WORLD-δ: CONTINUALLY INTEGRATING OBSERVATIONS FOR INCREMENTAL 3D SCENE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing 3D scene generation methods can synthesize visually compelling scenes from fixed inputs, yet struggle to continually incorporate new observations to coherently expand and revise an existing 3D world. Existing iterative approaches often fuse real observations and generated content into a single deterministic representation without explicitly modeling their provenance or uncertainty. Consequently, early generation errors can accumulate and become difficult to correct with subsequent observations. To address these limitations, we propose World-δ, an incremental 3D scene generation framework with uncertainty-aware persistent world modeling and delta-guided diffusion for continual scene expansion and correction. Our framework consists of three components: 1) Visual Evidence Encoding integrates geometric and semantic cues from incoming images into unified world tokens. 2) Observation-Calibrated World Modeling maintains a persistent latent world state that tracks geometry, appearance, provenance, observation confidence, and generative uncertainty, distinguishing observation-supported content from provisional completion hypotheses. It compares new evidence with the existing state to infer selective world deltas, enabling the revision of conflicting hypotheses while preserving well-supported content. 3) Incremental World Diffusion Generation derives update masks from these deltas and restricts diffusion to regions requiring generation or revision, while freezing unchanged regions. Only affected 3D chunks are decoded and assembled with the existing scene representation. Together, these components enable the world to expand with accumulating observations and allow uncertain completions to remain revisable, while reducing redundant generation and preserving scene consistency. Extensive experiments demonstrate the effectiveness of World-δ in incremental 3D scene generation, observation-driven correction, scene consistency, and update efficiency.

1 INTRODUCTION

One-shot 3D generation produces plausible geometry and appearance from fixed observations, completing unobserved regions. Yet continual exploration requires more than a fixed-input prediction: new images may reveal unseen space or contradict earlier completions. Generated worlds must expand and correct content as evidence accumulates, while preserving reliable history.

Geometric fusion incremental generation aligns and fuses synthesized views or completed regions into shared 3D representations (Fridman et al., 2023; Chung et al., 2025). However, this alone does not distinguish observations from provisional completions. Without provenance and confidence, early errors can persist, obscuring what to preserve or revise. Continual generation requires an observation-calibrated persistent state and a generation process that selectively realizes its changes.

Fig. 1 motivates two coupled designs for observation-driven incremental generation. First, provenance and confidence let new evidence confirm or revise hypotheses while preserving supported content. Second, new observations may reveal missing regions, expose conflicts, or corroborate existing content. A state delta identifies where generation is needed and where history can be retained.

We introduce World-δ, comprising three components. Visual Evidence Encoding fuses geometric and semantic cues into world tokens. Observation-Calibrated World Modeling compares incoming evidence with a persistent state to infer selective deltas, integrating new content and revising conflicting hypotheses. Incremental World Diffusion Generation converts deltas into update masks,

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.

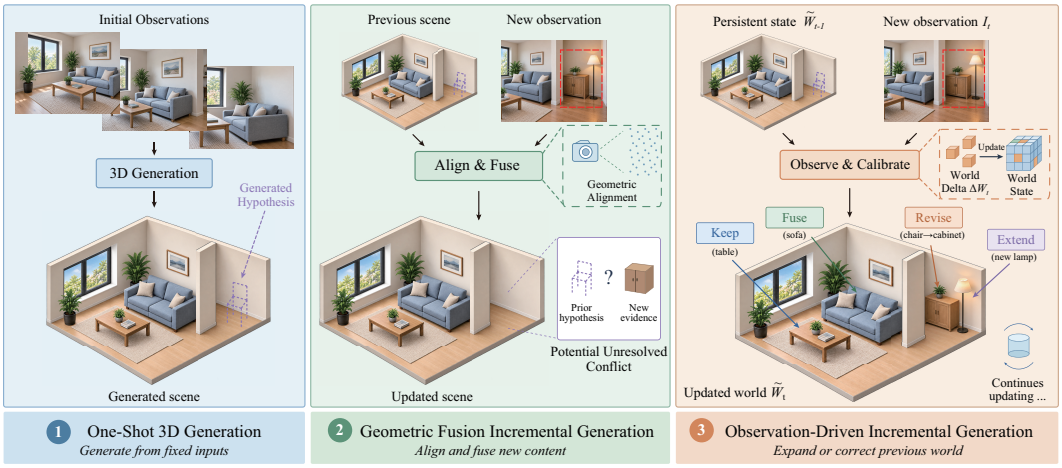

Figure 1: **Paradigm shift in incremental 3D generation.** One-shot 3D generation completes unseen content from fixed inputs. Geometric fusion incremental generation aligns and fuses new content but can retain conflicts with earlier completions. Our observation-driven incremental generation uses new observations to expand coverage and correct prior hypotheses. A persistent world state and selective deltas preserve supported content while integrating new evidence.

conditions diffusion on the integrated state, and freezes unaffected latents. Only affected 3D chunks are decoded and assembled with the existing explicit scene.

Our main contributions are:

- We introduce World-δ, a unified framework that continually expands and corrects a generated 3D world from incoming RGB observations, integrating new evidence while preserving reliable history and keeping provisional completions revisable.
- Observation-Calibrated World Modeling tracks provenance, observation confidence, and generative uncertainty in a persistent state, calibrating new evidence against history to infer selective updates that revise conflicting hypotheses while retaining supported content.
- Incremental World Diffusion Generation converts state deltas into update masks, freezes unaffected latents during conditional generation, and decodes only affected chunks to retain unchanged content and reduce redundant computation during continual world updates.
- We aggregate and adapt large-scale existing Gaussian datasets for incremental 3D Gaussian generation training. Experiments demonstrate improved rendering quality and a favorable balance of observation-driven correction, historical retention, and update cost.

2 Related Works

3D Generation. Recent 3D generation combines Gaussian representations with diffusion priors (Zhang et al., 2024; Lin et al., 2025). DreamScene (Li et al., 2025) optimizes Gaussian scenes under image diffusion guidance, whereas DiffSplat (Lin et al., 2025) generates Gaussian latents using adapted image diffusion. Despite high-quality outputs, fixed conditioning does not address revising earlier completions under new observations. Our revisable world representation reconciles new evidence with generative hypotheses for coherent expansion while preserving supported history.

Persistent World Modeling. Existing methods retain spatial history by evolving persistent states (Wang et al., 2025; Garcin et al., 2026) or retrieving stored observations (Xiao et al., 2025). These paradigms support continuous perception and long-term spatial consistency, but do not resolve what to preserve or revise under conflicting observations. Persistence and retrieval alone specify neither historical provenance nor the relative reliability of incoming and stored evidence. We therefore calibrate persistent states using provenance, observation confidence, and generative uncertainty, distinguishing supported history from provisional completions to infer selective updates.

