

BAT3R: Robust Online 3D Reconstruction with Bayesian Adaptive State Updates

Lijian Li^{1*} Yuanpeng He^{2,3,*†} Yudian Zheng¹
 Chung-ju Huang² Linyu Li^{2,3} Dongming Jin^{2,3}
 Wenpin Jiao^{2,3} Zhi Jin^{2,3} Chi-Man Pun^{1,†}

¹University of Macau, Macau SAR, China

²School of Computer Science, Peking University, Beijing, China

³Key Laboratory of High Confidence Software Technologies (Peking University),
 Ministry of Education, Beijing, China

Abstract

Feed-forward transformer models have driven remarkable progress in 3D vision, yet their quadratic complexity renders them impractical for long video sequences. Streaming 3D reconstruction addresses this bottleneck by processing frames sequentially with constant memory. However, existing recurrent architectures inevitably suffer from progressive degradation over long sequences, as they fail to account for the intrinsic reliability of observations and the accumulated confidence of the historical state, leading to severe error accumulation. To address these issues, we introduce BAT3R, a training-free Bayesian adaptation framework comprising two synergistic modules: State Innovation Gating (SIG) and Bayesian-guided Adaptive State Updating (BASU). SIG acts as a distributional regularizer that anchors state evolution to its geometric initialization, effectively arresting representational drift during long-term inference. Building on this, BASU reformulates state update as a recursive Bayesian inference process within a Kalman filtering framework. By jointly modeling the temporal reliability of observations via information-theoretic gain and spatial relevance via cross-attention maps, BASU achieves an asymmetric spatio-temporal state fusion: well-consolidated historical regions are protected from redundant updates, while genuinely novel scene content is aggressively incorporated. Extensive experiments on standard benchmarks (7-Scenes, TUM, ScanNet, KITTI, etc.) demonstrate that BAT3R consistently outperforms state-of-the-art online 3D reconstruction methods, with the most significant gains observed on long sequences and complex scenes.

1 Introduction

Real-time 3D scene understanding from monocular video streams, as illustrated in Fig. 1, is a foundational challenge in computer vision, with direct applications in robotic navigation and augmented reality. Recent feed-forward transformer models [45, 42, 57], such as DUST3R [45] and VGGT [42], have achieved remarkable reconstruction accuracy through dense global attention [38] that jointly reasons across all input frames; however, the computational and memory cost of these methods grows quadratically with sequence length, rendering them impractical for continuous perception over arbitrarily long videos. Streaming reconstruction methods [13, 14, 49, 54] address this scalability issue by incrementally processing frames and compressing historical context into a fixed-length

*Equal contribution.

†Corresponding authors: Yuanpeng He (heyuanpeng@stu.pku.edu.cn) and Chi-Man Pun (cmpun@um.edu.mo).

recurrent latent state, thereby achieving constant-memory spatial perception; yet existing recurrent frameworks [8, 14, 41] inevitably suffer from long-range performance decay, as the continuous update of scene states without revisiting prior observations causes cumulative structural drift and progressive distortion of global geometry.



Figure 1: **Robust Online 3D Reconstruction.** Given a stream of RGB images as input, our **BAT3R** framework enables dense 3D reconstruction in an online, continuous manner, estimating both camera parameters and dense 3D geometry with each incoming frame.

Test-Time Training (TTT) has recently emerged as a compelling paradigm for efficient, linear-complexity sequence modeling. In the context of spatial perception, TTT3R [10] pioneers this formulation by conceptualizing the persistent scene state as dynamic “fast weights” [32, 33] that are recursively updated during inference. However, while TTT3R employs an attention-based soft gating mechanism to modulate these updates, this heuristic approach lacks strict probabilistic bounds. Consequently, over long-horizon sequences, these unconstrained recursive updates inevitably suffer from progressive representational drift.

Mitigating such cumulative distortion requires a principled rethinking of state integration. On one hand, as evidenced by classical visual odometry [24, 47] and recent video generation frameworks [21, 30, 59], sequential drift can be explicitly bounded by anchoring recurrent updates to a reliable global reference (e.g., the initial frame), which prevents the latent state from deviating outside the valid geometric manifold. On the other hand, from a probabilistic standpoint, heuristic gating fundamentally deviates from the rigorous principles of Bayesian state estimation [2]. This uncertainty-aware paradigm has recently demonstrated profound success in suppressing drift across diverse vision tasks, including dynamic NeRFs [56], streaming 3D Gaussians [50], and video diffusion [52]. By reformulating the update process within a Bayesian framework, the integration of novel information is no longer arbitrary; rather, it is strictly governed by an optimal balance between the prior covariance, which encodes the accumulated confidence of the historical state, and the measurement noise, which quantifies the intrinsic uncertainty of incoming observations.

While reformulating the recurrent update as Bayesian inference offers a principled solution to sequential drift, calibrating the requisite noise models conventionally demands domain-specific training. However, we observe that spatial and temporal uncertainty signals are already inherently encoded within the pretrained transformer. Specifically, cross-attention alignment implicitly captures spatial visibility, while the prediction head’s confidence maps provide per-pixel proxies for observation reliability. Building upon this insight, we propose BAT3R, a training-free Bayesian adaptation framework for uncertainty-aware state updating in streaming 3D reconstruction, which comprises two synergistic modules (see Fig. 2). First, State Innovation Gating (SIG) acts as a distributional regularizer that explicitly anchors the recurrent state evolution to its geometric initialization, bounding spatial divergence by suppressing updates that fall outside the reliable latent manifold. Second, Bayesian-guided Adaptive State Updating (BASU) addresses optimal observation fusion by reformulating state integration as recursive Bayesian inference. By jointly modeling temporal observation reliability through information-theoretic gain and spatial relevance via cross-attention aggregation, BASU achieves an asymmetric spatio-temporal fusion, protecting well-consolidated memory from redundant updates while preserving full plasticity for novel scene content. Extensive experiments across seven standard benchmarks demonstrate that BAT3R substantially improves sequence length

generalization. BAT3R reduces camera pose ATE by up to **67.8%** and **32.3%** relative to CUT3R and TTT3R, respectively, while maintaining the exact similar computational cost as the baselines.

2 Related Works

2.1 Streaming Online Reconstruction Foundation Models.

To improve scalability, streaming methods progressively register local geometries onto a canonical 3D space, mitigating computational overhead from redundant multi-view observations. StreamVGGT [64] introduces a causal transformer for streaming 4D spatio-temporal reconstruction, employing temporal causal attention and KV-caching [28] as implicit memory for incremental historical context integration. However, it inherits the scaling bottlenecks of VGGT [42], incurring excessive resource overhead as frame count increases. STream3R [55] reformulates 3D perception as a decoder-only task, leveraging causal attention for scalable reconstruction in both static and dynamic environments. Spann3R [41] extends the DUST3R paradigm with an external spatial memory to track 3D context, enabling direct regression of pointmaps into a unified global coordinate system. Point3R [48] uses hierarchical 3D position embeddings and an explicit pointer-based memory to consolidate streaming visual information into a persistent scene representation. Recurrent architectures [8, 14, 41], exemplified by CUT3R, maintain a persistent latent memory to reconcile historical context with new observations recursively. While they stabilize inference memory usage, the inherent forgetting problem [11, 22] in recurrent units compromises long-term consistency, causing significant performance degradation as input views accumulate. More relevant models are provided in A.1.

2.2 Test-Time Training for 3D Reconstruction.

The $O(N^2)$ complexity of standard Transformers [40] has motivated the development of linear-complexity architectures. Modern RNNs and Linear Transformers [25], such as Mamba [20, 16] and DeltaNet [32], maintain a finite-sized hidden state as an efficient alternative. To counter information decay, these models incorporate advanced gating mechanisms, including input-dependent selection [20] and delta-rule updates [53], which regulate information flow across extended sequences. Paradigms such as Titans [4] and TTT [37, 60] reframe state updates as test-time optimization, enabling hidden states to function as fast weights [37, 43] for rapid online adaptation [29]. Recent works have ported the TTT paradigm to 3D reconstruction, achieving linear-time inference by compressing historical context [17] into fixed-capacity states. TTT3R [10] reinterprets recurrent state update as an online learning problem, using test-time optimization to iteratively align historical geometric memory with incoming observations. ZipMap [23] employs a stateful architecture driven by test-time training to continuously compress dense pointmap priors, enabling scalable multi-view association across extended video streams. To mitigate catastrophic forgetting in recurrent frameworks, subsequent research has introduced training-free adaptive mechanisms. TTSA3R [62] dynamically modulates update intensity based on temporal scene evolution, while MeMix [18] partitions the recurrent state into independent patches to prevent transient noise from corrupting stable structural components. In contrast, BAT3R replaces heuristic updates with a principled Bayesian filter that mathematically models the confidence of spatial observations, achieving robust probabilistic state updates.

