AMOR: A MEMORY-GUIDED OMNIDIRECTIONAL RGB-D DIFFUSION MODEL FOR GEOMETRY-CONSISTENT NEXT-VIEW SYNTHESIS

Anonymous authors
Paper under double-blind review



Figure 1: **AMOR generates panoramic RGB-D views and reconstructs a 3D scene from a single RGB panorama.** Top: input RGB panorama and detail comparisons with HY-World 2.0 and Matrix-3D. Bottom: fused point clouds and meshes at matched viewpoints, with the corresponding generated RGB-D panoramas below.

ABSTRACT

Visual world models aim to support viewpoint exploration while preserving a coherent environment. Repeatedly extending video clips can compound geometric drift and visual degradation. For static scenes, panoramas cover more space from sparse viewpoints, but each generated view must be internally consistent in appearance and depth, and anchored to prior observations. We present AMOR, a memory-guided paradigm that progressively expands a persistent, explicit 3D scene through joint omnidirectional RGB-D generation at sparse viewpoints. Surface-guided memory readout conditions generation on target-aligned evidence from previously observed surfaces, while conservative updates integrate geometrically compatible predictions into the shared 3D scene. Beyond generating new views, AMOR's joint modeling of appearance and geometry also enables depth estimation conditioned on observed RGB. Conservative memory updates reduce final-map geometric accuracy error by 21% relative to direct fusion. For novel-view synthesis, AMOR achieves the lowest LPIPS and FID on each of three public benchmarks and the highest average PSNR and SSIM. For depth estimation from observed RGB, it achieves the lowest RMSE and highest δ_1 on TartanGround among the evaluated methods.

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.

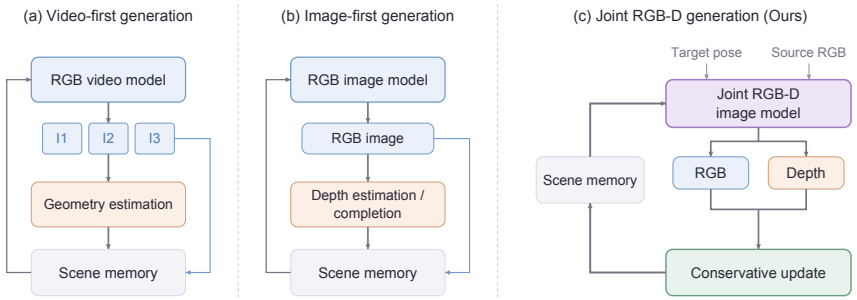


Figure 2: **Representative RGB-D generation pathways.** (a) RGB video generation followed by geometry estimation. (b) Per-view RGB synthesis followed by depth estimation or completion. (c) AMOR's joint RGB-D generation loop with conservative memory updates. These pathways are illustrative rather than exhaustive.

1 Introduction

Visual world modeling has gained substantial momentum in recent years, driven by advances in generative models for controllable scene exploration (Ren et al., 2025; Team HY-World, 2026). For static environments, a useful spatial world model must complete unseen regions while preserving previously observed appearance and geometry. Explicit representations such as point clouds and meshes make this geometry accessible beyond a particular sequence of rendered views, supporting reuse across robotics, simulation, and interactive graphics (Zhou et al., 2018; Savva et al., 2019). This motivates generating new observations as part of a persistent 3D scene.

Generating new views of a static scene requires high-fidelity appearance and corresponding geometry. Pano2Room and WonderWorld generate RGB images before recovering depth (Pu et al., 2024; Yu et al., 2025), while GEN3C and HY-World combine RGB generation with geometric reconstruction (Ren et al., 2025; Team HY-World, 2026) (Fig. 2). Joint RGB-D models instead learn appearance and geometry together (Byung-Ki et al., 2025; Duisterhof et al., 2026). Across viewpoints, generation additionally needs access to previously reconstructed surfaces (Ren et al., 2025; Huang et al., 2025), motivating an explicit geometric memory alongside joint generation.

Although joint RGB-D generation models appearance and geometry together, incorporating depth into a generative model pretrained on RGB requires a representation that preserves depth's numerical accuracy. Encoding depth as RGB enables reuse of an image VAE (Gabeur et al., 2026), but our panoramic codec diagnostic shows that VAE reconstruction distorts the recovered depth values (Appendix C). We therefore jointly generate RGB and pixel-space scalar depth, retaining the pretrained RGB VAE for appearance and a dedicated pathway for depth (Duisterhof et al., 2026).

We introduce AMOR, a memory-guided omnidirectional RGB-D diffusion model that progressively expands a 3D scene from a single RGB panorama. Its memory stores a 3D point cloud and associated RGB-D observations. **Surface-guided memory readout** identifies stored surfaces visible from the target camera pose and reprojects their RGB and depth into partial panoramas that guide generation. **Conservative memory updates** align newly predicted depth with existing geometry and selectively incorporate compatible regions. The updated memory provides scene content for subsequent views.

Across three public benchmarks, AMOR achieves the lowest LPIPS and FID on each dataset and the highest average PSNR and SSIM, while also supporting depth estimation from observed RGB. Controlled ablations clarify the complementary roles of memory readout and update. Using the same model architecture, surface-guided RGB readout improves generated-view quality over point projection. Conservative updates reduce final-map geometric accuracy error by 21% relative to direct fusion and better preserve previously observed appearance; the precision benefit persists when raw predictions are held fixed. These findings suggest that effective geometric memory depends on both how scene evidence is presented to the generator and how predictions are integrated for future use.

Our main contributions are as follows.

- We introduce AMOR, a memory-guided omnidirectional RGB-D diffusion model that jointly generates high-fidelity RGB panoramas and scalar depth to progressively expand a 3D scene from a single RGB panorama.
- We develop surface-guided memory readout that reprojects visible reconstructed surfaces into RGB and depth conditions at the target viewpoint, improving generated-view quality over point projection.
- We propose conservative memory integration that selectively incorporates geometrically compatible predictions. Sequential generation, fixed-prediction replay, and revisit experiments demonstrate improved map accuracy and better preservation of previously observed appearance relative to direct accumulation.

2 Related Work

Joint appearance and geometry generation. JointDiT (Byung-Ki et al., 2025) uses modality-dependent noise scheduling for joint RGB-D generation, depth estimation, and depth-conditioned image synthesis. Modality Forcing (Duisterhof et al., 2026) scales joint post-training on images and depth with sparse depth supervision, separate noise levels, and modality-specific decoders. MVGD (Guizilini et al., 2025) generates novel-view images and depth through raymap conditioning and task embeddings. AMOR builds on the Modality Forcing backbone and scalar depth pathway, adding source observations and target-aligned geometric evidence for panoramic viewpoint changes.

Geometry-aware world models and camera control. Aether (Zhu et al., 2025) unifies geometric reconstruction, action-conditioned video prediction, and visual planning, while FantasyWorld (Dai et al., 2026) couples video generation with a geometric branch. Voyager (Huang et al., 2025) jointly generates RGB-D videos and maintains a world cache with point culling. GEN3C (Ren et al., 2025) conditions video synthesis on point-cloud-cache renderings; WorldMem (Xiao et al., 2025) retrieves pose- and time-associated observations; and Wonder (Xu et al., 2026) combines spatial camera conditioning, sparse memory attention, and distillation. Building on joint generation and memory, AMOR combines surface-guided readout with conservative updates that align depth and admit geometrically compatible regions before they condition later views.

Panoramic world generation. Pano2Room (Pu et al., 2024) combines panoramic inpainting and depth recovery for Gaussian reconstruction. Matrix-3D (Yang et al., 2025) couples trajectory-guided panoramic video generation with 3D reconstruction, while HY-World 2.0 (Team HY-World, 2026) integrates panorama generation, trajectory planning, video-based expansion, and Gaussian reconstruction. Three distinct PanoWorld works explore spherical adaptation with depth and trajectory consistency (Jiang et al., 2026), whole-house synthesis using a floorplan-derived shell and Gaussian cache (Jia et al., 2026), and panoramic camera control with geometry-aware memory augmentation (Li et al., 2026). AMOR uses joint omnidirectional RGB-D synthesis at sparse queried poses to expand a persistent explicit scene.

3 Method

AMOR combines high-quality image synthesis with geometric memory for omnidirectional RGB-D generation. The core idea is to jointly generate appearance and depth using a pretrained image generative model, with geometric memory providing scene evidence to guide each new view. Generated RGB-D observations are then selectively integrated into memory to support subsequent generation. We first formulate the generation task in Sec. 3.1. We then describe joint RGB-D generation with a dedicated depth pathway in Sec. 3.2, followed by surface-guided memory readout and conservative updates in Sec. 3.3. Figure 3 presents an overview.