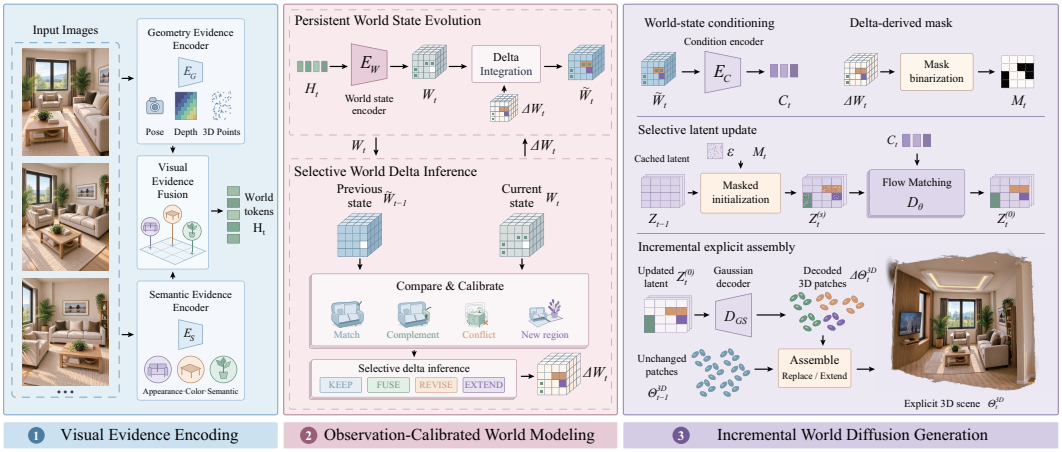

Figure 2: **Overview of World-δ.** Spatially anchored evidence $\mathcal{W}_t$ is compared with history $\widetilde{\mathcal{W}}_{t-1}$ to infer selective deltas and update $\widetilde{\mathcal{W}}_t$. Conditioned on this state, masked flow matching updates selected latents, freezing others. Only affected chunks are decoded and merged with cached chunks.

Incremental Diffusion Generation. Incremental 3D diffusion must update scenes while retaining unaffected content. Current pipelines generate complete samples under fixed conditioning (Lin et al., 2025) or inpaint missing regions in rendered views before 3D fusion (Lei et al., 2023; Chung et al., 2025). Hole filling can overlook occupied-region errors, while broad resampling disturbs reliable history and repeats computation. The key is therefore to align diffusion's writable support with evidence-driven state changes. World-δ derives masks from observation-calibrated deltas, freezes unaffected latents, and decodes only affected chunks for efficient, targeted correction.

3 Method

3.1 Overview

World-δ expands and corrects a 3D world from RGB observations $\mathcal{I}_t = \{I_{t,j}\}_{j=1}^{N_t}$, where t indexes observation rounds. Fig. 2 summarizes three stages.

Visual Evidence Encoding fuses geometric and semantic features, H_t^G and H_t^S, into world tokens using predicted poses Π_t, depths D_t, 3D points X_t^{3D}, and geometric reliability Q_t^G:

$$H_t = F_{\text{vis}}(H_t^G, H_t^S; \Pi_t, D_t, X_t^{3D}, Q_t^G). \tag{1}$$

Observation-Calibrated World Modeling encodes current evidence as $\mathcal{W}_t$, compares it with history $\widetilde{\mathcal{W}}_{t-1}$, and integrates a selective delta into the persistent state:

$$\begin{aligned} \Delta\mathcal{W}_t &= F_\Delta(\widetilde{\mathcal{W}}_{t-1}, \mathcal{W}_t), \\ \widetilde{\mathcal{W}}_t &= F_{\text{state}}(\widetilde{\mathcal{W}}_{t-1}, \mathcal{W}_t, \Delta\mathcal{W}_t), \quad t \geq 1. \end{aligned} \tag{2}$$

Incremental World Diffusion Generation conditions masked flow matching on the integrated state, refreshes selected latents as $Z_t = Z_t^{(0)}$, and decodes affected chunks into Θ_t^{3D}.

3.2 Visual Evidence Encoding

Geometry and Semantic Evidence Encoders. LingBot-Map (Chen et al., 2026a) yields $(H_t^G, \Pi_t, D_t, X_t^{3D}, Q_t^G) = E_G(\mathcal{I}_t)$, with back-projected anchors and depth confidence calibrated by visibility and geometric consistency. Frozen DINOv2 ViT-B/14 (Dosovitskiy et al., 2021; Oquab et al., 2023) supplies appearance and semantic features $H_t^S = E_S(\mathcal{I}_t)$.

Visual Evidence Fusion. F_{vis} fuses geometry and semantics at 3D anchors into $H_t = \{(h_{t,n}, a_{t,n}, p_{t,n}, q_{t,n}^{\text{obs}})\}_{n=1}^{L_t}$: features, anchors, image/round provenance, and confidence.

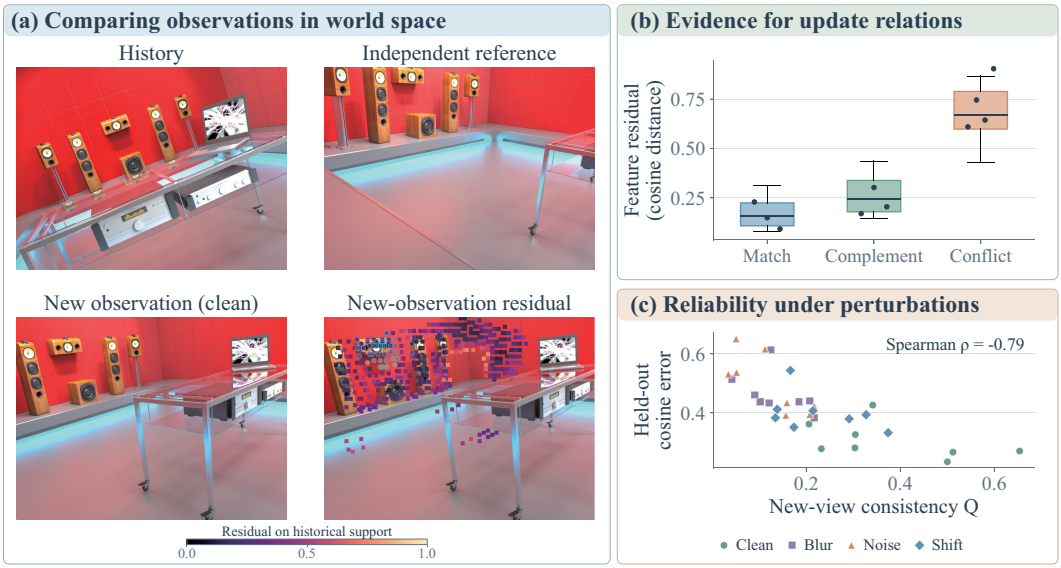

Figure 3: **Evidence reliability for observation-calibrated updates.** Using frozen DINOv2 features on four Hypersim scenes, we compare aligned observations and residuals on historical support (a), examine residual distributions for matching, complementary, and conflicting evidence (b), and relate new-view consistency to held-out feature error under perturbations (c). The analysis motivates considering content disagreement together with observation reliability before selecting world updates.

3.3 Observation-Calibrated World Modeling

A geometry-assisted probe on four Hypersim scenes (Roberts et al., 2021) with frozen DINOv2 yields larger residuals for controlled history conflicts (Fig. 3(b)). Higher new-view consistency Q corresponds to lower held-out error (Fig. 3(c)). Both cues inform update selection.