3 Method

3.1 Preliminaries

TTT3R [10] builds on CUT3R [44] and treats streaming 3D reconstruction as test-time training: the recurrent state $\mathbf{S}_{t-1} \in \mathbb{R}^{N \times D}$ acts as fast weights [37, 32], one vector per spatial token, updated at inference while the frozen encoder-decoder maps each frame. Each image I_t and ray map R_t are encoded into image tokens $\mathbf{X}_t \in \mathbb{R}^{M \times D}$ (with M patches) and Plücker ray embeddings P_t , which are fed into the decoder together. A single recurrent decoder maps $(\mathbf{S}_{t-1}, \mathbf{X}_t)$ through coupled streams: image-side tokens $\mathbf{Y}_t$ pass through a detokenizer to obtain the dense pointmap $\tilde{O}_t \in \mathbb{R}^{H \times W \times 3}$, while the state stream produces a novel state $\mathbf{S}_t$ aligned with the TTT interpretation of [10]. Abstractly, a test-time step subtracts a scaled pseudo-gradient from the previous fast weights:

$$\mathbf{S}_t = \text{Update}(\mathbf{S}_{t-1}, \mathbf{X}_t) = \mathbf{S}_{t-1} - \beta_t \nabla(\mathbf{S}_{t-1}, \mathbf{X}_t), \quad (1)$$

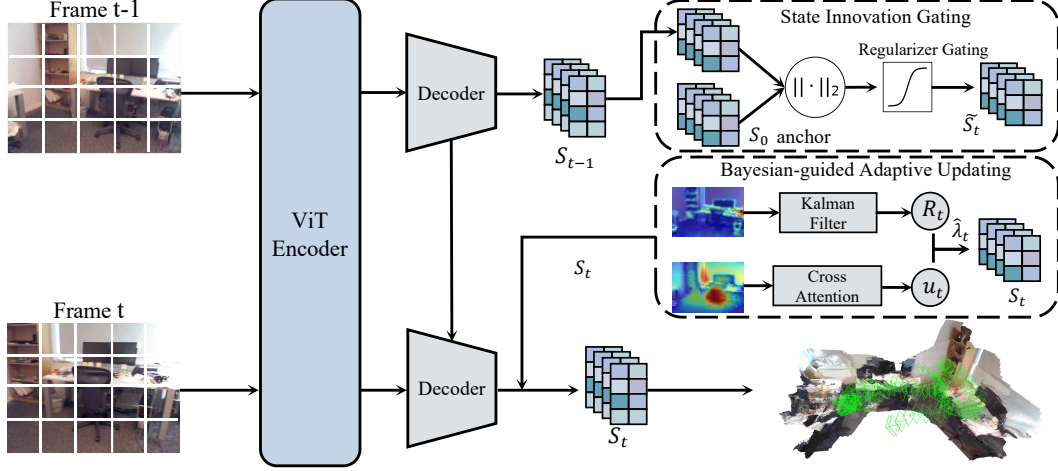


Figure 2: Overview of the BAT3R framework consisting of two modules State Innovation Gating (SIG) and Bayesian-guided Adaptive State Updating (BASU). SIG prevents representational drift by anchoring updates to the initial geometric state, while BASU optimizes observation fusion through a recursive Kalman filtering framework. By balancing observation reliability with historical confidence, the framework achieves robust, uncertainty-aware 3D scene reconstruction.

where $\nabla(\mathbf{S}_{t-1}, \mathbf{X}_t)$ denotes the pseudo-gradient of the fast-weight update given the previous state and the current observation, and β_t scales the descent step. TTT3R instantiates this term by cross-attention in which $\mathbf{S}_{t-1}$ supplies the queries and $\mathbf{X}_t$ supplies the keys and values:

$$\nabla(\mathbf{S}_{t-1}, \mathbf{X}_t) = \text{softmax}(\mathbf{Q}_{\mathbf{S}_{t-1}} \mathbf{K}_{\mathbf{X}_t}^\top) \mathbf{V}_{\mathbf{X}_t}, \quad (2)$$

where $\mathbf{Q}_{\mathbf{S}_{t-1}} = \mathbf{S}_{t-1} \mathbf{W}_Q$, $\mathbf{K}_{\mathbf{X}_t} = \mathbf{X}_t \mathbf{W}_K$, and $\mathbf{V}_{\mathbf{X}_t} = \mathbf{X}_t \mathbf{W}_V$ for learnable $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$. Instead of using a fixed scalar for β_t , TTT3R derives a per-token gate β_t via an element-wise sigmoid applied to the update signal, merging the candidate state with the previous state via a gated exponential moving average:

$$\beta_t = \sigma(\nabla(\mathbf{S}_{t-1}, \mathbf{X}_t)), \quad \mathbf{S}_t = (1 - \beta_t) \odot \mathbf{S}_{t-1} + \beta_t \odot \mathbf{S}_t. \quad (3)$$

Despite the soft-gating mechanism, the empirical updates in Eq. (3) lack explicit probabilistic bounds, leading to two critical bottlenecks over long horizons. First, without a global anchor, the unconstrained recursive updates cause the fast weights to progressively drift from their geometric initialization $\mathbf{S}_0$. Second, determining the gate β_t solely via spatial alignment ignores the intrinsic observation noise and historical state certainty, frequently overwriting well-consolidated memory.

3.2 State Innovation Gating

Error accumulation, as discussed above, stems from the absence of a fixed global reference in recurrent streaming, leading to a progressive departure of fast weights from the $\mathbf{S}_0$ initialization. Consequently, the latent state migrates outside the reliable regime of the frozen decoder, causing performance degradation. Validated by prior successes in visual odometry [24] and video modeling [30], we propose *State Innovation Gating* (SIG) to constrain state evolution. By anchoring each token of the recurrent state to the first-frame state $\mathbf{S}_0$, SIG measures the innovation residual and adaptively merges the decoder candidate $\mathbf{S}'_t$ with $\mathbf{S}_{t-1}$ via a gated convex combination.

For each token i , we formalize the innovation residual relative to the anchor $\mathbf{S}_0$ as:

$$\delta_t^{(i)} = \|\mathbf{S}_t^{(i)} - \mathbf{S}_0^{(i)}\|_2. \quad (4)$$

To maintain a consistent optimization landscape, we normalize this residual by the mean ℓ_2 norm of the previous state, r_t , thereby avoiding scale inflation induced by recursive drift:

$$r_t = \text{mean}_i(\|\mathbf{S}_{t-1}^{(i)}\|_2). \quad (5)$$

The adaptation process is modulated by a shifted sigmoid gate $g_t^{(i)}$:

$$g_t^{(i)} = \text{clip}\left(\sigma\left(-\kappa \cdot \frac{\delta_t^{(i)}}{r_t} + \tau\right), \eta, 1.0\right). \quad (6)$$

Specifically, large deviations drive the gate toward its lower bound η , favoring the historical state $\mathbf{S}_{t-1}$ to curb divergence, whereas small ratios permit the candidate $\mathbf{S}'_t$ to pass with minimal modification. The final state $\tilde{\mathbf{S}}_t$ is computed as:

$$\tilde{\mathbf{S}}_t = \mathbf{S}'_t \odot g_t + \mathbf{S}_{t-1} \odot (\mathbf{1} - g_t). \quad (7)$$

In the formulation, tokens that have drifted significantly from $\mathbf{S}_0$ are pulled back toward the previous state $\mathbf{S}_{t-1}$, while those remaining close to their initialization are permitted to pass through with minimal modification. The hyperparameter η ensures that no token is completely frozen, maintaining a baseline level of plasticity. SIG requires no learnable parameters beyond these heuristic constraints, and its output $\tilde{\mathbf{S}}_t$ serves as the refined observation for the subsequent BASU stage. The formal stability analysis and the proof of contraction toward the geometric anchor are detailed in [C.1](#).

3.3 Bayesian-guided Adaptive State Updating

Given the drift-regularized candidate $\tilde{\mathbf{S}}_t$, determining the optimal update magnitude for each state token becomes the central challenge. The heuristic gate β_t in TTT3R relies exclusively on attention alignment, overlooking crucial factors: it neither assesses the intrinsic reliability of the prediction nor accounts for the accumulated confidence of the existing state token, and fails to distinguish whether a given token is even visible in the current frame. We address these two concerns jointly.

We augment the deterministic state with a per-token uncertainty $\Sigma_t \in \mathbb{R}^N$, establishing a Gaussian belief $q_t = \mathcal{N}(\mu_t, \text{diag}(\Sigma_t))$ over the N state tokens. Here $\mathbf{S}_0 \in \mathbb{R}^{N \times D}$ denotes the scene state obtained by encoding the first frame through the image encoder and projecting into the state space, and $\Sigma_0 = \sigma_0^2 \mathbf{1}_N$ with $\sigma_0^2 = 10^3$, reflecting complete prior ignorance. Σ_t functions as a historical ledger of observation quality, progressively decreasing as reliable frames are accumulated over time.