3.1 Problem Formulation

Given a single RGB panorama I_0, its camera-to-world pose T_0, and a target camera pose T_t, we aim to generate an RGB panorama I_t and a radial-depth map D_t of the same static scene. Panoramas represent omnidirectional views in equirectangular projection (ERP), and radial depth measures

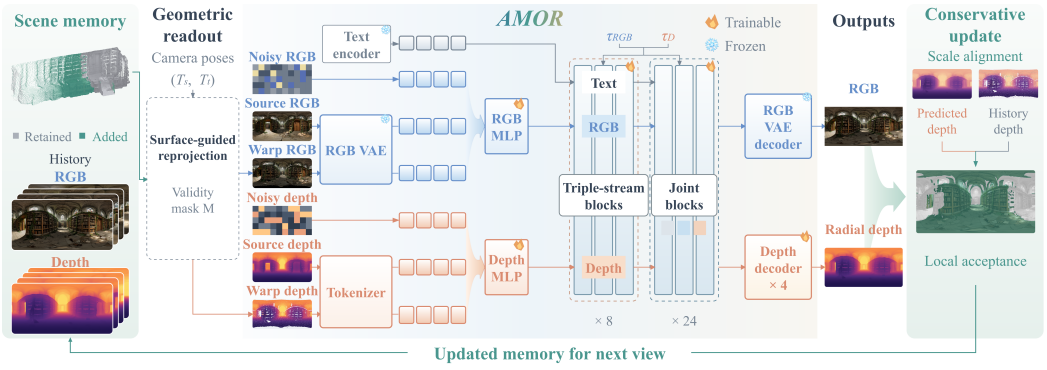


Figure 3: **AMOR overview.** Source observations and surface-guided reprojections condition a shared transformer with separate RGB and scalar-depth pathways. Aligned, geometrically compatible predictions update memory for subsequent queries. Teal denotes newly added points.

the distance from the camera center to the scene surface. As additional views are generated, their appearance and geometry should remain consistent with both the input panorama and previously generated views. When the RGB panorama at the target pose is observed, we instead predict its radial-depth map while keeping the observed RGB fixed. Thus, the model supports both RGB-D next-view synthesis and depth estimation from observed RGB.

3.2 Geometry-Conditioned Joint RGB-D Generation

Dedicated Scalar Depth Pathway. To enable pretrained image generative models to jointly model RGB and depth, some methods, such as Vision Banana (Gabeur et al., 2026), map depth values to RGB colors and reuse the pretrained RGB VAE. Quantizing the color representation can introduce numerical error, while lossy VAE reconstruction can further alter the encoded values. Our reconstruction experiment shows that floating-point color mapping and inversion alone introduce negligible error, whereas adding the RGB VAE increases depth error (Appendix C). AMOR therefore follows Modality Forcing (Duisterhof et al., 2026) in representing depth directly as one scalar value per pixel. We normalize radial depth and compress its numerical range before projecting depth patches into the shared transformer. A dedicated depth decoder maps the resulting tokens back to depth patches, from which we reconstruct normalized radial depth. The predicted depth is subsequently aligned to the existing scene geometry during memory updates (Sec. 3.3). Appendix B details the depth mapping, tokenization, and decoding procedure.

Joint RGB-D Generation. The generator uses a source RGB panorama I_s and its depth map D_s, together with RGB $\widetilde{I}_t$, radial depth $\widetilde{D}_t$, and a validity mask M_t reprojected from memory to the target camera view (Sec. 3.3). The source view supplies global appearance and layout; the reprojected inputs specify previously observed content at target-view pixels. AMOR jointly generates appearance and geometry by organizing target, source, and reprojected observations into six streams,

$$X^{\mathrm{RGB}} = [z_\tau^{\mathrm{RGB}}, z_s^{\mathrm{RGB}}, \widetilde{z}_t^{\mathrm{RGB}}], \tag{1}$$

$$X^{\mathrm{D}} = [z_\tau^{\mathrm{D}}, z_s^{\mathrm{D}}, \widetilde{z}_t^{\mathrm{D}}]. \tag{2}$$

Here z_τ^m denotes the target tokens at diffusion time τ_m, z_s^m the clean source tokens from the selected source view s, and $\widetilde{z}_t^m$ the clean reprojected conditioning tokens aligned to target viewpoint t, where $m \in \{\mathrm{RGB}, \mathrm{D}\}$ denotes the modality. Shared transformer attention couples target RGB and depth while incorporating both source-view context and reprojected RGB and depth, with each modality retaining its dedicated input and output pathways.

Within this representation, source and reprojected streams remain clean, while each unknown target modality follows

$$z_\tau^m = (1 - \tau_m) z_0^m + \tau_m \epsilon^m, \qquad m \in \{\mathrm{RGB}, \mathrm{D}\}, \quad \epsilon^m \sim \mathcal{N}(0, \mathbf{I}). \tag{3}$$

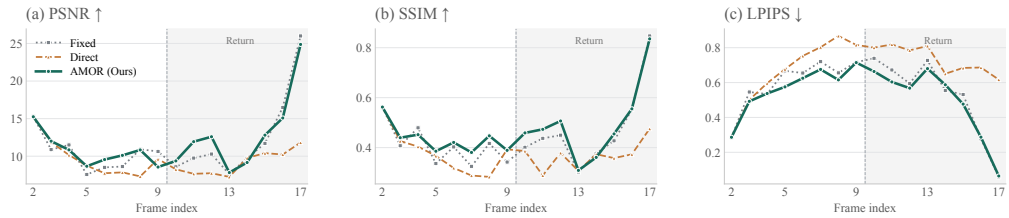


Figure 4: **Memory retention when revisiting previously seen regions.** PSNR, SSIM, and LPIPS of generated panoramas against GT. Green denotes AMOR (Ours), gray denotes Fixed, and orange denotes Direct. Shading marks the return (frames 10 to 17); frame 1 is the observed input and is excluded.

Separate timestep embeddings encode RGB and depth noise levels, and both are supplied to each branch so that joint denoising accounts for their different noise levels. Next-view generation denoises both target modalities; when target RGB is observed, $\tau_{\mathrm{RGB}} = 0$ keeps it fixed and only depth is denoised. Training mixes these two tasks using the same six-stream architecture. The backbone also permits target-only depth inference by omitting source and reprojection conditions; this separate diagnostic setting is used in Sec. 4.3.

Target supervision. Only target tokens are supervised, with the RGB head predicting velocity and the depth head predicting clean depth tokens. The objective combines these reconstruction terms with confidence-weighted log-depth, gradient, normal, and seam supervision,

$$\mathcal{L} = \mathcal{L}_{\mathrm{RGB}} + \lambda_D \mathcal{L}_D + \lambda_c \mathcal{L}_{\mathrm{conf}} + \lambda_g \mathcal{L}_{\mathrm{grad}} + \lambda_n \mathcal{L}_{\mathrm{normal}} + \mathcal{L}_{\mathrm{seam}}. \tag{4}$$

The geometric terms supervise local depth structure, while seam terms supervise appearance and depth near the panorama's wraparound boundary. Appendix B details the losses, task-specific masking, and augmentation of source-depth conditions.

3.3 Geometric Memory Readout and Conservative Update

AMOR maintains a geometric memory $\mathcal{M} = (\mathcal{P}, \mathcal{H})$. The 3D point cloud $\mathcal{P}$ represents accumulated scene geometry, while the observation history $\mathcal{H}$ stores RGB panoramas, their radial-depth maps, and camera poses. Each active point retains a link to its source observation and pixel: its 3D position determines visibility, and its associated observation supplies appearance and local surface geometry. We initialize memory from the input RGB panorama and its estimated depth.

Surface-Guided Memory Readout. Given a target camera pose T_t, readout produces a partial RGB panorama $\widetilde{I}_t$, a radial-depth map $\widetilde{D}_t$, and a validity mask M_t in the target view:

$$(\widetilde{I}_t, \widetilde{D}_t, M_t) = \mathcal{R}_{\mathrm{surf}}(\mathcal{M}, T_t). \tag{5}$$

The mask identifies pixels with valid reprojected RGB; the remaining pixels require generation. Each active point defines a finite surface patch whose tangent directions are estimated from depth-consistent neighbors of its source pixel in a stored RGB-D panorama. We use the four immediate neighbors, with the patch extending half a source-pixel step along each tangent. Target rays intersect these patches, and the nearest intersection determines visibility and radial depth. Appearance is retrieved from the associated observation only when surface-orientation and depth-continuity checks pass. This prevents foreground boundaries from exposing background colors. RGB is reprojected before VAE encoding, giving the generator condition $\mathcal{C}_t = (E_{\mathrm{RGB}}(\widetilde{I}_t), \widetilde{D}_t, M_t)$: encoded RGB, radial depth, and the validity mask. Figure 7(c) compares surface-guided and point-projected RGB conditions before encoding.