Persistent World State Evolution. For $t \geq 1$, E_W forms $\mathcal{W}_t = E_W(H_t; G_{t-1})$ in persistent coordinates. After generation, the integrated state stores generative latents Z_t, topology and accepted evidence attributes G_t, provenance P_t, observation confidence Q_t, and generative uncertainty U_t:

$$\widetilde{\mathcal{W}}_t = (Z_t, G_t, P_t, Q_t, U_t). \tag{3}$$

Confidence measures observational support; uncertainty estimates generation error. At $t = 0$, $\mathcal{W}_0 = E_W(H_0)$ establishes the coordinate frame, and $(\widetilde{\mathcal{W}}_0, \Omega_0^{\Delta}) = F_{\text{init}}(\mathcal{W}_0)$ marks all allocated locations for generation. Later rounds integrate each observation once via Eq. (2). Generation refreshes latents in the same state, which becomes subsequent history.

Selective World Delta Inference. We compare evidence with history using visibility, provenance, confidence, and uncertainty to assign $A_t(v) \in \{\text{KEEP}, \text{FUSE}, \text{REVISE}, \text{EXTEND}\}$. KEEP preserves content, FUSE integrates compatible evidence, REVISE corrects conflicting content, and EXTEND allocates new regions. Reliable observations can revise uncertain completions; weak or occluded evidence is downweighted. The structured edit is

$$\Delta\mathcal{W}_t = (\Delta Z_t, \Delta G_t, \Delta P_t, \Delta Q_t, \Delta U_t, \Omega_t^{\Delta}). \tag{4}$$

The delta specifies insertions, replacements, or removals. ΔZ_t requests regeneration; Ω_t^{Δ} covers changes requiring realization, including topology edits. Metadata-only fusion bypasses diffusion.

3.4 Incremental World Diffusion Generation

Fig. 4(b)–(d) shows that calibrated selection better recovers required update support than coverage-only or residual-only selection at limited budgets. We therefore derive editable regions from calibrated deltas and update chunks according to decoding dependencies.

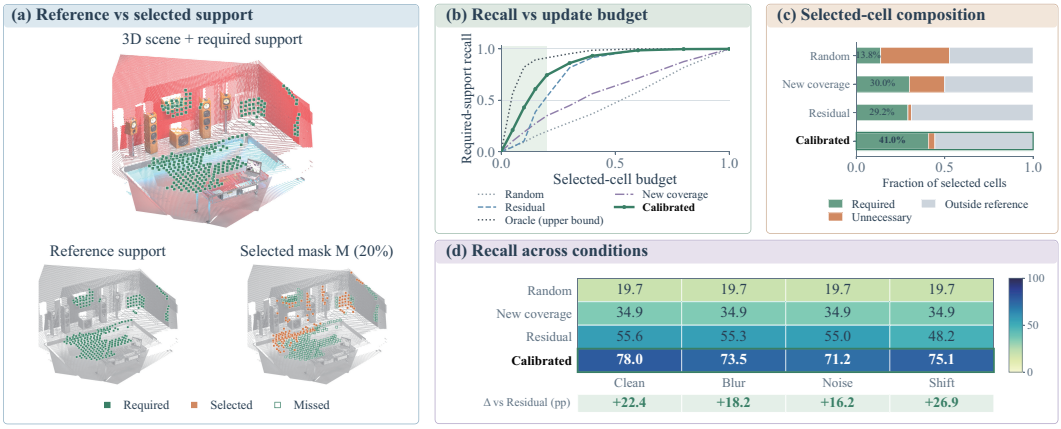

Figure 4: **Localization and allocation of incremental updates.** We compare calibrated, random, new-coverage, and residual-based selection under limited update budgets: reference support and a 20% mask (a), recall versus budget (b), selected-cell composition (c), and recall under observation perturbations (d). Cells outside the reference support lack valid held-out evidence. Calibrated selection improves support recall at matched budgets, motivating delta-derived generation masks.

World-State Conditioning and Update Masking. Conditioning uses $C_t = E_C(\widetilde{\mathcal{W}}_t)$. We form $M_t = R_{\text{mask}}(\Omega_t^{\Delta}; \mathcal{L}_{\text{diff}})$ on the diffusion lattice, including required boundary cells. Across channels, $M_t(v) = 1$ permits writes; zero preserves history. New slots and latent changes are masked; neighboring latents provide read-only context. Metadata-only edits can bypass diffusion.

Masked Latent Initialization. We align history to current topology as $Z_t^{\text{ref}} = \text{Align}(Z_{t-1}; G_t)$, preserving retained values. Let s be the number of integration steps and $k = s, \ldots, 0$ index sampling independently of t. Masked cells start from Gaussian noise:

$$Z_t^{(s)} = (1 - M_t) \odot Z_t^{\text{ref}} + M_t \odot \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, I). \tag{5}$$

New reference entries are zero-initialized and masked, so noise replaces them. At initialization, $Z_0^{\text{ref}} = 0$, $M_0 = 1$ on the allocated domain, and C_0 supplies observation conditioning.

Selective Latent Update with Flow Matching. We train D_θ by flow matching (Lipman et al., 2023). For a clean target $Z^\star$ and $\tau \in [0, 1]$, editable cells follow $(1-\tau)Z^\star + \tau\epsilon$ with target velocity $\epsilon - Z^\star$; others retain history. Sampling uses $\tau_k = k/s$ for $k = s, \ldots, 1$, from noise to clean latents:

$$\begin{aligned} V_t^{(k)} &= D_\theta(Z_t^{(k)}, \tau_k \mid C_t, M_t, Z_t^{\text{ref}}), \\ Z_t^{(k-1)} &= (1 - M_t) \odot Z_t^{\text{ref}} + M_t \odot \left[Z_t^{(k)} - \frac{1}{s} V_t^{(k)} \right]. \end{aligned} \tag{6}$$

The endpoint $Z_t = Z_t^{(0)}$ refreshes the state's latents. Provenance and uncertainty record realizations; topology and confidence retain calibrated values. Generation adds no observational support.

Incremental Explicit Assembly. Let $\mathscr{C}_t$ index chunks in G_t. Their decoding dependencies $\mathcal{R}_c$ cover neighborhoods and topology before and after edits. Affected chunks satisfy

$$\mathscr{C}_t^{\Delta} = \left\{ c \in \mathscr{C}_{t-1} \cup \mathscr{C}_t : \mathcal{R}_c \cap \left(\Omega_t^{\Delta} \cup \text{WorldSupp}(M_t) \right) \neq \varnothing \right\}, \quad t \geq 1. \tag{7}$$

WorldSupp maps the mask to world coordinates; $\mathscr{C}_0^{\Delta} = \mathscr{C}_0$ initializes all chunks. Given $Z_{t,c}$ with the required context, Gaussian decoder D_{GS} realizes affected chunks in the current topology:

$$\begin{aligned} \Delta\Theta_t^{3D} &= \{ D_{GS}(Z_{t,c}, G_t) : c \in \mathscr{C}_t^{\Delta} \cap \mathscr{C}_t \}, \\ \Theta_t^{3D} &= \text{Assemble}(\Theta_{t-1}^{3D}, \Delta\Theta_t^{3D}; \mathscr{C}_t^{\Delta}). \end{aligned} \tag{8}$$

Starting from $\Theta_{-1}^{3D} = \varnothing$, assembly replaces or adds affected patches, removes $\mathscr{C}_t^{\Delta} \setminus \mathscr{C}_t$, and reuses unaffected entries. Decoding reads a two-cell halo and writes only selected owners' outputs.