To execute the Bayesian update, we must quantify both the unreliability of the incoming observation and the dynamic instability of the current state. First, we leverage the prediction head’s confidence map C_t as a data-driven proxy for per-pixel depth reliability, which is inversely mapped to a per-token observation noise covariance:

$$\mathbf{R}_t = (C_t^\gamma + \epsilon)^{-1}, \quad (8)$$

so that well-textured regions yield low $\mathbf{R}_t$ (high observation confidence) and occluded or textureless areas yield high $\mathbf{R}_t$ (low observation confidence). The exponent γ softens the confidence-to-noise mapping, preventing overly aggressive down-weighting of moderately uncertain observations. An adaptive process noise proportional to the innovation magnitude encodes inter-frame motion:

$$\mathbf{Q}_t = e^{\alpha_Q} \cdot (1 + \alpha_1 \cdot \text{softplus}(\|\tilde{\mathbf{S}}_t - \mu_{t-1}\|)), \quad \Sigma_{t|t-1} = \Sigma_{t-1} + \mathbf{Q}_t, \quad (9)$$

so rapid motion inflates $\Sigma_{t|t-1}$ and loosens the filter, while static scenes keep it tight. From $\Sigma_{t|t-1}$ and $\mathbf{R}_t$, we compute the Kalman gain as the optimal fusion weight:

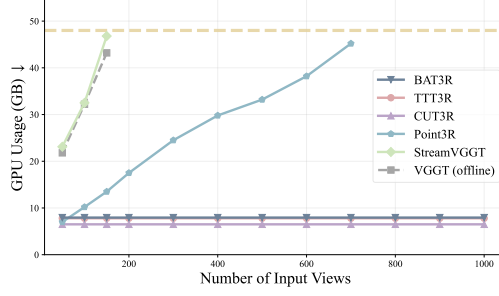
$$\mathbf{K}_t = \omega \cdot \frac{\Sigma_{t|t-1}}{\Sigma_{t|t-1} + \mathbf{R}_t} + (1 - \omega) \cdot \frac{\sqrt{\Sigma_{t|t-1}}}{\sqrt{\Sigma_{t|t-1}} + \sqrt{\mathbf{R}_t}}, \quad (10)$$

where $\omega \in [0, 1]$ is a weight that interpolates between the standard ratio form term and a square-root term. The standard ratio naturally follows from Gaussian Bayesian inference, while the square-root variant offers improved numerical stability when $\Sigma_{t|t-1}$ and $\mathbf{R}_t$ differ by orders of magnitude. $\mathbf{K}_t$ drives aggressive updates ($\mathbf{K}_t \rightarrow 1$) only when the state is highly uncertain and the observation is reliable; otherwise it conservatively preserves the existing memory. The posterior updates:

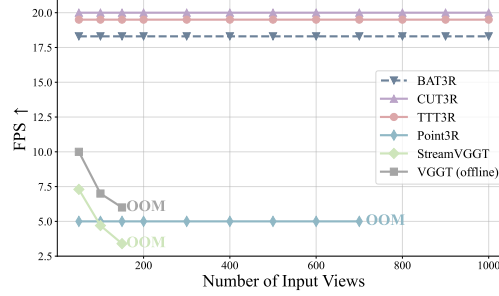
$$\mu_{t|t} = \mu_{t-1} + \mathbf{K}_t \cdot (\tilde{\mathbf{S}}_t - \mu_{t-1}), \quad (11)$$

$$\Sigma_{t|t} = (\mathbf{I} - \mathbf{K}_t) \Sigma_{t|t-1}, \quad \Sigma_{t|t} \geq \rho \Sigma_{t|t-1}. \quad (12)$$

Each reliable integration reduces $\Sigma_{t|t}$ toward its floor, making well-observed tokens progressively more resistant to future noisy updates. We provide theoretical justifications for the uncertainty boundedness and the optimality of this Bayesian fusion process in [C.2](#) and [C.3](#), respectively. We



(a) Comparison of GPU memory consumption (GB)



(b) Comparison of processing speed (FPS)

Figure 3: **Evaluation of computational efficiency on ScanNet.** (a) GPU memory and (b) inference speed across varying numbers of input views.

quantify the information gain of an update by computing the symmetric Kullback-Leibler (KL) divergence between the posterior and the prior:

$$\text{IG}_t = \frac{1}{2} \left[\frac{\Sigma_{t|t-1}}{\Sigma_{t|t}} + \frac{\Sigma_{t|t}}{\Sigma_{t|t-1}} - 2 \right] + \frac{(\mu_{t|t} - \mu_{t-1})^2}{2} (\Sigma_{t|t-1}^{-1} + \Sigma_{t|t}^{-1}). \quad (13)$$

The posterior statistics are squashed into a per-token reliability gate $\lambda_t \in (0, 1)$ via a pre-calibrated sigmoid function. Consequently, updates that are anomalously large (likely outliers) or negligibly small (redundant observations) are severely penalized, whereas genuinely informative updates are substantially integrated. However, even high-fidelity observations are strictly bounded by the current camera frustum; applying the temporal Bayesian gain λ_t uniformly across the entire state risks corrupting well-consolidated memory residing outside the active field of view. To circumvent this, we modulate λ_t with a spatial relevance weight derived from the decoder’s intrinsic cross-attention maps. Since the recurrent state assimilates novel geometry by querying incoming visual tokens, these attention affinities inherently capture spatial visibility. Formally, for a given state token i , we compute an aggregated visibility score $a_t^{(i)}$ by accumulating its cross-attention probabilities to all incoming image tokens across all L layers and H heads. The score is subsequently projected into a normalized spatial relevance weight $\mathbf{u}_t^{(i)} = \sigma(a_t^{(i)}) \in [0, 1]$. Ultimately, the temporal reliability gate and the spatial relevance weight are synergized into a unified update mask:

$$\hat{\lambda}_t = \lambda_t \odot (\alpha + (1 - \alpha) \cdot \mathbf{u}_t), \quad (14)$$

where $\alpha \in [0, 1]$ preventing complete token exclusion. The state is finally formulated as:

$$\mathbf{S}_t = \hat{\lambda}_t \odot \mu_{t|t} + (1 - \hat{\lambda}_t) \odot \mathbf{S}_{t-1}. \quad (15)$$

Tokens with high temporal reliability and high spatial coverage receive the strongest Bayesian update, whereas tokens that are either unreliably observed or outside the current frustum are safely preserved from $\mathbf{S}_{t-1}$. To close the recursive loop, the state covariance is updated as $\Sigma_t = \hat{\lambda}_t \Sigma_{t|t} + (1 - \hat{\lambda}_t) \Sigma_{t|t-1}$ and carried forward.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate BAT3R on three tasks across Seven benchmarks. For camera pose estimation, we evaluate BAT3R on TUM dynamics [36], ScanNet [15], and Sintel [7] datasets. TUM dynamics contains indoor RGB-D sequences captured with a hand-held camera, including both static scenes and dynamic scenes with moving people. ScanNet contains large-scale indoor RGB-D sequences captured across diverse room types with significant viewpoint variation. Sintel contains long synthetic sequences rendered from a 3D animated film, featuring complex motions, dynamic objects, and challenging lighting conditions. For video depth estimation, we evaluate BAT3R on Bonn [31], KITTI [19], and Sintel [7] datasets. Bonn contains highly dynamic indoor RGB-D sequences, featuring complex human activities, fast-moving objects, and varying camera motions. KITTI contains large-scale outdoor driving sequences captured from a moving vehicle, including real-world street views with dynamic entities such as pedestrians and cars. For 3D reconstruction, we evaluate on 7-Scenes [35] and NRGBD [3], which contains indoor RGB-D sequences of seven distinct room categories captured with a Kinect sensor, commonly used to assess geometric reconstruction accuracy.

Baselines. We evaluate BAT3R against state-of-the-art online 3D reconstruction baselines, including the recurrent CUT3R [44] and its long-sequence extensions, Point3R [48] and StreamVGGT [64], which rely on explicit spatial memory and KV-caching mechanisms. Furthermore, we compare against recent training-free Test-Time Training (TTT) frameworks, such as TTT3R [10] and TTSA3R [62], which utilize heuristic gating and spatio-temporal alignment for adaptive state updates.

4.2 Camera Pose Estimation

Consistent with established evaluation protocols [58], we assess camera pose estimation on the TUM dynamics [36] and ScanNet [15] datasets. Performance is measured using the standard Absolute Translation Error (ATE) following Sim(3) trajectory alignment [39] between the estimated and ground-truth trajectories. The long-sequence evaluation results are presented in Fig. 4. For a fair comparison, we adopt the experimental settings and baseline data established by TTT3R [10]. Empirical evaluations demonstrate that BAT3R consistently bounds representational drift, sustaining high-accuracy pose estimation across widely varying sequence lengths ranging from 50 to 1000 frames. When scaling to 1000-frame sequences, BAT3R reduces ATE by 67.8% on ScanNet and 66.1% on TUM dynamics compared to the CUT3R baseline. Even when compared to the competitive TTT3R, BAT3R achieves a further error reduction of 32.3% and 44.4% on ScanNet and TUM, respectively. Besides, we also conduct comprehensive evaluations of GPU memory consumption and inference speed on the ScanNet dataset in Fig. 3. As is shown in Fig. 3a, 3b, BAT3R achieves competitive GPU memory and inference speed compared with previous methods. Additional quantitative comparisons against offline methods on short sequences and qualitative visualizations of the estimated camera trajectories can be found in B.3 and B.4.