Conservative Memory Update. After generating RGB $\hat{I}_t$ and depth $\hat{D}_t$, AMOR aligns the predicted depth to the existing scene and checks geometric compatibility before inserting new points. Low-coverage regions identify where additional points are needed, while compatibility checks determine whether these candidates can be incorporated. Predicted depth is first aligned to supported

Table 1: **Panoramic NVS results by dataset.** Each group reports PSNR (dB), SSIM, LPIPS, and FID. Orange cells with bold text and blue cells with underlined text indicate the best and second-best generator values, respectively.

Method	Holo360D				Replica				TartanGround				Avg.			
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Pano2Room	15.42	0.5205	0.5011	140.59	17.15	0.6707	0.3969	124.71	13.24	0.3880	0.5577	183.67	15.27	0.5264	0.4852	149.66
HY-World 2.0	17.15	**0.5809**	0.4198	99.52	18.38	0.7832	0.2702	69.20	**17.44**	**0.4960**	0.3864	95.18	17.66	0.6200	0.3588	87.97
Matrix-3D	16.16	0.5722	0.4803	105.97	**21.05**	**0.8247**	0.2744	68.75	15.35	0.4419	0.4708	139.17	17.52	0.6130	0.4085	104.63
PanoWorld	14.41	0.4917	0.5395	107.59	15.20	0.7199	0.4492	75.19	12.55	0.3391	0.5315	155.22	14.06	0.5169	0.5067	112.67
AMOR (Ours)	**17.69**	0.5807	**0.3452**	**72.97**	20.10	0.8094	**0.2375**	**49.07**	16.99	0.4935	**0.2855**	**76.45**	**18.26**	**0.6279**	**0.2894**	**66.17**

historical geometry, yielding $D_t^\star = a_t \hat{D}_t + \delta_t$, where a_t aligns scale and δ_t is a local depth correction based on compatible boundary geometry. Updates without sufficient compatible support are rejected. For accepted pixels Ω_t, the active point set becomes

$$\mathcal{P}_t = (\mathcal{P}_{t-1} \setminus \mathcal{R}_t) \cup \mathcal{B}(\hat{I}_t, D_t^\star, T_t; \Omega_t), \tag{6}$$

where $\mathcal{B}$ back-projects radial depth and $\mathcal{R}_t$ contains visible old points retired within the accepted region. The RGB panorama, corrected depth map, and camera pose associated with an accepted update are stored in $\mathcal{H}$, while only pixels in Ω_t add points to $\mathcal{P}_t$. Later readout uses surviving active points and their links to these stored observations; retired points no longer determine visibility. To examine whether conservative memory updates preserve previously observed content while allowing scene expansion, we conduct a simple experiment in which the camera first moves outward and then retraces its trajectory through previously generated views. Compared with fixed memory and direct accumulation, our method achieves better average appearance fidelity on the return views, suggesting that selective integration helps limit the propagation of inconsistent predictions into subsequent generation. Figure 4 shows the metric curves, and Sec. 4.4 reports the quantitative results. Appendix D details surface construction, candidate selection, alignment, and reference management.

4 Experiments

4.1 Experimental Setup

Datasets and evaluation protocol. We evaluate AMOR's omnidirectional novel-view synthesis capability on Holo360D (Ou et al., 2026), Replica (Straub et al., 2019), and TartanGround (Patel et al., 2025). We evaluate 10 Holo360D scenes, all 18 Replica scenes, and four TartanGround scenes. Given a source panorama and a prescribed 81-frame camera trajectory, we evaluate target views at frames 20, 40, 60, and 80 at 1440×720 resolution. To evaluate depth accuracy against reliable ground truth, we conduct our depth experiments on TartanGround, a synthetic dataset built using Unreal Engine that provides renderer-derived ground-truth depth. We additionally examine multi-view point clouds and meshes reconstructed from generated RGB-D observations. Dataset composition and split details are provided in Appendix A.

Baselines and metrics. For omnidirectional novel-view synthesis, we compare against Pano2Room (Pu et al., 2024), HY-World 2.0 (Team HY-World, 2026), Matrix-3D (Yang et al., 2025), and the Real-World Panoramic Generation variant of PanoWorld (Li et al., 2026) using PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and FID (Heusel et al., 2017). PSNR, SSIM, and LPIPS are computed on complete ERP images without masking or latitude weighting. FID is computed from fixed perspective views shared by ground truth and all methods. We average image-level metrics within each dataset and compute FID from each dataset's view collection. Avg. denotes the arithmetic mean of the three dataset-level scores. For depth estimation, we compare against DepthMaster (Wang et al., 2026a), DAP (Lin et al., 2026), and DA360 (Jiang et al., 2025) using frame-averaged AbsRel, RMSE, and δ_1. Baseline-specific calibration details and metric definitions are provided in Appendix E, while the depth alignment and scoring protocol is described in Sec. 4.3.

Implementation details. We fine-tune the Modality Forcing backbone with source and reprojection conditioning, freezing the RGB VAE and text encoder. Training starts with joint RGB-D

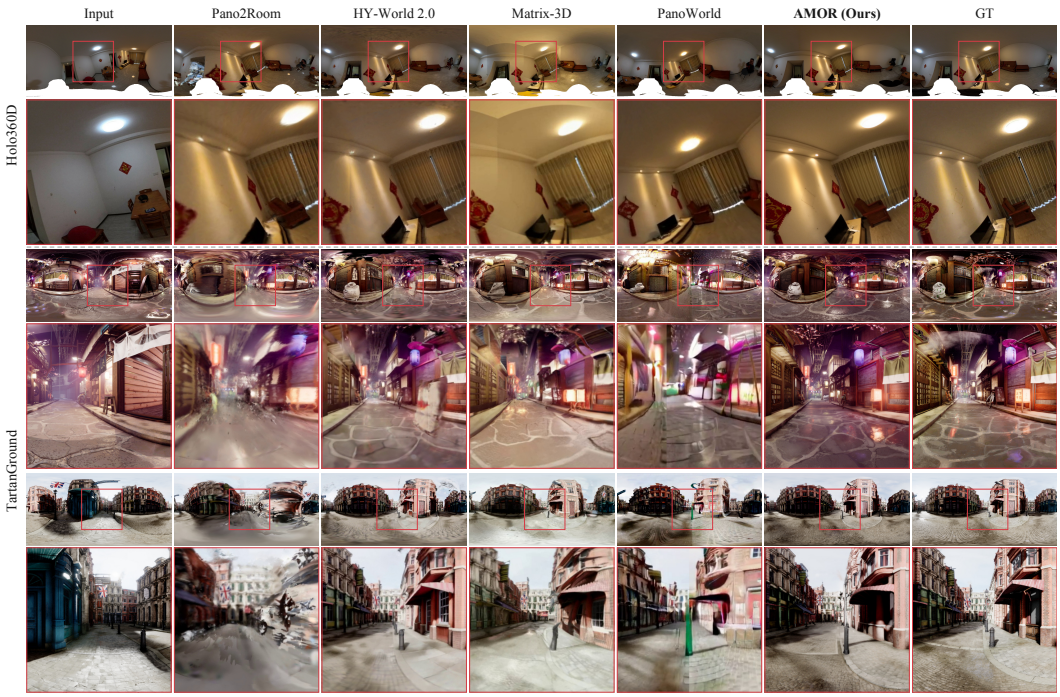


Figure 5: **Qualitative comparison of panoramic novel-view synthesis.** Columns show the input, four baselines, AMOR, and GT. Detail crops are shown below the panoramas; predictions and GT share the same target-view crop, while the input shows source-view context. Invalid Holo360D regions are shown in white.

generation before mixing it with observed-RGB depth prediction at a 7:3 ratio. We use 64 GPUs for seven days and 50 denoising steps at inference. For NVS, we initialize memory using MoGe-2 (Wang et al., 2026b) depth estimated from the source RGB panorama and calibrate its scale using source GT depth. Optimization and source-depth augmentation are detailed in Appendix B.

4.2 Panoramic World Generation

Novel-view synthesis. Table 1 compares novel-view synthesis quality across the three benchmarks. AMOR achieves the lowest LPIPS and FID on each dataset and the best average scores across datasets on all four metrics. PSNR and SSIM rankings vary across individual datasets. Figure 5 shows qualitative comparisons at matched target views with corresponding detail crops.

