consistency.

Method	WorldScore Three-step Extrapolation								Revisit Evaluation	
	3D Consist.	Photo Consist.	Style Consist.	Camera Ctrl.	Object Ctrl.	Content Align.	Subjective Qual.	Avg.	PSNR	Match Ratio
w/o filter	72.78	73.97	78.16	68.15	56.09	52.58	54.23	65.14	17.15	0.378
w/o BA	75.54	81.16	81.43	78.50	55.69	51.79	58.29	68.91	17.41	0.406
Ours	**77.50**	**84.51**	**84.36**	**85.17**	**56.25**	**52.82**	**60.34**	**71.56**	**19.85**	**0.481**

4.3 Ablation Study

Conditioning design. As shown in Table 4, each conditioning modality contributes to the final performance. Removing depth mainly degrades geometric consistency, reducing 3D Consistency from 90.83 to 87.13, while removing opacity also lowers the overall score by 0.69 points. Pose conditioning is important for controllability: removing pose reduces Camera Control from 99.07 to 95.58. Moreover, injecting pose together with the rendered conditions (w/o RRE-Attn) underperforms our separate RRE branch in both Camera Control and the average score. Removing both pose and opacity causes the largest overall drop, further showing their complementary roles. The full design provides the best combination of controllability, consistency, and generation quality.

Persistent world-state update. Table 5 evaluates the two key components of our persistent world-state update. Without inconsistency filtering, the multi-step average drops from 71.56 to 65.14, while revisit PSNR and Match Ratio decrease from 19.85/0.481 to 17.15/0.378. This indicates that filtering conflicting generated observations is important for preventing errors from propagating into subsequent scene updates. Removing cross-chunk bundle adjustment also consistently degrades both multi-step generation and revisit performance, reaching 68.91 average score, 17.41 PSNR, and 0.406 Match Ratio. These results show that inconsistency filtering and cross-chunk alignment are complementary for maintaining a reliable persistent world state.

5 Conclusion

GSExplorer shows that interactive world generation can benefit from separating long-term spatial state from local video synthesis. Rather than asking the generator to carry an increasingly long visual history, we maintain an explicit 3DGS world that is repeatedly queried, selectively updated, and geometrically aligned as exploration proceeds. This design keeps generation local while allowing spatial information to accumulate in a reusable scene representation. Our results on multi-step extrapolation and long-horizon revisit suggest that reliable world-state maintenance is a useful complement to increasingly capable video generators. A natural next step is to make the world update more robust to dynamic content and ambiguous observations, while improving the efficiency and scalability of persistent scene maintenance.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and writing polish, including improving grammar, clarity, and presentation of author-written text. They were not used to develop the research methodology, design experiments, generate or analyze experimental results, or make scientific conclusions. All AI-assisted edits were reviewed and revised by the authors to ensure technical accuracy, consistency with the conducted research, and appropriate attribution. The authors take full responsibility for the final content of this work, including all text, claims, results, and conclusions.

ETHICS STATEMENT

This work focuses on interactive world generation and does not involve human subjects or the collection of personally identifiable information. Our experiments use existing research datasets, including RealEstate10K, DL3DV, ACID, and ScanNet, and we use them in accordance with their respective licenses and terms of use. As with other generative models, the proposed system may produce inaccurate, biased, or synthetic visual content and could potentially be misused to create misleading imagery. The generated content should therefore not be interpreted as a faithful record of real-world events. Our work is intended for research on controllable world generation, 3D scene modeling, and related applications.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, training procedure, data construction pipeline, evaluation protocols, and implementation settings to facilitate reproducibility. The main components of GSExplorer are described in Secs. 3.1–3.3, including the persistent 3DGS world update, condition-driven video generator, and training-pair construction. Experimental protocols, baselines, metrics, and training settings are provided in Sec. 4.1. Additional implementation details are given in the Appendix, including the reconstruction and inconsistency-filtering settings, multimodal conditioning and Relative Ray Encoding architecture and training configuration, and dataset preprocessing, filtering, and training-pair construction procedures.

REFERENCES

Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

Chong Cheng, Yu Hu, Sicheng Yu, Beizhen Zhao, Zijian Wang, and Hao Wang. Reggs: Unposed sparse views gaussian splatting with 3dgs registration. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8100–8109. IEEE, 2025a.