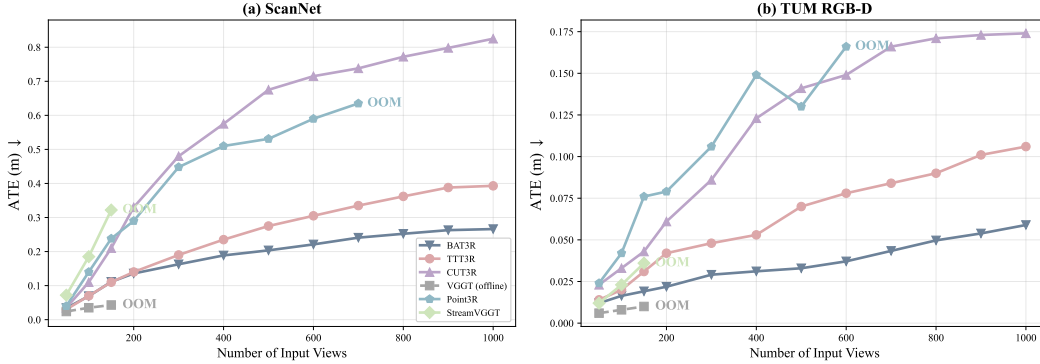


Figure 4: Comparison of Camera Pose Estimation. Results on ScanNet (left) and TUM-D (right). OOM denotes the method out-of-memory beyond this point.

4.3 Video Depth Estimation

Following standard evaluation protocols [58], we assess video depth estimation on the KITTI [19] and Bonn [31] datasets, which encompass a variety of indoor and outdoor driving environments. We adopt Absolute Relative Error (Abs Rel) and $\delta < 1.25$ as our primary performance metrics. Following the evaluation protocol established in TTT3R [10], relative depth accuracy and temporal consistency are gauged by aligning predictions to the ground truth using a per-sequence scale, while metric-scale performance is assessed directly. As illustrated in the long-sequence evaluations in Fig. 5 and Fig. 9, BAT3R method consistently outperforms existing recurrent baselines. On the Bonn dataset, BAT3R achieves a substantial improvement, reducing the Abs Rel by approximately 15.5% compared to the CUT3R baseline at a sequence length of 500 frames. Similarly, on the KITTI dataset, BAT3R effectively bounds progressive state drift, yielding a 24.5% and 13.6% reduction in Abs Rel relative to the CUT3R and TTT3R baselines at a sequence length of 500 frames. Furthermore, as the sequence length increases, BAT3R mitigates the temporal degradation in the $\delta < 1.25$ metric, sustaining a more stable and consistently higher $\delta < 1.25$ across all sequence lengths.

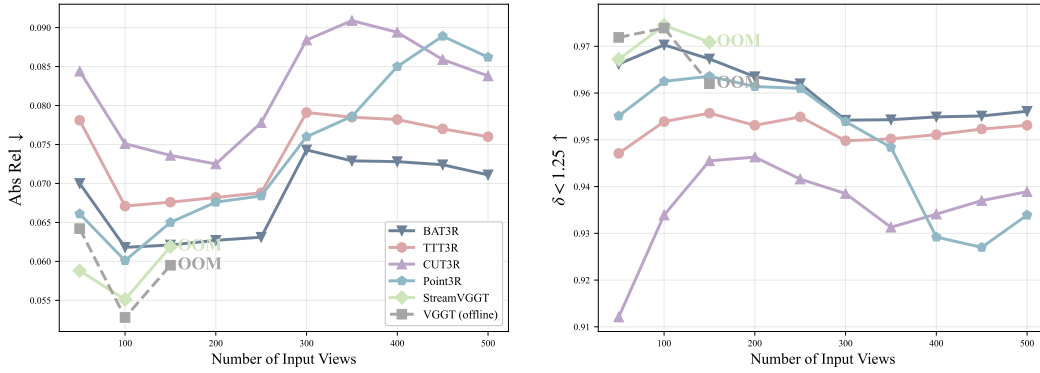


Figure 5: Scale-invariant relative depth evaluation on Bonn dataset.

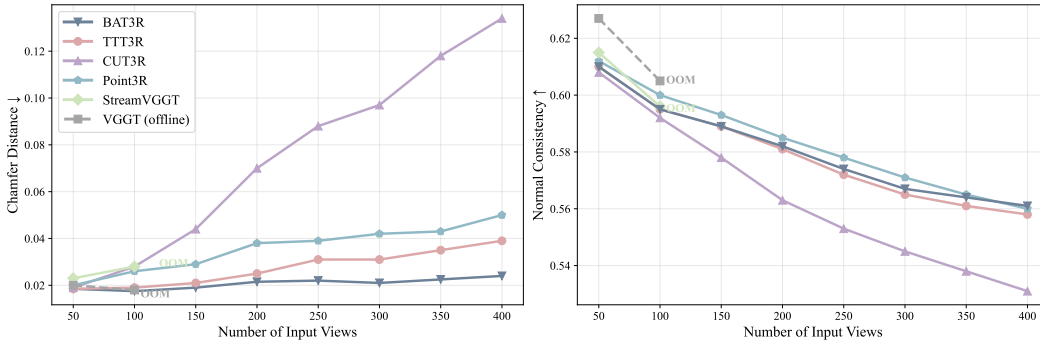


Figure 6: Comparison of 3D Reconstruction on 7-scene.

4.4 3D Reconstruction

Following established protocols [41, 44, 12], we evaluate 3D reconstruction on the 7-Scenes [35] dataset. Geometric accuracy is quantified through Chamfer Distance which denotes the average of accuracy and completeness, and Normal Consistency (NC). In contrast to the sparse sampling, we conduct evaluations on extended image sequences to rigorously assess the models' capacity for long-term spatial memorization. As illustrated in the long-sequence evaluations in Fig. 6, our BAT3R method consistently outperforms existing recurrent baselines. On the 7-Scenes benchmark, BAT3R achieves a substantial improvement, reducing the Chamfer Distance (from 0.134 to 0.024) by approximately 82.1% compared to the CUT3R baseline at a sequence length of 400 frames. Furthermore, when compared to the highly competitive TTT3R, BAT3R achieves a further reduction of approximately 38.5% (from 0.039 to 0.024) while maintaining a slight but consistent lead in Normal

Consistency (e.g., 0.561 vs. 0.558). We provide qualitative comparisons of these reconstruction results in Fig. 7. Furthermore, the detailed quantitative results on the 7-Scenes and NRGBD benchmarks across various sequence lengths are provided in B.5, demonstrating that BAT3R consistently yields substantial error reductions compared to SOTA baselines (TTSA3R, TTT3R, and CUT3R).

Table 1: Comprehensive Ablation Study on Camera Pose and Video Depth Estimation. We evaluate the impact of different components across four standard benchmarks.

Setting	Camera Pose Estimation						Video Depth Estimation			
	TUM-dynamics (90 frames)			ScanNet (90 frames)			Bonn (100 frames)		KITTI (100 frames)	
	ATE ↓	RPE trans ↓	RPE rot ↓	ATE ↓	RPE trans ↓	RPE rot ↓	Abs Rel ↓	$\delta < 1.25 \uparrow$	Abs Rel ↓	$\delta < 1.25 \uparrow$
BAT3R (Full)	0.016	0.008	0.287	0.058	0.022	0.852	0.061	97.0	0.100	89.2
w/o BRU	0.024	0.008	0.288	0.099	0.021	0.562	0.062	96.5	0.106	87.9
w/o SIG	0.016	0.008	0.289	0.061	0.025	1.163	0.065	95.6	0.105	88.8
w/o Attention	0.017	0.008	0.260	0.060	0.024	0.988	0.063	96.8	0.105	88.1
w/o Q	0.017	0.009	0.300	0.077	0.028	1.206	0.061	96.7	0.104	88.5

4.5 Ablation Studies

Effect of Each Component. To evaluate the individual components of BAT3R, we conduct comprehensive ablation studies on camera pose and depth estimation (Table 1). Removing the Bayesian Recursive Update (BRU) causes the most severe degradation (ScanNet ATE surging from 0.058 to 0.099), validating its necessity in optimally balancing historical confidence against observation noise. Omitting the State Innovation Gating (SIG) increases Abs Rel from 0.061 to 0.065 on Bonn, confirming its critical role as a distributional regularizer to bound representational drift. Furthermore, excluding cross-attention compromises spatial awareness, dropping the KITTI $\delta < 1.25$ metric from 89.2 to 88.1. Finally, the absence of adaptive process noise (Q_t) degrades performance (ATE increasing to 0.077), verifying that motion-aware uncertainty dynamically prevents the Kalman filter from becoming overly restrictive.

Parameter Analysis. Furthermore, we conduct hyperparameter sensitivity experiments on the TUM-dynamics dataset to assess the impact of the update floor η , scaling factor κ , and threshold τ . The detailed experimental results and analysis are provided in Appendix. As illustrated in Fig. 10, the Average ATE is least sensitive to the floor η , whereas κ and τ both exhibit a sharp initial performance gain that plateaus once they exceed 1.0 and 5.0, respectively. Furthermore, additional exhaustive analysis of both SIG and BASU hyperparameters, including the scaling factors for process noise and interpolation weights, is detailed in B.6.

5 Conclusion

In this work, we present BAT3R, a training-free Bayesian adaptation framework that mitigates catastrophic forgetting and representational drift in streaming 3D reconstruction. BAT3R comprises two synergistic modules: State Innovation Gating (SIG), which acts as a distributional regularizer to anchor state evolution to its geometric initialization, and Bayesian-guided Adaptive State Updating (BASU), which reformulates state integration as a recursive Bayesian inference process within a Kalman filtering framework. By dynamically balancing the reliability of incoming observations against the accumulated confidence of historical memory, BAT3R achieves an asymmetric spatio-temporal state fusion that protects well-consolidated regions from redundant updates while preserving plasticity for genuinely novel content. Extensive experiments across seven standard benchmarks demonstrate consistent improvements over state-of-the-art online methods, with particularly strong gains on long sequences. BAT3R achieves these results without any training or fine-tuning, making it immediately applicable to any pretrained streaming reconstruction model.