Depth accuracy of generated views. We evaluate depth at the prescribed NVS targets on TartanGround, using each method's own generated RGB. Table 2 compares the depth produced by AMOR and three baseline pipelines at generated viewpoints. AMOR achieves the best scores across all three metrics on their common valid regions (Appendix E), indicating more accurate generated geometry. Together with the RGB results, these findings demonstrate AMOR's ability to generate both high-quality appearance and accurate depth at novel viewpoints.

Scene reconstruction from generated RGB-D. We compare point clouds and meshes from HY-World 2.0 (Team HY-World, 2026), Matrix-3D (Yang et al., 2025), and AMOR on Holo360D and TartanGround under matched source observations and camera trajectories, using a shared visualization viewpoint and mask (Appendix E). Baseline meshes use TSDF fusion (Curless & Levoy, 1996), while AMOR stitches boundaries across its active geometric memory. Figure 7(a) shows clearer desk and shelf boundaries in AMOR's point clouds. Its meshes retain texture detail but exhibit small cracks and holes, whereas TSDF produces smoother surfaces with some blurred detail. These comparisons reflect the complete generation and reconstruction pipelines rather than isolated fusion quality.

Table 2: **Depth accuracy of generated views on TartanGround.**

Method	AbsRel ↓	RMSE (m) ↓	δ_1 ↑
Pano2Room	0.2873	6.151	47.86%
Matrix-3D	0.1276	4.144	83.34%
HY-World 2.0	0.0769	3.542	94.39%
AMOR (Ours)	**0.0710**	**3.324**	**95.14%**

Table 3: **Depth estimation with observed RGB on TartanGround.**

Method	AbsRel ↓	RMSE (m) ↓	δ_1 ↑
DepthMaster	**0.0536**	2.187	96.32%
DAP	0.1088	4.203	93.54%
DA360	0.0655	5.424	97.60%
AMOR (Ours)	0.0619	**2.026**	**97.71%**

Table 4: **Ablation of RGB readout for novel-view synthesis.**

RGB readout	Holo360D			TartanGround			Overall		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Point projection	16.40	0.5350	0.3419	11.80	0.3522	0.4039	14.36	0.4537	0.3695
Surface-guided (Ours)	**17.33**	**0.5592**	**0.3320**	**16.91**	**0.4696**	**0.2889**	**17.14**	**0.5194**	**0.3129**

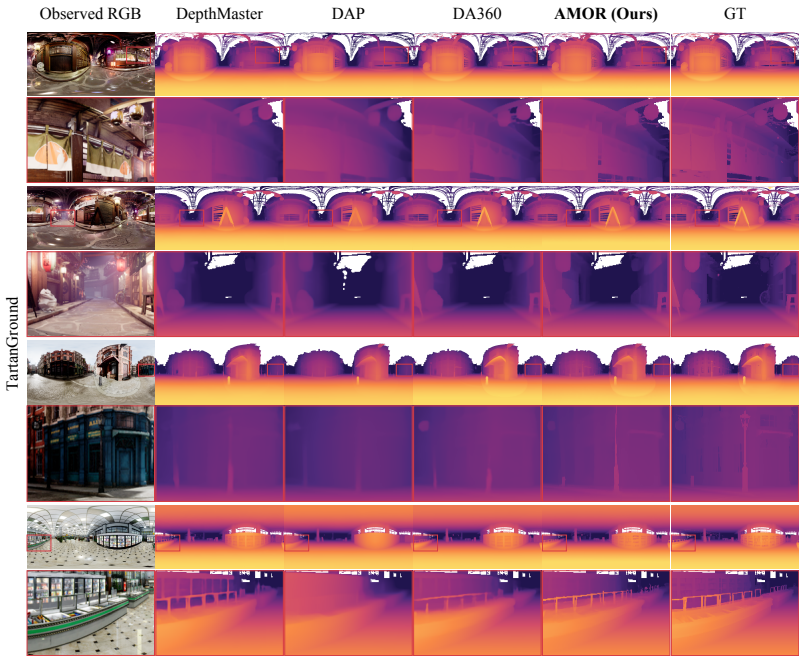


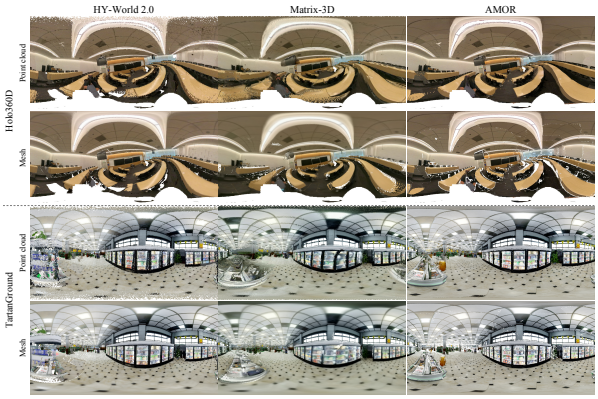
Figure 6: **Qualitative depth comparisons on TartanGround.** Each example shows observed RGB, three baselines, AMOR, and GT, with matched red-box regions enlarged below. Depth is displayed on a logarithmic scale with a shared GT-derived range within each example; white denotes invalid depth.

4.3 Depth Estimation from Observed RGB

As an additional capability, AMOR predicts depth conditioned on observed RGB. All methods use observed RGB and share camera poses and alignment, admission, and fusion procedures, while maintaining independent maps. AMOR uses target RGB alone, with source and reprojection streams disabled. Only the first frame of each sequence uses GT scale calibration; later predictions align to that method's historical geometry. Historical depth supplies a post-prediction scale reference, not model conditioning. We evaluate aligned depth before local depth correction, including predictions rejected from geometric memory. Validity rules and metric definitions are provided in Appendix E.

AMOR achieves the lowest RMSE and highest δ_1 on TartanGround, while DepthMaster has lower AbsRel (Table 3). Figure 6 shows lower AMOR AbsRel in selected structural regions, without implying uniform superiority across frames.

(a) Point clouds and meshes



(b) Memory-update ablation

Metric	Direct	Ours
Map accuracy error (m)↓	0.672	**0.533**
Initial-view SSIM↑	0.693	**0.873**
RGB PSNR (dB)↑	16.570	**16.614**
RGB SSIM↑	0.4670	**0.4713**
RGB LPIPS↓	0.3291	**0.3257**

(c) RGB conditions

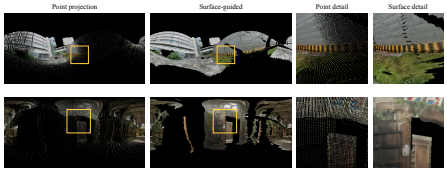


Figure 7: **Scene geometry and memory analysis.** (a) Point clouds and meshes at matched viewpoints; white denotes missing or masked geometry. (b) Memory-update ablation: map metrics average nine scenes, and generated-view metrics average 36 views. Initial-view SSIM measures retained appearance at the input pose. (c) Point-projected and surface-guided RGB conditions before generation. Yellow boxes mark matched crops; black denotes missing evidence.

4.4 Ablation Studies

Conservative memory updates. We compare conservative memory updates with direct accumulation of aligned predictions on scenes from Holo360D and TartanGround under matched settings (Appendix F). Each policy generates subsequent views from its own memory. Figure 7(b) summarizes the comparison with Direct: AMOR reduces final-map accuracy error by 20.7% and raises initial-view rendering SSIM by 0.180. Accuracy and initial-view retention improve in all nine scenes. The precision benefit persists when identical raw predictions are replayed with policy-specific alignment (Appendix F), supporting the integration policy beyond differences in generated inputs. Its main benefit is a more accurate accumulated map with better retained appearance; improvements in generated RGB are modest. Direct retains higher F-score, reflecting a precision-completeness tradeoff.

Revisiting previously seen regions. On the Hospital return trajectory, AMOR improves all three mean RGB metrics over Fixed and Direct (Fig. 4), reaching 12.96 dB PSNR versus 12.52 and 9.14 dB, respectively. This single-scene result supports appearance retention during revisitation; per-frame gains vary. Full numerical results are reported in Appendix F.

Surface-guided RGB conditions. We compare point-projected and surface-guided RGB conditions using the same model architecture on 36 Holo360D and TartanGround views. Surface-guided readout improves all three metrics on both datasets, gaining 2.78 dB PSNR and reducing LPIPS by 0.0566 overall (Table 4). These gains highlight the importance of how stored appearance is supplied to the generator. Paired visualizations show more continuous RGB evidence, consistent with surface-guided readout making existing scene content more useful for synthesis. This complements conservative updates, which primarily improve the accumulated map. The single-seed protocol and visualizations are detailed in Appendix F.3.