Table 1: Single-image 3D scene generation comparison.

Method	RealEstate10K				Tanks and Temples			
	PSNR(dB)↑	SSIM↑	LPIPS↓	Time(s)↓	PSNR(dB)↑	SSIM↑	LPIPS↓	Time(s)↓
LucidDreamer	16.4	0.578	0.322	800.2	14.5	0.357	0.407	825.3
ZeroNVS	16.6	0.578	0.360	10,827.1	14.7	0.383	0.477	11,326.4
Lyra-1	21.8	0.749	0.221	159.8	19.3	0.573	0.342	164.9
InfiniSplat (RGB)	18.8	0.698	0.291	**68.3**	16.9	**0.679**	0.325	**73.6**
World-δ (Ours)	**22.3**	**0.767**	**0.202**	82.7	**19.8**	0.602	**0.313**	94.6

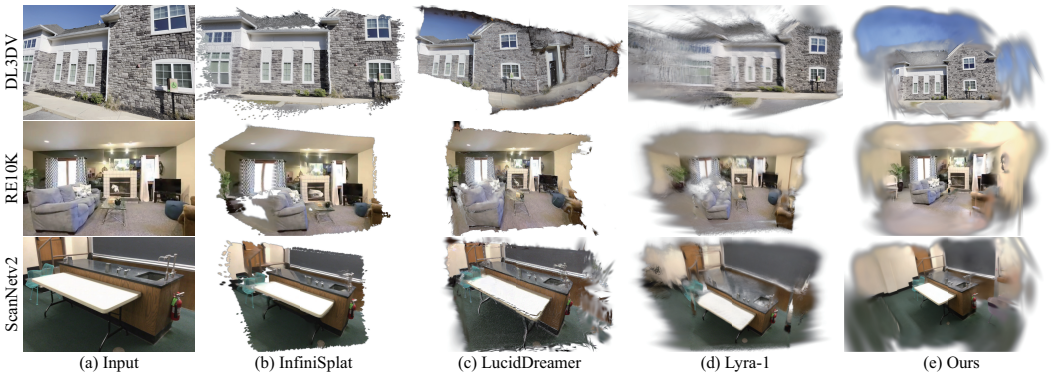

Figure 5: Visualization of single-image 3D scene generation.

3.5 Training Objectives

Gaussian Reconstruction Loss. The codec uses reconstruction and a standard normal prior:

$$\mathcal{L}_{\text{codec}} = \mathcal{L}_{\text{slot}} + \mathcal{L}_{\text{render}} + \beta_{\text{KL}}\mathcal{L}_{\text{KL}}. \tag{9}$$

$\mathcal{L}_{\text{slot}}$ supervises occupancy and attributes; $\mathcal{L}_{\text{render}}$ combines RGB L_1, SSIM, depth, and coverage.

State Calibration Loss. State learning combines weighted action cross-entropy, binary cross-entropy for reliability and free space, squared uncertainty error, and feature cosine distance:

$$\mathcal{L}_{\text{state}} = \mathcal{L}_{\text{action}} + \mathcal{L}_q + \mathcal{L}_U + 0.2\mathcal{L}_{\text{feature}} + 0.5\mathcal{L}_{\text{free}}. \tag{10}$$

Masked Flow-Matching Loss. On the linear path above, V_θ is the predicted velocity and T_{valid} marks valid targets. Using $S_t = M_t \odot T_{\text{valid}}$, we minimize

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\tau,\epsilon}\left[\frac{\|S_t \odot (V_\theta - (\epsilon - Z^\star))\|_2^2}{d_z \sum_v S_t(v) + 10^{-8}}\right], \qquad d_z = 64. \tag{11}$$

We average all terms over valid targets. With the codec frozen, we apply $0.05\mathcal{L}_{\text{render}}$ to decoded clean-latent estimates every ten steps; target validity affects only supervision.

4 Experiments

4.1 Implementation Details

Network configuration. Evidence is encoded as 256-dimensional world tokens. Each chunk has 8^3 64-channel latent cells and up to 4,096 Gaussians. A six-block Transformer predicts velocities for 50 Euler steps. Two valid $3 \times 3 \times 3$ decoder convolutions yield the dependency radius in Eq. (7).

Training data. We aggregate training splits from SceneSplat-49K (Ma et al., 2025): 3RScanGS, ARKitScenesGS, Matterport3DGS, HyperSimGS, ReplicaGS, ScanNetPPGS-V1/V2, ScanNetGS, DL3DV-GS, HoliCityGS, AriaSyntheticEnvGS, and crowdsourced scenes. We normalize coordinates and convert scenes into canonical chunk targets, pairing RGB prefixes with Gaussian and latent supervision for initialization, expansion, and revisiting. Reference Gaussians and available geometric labels provide offline supervision; inference uses only RGB prefixes.

Table 2: Continual RGB world update comparison.

Method	ScanNet v2				Replica			
	PSNR(dB)↑	SSIM↑	LPIPS↓	Update(s)↓	PSNR(dB)↑	SSIM↑	LPIPS↓	Update(s)↓
MonoGS (RGB)	18.2	0.692	0.602	65.1	29.2	0.880	0.238	58.1
Splat-SLAM (RGB)	26.8	0.812	0.231	17.9	33.6	0.933	0.092	15.7
S3PO-GS (RGB)	17.7	0.726	0.627	24.1	32.2	0.938	0.147	21.9
OF3GS (RGB)	20.6	0.692	0.362	12.7	30.8	0.909	0.172	10.3
World-δ (Ours)	**27.4**	**0.827**	**0.211**	**6.8**	**34.1**	**0.942**	**0.081**	**5.4**

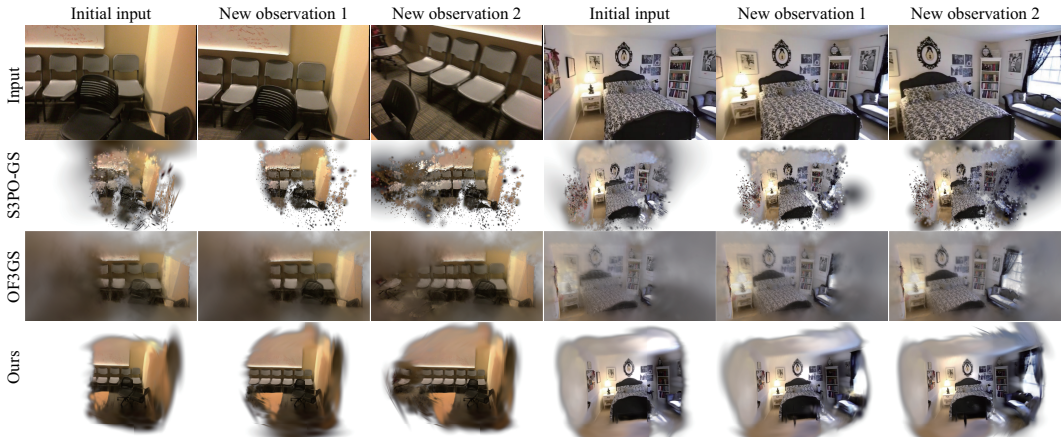

Figure 6: Visualization of continual 3D world updates.