Acknowledgments

This work was supported in part by the Science and Technology Development Fund, Macau SAR, under Grant 0193/2023/RIA3, 0079/2025/AFJ and 0052/2026/RIA1.

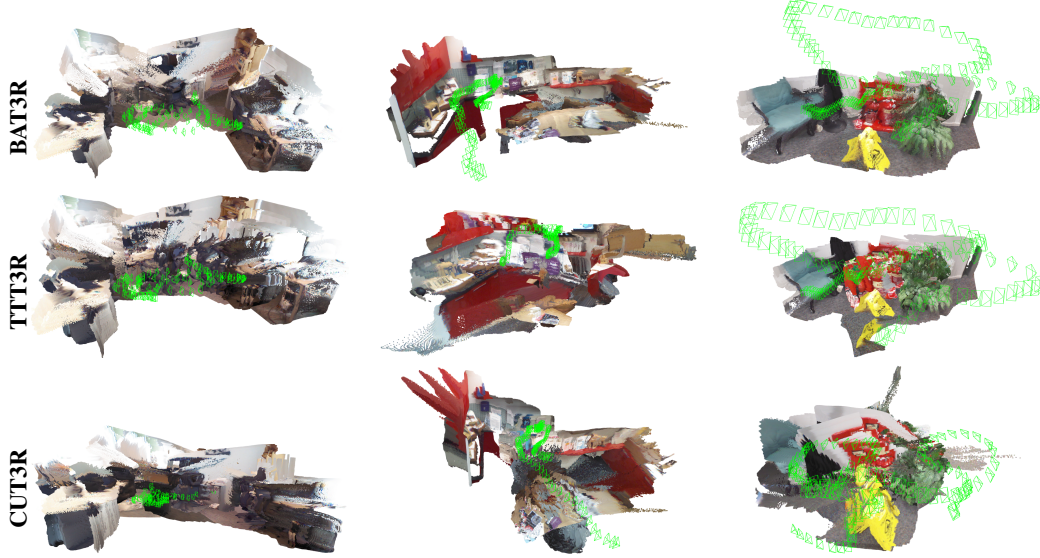


Figure 7: **Qualitative comparison of 3D reconstruction.** We compare BAT3R with TTT3R and CUT3R across three challenging sequences in the 7-Scene dataset.

References

- [1] Honggyu An, Jin Hyeon Kim, Seonghoon Park, Jaewoo Jung, Jisang Han, Sunghwan Hong, and Seungryong Kim. Cross-view completion models are zero-shot correspondence estimators. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1103–1115, 2025.
- [2] Peter Anderson, Ayush Shrivastava, Devi Parikh, Dhruv Batra, and Stefan Lee. Chasing ghosts: Instruction following as bayesian state tracking. *Advances in neural information processing systems*, 32, 2019.
- [3] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6290–6301, 2022.
- [4] Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*, 2024.
- [5] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6684–6692, 2017.
- [6] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Aron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. In *European Conference on Computer Vision*, pages 421–440. Springer, 2024.
- [7] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pages 611–625. Springer, 2012.
- [8] Yohann Cabon, Lucas Stofl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1050–1060, 2025.

- [9] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Easi3r: Estimating disentangled motion from dust3r without training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9158–9168, 2025.
- [10] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025.
- [11] Yingfa Chen, Xinrong Zhang, Shengding Hu, Xu Han, Zhiyuan Liu, and Maosong Sun. Stuffed mamba: Oversized states lead to the inability to forget. *arXiv preprint arXiv:2410.07145*, 2024.
- [12] Yue Chen, Xingyu Chen, Anpei Chen, Gerard Pons-Moll, and Yuliang Xiu. Feat2gs: Probing visual foundation models with gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6348–6361, 2025.
- [13] Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3r: Everyone everywhere all at once. *arXiv preprint arXiv:2510.06219*, 2025.
- [14] Zhuoguang Chen, Minghui Qin, Tianyuan Yuan, Zhe Liu, and Hang Zhao. Long3r: Long sequence streaming 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5273–5284, 2025.
- [15] 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*, pages 5828–5839, 2017.
- [16] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.
- [17] Benoit Dherin, Michael Munn, Hanna Mazzawi, Michael Wunder, and Javier Gonzalvo. Learning without training: The implicit dynamics of in-context learning. *arXiv preprint arXiv:2507.16003*, 2025.
- [18] Jiacheng Dong, Huan Li, Sicheng Zhou, Wenhao Hu, Weili Xu, and Yan Wang. Memix: Writing less, remembering more for streaming 3d reconstruction.
- [19] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013.
- [20] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.
- [21] Xuehan Hou, Meng Fan, Pengchong Qiao, Zesen Cheng, Yian Zhao, Lei Zhu, Kaiwen Cheng, Chang Liu, and Jie Chen. Anchor frame bridging for coherent first-last frame video generation. In *The Fourteenth International Conference on Learning Representations*.
- [22] Samy Jelassi, David Brandfonbrener, Sham M Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. *arXiv preprint arXiv:2402.01032*, 2024.
- [23] Haian Jin, Rundi Wu, Tianyuan Zhang, Ruiqi Gao, Jonathan T Barron, Noah Snavely, and Aleksander Holynski. Zipmap: Linear-time stateful 3d reconstruction with test-time training. *arXiv preprint arXiv:2603.04385*, 2026.
- [24] Graeme Jones. Accurate and computationally-inexpensive recovery of ego-motion using optical flow and range flow with extended temporal support. In *BMVC*, 2013.
- [25] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020.
- [26] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1621, 2021.

- [27] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pages 71–91. Springer, 2024.
- [28] Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.
- [29] Bo Liu, Rui Wang, Lemeng Wu, Yihao Feng, Peter Stone, and Qiang Liu. Longhorn: State space models are amortized online learners. *arXiv preprint arXiv:2407.14207*, 2024.
- [30] Zichi Liu, Yinggui Wang, Tao Wei, and Chao Ma. Anchorsync: Global consistency optimization for long video editing. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4494–4503, 2025.
- [31] Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguere, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7855–7862. IEEE, 2019.
- [32] Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International conference on machine learning*, pages 9355–9366. PMLR, 2021.
- [33] Jürgen Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992.
- [34] You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. Fastvggt: Training-free acceleration of visual geometry transformer. *arXiv preprint arXiv:2509.02560*, 2025.
- [35] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013.
- [36] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012.
- [37] Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, et al. Learning to (learn at test time): Rnns with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024.
- [38] Lorenzo Torresani, Aaron Hertzmann, and Chris Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE transactions on pattern analysis and machine intelligence*, 30(5):878–892, 2008.
- [39] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4):376–380, 2002.
- [40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [41] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pages 78–89. IEEE, 2025.
- [42] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.
- [43] Ke Alexander Wang, Jiaxin Shi, and Emily B Fox. Test-time regression: a unifying framework for designing sequence models with associative memory. *arXiv preprint arXiv:2501.12352*, 2025.

- [44] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025.
- [45] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20697–20709, 2024.
- [46] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3 : Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025.
- [47] Brian Williams and Ian Reid. On combining visual slam and visual odometry. In *2010 IEEE International Conference on Robotics and Automation*, pages 3494–3500. IEEE, 2010.
- [48] Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. *arXiv preprint arXiv:2507.02863*, 2025.
- [49] Haodong Xiang, Xinghui Li, Kai Cheng, Xiansong Lai, Wanting Zhang, Zhichao Liao, Long Zeng, and Xueping Liu. Gaussianroom: Improving 3d gaussian splatting with sdf guidance and monocular cues for indoor scene reconstruction. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2686–2693. IEEE, 2025.
- [50] Fuchen Yan, Luyang Tang, Shihe Shen, and Ronggang Wang. Online training of 3d gaussians for streamable dynamic scenes via kalman-guided state estimation. In *Proceedings of the 2025 7th International Conference on Image, Video and Signal Processing*, pages 1–8, 2025.
- [51] Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21924–21935, 2025.
- [52] Liudi Yang, George Eskandar, Fengyi Shen, Mohammad Altillawi, Yang Bai, Chi Zhang, Ziyuan Liu, and Abhinav Valada. Confctrl: Enabling precise camera control in video diffusion via confidence-aware interpolation. *arXiv preprint arXiv:2603.09819*, 2026.
- [53] Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. *arXiv preprint arXiv:2412.06464*, 2024.
- [54] Yuheng Yuan, Qiuhong Shen, Shizun Wang, Xingyi Yang, and Xinchao Wang. Test3r: Learning to reconstruct 3d at test time. *arXiv preprint arXiv:2506.13750*, 5, 2025.
- [55] LAN Yushi, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Bo Dai, Shuai Yang, Chen Change Loy, and Xingang Pan. Stream3r: Scalable sequential 3d reconstruction with causal transformer. In *The Fourteenth International Conference on Learning Representations*, 2026.
- [56] Yifan Zhan, Zhuoxiao Li, Muyao Niu, Zhihang Zhong, Shohei Nobuhara, Ko Nishino, and Yinqiang Zheng. Kfd-nerf: Rethinking dynamic nerf with kalman filter. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024.
- [57] Jiahui Zhang, Yuelei Li, Anpei Chen, Muyu Xu, Kunhao Liu, Jianyuan Wang, Xiao-Xiao Long, Hanxue Liang, Zexiang Xu, Hao Su, et al. Advances in feed-forward 3d reconstruction and view synthesis: A survey. *arXiv preprint arXiv:2507.14501*, 2025.
- [58] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024.
- [59] Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *arXiv preprint arXiv:2504.12626*, 2025.