5 Conclusion

We presented AMOR, a memory-guided omnidirectional RGB-D diffusion model for high-fidelity, geometry-consistent next-view synthesis. Joint appearance-depth generation, surface-guided memory readout, and conservative updates support progressive expansion of a persistent 3D scene. Experiments demonstrate strong novel-view synthesis and observed-RGB depth estimation, with conservative updates improving geometric accuracy and appearance retention at a completeness cost. Consistency over longer exploration sequences remains to be evaluated.

AI USE STATEMENT

Generative AI tools were used to assist with method implementation and code development. ChatGPT/Codex also assisted with literature retrieval and discovery of related work, as well as manuscript drafting and polishing, including the written presentation of existing methods and experiments, interpretation of author-provided experimental results, preparation of comparison figures, and LaTeX editing. The authors are responsible for the accuracy, originality, and reproducibility of the final manuscript, including all AI-assisted text, interpretations, and artifacts.

ETHICS STATEMENT

AMOR synthesizes unobserved appearance and geometry; generated regions should not be interpreted as measurements of a real environment. Plausible outputs can contain structural errors or biases inherited from training data and pretrained models. Deployment for navigation, mapping, or other consequential decisions therefore requires independent geometric validation. Panoramic observations may also contain personal or sensitive environmental information, which should be removed or protected when preparing data and demonstrations. Use and distribution of data, pretrained weights, and generated artifacts should respect applicable licenses and consent requirements.

REPRODUCIBILITY STATEMENT

Section 3 describes the method. Section 4.1 specifies inputs, scale calibration, masks, and aggregation, while Sec. 4.3 and Appendix E specify the observed-RGB depth protocol and scoring. Appendix E records baseline settings and the point-cloud and mesh comparison, while Appendix F details the completed sequential memory ablations, fixed-sequence replay, and condition visualization provenance. Appendix B records training data, optimization settings, and loss definitions.

REFERENCES

Black Forest Labs. FLUX.2: Frontier visual intelligence, 2025. URL https://bfl.ai/blog/flux-2.

Kwon Byung-Ki, Qi Dai, Lee Hyoseok, Chong Luo, and Tae-Hyun Oh. Jointdit: Enhancing rgb-depth joint modeling with diffusion transformers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25261–25271. IEEE, 2025.

Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of SIGGRAPH*, 1996. URL https://graphics.stanford.edu/papers/volrange/.

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.

Bardienus Pieter Duisterhof, Deva Ramanan, Jeffrey Ichnowski, Justin Johnson, and Keunhong Park. Modality forcing for scalable spatial generation. *arXiv preprint arXiv:2606.13676*, 2026.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, et al. Image generators are generalist vision learners. *arXiv preprint arXiv:2604.20329*, 2026. URL https://arxiv.org/abs/2604.20329.

Vitor Guizilini, Muhammad Zubair Irshad, Dian Chen, Greg Shakhnarovich, and Rares Ambrus. Zero-shot novel view and depth synthesis with multi-view geometric diffusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 764–776. IEEE, 2025.

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*, 2017. URL https://arxiv.org/abs/1706.08500.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson W. H. Lau, Wangmeng Zuo, and Chunchao Guo. Voyager: Long-range and world-consistent video diffusion for explorable 3D scene generation. *ACM Transactions on Graphics*, 2025. doi: 10.1145/3763330. URL https://doi.org/10.1145/3763330.

Jinrang Jia, Zhenjia Li, Yijiang Hu, and Yifeng Shi. PanoWorld: A generative spatial world model for consistent whole-house panorama synthesis. *arXiv preprint arXiv:2605.17916*, 2026. URL https://arxiv.org/abs/2605.17916.

Hualie Jiang, Ziyang Song, Zhiqiang Lou, Rui Xu, and Minglang Tan. Depth anything in 360°: Towards scale invariance in the wild. *arXiv preprint arXiv:2512.22819*, 2025. URL https://arxiv.org/abs/2512.22819.

Le Jiang, Xiangyu Bai, Bishoy Galoaa, Shayda Moezzi, Caleb James Lee, Tooba Imtiaz, Edmund Yeh, Jennifer Dy, Yanzhi Wang, and Sarah Ostadabbas. PanoWorld: Geometry-consistent panoramic video world modeling. *arXiv preprint arXiv:2605.15391*, 2026. URL https://arxiv.org/abs/2605.15391.

Haoyuan Li, Dizhe Zhang, Yuemei Zhou, Xiangkai Zhang, Haoran Feng, Xiaofan Lin, Wenjie Jiang, Bo Du, Ming-Hsuan Yang, and Lu Qi. PanoWorld: Real-world panoramic generation. *arXiv preprint arXiv:2607.09661*, 2026. URL https://arxiv.org/abs/2607.09661.

Xin Lin, Meixi Song, Dizhe Zhang, Wenxuan Lu, Haodong Li, Bo Du, Ming-Hsuan Yang, Truong Nguyen, and Lu Qi. Depth any panoramas: A foundation model for panoramic depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26835–26844, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Jing Ou, Zidong Cao, Yinrui Ren, Zhuoxiao Li, Jinjing Zhu, Tongyan Hua, Shuai Zhang, Hui Xiong, and Wufan Zhao. Holo360D: A large-scale real-world dataset with continuous trajectories for advancing panoramic 3D reconstruction and beyond. *arXiv preprint arXiv:2604.22482*, 2026. URL https://arxiv.org/abs/2604.22482.

Manthan Patel, Fan Yang, Yuheng Qiu, Cesar Cadena, Sebastian Scherer, Marco Hutter, and Wenshan Wang. Tartanground: A large-scale dataset for ground robot perception and navigation. *arXiv preprint arXiv:2505.10696*, 2025. URL https://arxiv.org/abs/2505.10696.

Guo Pu, Yiming Zhao, and Zhouhui Lian. Pano2Room: Novel view synthesis from a single indoor panorama. In *SIGGRAPH Asia 2024 Conference Papers*, 2024. doi: 10.1145/3680528.3687616. URL https://arxiv.org/abs/2408.11413.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: international conference for high performance computing, networking, storage and analysis*, pp. 1–16. IEEE, 2020.

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)*, 2025. URL https://arxiv.org/abs/2503.03751.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. URL https://arxiv.org/abs/1904.01201.

SpatialVerse Research Team, Manycore Tech Inc. InteriorGS: A 3D gaussian splatting dataset of semantically labeled indoor scenes, 2025. URL https://huggingface.co/datasets/spatialverse/InteriorGS.

Julian Straub, Thomas Whelan, Lingni Ma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. URL https://arxiv.org/abs/1906.05797.

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.

Pengfei Wang, Shihao Wang, Liyi Chen, Zhiyuan Ma, Guowen Zhang, and Lei Zhang. Depthmaster: Unified monocular depth estimation for perspective and panoramic images. *arXiv preprint arXiv:2606.12368*, 2026a. URL https://arxiv.org/abs/2606.12368.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *Advances in Neural Information Processing Systems*, 38:35928–35959, 2026b.

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.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025. URL https://arxiv.org/abs/2504.12369.

Jiacong Xu, Hanwen Jiang, Zhixin Shu, Kalyan Sunkavalli, Vishal M. Patel, and Yiqun Mei. Wonder: Video world model done better. *arXiv preprint arXiv:2607.26037*, 2026. URL https://arxiv.org/abs/2607.26037.

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. URL https://arxiv.org/abs/2508.08086.

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)*, 2025. URL https://arxiv.org/abs/2406.09394.

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*, 2018. URL https://arxiv.org/abs/1801.03924.

Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:1801.09847*, 2018. URL https://arxiv.org/abs/1801.09847.

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 Datasets and Evaluation Splits

Panoramic novel-view synthesis. Dataset names in the main text refer to evaluation subsets rather than the complete datasets. The evaluation comprises 32 scenes and 128 target frames. Holo360D contributes 10 scenes and 40 frames, Replica 18 scenes and 72 frames, and TartanGround 4 scenes and 16 frames. All 18 Replica scenes are included. Each scene contributes frames 20, 40, 60, and 80 from an 81-frame trajectory.

Observed-RGB depth subset. The depth evaluation uses 230 frames from four TartanGround sequences: Hospital_P0000 (33), Supermarket_P0000 (40), JapaneseAlley_P0000 (81), and VictorianStreet_P0000 (76). Scoring protocols are given in Appendix E.

Split provenance. Hospital_P0000, Supermarket_P0000, JapaneseAlley_P0000, and VictorianStreet_P0000 are excluded from training before pair sampling. Replica is not used for training; Holo360D training uses its training split.