Training configuration. Training covers codec, state, flow, and sequential calibration stages. Non-codec stages use AdamW with learning rate 10^{-4}, weight decay 0.01, and 16-chunk batches; adaptation uses 10^{-5}. Evidence backbones remain frozen. Codec pretraining ramps β_{KL} to 10^{-4}.

4.2 Comparison Experiments

We evaluate single-image generation and continual RGB updates with identical observations and query views. We report PSNR, SSIM, and LPIPS (Zhang et al., 2018), measuring generation and update times separately. Ground-truth geometry is used only for supervision or evaluation.

3D Generation Comparison. We compare with LucidDreamer (Chung et al., 2025), ZeroNVS (Sargent et al., 2024), Lyra-1 (Bahmani et al., 2026), and RGB-only InfiniSplat (Wang et al., 2026) on RealEstate10K (Zhou et al., 2018) and Tanks and Temples (Knapitsch et al., 2017). World-δ achieves the best PSNR and LPIPS on both datasets (Table 1), while InfiniSplat offers faster generation and higher Tanks and Temples SSIM. Fig. 5 shows broader coverage on DL3DV (Ling et al., 2024), RealEstate10K, and ScanNet v2, although some completions remain smooth.

3D Incremental Generation Comparison. We compare RGB variants of MonoGS (Matsuki et al., 2024), Splat-SLAM (Sandström et al., 2025), S3PO-GS (Cheng et al., 2025), and OF^{3}GS (Chen et al., 2026b) on ScanNet v2 (Dai et al., 2017) and Replica (Straub et al., 2019). Reconstruction baselines test observation integration; rendering scores alone do not assess unseen-region generation. Under the shared update protocol, World-δ achieves the best rendering scores and lowest update latency (Table 2). Fig. 6 illustrates continual expansion and correction as new observations arrive, particularly around chairs and bedroom structures.

4.3 Ablation Study

We ablate three components on ScanNet v2 using LPIPS, surface F-score and coverage at 5 cm, correction gain $G_{\mathrm{corr}} = E^{-}_{\mathrm{corr}} - E^{+}_{\mathrm{corr}}$, and retention error increase $F_{\mathrm{keep}} = E^{+}_{\mathrm{keep}} - E^{-}_{\mathrm{keep}}$. Superscripts $-/+$ mark before/after updates. Correction uses common newly revealed support; retention

Table 3: Visual evidence encoding ablation.

Variant	LPIPS↓	F-score@5cm↑	Coverage@5cm↑	G_{corr} ↑	F_{keep} ↓
w/o Semantic Features	0.232	72.4%	79.8%	0.067	0.011
w/o Geometry Features	0.267	63.0%	71.6%	0.041	0.023
Full	**0.211**	**74.3%**	**80.8%**	**0.081**	**0.008**

Table 4: Observation-calibrated world modeling ablation.

Variant	LPIPS↓	E^{+}_{corr} ↓	G_{corr} ↑	F_{keep} ↓
w/o Provenance P	0.236	0.116	0.044	0.018
w/o Confidence Q	0.228	0.104	0.056	0.025
w/o Uncertainty U	0.219	0.097	0.063	0.013
Unconditional Fusion	0.252	0.123	0.037	0.032
Full	**0.211**	**0.079**	**0.081**	**0.008**

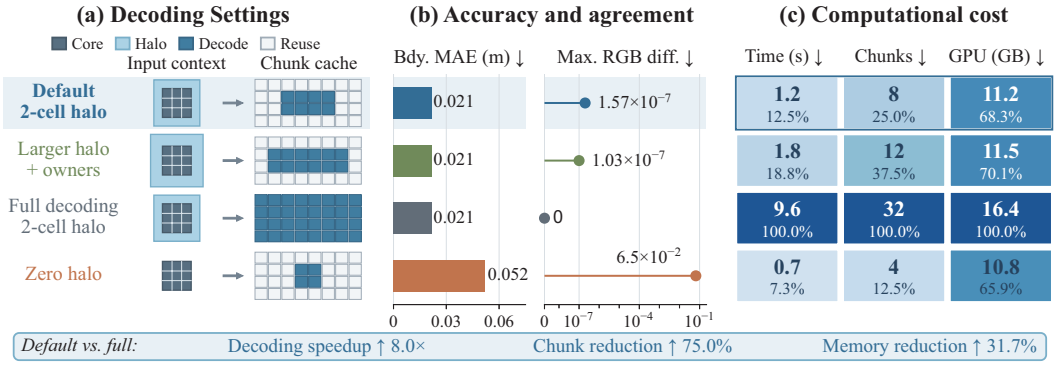

Figure 7: Analysis of dependency-aware incremental decoding.

uses reliable historical support. We average relative depth errors clipped at one; missing predictions count as one. Both metrics are absolute differences; higher G_{corr} and lower F_{keep} are better.

Visual Evidence Encoding. In Table 3, we ablate geometric and semantic features to assess their roles in world updates. Removing semantics degrades perceptual quality and correction; removing geometry causes larger losses in reconstruction accuracy and coverage.

Observation-Calibrated World Modeling. In Table 4, we ablate state attributes and fusion to assess observation calibration. Pre-update correction error E^{-}_{corr} is 0.160 across ScanNet v2 variants. Removing provenance weakens correction, removing confidence increases historical drift, and removing uncertainty reduces correction gain. Unconditional fusion increases retention error most.

Incremental World Diffusion Generation. In Table 5, we vary masking, hard freezing, and decoding on sampled ScanNet v2 scenes to evaluate incremental generation. Visibility-only masking corrects least. Global diffusion slightly improves LPIPS but increases historical drift and cost. Removing hard freezing increases drift. Full decoding matches quality and retention at higher cost.

4.4 Discussion

Correction and Retention under New Evidence. In Table 6, we compare update strategies on ScanNet v2 and Replica for correction and retention. Initial states share $E^{-}_{\text{corr}} = 0.160$ and 0.120, respectively. Correction targets valid, incorrect predictions. Extend-only retains history but leaves conflicts; Batch-Rebuild corrects more but increases drift. World-δ balances both by revision.

Selection under a Matched Update Budget. In Table 7, we evaluate update allocation on sampled ScanNet v2 scenes with 20% of latent cells writable. Calibrated selection improves quality and correction over residual-only and random selection. Removing U weakens correction; visibility-

Table 5: Incremental world diffusion generation ablation.

Variant	LPIPS↓	G_{corr} ↑	F_{keep} ↓	Update(s)↓	Decode(s)↓	GPU(GB)↓
Visibility-only Mask	0.254	0.026	**0.007**	**4.7**	**1.1**	**11.1**
Global Diffusion	**0.208**	**0.085**	0.042	22.8	9.6	16.8
w/o Hard Freezing	0.224	0.082	0.034	23.2	9.6	16.8
Full Decoding	0.211	0.081	0.008	13.2	9.6	16.4
Full	0.211	0.081	0.008	4.8	1.2	11.2

Table 6: Correction and retention under new evidence.