- [60] Tianyuan Zhang, Sai Bi, Yicong Hong, Kai Zhang, Fujun Luan, Songlin Yang, Kalyan Sunkavalli, William T Freeman, and Hao Tan. Test-time training done right. *arXiv preprint arXiv:2505.23884*, 2025.
- [61] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T Freeman. Structure and motion from casual videos. In *European Conference on Computer Vision*, pages 20–37. Springer, 2022.
- [62] Zhijie Zheng, Xinhao Xiang, and Jiawei Zhang. Ttsa3r: Training-free temporal-spatial adaptive persistent state for streaming 3d reconstruction. *arXiv preprint arXiv:2601.22615*, 2026.
- [63] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8535–8546, 2025.
- [64] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025.

Appendix

A	Related Works	16
A.1	Offline Reconstruction Foundation Models.	16
B	Additional Results	16
B.1	Experimental Settings	16
B.2	Baselines	16
B.3	Camera Pose Estimation	18
B.4	Video Depth Estimation	18
B.5	3D Reconstruction	20
B.6	Parameter Analysis	20
C	Theoretical Analysis	21
C.1	Stability Analysis of SIG	22
C.2	Geometric Manifold Consistency	22
C.3	Bayesian Fusion Optimality	22
C.4	Information-Theoretic Gain and Update	23

A Related Works

A.1 Offline Reconstruction Foundation Models.

Recent advances in 3D vision have been driven by feed-forward foundation models that streamline the traditional reconstruction pipeline. DUST3R [45] pioneers this shift with an end-to-end framework that directly regresses pixel-aligned pointmaps [5, 6, 35] from image pairs, bypassing explicit feature matching [1, 9] and camera calibration. MAST3R [27] extends this with cross-view correspondences for 3D matching, though it remains limited to pairwise inputs and requires heavy global optimization for multi-view integration. Fast3R [51] leverages global self-attention to jointly reason across all input frames in a single forward pass, achieving strong accuracy in camera pose and 3D reconstruction without iterative alignment. VGGT [42] introduces a feed-forward framework that reconstructs 3D attributes from an arbitrary number of views in one pass. π^3 [46] employs a permutation-equivariant architecture to predict camera poses and point maps in a reference-free manner, ensuring robustness to input ordering. Despite inference acceleration through token merging or sparse attention [34], the $O(N^2)$ complexity of full attention remains a fundamental bottleneck: it routinely causes out-of-memory errors on consumer GPUs, limiting the maximum sequence length that can be processed.

B Additional Results

B.1 Experimental Settings

Evaluation Settings. To thoroughly evaluate our model, we establish two primary evaluation setups: (1) *Long-sequence evaluation* (results in the main paper), which challenges the recurrent state’s ability to maintain long-term memory and geometric consistency over hundreds of frames against other online methods. (2) *Short-sequence evaluation* (detailed in this supplement), which allows for benchmarking against offline or memory-intensive baselines that would otherwise fail due to out-of-memory (OOM) errors on longer videos.

Implementation Details. In all experiments, inference is conducted on a single NVIDIA RTX 4090 GPU (24 GB). BAT3R is built upon the pretrained CUT3R checkpoint and operates strictly without any fine-tuning or additional training. The underlying architecture consists of a ViT-Large encoder (24 layers, embedding dimension of 1024, 16 attention heads) coupled with a 12-layer recurrent decoder (embedding dimension of 768, 12 attention heads). At the beginning of each sequence, the Bayesian state variance is initialized to $1e^3$ per patch, reflecting a state of high initial uncertainty. Consistent with prior training-free adaptation frameworks [10, 62], we do not perform any test-time gradient updates; instead, continuous state adaptation is achieved entirely through our proposed Bayesian state update rules.

B.2 Baselines

We first compare BAT3R against pairwise 3D reconstruction foundation models (e.g., DUST3R [45], MAST3R [27], MonST3R [58], and Easi3R [9]), which take view pairs as input and require an extra global alignment stage to integrate their predictions. We also evaluate against AETHER [63] and VGGT [42], which simultaneously predict all pointmaps to achieve state-of-the-art 3D point and camera pose reconstruction. However, constrained by their reliance on global alignment or full-attention mechanisms, these methods are restricted to offline processing of short sequences, necessitating the re-inference of the entire sequence whenever a new frame arrives.

For online methods, we compare BAT3R with Spann3R [41] and CUT3R [44], which leverage RNN-based architectures to process streaming images on the fly. Additionally, we compare against the most relevant concurrent works, Point3R [48] and StreamVGGT [64]. To handle long sequences, these approaches fine-tune CUT3R and VGGT by introducing explicit pointmap memory and KV caching as their respective state representations. In contrast to these methods, TTT3R reformulate CUT3R from a Test-Time Training (TTT) perspective. Build on TTT3R, BAT3R achieves robust online perception by integrating State Innovation Gating (SIG) to constrain representational drift and Bayesian-guided Adaptive State Updating (BASU) for optimal, uncertainty-aware fusion, entirely eliminating the need to fine-tune CUT3R or train additional parameterized components.

Table 2: **Evaluation on Camera Pose Estimation- Short Sequence** on Sintel [7], TUM-dynamic [36], and ScanNet [15] datasets. TTT3R achieves the best overall performance among online methods, while its accuracy has not yet matched strong offline methods (e.g., VGGT), where full attention, despite being slower and more memory-demanding, preserves the entire history context.

Method	Online	Sintel (50 frames)			TUM-dynamics (90 frames)			ScanNet (90 frames)		
		ATE ↓	RPE trans ↓	RPE rot ↓	ATE ↓	RPE trans ↓	RPE rot ↓	ATE ↓	RPE trans ↓	RPE rot ↓
Robust-CVD [26]	✗	0.360	0.154	3.443	0.153	0.026	3.528	0.227	0.064	7.374
CasualSAM [61]	✗	0.141	0.035	<u>0.615</u>	0.071	0.010	1.712	0.158	0.034	1.618
DUSi3R [45]	✗	0.417	0.250	5.796	0.083	0.017	3.567	0.081	0.028	0.784
MAST3R [27]	✗	0.185	0.060	1.496	<u>0.038</u>	<u>0.012</u>	<u>0.448</u>	0.078	0.020	<u>0.475</u>
MonST3R [58]	✗	<u>0.111</u>	0.044	0.869	0.098	0.019	0.935	0.077	0.018	0.529
Easi3R [9]	✗	0.110	0.042	0.758	0.105	0.022	1.064	<u>0.061</u>	<u>0.017</u>	0.525
AETHER [63]	✗	0.189	0.054	0.694	0.092	<u>0.012</u>	1.106	0.176	0.028	1.204
VGGT [42]	✗	0.172	0.062	0.471	0.012	0.010	0.310	0.035	0.015	0.377
Spann3R [41]	✓	0.329	0.110	4.471	0.056	0.021	0.591	0.096	0.023	0.661
CUT3R [44]	✓	<u>0.213</u>	<u>0.066</u>	<u>0.621</u>	0.046	0.015	0.473	0.099	<u>0.022</u>	<u>0.600</u>
Point3R [48]	✓	0.351	0.128	1.822	0.075	0.029	0.642	0.106	0.035	1.946
StreamVGGT [64]	✓	0.251	0.149	1.894	0.061	0.033	3.209	0.161	0.057	3.647
TTT3R [10]	✓	0.201	0.063	0.61	<u>0.028</u>	<u>0.012</u>	<u>0.379</u>	<u>0.064</u>	0.021	0.592
BAT3R	✓	0.219	0.089	0.84	0.016	0.008	0.287	0.058	0.022	0.852