Chong Cheng, Sicheng Yu, Zijian Wang, Yifan Zhou, and Hao Wang. Outdoor monocular slam with global scale-consistent 3d gaussian pointmaps. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26035–26044. IEEE, 2025b.

Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2432–2443. IEEE, 2017.

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*, volume 2026, pp. 103603–103622, 2026.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *2018 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 337–33712. IEEE, 2018.

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*, pp. 27713–27724, 2025.

Georgios D Evangelidis and Emmanouil Z Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE transactions on pattern analysis and machine intelligence*, 30(10):1858–1865, 2008.

Qiwen Gu, Bingjie Gao, Rui Chen, Geng Li, Jifan Li, Qishuai Wen, Li Niu, Jing Tang, Xiangxiang Chu, and Junqiao Zhao. R2m-bench: Evaluating revisit memory via relative consistency in interactive video world models. *arXiv preprint arXiv:2608.27328*, 2026.

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.

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, 2025.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Guanghao Li, Kerui Ren, Linning Xu, Zhewen Zheng, Changjian Jiang, Xin Gao, Bo Dai, Jian Pu, Mulin Yu, and Jiangmiao Pang. Artdeco: Towards efficient and high-fidelity on-the-fly 3d reconstruction with structured scene representation. *arXiv preprint arXiv:2510.08551*, 2025a.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *Advances in Neural Information Processing Systems*, 38:15984–16009, 2026.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25690–25699. IEEE, 2025b.

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. *arXiv preprint arXiv:2511.10647*, 2025.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 17581–17592. IEEE, 2023.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169. IEEE, 2024.

Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the ieee/cvf international conference on computer vision*, pp. 14458–14467, 2021.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132. IEEE, 2025.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world model. *arXiv preprint*, 2025.

DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, et al. Dreamx-world 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026.

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.

Weijie Wang, Haoyu Zhao, Yifan Yang, Feng Chen, Zeyu Zhang, Yefei He, Zicheng Duan, Donny Y Chen, Yuqing Yang, and Bohan Zhuang. Latent spatial memory for video world models. *arXiv preprint arXiv:2606.09828*, 2026.

Zhou Wang, Alan C. Bovik, Hamid R. 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.

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. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Sibo Wu, Congrong Xu, Binbin Huang, Andreas Geiger, and Anpei Chen. Genfusion: Closing the loop between reconstruction and generation via videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6078–6088. IEEE, 2025.

Haibo Yang, Yang Chen, Yingwei Pan, Zhineng Chen, Ting Yao, and Tao Mei. Dreamworld: Geometry-grounded video diffusion for 3d-consistent world modeling. In *ECCV*, 2026a.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos. *arXiv preprint arXiv:2601.00393*, 2026b.

Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6658–6667. IEEE, 2024a.

Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5916–5926. IEEE, 2025a.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025b.

Sicheng Yu, Chong Cheng, Yifan Zhou, Xiaojun Yang, and Hao Wang. Rgb-only gaussian splatting slam for unbounded outdoor scenes. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11068–11074. IEEE, 2025c.

Sicheng Yu, Dongxu Shen, Beizhen Zhao, Guanzhi Ding, and Hao Wang. Robust and efficient monocular 3d gaussian slam for kilometer-scale outdoor scenes. *arXiv preprint arXiv:2606.30436*, 2026.

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*, 2024b.

Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Gambardella, Dinh Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38027–38037, 2026.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Shengjun Zhang, Jinzhao Li, Xin Fei, Hao Liu, and Yueqi Duan. Scene splatter: Momentum 3d scene generation from single image with video diffusion model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6089–6098. IEEE, 2025.

Beizhen Zhao, Sicheng Yu, Guanzhi Ding, Yu Hu, and Hao Wang. Pseudo-view enhancement via confidence fusion for unposed sparse-view reconstruction. *arXiv preprint arXiv:2602.21535*, 2026a.

Beizhen Zhao, Sicheng Yu, Ziran Yin, Dongxu Shen, and Hao Wang. Mmgs: 10

$$\backslash$$

times $compressed3dgsthroughoptimaltransportaggregationbasedonmulti-viewranking$. *arXiv preprint arXiv:2605.19304*, 2026*b*.