B Training Configuration and Objectives

Depth mapping, tokenization, and decoding. For a radial-depth map d, we compute its mean μ over valid pixels and normalize it as $r = d/\mu$. We then compress its numerical range through

$$u = c(r) = \begin{cases} r, & r \le 1, \\ 2 - 1/r, & r > 1, \end{cases} \tag{7}$$

which maps positive normalized depth values to $(0, 2)$ while retaining an explicit inverse. The resulting map is divided into 16×16 patches. Each patch is flattened and mapped to a depth token by a learned input projection, allowing depth and RGB tokens to interact within the shared transformer without encoding depth through the RGB VAE.

On the output side, decoder blocks dedicated to the depth branch process the depth tokens produced by the shared transformer. A depth output head maps each resulting token to the values of a 16×16 patch. At the end of diffusion sampling, the generated patches are rearranged into a dense map $\hat{u}$, and the inverse mapping

$$\hat{r} = c^{-1}(\hat{u}) = \begin{cases} \hat{u}, & \hat{u} \le 1, \\ 1/(2 - \hat{u}), & \hat{u} > 1 \end{cases} \tag{8}$$

recovers normalized radial depth. This inverse mapping reverses the range compression but does not restore the absolute scale removed by mean normalization. We therefore align the predicted depth to the existing scene geometry during memory updates (Sec. 3.3).

Training data and conditioning. We train on posed RGB-D pairs from the training splits of Holo360D (Ou et al., 2026), InteriorGS (SpatialVerse Research Team, Manycore Tech Inc., 2025), and TartanGround (Patel et al., 2025). Pairs come from the same scene, with camera translation between 0.2 and 5 m, source-warp coverage between 0.2 and 0.8, and a warp-to-target RGB discrepancy threshold of $40/255$. Dataset-specific depth validity filters remove unreliable supervision. Training/evaluation separation is specified in Appendix A.

The target and reprojected conditions have resolution 1440×720, while complete source RGB-D is resized to 736×368. Surface-reprojected RGB passes through the frozen RGB VAE, and radial depth uses the scalar pathway. With probability 0.5, cached MoGe-2 source depth (Wang et al., 2026b) replaces source GT depth after robust median log-scale alignment; failed alignment falls back to GT. Reprojection is recomputed from the selected source depth, while target depth supervision remains unchanged. In the mixed-task stage, training samples joint RGB-D generation with probability 0.7 and geometry-conditioned observed-RGB depth prediction with probability 0.3. Source and reprojected tokens remain clean in both modes; target-only depth inference is supported by the architecture but is not sampled in this full-parameter training stage.

Optimization. All diffusion-transformer parameters are trainable, including the dedicated depth pathway and confidence head; the RGB VAE and text encoder remain frozen. We use PyTorch AdamW (Loshchilov & Hutter, 2019) with a base learning rate of 10^{-5} and weight decay 0.01; other optimizer hyperparameters retain their PyTorch defaults. The learning rate is constant from the first optimizer update, with no learning-rate warmup or decay. The depth decoder blocks, decoder modulation, and final depth projection use a learning rate of 10^{-6} (0.1 times the base rate), while the remaining trainable parameters, including the confidence head, use 10^{-5}. Trainable parameters and optimizer state use FP32, with BF16 autocast for eligible forward operations, ZeRO-2 optimizer sharding (Rajbhandari et al., 2020), and activation recomputation. The RGB VAE uses horizontal circular padding. We use seed 0 for NVS sampling.

Target reconstruction and confidence. Target tokens follow the modality-specific noise schedule in Eq. equation 3; RGB and depth noise levels are sampled independently for joint generation.

Table 5: **Loss weights for full-parameter training.** The three seam terms together form $\mathcal{L}_{\text{seam}}$ in the main objective.

Loss term	Weight
RGB velocity MSE	1.0
Contracted-depth token MSE	1.0
Confidence-weighted relative log-depth	0.1
Periodic relative log-depth gradient	0.1
Surface-normal angular error	0.02
Decoded RGB seam reconstruction	0.1
Decoded RGB seam gradient	0.05
Relative log-depth seam gradient	0.05

Setting $\tau_{\text{RGB}} = 0$ keeps target RGB observed for geometry-conditioned depth prediction. The RGB head predicts velocity, whereas the depth head predicts clean contracted-depth tokens:

$$\mathcal{L}_{\text{RGB}} = \mathbb{E}\Big[\big\|v_\theta^{\text{RGB}} - (\epsilon^{\text{RGB}} - z_0^{\text{RGB}})\big\|^2_{M_{\text{RGB}}}\Big], \tag{9}$$

$$\mathcal{L}_D = \mathbb{E}\Big[\big\|\hat{z}_0^{\text{D}} - z_0^{\text{D}}\big\|^2_{M_{\text{D}}}\Big]. \tag{10}$$

Masked norms average over valid target elements. Source and reprojected blocks are conditions rather than reconstruction targets. RGB velocity and RGB seam losses are disabled when target RGB is observed, while depth losses remain active on valid target depth.

Let ℓ and $\hat{\ell}$ denote reference and predicted relative log-depth. For valid pixels V, the confidence head predicts $c(u) \in [1, 100]$, and

$$\mathcal{L}_{\text{conf}} = \frac{1}{|V|}\sum_{u\in V}\Big[c(u)|\hat{\ell}(u) - \ell(u)| - \log c(u)\Big]. \tag{11}$$

The gradient loss averages horizontal and vertical absolute differences between predicted and reference log-depth gradients, with periodic horizontal neighbors and nonperiodic vertical neighbors. Confidence weights regulate depth fitting; they are distinct from geometric coverage and memory-admission support labels.

Normal and seam supervision. For surface normals, relative radial depth is back-projected along ERP rays. Four incident triangle normals provide local orientation estimates, and the loss averages squared angular errors using cosine-latitude area weights. Validity is determined from GT depth and normal stencils, excluding vertical image borders, near-pole pixels with $\cos\phi < 0.1$, and neighboring GT-depth ratios exceeding 1.1. Horizontal stencils wrap across the seam. Predicted confidence and memory-admission masks do not gate this loss.

Seam supervision uses the leftmost and rightmost 64 columns. It combines an RGB reconstruction L1 loss with horizontal-gradient L1 losses for decoded RGB and relative log-depth, including the last-to-first-column pair. These terms match GT values and gradients rather than forcing the two boundary columns to be equal. The frozen RGB decoder remains differentiable so that RGB seam losses update the transformer. Table 5 gives all nonzero loss weights; the normal weight ramps linearly from zero to its listed value over the first 500 optimizer steps. Training uses paired views and target-space supervision, without sequential memory rollouts or a world-space rendering loss.

C Depth Representation Diagnostic

We evaluate representation fidelity on five 1920×3840 panoramic depth maps, treating the existing predicted depths as numerical references. For each color representation, we compare direct floating-point mapping and inversion with a path that inserts the same frozen FLUX.2 RGB VAE (Black Forest Labs, 2025) before inversion. Table 6 shows that both direct mappings preserve depth to numerical precision, while RGB VAE reconstruction introduces additional depth error. These results motivate the dedicated scalar pathway in Sec. 3.2; they do not measure trained AMOR depth-prediction accuracy.

Table 6: **RGB VAE reconstruction introduces depth error for both tested color representations.** Mapping-only rows encode and invert floating-point colors directly; VAE rows insert the frozen FLUX.2 RGB VAE between the same mapping and inverse. Errors are measured against the input predicted depth maps, with spherical weighting and an average over five panoramas. Orange/bold and blue/underlined cells mark the better and worse RGB VAE representations, respectively; mapping-only controls are unranked.

Depth representation	Processing path	AbsRel ↓	MAE (mm) ↓
Color curve	Mapping only	≈ 0	≈ 0
Color curve	Mapping + RGB VAE	**0.010609**	**20.863**
Log-depth grayscale	Mapping only	≈ 0	≈ 0
Log-depth grayscale	Mapping + RGB VAE	<u>0.040781</u>	<u>52.536</u>

Mapping-only errors are numerically negligible. Primary metrics cap reconstructed depth at 1000 m and retain non-finite predictions as penalties.