Variant	ScanNet v2				Replica			
	E^{+}_{corr} ↓	G_{corr} ↑	E^{+}_{keep} ↓	F_{keep} ↓	E^{+}_{corr} ↓	G_{corr} ↑	E^{+}_{keep} ↓	F_{keep} ↓
Extend-only	0.148	0.012	**0.083**	**0.003**	0.112	0.008	**0.053**	**0.001**
Batch-Rebuild	**0.075**	**0.085**	0.115	0.035	**0.049**	**0.071**	0.083	0.031
Full	0.079	0.081	0.088	0.008	0.054	0.066	0.058	0.006

Table 7: Selection under a matched update budget.

Selection rule	LPIPS↓	G_{corr} ↑	F_{keep} ↓	Update(s)↓
w/o U in Selection	0.222	0.061	0.014	5.2
Residual-only	0.232	0.059	0.019	6.2
Visibility-only	0.254	0.026	**0.007**	**4.7**
Random	0.271	0.019	0.026	8.9
Calibrated Delta	**0.211**	**0.081**	0.008	4.8

only retains history slightly better but resolves fewer errors. Runtime depends on which chunks are updated. With the writable budget fixed, these results show the benefit of using calibrated evidence to prioritize corrections within the available update budget.

Dependency-Aware Incremental Decoding. In Fig. 7, we compare decoder neighborhoods on ScanNet v2 with fixed latents and a fixed boundary region to assess agreement and efficiency. Boundary MAE measures depth error against reference geometry; maximum RGB discrepancy measures deviation from full decoding. The default matches full decoding's reported MAE with negligible RGB discrepancy and lower runtime and memory use. Enlarging the halo and owner set increases cost without lowering MAE; zero halo increases depth error and RGB discrepancy. These results support selecting owners by network dependencies and reusing unaffected chunks from the cache. Read-only halo cells supply context without enlarging the writable latent mask.

Limitations. World-δ assumes a largely static scene, so object motion can be confused with errors in earlier completions. Fusing observations from different motion states may leave stale geometry and inconsistent appearance. Future work could explore motion-aware masks to separate dynamic regions from background updates and object tracking to maintain evolving object states.

5 Conclusion

We present World-δ, an observation-driven framework for continually expanding and correcting generated 3D worlds. By distinguishing observed evidence from provisional completions, our method keeps uncertain content revisable while preserving reliable history. We encode geometric and semantic evidence in a persistent state that tracks provenance, confidence, and generative uncertainty. Calibrated state deltas then guide masked latent diffusion and dependency-aware decoding, restricting generation and explicit assembly to affected regions. Experiments on single-image generation and continual world updates demonstrate improved rendering quality and a favorable balance of observation-driven correction, historical retention, and computational cost.

AI USE STATEMENT

No generative AI tools were used in the research, implementation, experimentation, analysis, or preparation of this work. The authors take full responsibility for all content presented in this work.

REFERENCES

Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David Lindell, Zan Gojcic, Sanja Fidler, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra: Generative 3D scene reconstruction via video diffusion model self-distillation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/85e446f011bd3b6ee382ef2a7e47f20e-Abstract-Conference.html.

Lin-Zhuo Chen, Jian Gao, Shangzhan Zhang, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, and Yinghao Xu. LingBot-Map: Geometric context transformer for streaming 3D reconstruction. *arXiv preprint arXiv:2604.14141*, 2026a. doi: 10.48550/arXiv.2604.14141. URL https://arxiv.org/abs/2604.14141. Accepted to ECCV 2026.

Ruiyang Chen, Feiran Li, Chu Zhou, Zonglin Li, Zhanyu Ma, and Heng Guo. OF^{3}GS: On-the-fly feed-forward 3D gaussian splatting from unposed images. *arXiv preprint arXiv:2606.03254*, 2026b. URL https://arxiv.org/abs/2606.03254.

Chong Cheng, Sicheng Yu, Zijian Wang, Yifan Zhou, and Hao Wang. Outdoor monocular SLAM with global scale-consistent 3D gaussian pointmaps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26035–26044, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Cheng_Outdoor_Monocular_SLAM_with_Global_Scale-Consistent_3D_Gaussian_Pointmaps_ICCV_2025_paper.html.

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

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5828–5839, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Dai_ScanNet_Richly-Annotated_3D_CVPR_2017_paper.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. SceneScape: Text-driven consistent scene generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 39897–39914, 2023. doi: 10.52202/075280-1734. URL https://papers.neurips.cc/paper_files/paper/2023/hash/7d62a85ebfed2f680eb5544beae93191-Abstract-Conference.html.

Samuel Garcin, Thomas Walker, Steven McDonagh, Tim Pearce, Hakan Bilen, Tianyu He, Kaixin Wang, and Jiang Bian. Beyond pixel histories: World models with persistent 3D state. In *Forty-third International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.03482.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017. URL https://www.tanksandtemples.org/.

Jiabao Lei, Jiapeng Tang, and Kui Jia. RGBD2: Generative scene synthesis via incremental view inpainting using RGBD diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8422–8434, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Lei_RGBD2_Generative_Scene_Synthesis_via_Incremental_View_Inpainting_Using_RGBD_CVPR_2023_paper.html.

Haoran Li, Haolin Shi, Wenli Zhang, Wenjun Wu, Yong Liao, Lin Wang, Lik-Hang Lee, and Peng Yuan Zhou. DreamScene: 3D gaussian-based text-to-3D scene generation via formation pattern sampling. In *Computer Vision – ECCV 2024*, volume 15132 of *Lecture Notes in Computer Science*, pp. 214–230. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-72904-1_13. URL https://doi.org/10.1007/978-3-031-72904-1_13.

Chenguo Lin, Panwang Pan, Bangbang Yang, Zeming Li, and Yadong Mu. DiffSplat: Repurposing image diffusion models for scalable gaussian splat generation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=eajZpoQkGK.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ling_DL3DV-10K_A_Large-Scale_Scene_Dataset_for_Deep_Learning-based_3D_Vision_CVPR_2024_paper.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2023. doi: 10.48550/arXiv.2210.02747. URL https://arxiv.org/abs/2210.02747.

Mengjiao Ma, Qi Ma, Yue Li, Jiahuan Cheng, Runyi Yang, Bin Ren, Nikola Popovic, Mingqiang Wei, Nicu Sebe, Luc Van Gool, Theo Gevers, Martin R. Oswald, and Danda Pani Paudel. SceneSplat++: A large dataset and comprehensive benchmark for language gaussian splatting. *arXiv preprint arXiv:2506.08710*, 2025. doi: 10.48550/arXiv.2506.08710. URL https://arxiv.org/abs/2506.08710.

Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison. Gaussian splatting SLAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18039–18048, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Matsuki_Gaussian_Splatting_SLAM_CVPR_2024_paper.html.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://github.com/apple-aiml-research/ml-hypersim.

Erik Sandström, Ganlin Zhang, Keisuke Tateno, Michael Oechsle, Michael Niemeyer, Youmin Zhang, Manthan Patel, Luc Van Gool, Martin Oswald, and Federico Tombari. Splat-SLAM: Globally optimized RGB-only SLAM with 3D gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1686–1697, 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/VOCVALC/html/Sandstrom_Splat-SLAM_Globally_Optimized_RGB-only_SLAM_with_3D_Gaussians_CVPRW_2025_paper.html.