Jinjing Zhao, Fangyun Wei, Zhening Liu, Hongyang Zhang, Chang Xu, and Yan Lu. Spatia: Video generation with updatable spatial memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4245–4257, 2026c.

Haiyang Zhou, Wangbo Yu, Chaoran Feng, Xunyu Zhou, Yonghong Tian, and Li Yuan. Uniworld-view: Large-baseline view synthesis via video diffusion models. *arXiv preprint arXiv:2608.04701*, 2026.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

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

We provide additional implementation details of the persistent 3DGS world update and the condition-driven video generator in Appendix A, followed by details of the training-pair construction pipeline in Appendix B. Appendix C discusses the current limitations of GSExplorer, and Appendix D presents additional qualitative results. The supplementary material also includes an MP4 video with extended multi-step exploration and long-horizon revisit examples.

A.1 Additional Implementation Details

A.1.1 Persistent 3DGS World Update

3DGS-SLAM backend. Our world state must support both viewpoint-dependent rendering and incremental updates as new generated videos arrive. A 3DGS-SLAM system (Yu et al., 2025c; Cheng et al., 2025b) naturally provides these capabilities by maintaining a continuously renderable Gaussian scene (Zhao et al., 2026b;a) while estimating and refining camera states from streaming observations (Cheng et al., 2025a; Yu et al., 2026). We therefore build our world update on an ARTDECO-based (Li et al., 2025a) 3DGS reconstruction backend, using MASt3R (Leroy et al., 2024) correspondences and a Depth Anything 3 (DA3) (Lin et al., 2025) geometry prior.

The same Gaussian scene, historical mapping views, camera graph, and retrieval database are retained throughout a rollout. Each 49-frame generation chunk shares its first frame with the preceding chunk and contributes 48 new observations. DA3 jointly processes the shared boundary frame and the new frames to initialize depth and cameras, with depth aligned to the existing reconstruction using a scale-only transformation. Image correspondences and retrieved historical matches then connect current and previous keyframes for cross-chunk camera refinement. The refined camera states are propagated to Gaussian reconstruction before the updated world state is rendered for the next generation step.

Natural tracking keyframes are selected with a match-fraction threshold of 0.45 and a minimum separation of four global frames. Up to 24 ordinary mapping frames are retained per chunk. Frames with insufficient geometric support are discarded before entering the mapping stage.

Inconsistency-aware scene update. Sec. 3.1 defines the historical consistency score and reconstruction support mask. In our implementation, pixels with historical accumulated opacity above 0.95 are treated as covered. For a covered pixel with RGB discrepancy $e_i(u)$, the consistency score is

$$c_i(u) = \exp\left(-\frac{\log 2}{0.15} e_i(u)\right), \tag{7}$$

and pixels with $c_i(u) \geq 0.5$ are retained. Geometry validity additionally requires finite 3D points with positive depth and finite RGB values.

The resulting binary support is used for Gaussian insertion and scene optimization. Before inserting new Gaussians, rejected regions are conservatively expanded with a 7×7 neighborhood to avoid mixing conflicting colors near their boundaries. During scene refinement, the support is applied to RGB, SSIM, and inverse-depth supervision. Historical mapping views retain their consistency masks when reused in later optimization, allowing previously reconstructed content to remain protected while still being refined by consistent observations.

A.1.2 Condition-Driven Video Generation

Multimodal condition adapter. Rendered RGB and normalized depth are independently encoded by the same frozen Wan VAE, producing 48-channel latent features for each modality. Accumulated opacity and the binary hole mask are concatenated and processed by a lightweight 3D convolutional encoder with 64 output channels. The resulting

$$48 + 48 + 64 = 160$$

channels are fused and projected to the 3072-dimensional DiT hidden space.

The projected condition is injected once at the DiT input and after five selected blocks. Using zero-based indexing, these block-output injections follow blocks

$$\{5, 10, 15, 20, 25\}.$$

The six conditioning projections are zero-initialized, allowing the pretrained backbone to begin from its original behavior while gradually learning the additional scene conditions.