D Geometric Memory Implementation Details

Resolving visible evidence. We construct finite tangent patches using the four immediate horizontal and vertical neighbors of each source pixel. A neighbor is depth-consistent when $|d_i - d_j| \leq 0.02\,\mathrm{m} + 0.05 \min(d_i, d_j)$. Horizontal neighborhoods wrap across the ERP seam; vertical neighborhoods do not wrap. For each axis, we average forward and backward 3D differences when both neighbors are consistent, and otherwise use the available one-sided difference. These tangent vectors define a parallelogram extending half a source-pixel step in each direction around the active point. Both axes require valid support; appearance reuse additionally requires all four neighbors to pass the continuity check. We intersect target rays with these finite patches. The nearest intersected surface determines visibility and radial depth; surface-orientation and depth-discontinuity checks determine whether it can supply appearance. In particular, a boundary surface can block farther content without providing a valid source RGB pixel. This distinction prevents an unreliable foreground match from exposing background colors through an occluder.

Each active point retains its source-frame and pixel indices. At inference, its current position determines the patch center, while cached source geometry provides the local tangent estimate. Visibility is resolved jointly across surviving points from accepted observations; retired points no longer participate. The resulting pixel-level assignments link visible target surfaces to their original observations and provide a common basis for target depth, RGB reprojection, and validity.

Reprojecting RGB before encoding. Using the visibility assignments above, we render a partial RGB panorama together with radial depth and a validity mask at the target pose as in equation 5, where $\mathcal{R}_{\mathrm{surf}}$ denotes surface-guided rendering from the active geometry and its associated observations. Target pixels without a valid source RGB pixel are filled with zero. We then apply the shared frozen RGB VAE to the complete source image and the rendered target-view image, obtaining $Z_s = E_{\mathrm{RGB}}(I_s)$ and $\widetilde{Z}_t = E_{\mathrm{RGB}}(\widetilde{I}_t)$. The target-aligned condition is

$$\mathcal{C}_t = (\widetilde{Z}_t, \widetilde{D}_t, M_t). \tag{12}$$

The RGB features use the backbone's standard token packing and normalization. Projected radial depth is encoded together with M_t using the native depth representation. Pairwise training obtains surfaces and colors from one source observation; progressive inference applies the same rendering-and-encoding procedure to the accumulated memory.

Completion candidates. We project active point centers into the target view and measure local occupancy over valid predicted depth as

$$\rho_t(u) = \frac{\sum_{v \in \mathcal{N}(u)} M_t^{\mathrm{point}}(v) V_t(v)}{\sum_{v \in \mathcal{N}(u)} V_t(v)}, \tag{13}$$

where M_t^{point} denotes point occupancy, V_t denotes valid generated depth, and an empty denominator gives zero density. Completion candidates are sufficiently large connected components of $\{u :$

$V_t(u) = 1, \rho_t(u) < \tau_\rho\}$, using periodic horizontal connectivity. Point occupancy measures the need for additional samples; the surface query above provides geometric support. A region can therefore have usable surface evidence while still needing denser point coverage.

Alignment before admission. Candidate depth must agree in scale with supported historical geometry. We select dense reference overlap away from depth boundaries and reject log-depth-ratio outliers using the median absolute deviation. Retained anchors A determine

$$a_t = \exp\left[\text{median}_{u\in A} \log \frac{D_{\text{ref}}(u)}{\hat{D}_t(u)}\right], \qquad D_t^\star = a_t \hat{D}_t + \delta_t. \tag{14}$$

Insufficient anchors or an inadmissible scale reject the update. The correction δ_t uses compatible retained geometry along insertion boundaries and decays over a bounded distance through same-surface neighbors, without crossing depth discontinuities. Missing or inconsistent boundary evidence is recorded as seam uncertainty.

Controlling future reference quality. For accepted insertion pixels Ω_t, we update according to equation 6, where $\mathcal{B}$ back-projects radial depth and $\mathcal{R}_t$ contains only visible old points retired inside the accepted region. Other geometry is retained. New points preserve observation provenance and are labeled supported or provisional according to reference consistency. Provisional or seam-uncertain points cannot automatically become scale anchors. A new point can inherit scale support only from a compatible visible reference point; we limit the number of successive inheritance steps, after which the point cannot serve as a scale reference. Admission thus permits local completion without giving every inserted prediction authority over later scales. These labels express internal geometric support, not GT correctness. Accepted observations enter the observation history $\mathcal{H}$ for subsequent memory readout.

E Evaluation Metrics and Protocols

Baseline settings. Pano2Room, HY-World 2.0, and Matrix-3D use source GT depth to calibrate trajectory scale, without receiving it as dense conditioning; PanoWorld assumes one world unit per metre. Pano2Room uses its final Gaussian-rendered RGB output with scene-specific SDFT disabled. HY-World uses the benchmark trajectory and generates 832×480 perspective views before ERP rendering. Matrix-3D RGB is taken directly from its 1920×960 panoramic video, with trajectory rescaling compensated by the adapter. PanoWorld uses the Real-World Panoramic Generation Action model at 960×480; fixed-heading outputs are rotated to the target orientations by spherical resampling. These comparisons retain differences in native resolution and output stage.

Point-cloud and mesh comparison. Figure 7(a) compares clouds before Gaussian fitting in a common metric frame, using source-scale calibration without target-GT fitting. HY-World retains its native foreground observations, Matrix-3D uses optimized panoramic depth, and AMOR uses its active geometric memory; observation and point counts are not equalized. Baseline meshes use Open3D TSDF fusion (Zhou et al., 2018; Curless & Levoy, 1996), whereas AMOR triangulates source-grid patches and stitches compatible cross-frame boundaries. All results share panoramic viewpoints and display masks, without smoothing or hole filling; display masks do not affect benchmark scores. The comparison evaluates complete pipelines rather than an isolated meshing or fusion algorithm.

Depth visualizations. Figure 6 uses a shared logarithmic color range within each example, set by the 2nd and 98th percentiles of valid GT depth; invalid pixels are white, and DA360 follows the scoring cap below. Frames and crops are selected for favorable AMOR RMSE and local AbsRel under common valid support, rather than representing average performance.

Full-image RGB metrics. PSNR, SSIM, and LPIPS are computed on complete ERP images without validity masks or latitude weighting. For RGB values in $[0, 1]$,

$$\text{PSNR} = -10 \log_{10} \frac{\sum_{u\in\Omega} \|\hat{I}_t(u) - I_t(u)\|_2^2}{3|\Omega|}, \tag{15}$$

where Ω contains all image pixels. SSIM uses a 7×7 uniform window and sample covariance, excluding the outer three-pixel border from reduction and without horizontal padding. Aggregation follows Sec. 4.1.

FID evaluation. Each panorama is converted into ten 299×299 perspective views with a 90° field of view and elevations within $\pm 45^\circ$. Ground truth and all methods share the same fixed random viewing directions. FID is computed from the resulting view collection for each dataset.

Depth evaluation and scoring. The observed-RGB depth subset is specified in Appendix A. We score aligned depth before local depth correction and retain all frames, including initialization and predictions rejected from the map. Only DA360 predictions are capped at 100 m immediately before scoring; GT remains uncapped. This cap does not affect historical alignment or point-cloud construction. Metrics are computed per frame and then averaged over all evaluated frames. AbsRel and RMSE (m) use valid GT/prediction intersections; δ_1 uses all valid GT pixels as its denominator, counting invalid predictions as incorrect, with threshold $\max(\hat{d}/d, d/\hat{d}) < 1.25$. Valid GT requires the benchmark mask and finite positive radial depth.

Generated-view depth evaluation. Table 2 uses the 16 TartanGround NVS targets, separate from the observed-RGB depth evaluation. AMOR uses jointly generated depth restored to metric scale using historical geometry and scored before local correction. Matrix-3D uses the MoGe-based optimization route with a depth-estimation interval of 10 frames; native foreground masks exclude background sentinel values. We evaluate aligned depth before 3DGS optimization, not the LRM depth head. HY-World uses WorldMirror with its native depth alignment on generated perspective videos. We retain the original temporal samples and add the evaluated keyframes, then project perspective Z-depth into radial ERP depth, selecting the most central valid view where projections overlap. Source-derived scene calibration converts baseline outputs into metres; target GT is never used for scale fitting. Pano2Room follows its official geometry-generation pipeline, including greedy viewpoint selection, RGB inpainting, history-conditioned depth prediction, geometric admission, and mesh updates. This stage runs at 2048×1024 along the benchmark camera trajectory, with floating-point radial depth exported before mesh insertion. We rotate the world-aligned panoramas to the evaluated camera orientations and convert depth into metres using calibration from its source prediction only. GeoCheck admission masks do not restrict the scored predictions. Depth is evaluated before 3DGS training, whereas the RGB table uses final 3DGS renders. For all three metrics, the per-frame scoring mask intersects valid non-sky GT with finite positive predictions from all four methods. HY-World's perspective support covers about 44.7% of valid GT, so the common-support scores do not measure full-panorama coverage. We compute unweighted ERP pixel metrics in radial metres and average over the 16 frames; δ_1 is the fraction satisfying $\max(\hat{d}/d, d/\hat{d}) < 1.25$ on this same mask. These pipeline-level results support generated geometry quality but do not constitute a controlled joint-versus-sequential ablation.