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. ZeroNVS: Zero-shot 360-degree view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9420–9429, 2024. URL

https://openaccess.thecvf.com/content/CVPR2024/html/Sargent_ZeroNVS_Zero-Shot_360-Degree_View_Synthesis_from_a_Single_Image_CVPR_2024_paper.html.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. URL https://arxiv.org/abs/1906.05797.

Jiawei Wang, Hao Yu, Yongzhen Hu, Xinyi Yang, Tao Ni, Xin Zhan, Junbo Chen, Xiaowei Zhou, Ruizhen Hu, and Sida Peng. InfiniSplat: Implicit gaussian decoding for large-baseline monocular view synthesis. *ACM Transactions on Graphics*, 2026. URL https://arxiv.org/abs/2608.02437. Accepted to SIGGRAPH Asia 2026 (Journal Track).

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Continuous_3D_Perception_Model_with_Persistent_State_CVPR_2025_paper.html.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term consistent world simulation with memory. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1659. URL https://papers.neurips.cc/paper_files/paper/2025/hash/470629a47e2d65ce0606c40055df5d26-Abstract-Conference.html.

Bowen Zhang, Yiji Cheng, Jiaolong Yang, Chunyu Wang, Feng Zhao, Yansong Tang, Dong Chen, and Baining Guo. GaussianCube: A structured and explicit radiance representation for 3D generative modeling. In *Advances in Neural Information Processing Systems*, volume 37, pp. 97445–97475, 2024. doi: 10.52202/079017-3091. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b0b750c4189f19d0cd71375e9e17f83f-Abstract-Conference.html.

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

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4):65:1–65:12, 2018. doi: 10.1145/3197517.3201323. URL https://tinghuiz.github.io/projects/mpi/.

APPENDIX

Table A1: Long-horizon expansion and revisiting.

Variant	Updates	ScanNet v2				
		Old-view LPIPS↓	Depth MAE(m)↓	G_{corr} ↑	Total time(s)↓	State(MB)↓
Batch-Rebuild	8	0.251	0.144	**0.076**	120.1	160
Extend-only	8	0.246	0.168	0.012	**32.1**	**128**
Full	8	**0.221**	**0.132**	0.072	38.4	**128**
Batch-Rebuild	16	0.268	0.154	**0.085**	479.9	320
Extend-only	16	0.248	0.171	0.012	**63.9**	**256**
Full	16	**0.214**	**0.120**	0.081	76.8	**256**
Batch-Rebuild	32	0.301	0.169	**0.083**	1,919.9	640
Extend-only	32	0.251	0.173	0.011	**141.0**	**512**
Full	32	**0.209**	**0.112**	0.078	169.8	**512**
Batch-Rebuild	64	0.328	0.188	**0.081**	7,680.2	1,280
Extend-only	64	0.253	0.176	0.010	**319.7**	**1,024**
Full	64	**0.203**	**0.108**	0.074	390.3	**1,024**

Variant	Updates	Replica				
		Old-view LPIPS↓	Depth MAE(m)↓	G_{corr} ↑	Total time(s)↓	State(MB)↓
Batch-Rebuild	8	0.106	0.101	**0.064**	111.9	140
Extend-only	8	0.112	0.121	0.009	**28.8**	**112**
Full	8	**0.092**	**0.092**	0.061	35.2	**112**
Batch-Rebuild	16	0.112	0.107	**0.071**	448.2	280
Extend-only	16	0.114	0.124	0.008	**57.6**	**224**
Full	16	**0.086**	**0.089**	0.066	70.4	**224**
Batch-Rebuild	32	0.121	0.114	**0.068**	1,792.2	560
Extend-only	32	0.116	0.126	0.007	**127.7**	**448**
Full	32	**0.083**	**0.086**	0.065	153.6	**448**
Batch-Rebuild	64	0.134	0.126	**0.066**	7,167.8	1,120
Extend-only	64	0.117	0.129	0.006	**288.2**	**896**
Full	64	**0.081**	**0.083**	0.062	351.7	**896**

A IMPLEMENTATION AND EVALUATION DETAILS

State layout and learning. The model uses chunks of side length 3.2 in normalized scene coordinates and cells of side length 0.4. These are scene units, not metric distances. Each cell has eight Gaussian slots and a 64-dimensional latent. The condition encoder produces 256-dimensional tokens. The flow network uses six Transformer blocks with a hidden width of 384, six attention heads, and a feed-forward width of 1,536. It predicts the velocity $\epsilon - Z^\star$ along a linear path from clean latents at $\tau = 0$ to Gaussian noise at $\tau = 1$. Inference integrates from $\tau = 1$ to $\tau = 0$ using 50 Euler steps. Historical latents outside the update mask are restored after every integration step, and all chunks at a given step read the same state snapshot. The two valid spatial convolutions give the decoder a two-cell latent support; channel-wise normalization does not enlarge it.

The training schedule allocates 100k steps to the codec, 50k to evidence and delta learning, 200k to flow matching, 20k to calibration, and 30k to sequential adaptation. Non-codec optimization uses 2k warm-up steps. Training begins with the codec and latent normalization. It then proceeds to evidence and delta prediction, conditional flow matching, and state calibration using sequential rollouts. Reference Gaussians and available RGB-D/camera labels supervise these offline stages and are not used as online inputs. All source datasets have training and test splits; learning uses only training data. Sequential rollouts use eight observation rounds and four-round truncated adaptation.

Table A2: Paired hard-freezing diagnostic.

Variant	$E^-_{\text{keep}} \downarrow$	$E^+_{\text{keep}} \downarrow$	$F_{\text{keep}} \downarrow$	RGB drift↓	Depth drift(m)↓	LPIPS↓
w/o Hard Freezing	**0.080**	0.114	0.034	0.048	0.029	0.224
Hard Freezing	**0.080**	**0.088**	**0.008**	**0.012**	**0.008**	**0.211**

Table A3: Decoder-neighborhood sensitivity.

Decoding setting	Chunks	T_{dec}(s)↓	GPU(GB)↓	Max. RGB diff.↓	Bdy. MAE(m)↓
Larger Halo + Owners	12	1.8	11.5	1.03E-07	**0.021**
Full Decoding	32	9.6	16.4	**0.0E+00**	**0.021**
Zero Halo (negative control)	4	**0.7**	**10.8**	6.5E-02	0.052
Default Halo (2 cells)	8	1.2	11.2	1.57E-07	**0.021**

Loss implementation. Canonical chunk targets define the Gaussian-slot correspondence. We use $\mathcal{L}_{\text{slot}} = \mathcal{L}_{\text{occ}} + \mathcal{L}_{\mu} + 0.2\mathcal{L}_s + 0.1\mathcal{L}_R + 0.5\mathcal{L}_c$, combining balanced occupancy binary cross-entropy, normalized center L_1, log-scale L_1, quaternion alignment, and color L_1. Rendering uses $0.8L_1^{\text{RGB}} + 0.2(1-\text{SSIM}) + 0.1L_1^{\text{depth}}/3$, with an additional 0.1 opacity-coverage penalty on valid reference pixels. Unobserved regions are treated as unknown rather than as free-space targets. The KL coefficient increases from zero to 10^{-4} over the first 10k codec steps. Action cross-entropy weights for KEEP/FUSE/REVISE/EXTEND are $1/2/3/2$; uncertainty supervision covers both pre-update and realized states. Writing X_τ for the masked linear-path input, auxiliary flow rendering decodes $X_\tau - \tau V_\theta$ inside the update mask and retains Z^{ref} elsewhere. We do not backpropagate through discrete state edits or sequential rollout sampling.