Relative Ray Encoding. We use the Relative Ray Encoding (RRE) component of UCPE (Zhang et al., 2026) without its Absolute Orientation Encoding. Camera intrinsics and camera-to-world poses define a local viewing-ray frame for each latent token. Each RRE module projects the 3072-dimensional hidden state to three attention heads of width 128. Geometry-aware transformations are applied to the query, key, and value features in their corresponding ray frames, followed by the inverse geometric transform on the attention output.

The RRE branch is instantiated in every DiT block and operates in parallel with the original self-attention. Its output is projected back to the 3072-dimensional hidden space through a zero-initialized projection and combined with the original attention output before the remaining cross-attention and feed-forward operations.

Geometric normalization. Depth and camera translation share a clip-level normalization scale. We compute the 98th percentile of valid depth values over the 49-frame clip, where valid pixels have accumulated opacity of at least 0.95, finite depth, and depth within the configured valid range. Depth is divided by this scale and clipped to $[0, 1]$, while camera translations relative to the first frame are divided by the same scale. The binary hole mask is

$$M_{\text{hole}}(u) = \mathbf{1}[A(u) < 0.95], \tag{8}$$

while accumulated opacity remains a continuous condition.

Table 6: Composition of the final training set.

Dataset	Training pairs
RealEstate10K	231,644
DL3DV	14,279
ACID	9,983
ScanNet	38,037
Total	**293,943**

Training details. We train the generator for 30K iterations on 32 NVIDIA H20 GPUs with a global batch size of 32 using BF16 precision. Training clips contain 49 frames at 640×352, while ScanNet clips use 640×480. We use AdamW with zero weight decay and cosine learning-rate schedules.

During the first 10K iterations, the DiT backbone is frozen and the multimodal adapter and RRE branch are trained with a peak learning rate of 1×10^{-4}. During the remaining 20K iterations, the DiT, multimodal adapter, and RRE branch are jointly optimized with peak learning rates of

$$1 \times 10^{-7}, \qquad 1 \times 10^{-5}, \qquad 1 \times 10^{-5},$$

respectively. The VAE and text encoder remain frozen throughout training.

We apply clean-reference dropout with probability 0.5. When the reference is retained, its clean latent replaces the first noisy temporal slice and that slice is excluded from the flow-matching loss. When the reference is dropped, all temporal slices follow the standard noisy training path and contribute to the loss. The clean reference improves continuity across consecutive generation chunks, while the dropout prevents the model from over-relying on the reference view and preserves the influence of the multimodal scene and camera conditions. We found this balance useful for maintaining both cross-chunk continuity and condition controllability. Text prompts are independently dropped with probability 0.1.

For the weighted flow-matching objective in Sec. 3.2, the training scheduler contains 1000 noise levels with shift 5. For scheduler noise level σ_k, we use

$$a_k = \exp\left[-2(\sigma_k - \tfrac{1}{2})^2\right], \qquad w_k = \frac{a_k - \min_j a_j}{\frac{1}{L}\sum_{j=1}^{L}(a_j - \min_r a_r)}. \tag{9}$$

No additional geometric, reprojection, perceptual, temporal-consistency, or RGB reconstruction losses are used for generator training.

A.2 Training Pair Construction Details

A.2.1 Dataset Statistics

After preprocessing, reconstruction, and alignment filtering, the final training set contains 293,943 video pairs. Table 6 summarizes their composition.

A.2.2 Data Sampling and Filtering

All datasets are converted to a common 57-frame representation containing 8 context frames and 49 target frames, with RGB images, intrinsics, and camera poses sampled at identical indices. DL3DV and ScanNet use consecutive 57-frame windows with a stride of 28. For RealEstate10K, the eight context frames are consecutive while target frames are sampled over the remaining valid interval. ACID sequences are sampled approximately uniformly over temporally and geometrically valid intervals.

We screen candidate clips using bidirectional depth-reprojection coverage and relative camera rotation. Given the directional coverage $c_{i\to j}$, the symmetric overlap is

$$o_{ij} = \min(c_{i\to j}, c_{j\to i}). \tag{10}$$

DL3DV and ScanNet retain anchor views with overlap at least 0.35 and relative rotation below $50°$. RealEstate10K and ACID additionally use geometry estimated by DA3 to truncate or split sequences with excessive viewpoint change.