F Memory Ablations and Additional Results

F.1 Conservative Memory Updates

Inputs and controlled variants. The ablation uses Holo360D Indoor_002, Indoor_018, Indoor_053, Outdoor_001, and Outdoor_004, together with TartanGround Hospital_P0000, JapaneseAlley_P0000, Supermarket_P0000, and VictorianStreet_P0000. Both variants use the same model architecture, 50 denoising steps, seed 0, and at most one complete source RGB-D observation selected for each target view. Both variants reuse the same initial RGB, depth, and validity mask. Generation is evaluated at trajectory indices 20, 40, 60, and 80, giving four successive generation queries per scene; the initial observation is index 0. Both policies use surface-guided RGB readout.

- **Direct:** retain trusted scale-reference management, but remove local geometric admission, local depth correction, old-point retirement, and residual rejection. Append finite, positive, non-sky predictions after global scale alignment with available anchors. This compares conservative integration with aggressive accumulation as complete policies.

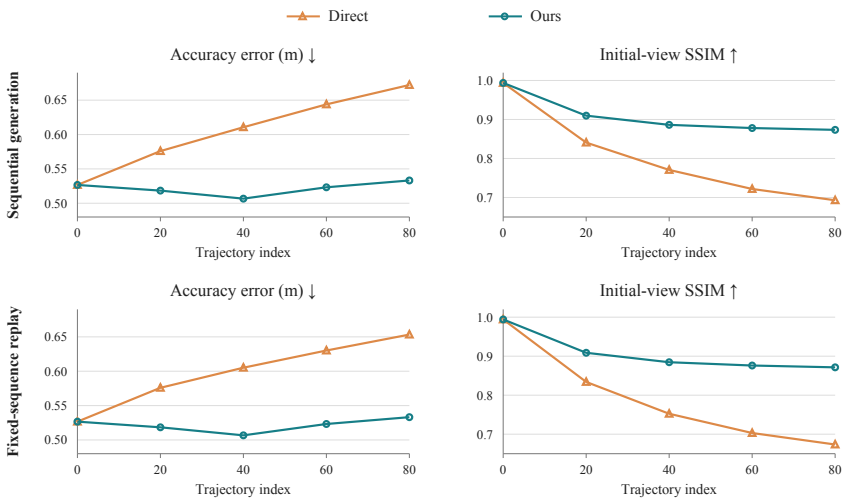


Figure 8: **Geometric accuracy and appearance retention for Direct and Ours.** Top: sequential generation; bottom: fixed-sequence integration replay. Columns show geometric accuracy error and initial-view SSIM, averaged over nine scenes. Initial-view SSIM measures retained appearance at the input pose.

- **Ours:** use the complete update procedure in Appendix D, including compatibility-based admission, local boundary correction, retirement of visible old points within accepted regions, and controlled eligibility of new points as scale references.

Geometry and initial-view retention. The common geometric reference is the union of GT depth back-projections at indices 0, 20, 40, 60, and 80. Predicted and reference clouds are sampled with 5 cm voxels in metric world coordinates, without ICP or fitted scale. Accuracy is the mean prediction-to-reference nearest-neighbor distance; completeness reverses this direction. Precision, recall, and F-score use a 5 cm threshold. Completeness concerns GT-observed surfaces, not a full-scene mesh; generated surfaces absent from the observed reference may contribute to accuracy error. For retention, each current map is rendered at the initial pose and compared with the initial RGB on a fixed initial RGB-D validity mask, without exempting missing rendered pixels. This is a memory-rendering diagnostic, not diffusion-based loop closure. Geometry and retention values are equal-weight scene means. Generated RGB uses the full-image metrics defined in Appendix E; each of the 36 evaluated views has equal weight here. The experiment uses one seed and does not establish statistical significance.

Sequential generation and fixed-sequence replay. Figure 8 distinguishes two questions. Sequential runs measure the complete feedback process. Each policy maintains an independent map, whose readout conditions later generation, so both predictions and integrated geometry can diverge. Fixed-sequence replay instead applies both integration policies to the same raw RGB, relative-depth predictions, and sky masks produced by our method, without additional diffusion sampling. Each policy still re-estimates alignment from its own evolving memory; replay controls the raw predictions, not the aligned metric depths or scale anchors.

Final replay accuracy error is 0.653 m for Direct and 0.533 m for Ours, consistent with the sequential comparison of 0.672 and 0.533 m. The precision advantage thus persists when raw generated inputs are held fixed, supporting the combined integration policy, including its alignment behavior.

F.2 Revisiting Previously Seen Regions

Fixed retains only the initial RGB-D observation without updating memory. Direct and Ours use the update policies defined above. On the Hospital trajectory, nine outgoing poses are followed by eight return frames; only frame 1 is observed. Across return frames 10 to 17, AMOR achieves

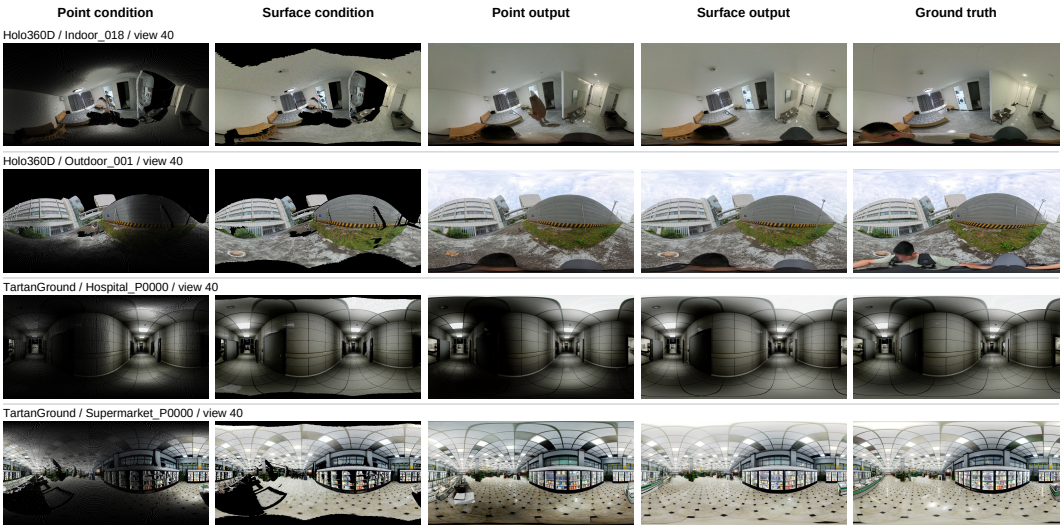


Figure 9: **Paired RGB conditions and final generated panoramas.** All examples use target index 40 with fixed initial memory. Columns show the point and surface RGB conditions before VAE encoding, their corresponding raw generated RGB images, and the target ground truth. Each row shares the same source RGB-D, target pose, depth condition, and sampling seed. Black in the conditioning images denotes missing evidence.

12.96 dB PSNR, 0.4942 SSIM, and 0.4913 LPIPS, versus 12.52 dB, 0.4741, and 0.5194 for Fixed, and 9.14 dB, 0.3676, and 0.7314 for Direct (Fig. 4). These RGB metrics measure revisit appearance fidelity rather than 3D accuracy.

F.3 RGB Readout Ablation

We compare point-projected and surface-guided RGB readout on the nine scenes listed above, using the same initial RGB-D observation and target poses at indices 20, 40, 60, and 80. Initial memory remains fixed throughout both runs. The depth condition and geometric support are unchanged; only the reprojected RGB supplied to the RGB VAE differs. Both runs use 50 denoising steps and seed 0. All 36 raw generated RGB images per variant are scored with the full-image PSNR, SSIM, and LPIPS definitions in Appendix E, without filtering views by downstream geometric admission. Overall scores in Table 4 give each of the 36 views equal weight (20 Holo360D and 16 TartanGround). This is an inference-time readout intervention, not a comparison of separately trained readout variants, and uses a single sampling seed.

Condition visualizations. Figure 7(c) shows RGB warps before VAE encoding, with matched inputs and crops within each row and no hole filling. The top row is Holo360D Outdoor_001 from the readout ablation. The bottom row is the internal example panorama2 at index 80, using a stored full-pixel depth-buffered point splat rather than the single-pixel baseline used in the quantitative ablation. It is a qualitative rendering diagnostic only and does not enter benchmark scores.