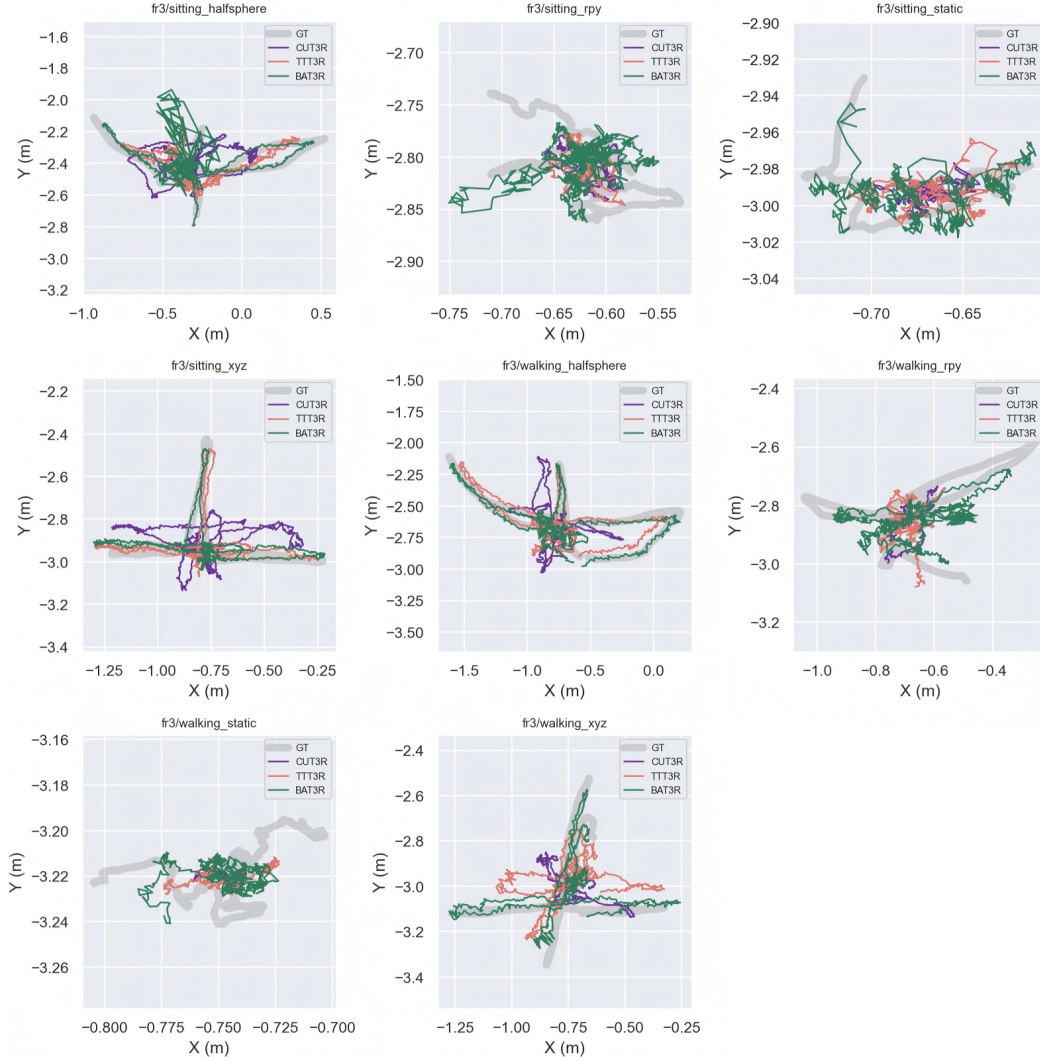


Figure 8: Qualitative comparison of eight estimated camera trajectories on the TUM RGB-D dataset.

B.3 Camera Pose Estimation

Experimental Settings: Following established protocols [44, 58], we evaluate our method on the Sintel, TUM-dynamics, and ScanNet datasets. After applying Sim(3) alignment to the estimated trajectories, we adopt Absolute Translation Error (ATE), Relative Translation Error (RTE), and Relative Rotation Error (RRE) as our evaluation metrics. The compared baselines encompass mainstream 3D reconstruction foundation models, as well as per-sequence optimization approaches, such as RobustCVD [26] and CasualSAM [61], which jointly optimize camera parameters and dense depth maps.

Quantitative Analysis: The experimental results for the long-sequence evaluation are illustrated in Fig. 4. Since many baseline methods are restricted to short sequences due to out-of-memory (OOM) constraints, we additionally compile the short-sequence comparisons in Table 2, explicitly categorizing the baselines into offline methods (which require global attention or extra optimization) and online methods. The quantitative results indicate that our method achieves leading performance among all online approaches on the TUM-dynamics and ScanNet datasets. Specifically, on the TUM-dynamics dataset, BAT3R achieves an Absolute Translation Error (ATE) of 0.016, representing a significant 65.2% error reduction compared to the CUT3R baseline (0.046) and a 42.9% improvement over the highly competitive TTT3R (0.028). Furthermore, BAT3R dominates all online methods in rotational and translational precision on this dataset (RPE trans of 0.008 and RPE rot of 0.287). Similarly, on the ScanNet dataset, BAT3R secures the lowest ATE among online methods at 0.058, yielding a 41.4% reduction relative to CUT3R (0.099) and a 9.4% reduction compared to TTT3R (0.064). Despite these substantial gains, we observe a remaining performance gap compared to the upper bound established by strong offline methods (e.g., VGGT), which preserve the entire historical context through full attention at the cost of significantly higher memory demands and slower processing speeds.

Qualitative Analysis: Visual comparisons of the camera trajectories in Fig. 8 demonstrate that BAT3R yields more accurate and robust pose estimation than TTT3R [10] and CUT3R [44]. Specifically, BAT3R (green) consistently maintains high fidelity to the ground truth across various motion patterns, effectively suppressing the cumulative drift observed in the baselines. This performance underscores the efficacy of our BASU module in ensuring global consistency during long-term reconstruction.

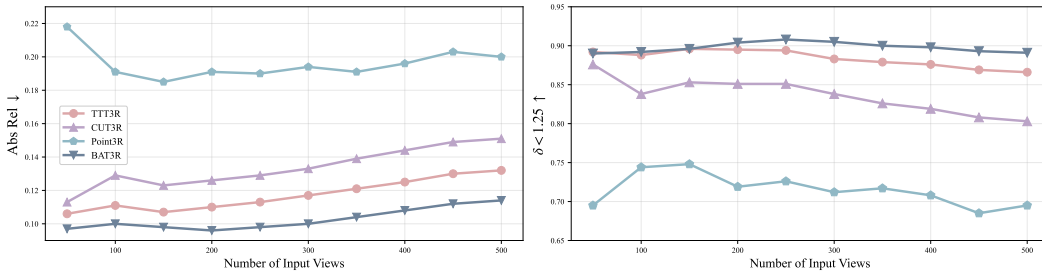


Figure 9: Metric depth evaluation on KITTI, excluding VGGT-based methods that don’t support metric depth.

B.4 Video Depth Estimation

Experimental Settings: Following standard protocols [44, 58], we assess video depth estimation across the KITTI, Sintel, and Bonn datasets, which encompass a diverse mix of dynamic, static, indoor, and outdoor environments. Our primary evaluation metrics are the Absolute Relative Error (Abs Rel) and the $\delta < 1.25$ threshold accuracy. To evaluate relative depth accuracy, capturing both per-frame fidelity and inter-frame consistency, we align the predicted depth maps with the ground truth using a per-sequence scale factor. Additionally, for models capable of generating metric pointmaps (i.e., absolute scale in meters), we report unaligned metrics to rigorously evaluate real-world absolute-scale accuracy.

Quantitative Results: The short-sequence evaluations presented in Table 3 confirm that our approach consistently maintains state-of-the-art or highly competitive performance among online frameworks

Table 3: **Evaluation of Video Depth Estimation - Short Sequence.** We report scale-invariant relative depth (aligned by a per-sequence scale) and metric scale absolute depth accuracy on Sintel [7], KITTI [19] and Bonn [31] datasets. BAT3R achieves competitive performance among online methods.

Alignment	Method	Online	Sintel (50 frames)		BONN (110 frames)		KITTI (110 frames)	
			Abs Rel ↓	$\delta < 1.25 \uparrow$	Abs Rel ↓	$\delta < 1.25 \uparrow$	Abs Rel ↓	$\delta < 1.25 \uparrow$
Per-sequence Scale	DUST3R [45]	✗	0.656	45.2	0.155	83.3	0.144	81.3
	MAST3R [27]	✗	0.641	43.9	0.252	70.1	0.183	74.5
	MonST3R [58]	✗	0.378	55.8	0.067	96.3	0.168	74.4
	Easi3R [9]	✗	0.377	55.9	0.059	97.0	0.102	91.2
	AETHER [63]	✗	0.324	50.2	0.273	59.4	0.056	97.8
	VGGT [42]	✗	0.287	66.1	0.055	97.1	<u>0.070</u>	<u>96.5</u>
	Spann3R [41]	✓	0.622	42.6	0.144	81.3	0.198	73.7
	CUT3R [44]	✓	0.421	47.9	0.078	93.7	0.118	88.1
	Point3R [48]	✓	0.452	48.9	0.060	96.0	0.136	84.2
	StreamVGGT [64]	✓	0.323	65.7	0.059	97.2	0.173	72.1
Metric Scale	TTT3R [10]	✓	0.404	50.0	0.068	95.4	<u>0.113</u>	<u>90.4</u>
	BAT3R (Ours)	✓	<u>0.401</u>	<u>50.1</u>	<u>0.063</u>	<u>96.8</u>	0.111	91.1
	MAST3R [27]	✗	1.022	14.3	0.272	70.6	0.467	15.2
	CUT3R [44]	✓	1.029	23.8	0.103	88.5	0.122	85.5
	Point3R [48]	✓	0.777	17.1	0.137	94.7	0.191	73.8
	TTT3R [10]	✓	0.977	24.5	0.090	<u>94.2</u>	<u>0.110</u>	<u>89.1</u>
	BAT3R (Ours)	✓	<u>0.930</u>	24.5	0.105	88.6	0.099	89.4

Table 4: 3D Reconstruction Results on 7-Scenes and NRGBD.