We further reject clips containing invalid cameras, severe image degradation, temporal discontinuities, or dominant dynamic content. Dataset-specific checks are applied where necessary: DL3DV and ScanNet additionally use DOVER-based quality filtering, while RealEstate10K includes transition checks and ACID includes dedicated checks for persistent borders, exposure changes, and highly dynamic foreground regions.

A.2.3 Pair Construction and Alignment Filtering

For all datasets, pair construction follows the procedure described in Sec. 3.3. Full-sequence reconstruction is performed over all 57 selected frames, and the first target frame (index 8) is explicitly used for Gaussian insertion. Each Gaussian records its birth frame, while the retained context map contains only primitives satisfying

$$0 \leq b_g < 8. \tag{11}$$

The retained primitives keep their parameters after full-sequence optimization, preserving the improved camera and geometric alignment while removing scene coverage introduced by target-frame Gaussians.

The retained context map is rendered at the 49 estimated target poses to obtain RGB, depth, and accumulated opacity. We additionally apply translation-only ECC registration between the valid region of the final rendered frame and its real target. The valid region is determined from accumulated opacity and conservatively eroded before registration. Clips with an estimated translation magnitude larger than 4 pixels are discarded. The ECC transformation is used only for alignment verification and does not modify the rendered training conditions.

A.3 Limitations

GSExplorer currently introduces additional runtime compared with purely feed-forward video generation because each generation chunk is followed by camera estimation and 3DGS scene refinement. The current formulation is also primarily designed for static or predominantly static environments, where persistent geometric conditions remain meaningful across exploration. Extending the world state to explicitly model dynamic objects would broaden its applicability to more interactive scenes. Finally, repeated generation and reconstruction may soften very fine textures and small indoor details over long rollouts. Improving high-frequency appearance preservation and reconstruction efficiency are promising directions for future work.

A.4 Additional Qualitative Results

Figure 6 provides additional examples of text- and camera-controlled world generation on WorldScore. Figure 7 presents additional long-horizon leave-and-return comparisons. More complete rollouts are included in the supplementary MP4 video.



Figure 6: Additional qualitative results on WorldScore. GSExplorer follows the specified text and camera controls while maintaining coherent scene appearance and geometry during extrapolation.

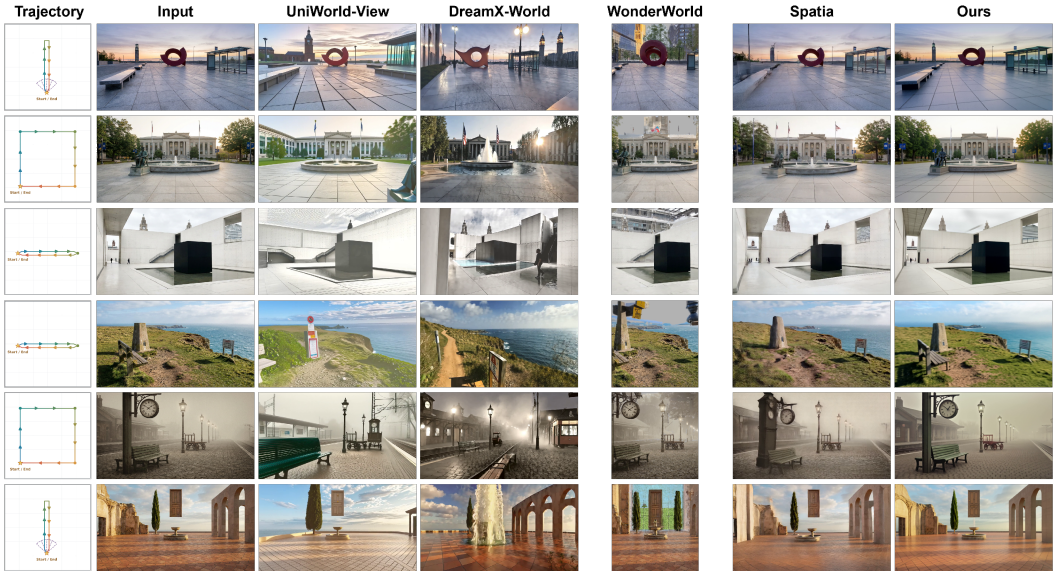

Figure 7: Additional qualitative comparison of long-horizon revisit results. After extended leave-and-return exploration, GSExplorer better preserves the scene content and spatial layout of the initial view.