Rendering and timing. Single-image generation uses the initial RGB observation and is evaluated at held-out cameras. For continual updates, we evaluate the current scene after sequential updates. All methods receive identical observation prefixes, and future images are excluded from conditioning. We report initial generation and subsequent update times separately. Update time covers evidence encoding, state calibration, flow integration, decoding, and assembly but excludes model loading. Renderer queries evaluate the model without refining it against held-out target images.

Correction and retention. Let S_{corr} be a common newly revealed reference surface and let S_{keep} contain reliable historical locations. Missing predictions on the revealed support are included and penalized, so the general correction score reflects both filling missing content and improving existing estimates. Only the paired correction-and-retention diagnostic in Table 6 restricts S_{corr} to locations with valid but incorrect pre-update predictions. This restriction isolates revision. The supports and probe views are shared between variants and are not defined by each variant's own update mask. For a valid reference depth d^*, the per-location error is the truncated relative depth difference

$$e(d, d^*) = \min\left(\frac{|d - d^*|}{\max(d^*, \varepsilon_{\text{d}})}, 1\right), \qquad \mathcal{E}(S) = \frac{1}{|S|}\sum_{i \in S} e(d_i, d_i^*). \tag{12}$$

Here, $\varepsilon_{\text{d}} > 0$ stabilizes the depth denominator, and S is a nonempty evaluation support. Invalid predicted depths receive the maximum error. Writing $E_{\text{corr}} = \mathcal{E}(S_{\text{corr}})$ and $E_{\text{keep}} = \mathcal{E}(S_{\text{keep}})$, the reported differences are $G_{\text{corr}} = E^-_{\text{corr}} - E^+_{\text{corr}}$ and $F_{\text{keep}} = E^+_{\text{keep}} - E^-_{\text{keep}}$. These absolute differences are dimensionless; depth MAE is reported in meters. F-score is the harmonic mean of surface precision and recall at 5 cm; coverage is recall under the same alignment and support.

B Additional Experiments

Long-horizon expansion and revisiting. Table A1 extends the comparison on ScanNet v2 and Replica to 8, 16, 32, and 64 update rounds. World-δ maintains low LPIPS on historical views as the world grows, while Batch-Rebuild incurs substantially higher cumulative cost. Extend-only is less expensive but achieves only small correction gains. State size still increases as scene content accumulates, so selective updates do not imply constant memory use.

Table A4: Robustness to observation degradation.

Pose	Factor	Level	Variant	LPIPS↓	G_{corr} ↑	F_{keep} ↓	Failure(%)↓
Predicted	Blur	Gaussian σ=1.5 px	w/o Q	0.277	0.033	0.035	15.2%
Predicted	Blur	Gaussian σ=1.5 px	**Full**	**0.241**	**0.054**	**0.018**	**11.2%**
Predicted	Noise	RGB σ=0.05	w/o Q	0.273	0.034	0.039	14.4%
Predicted	Noise	RGB σ=0.05	**Full**	**0.237**	**0.058**	**0.022**	**10.4%**
Predicted	Frame Drop	0%	w/o Q	0.228	0.056	0.025	3.9%
Predicted	Frame Drop	0%	**Full**	**0.211**	**0.081**	**0.008**	**0.0%**
Predicted	Frame Drop	25%	w/o Q	0.265	0.040	0.028	12.8%
Predicted	Frame Drop	25%	**Full**	**0.229**	**0.064**	**0.011**	**8.8%**
Predicted	Frame Drop	50%	w/o Q	0.293	0.021	0.033	18.4%
Predicted	Frame Drop	50%	**Full**	**0.257**	**0.044**	**0.016**	**14.4%**
Predicted	View Overlap	20%	w/o Q	0.302	0.015	0.041	20.2%
Predicted	View Overlap	20%	**Full**	**0.266**	**0.039**	**0.023**	**16.2%**
GT (diagnostic)	Blur	Gaussian σ=1.5 px	w/o Q	0.264	0.036	0.032	13.2%
GT (diagnostic)	Blur	Gaussian σ=1.5 px	**Full**	**0.228**	**0.060**	**0.015**	**9.2%**
GT (diagnostic)	Noise	RGB σ=0.05	w/o Q	0.261	0.040	0.036	12.4%
GT (diagnostic)	Noise	RGB σ=0.05	**Full**	**0.224**	**0.064**	**0.019**	**8.4%**
GT (diagnostic)	Frame Drop	0%	w/o Q	0.214	0.062	0.022	2.2%
GT (diagnostic)	Frame Drop	0%	**Full**	**0.197**	**0.086**	**0.004**	**0.0%**
GT (diagnostic)	Frame Drop	25%	w/o Q	0.252	0.046	0.025	10.8%
GT (diagnostic)	Frame Drop	25%	**Full**	**0.216**	**0.071**	**0.008**	**6.8%**
GT (diagnostic)	Frame Drop	50%	w/o Q	0.280	0.027	0.031	16.4%
GT (diagnostic)	Frame Drop	50%	**Full**	**0.244**	**0.051**	**0.013**	**12.4%**
GT (diagnostic)	View Overlap	20%	w/o Q	0.289	0.021	0.037	18.2%
GT (diagnostic)	View Overlap	20%	**Full**	**0.253**	**0.044**	**0.021**	**14.2%**

Hard freezing and decoder neighborhoods. Table A2 compares hard-freezing variants from paired initial states on ScanNet v2. Table A3 compares decoder neighborhoods on ScanNet v2. Default, enlarged-neighborhood, and full decoding match in reported boundary MAE; the default uses less time and memory. The zero-halo setting increases both MAE and RGB discrepancy, supporting the use of neighboring latent context for incremental decoding.

Observation degradation. Table A4 evaluates the full model and the variant without observation confidence Q under the specified observation perturbations on ScanNet v2. The paired conditions test whether observation confidence remains useful when evidence becomes less accurate. The main comparisons use predicted cameras. In the diagnostic rows, ground-truth cameras replace only the camera estimates. The RGB observations remain unchanged, allowing this comparison to isolate registration error. We report failure rates alongside rendering and correction metrics so that quality scores alone do not hide unsuccessful update sequences.

C Additional Qualitative Results

Single-image 3D Scene Generation. Fig. A1 presents additional comparisons on DL3DV, RealEstate10K, and ScanNet v2. The examples show scene coverage and appearance across varied layouts, complementing the single-image generation results in Fig. 5.

Continual 3D World Updates. Fig. A2 compares World-δ with S3PO-GS and OF^{3}GS on two additional continual-update sequences. Successive observations reveal more scene content. The examples illustrate world expansion and revisions to furniture and the surrounding room structure.



Figure A1: Additional visualizations of single-image 3D scene generation.



Figure A2: Additional visualizations of continual 3D world updates.