Model	Input	7-Scenes						NRGBD					
		Acc. ↓		Comp. ↓		NC ↑		Acc. ↓		Comp. ↓		NC ↑	
		Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
VGGT (Offline) [42]	300	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>
	400	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>
	500	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>
StreamVGGT [64]	300	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>
	400	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>
	500	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>	<i>OOM</i>
CUT3R [44]	300	0.141	0.096	0.076	0.034	0.543	0.564	0.234	0.139	0.074	0.018	0.575	0.614
	400	0.178	0.121	0.115	0.069	0.532	0.546	0.342	0.227	0.127	0.067	0.561	0.591
	500	0.190	0.138	0.090	0.033	0.530	0.543	0.359	0.264	0.173	0.081	0.560	0.591
TTT3R [10]	300	0.040	0.025	0.024	0.005	0.567	0.602	0.101	0.044	0.025	0.005	0.610	0.678
	400	0.052	0.031	0.027	0.005	0.558	0.588	0.143	0.065	0.071	0.012	0.600	0.658
	500	0.066	0.039	0.031	0.006	0.551	0.577	0.166	0.092	0.087	0.021	0.593	0.647
TTSA3R [62]	300	0.036	0.020	0.035	0.006	0.566	0.600	0.090	0.036	0.020	0.004	0.620	0.696
	400	0.036	0.019	0.024	0.004	0.561	0.592	0.104	0.045	0.035	0.006	0.618	0.692
	500	0.042	0.021	0.024	0.004	0.556	0.585	0.121	0.054	0.050	0.006	0.613	0.684
BAT3R	300	0.021	0.010	0.021	0.004	0.567	0.602	0.082	0.028	0.014	0.004	0.625	0.705
	400	0.025	0.012	0.023	0.004	0.561	0.592	0.093	0.037	0.027	0.004	0.620	0.697
	500	0.026	0.012	0.023	0.004	0.558	0.588	0.097	0.037	0.032	0.004	0.615	0.687

entirely without fine-tuning. Specifically, BAT3R secures the leading position among online methods on the KITTI dataset across both metric and scale-invariant settings. Under the metric scale evaluation on KITTI, BAT3R achieves an Abs Rel of 0.099, representing a significant 18.9% error reduction compared to the CUT3R baseline (0.122) and an additional 10.0% reduction over the highly competitive TTT3R (0.110). Furthermore, it demonstrates robust generalization on the Sintel and Bonn datasets by frequently ranking in the top two. For instance, in the scale-invariant setting on the Bonn dataset, BAT3R reduces the Abs Rel to 0.063, a 19.2% relative improvement over CUT3R (0.078). Similarly, under Sintel’s metric scale evaluation, BAT3R decreases the absolute relative error by 9.6% against CUT3R (0.930 vs. 1.029), while simultaneously tying for the highest $\delta < 1.25$ threshold accuracy (24.5).

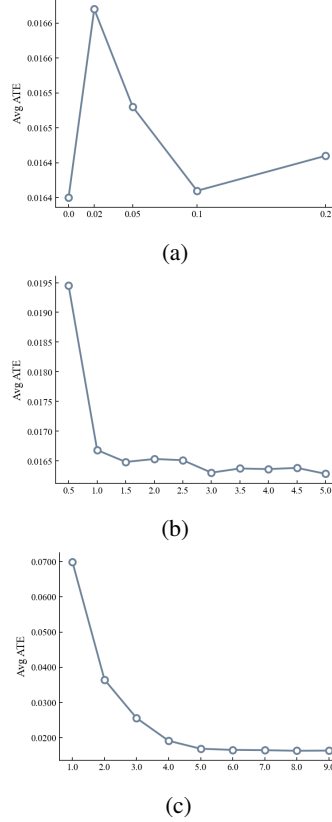


Figure 10: **Hyperparameter sensitivity analysis** defined in Eq. 6. We report the Average ATE to evaluate the impact of (a) the minimum update floor η , (b) the scaling factor κ controlling the transition slope, and (c) the shifting threshold τ .

B.5 3D Reconstruction

To provide a comprehensive quantitative analysis, we evaluate the models across multiple geometric dimensions, including Accuracy, Completeness, and Normal Consistency on the 7-Scenes [35] and NRGBD [3] benchmarks. As demonstrated in the results, BAT3R consistently outperforms existing recurrent baselines across all metrics and sequence lengths. On the 7-Scenes dataset at 500 frames, BAT3R achieves a Mean Accuracy error of 0.026, which represents an approximate 86.3% error reduction compared to the CUT3R baseline (0.190) and a 60.6% reduction compared to the highly competitive TTT3R (0.066). Furthermore, BAT3R demonstrates exceptional performance on the NRGBD dataset at 500 frames, achieving a Mean Accuracy of 0.097 and a Mean Completeness of 0.032. This corresponds to a 41.6% decrease in accuracy error and a 63.2% decrease in completeness error relative to TTT3R. Even when compared to the recent TTSA3R [62] method, BAT3R yields robust error reductions across accuracy and completeness while simultaneously securing the highest Normal Consistency scores (e.g., 0.615 on NRGBD at 500 frames) across both datasets.

B.6 Parameter Analysis

To comprehensively evaluate the robustness of the BAT3R framework, we conduct detailed hyperparameter sensitivity experiments on the TUM-dynamics dataset, utilizing Average ATE as the primary evaluation metric. We systematically analyze the critical parameters within both the State Innovation Gating (SIG) and Bayesian-guided Adaptive State Updating (BASU) modules.

SIG Hyperparameters: We assess the impact of the minimum update floor (η), the scaling factor (κ), and the shifting threshold (τ) defined in Eq. 6. As illustrated in Fig. 10(a), the Average ATE is relatively insensitive to the update floor η , maintaining highly stable performance with only minor fluctuations across the evaluated range. In contrast, the scaling factor κ (Fig. 10(b)) and the threshold

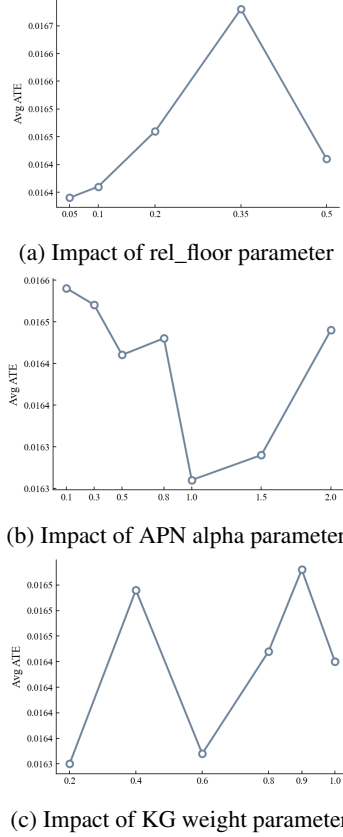


Figure 11: Ablation studies on additional hyperparameters. We report the Average ATE for (a) relative floor, (b) APN alpha, and (c) KG weight settings to evaluate their sensitivity.

τ (Fig. 10(c)) both exhibit a sharp initial performance gain. Specifically, the ATE drops significantly and then plateaus once κ exceeds 1.0 and τ exceeds 5.0. This indicates that the gating mechanism requires a sufficiently steep slope and a wide margin to effectively bound representational drift.

BASU Hyperparameters: We further evaluate the additional hyperparameters within the BASU module, including the relative covariance floor, the Adaptive Process Noise (APN) scaling parameter (α), and the Kalman Gain (KG) interpolation weight (ω), as shown in Fig. 11. First, Fig. 11(a) demonstrates the impact of the minimum threshold bounding the state covariance; the model performs optimally at an extremely low value (0.0), allowing the covariance to shrink organically as confident observations accumulate, whereas setting it marginally higher (e.g., 0.02) artificially inflates the lower bound and causes a temporary spike in ATE before stabilizing near 0.1. Second, the APN α parameter dictates how aggressively process noise is inflated in response to inter-frame motion. As depicted in Fig. 11(b), the model achieves optimal performance precisely at 1.0, since lower values fail to accommodate rapid motion, while higher values inject excessive uncertainty into the latent state. Finally, Fig. 11(c) evaluates the KG interpolation weight (ω) between the standard ratio and square-root Kalman gain formulations. The performance shows minor fluctuations, yielding the best results at 0.2 and 0.6, which suggests that blending the two formulations—leaning slightly towards the square-root variant—provides optimal numerical stability when observation noise and state covariance differ significantly.

C Theoretical Analysis

In this section, we provide the formal theoretical foundation for the BAT3R framework, specifically focusing on the stability and optimality of our Bayesian-guided state transition rules.

C.1 Stability Analysis of SIG

The SIG module acts as a distributional regularizer by penalizing states that deviate excessively from the initial geometric embedding $\mathbf{S}_0$. We prove that this mechanism ensures the latent state remains within a bounded region of the decoder’s reliable latent manifold.

Theorem 1 (Contraction toward Anchor). *Let $\delta_t = \|\mathbf{S}'_t - \mathbf{S}_0\|_2$ be the innovation residual at time t . Under the SIG update rule $\tilde{\mathbf{S}}_t = g_t \mathbf{S}'_t + (1 - g_t) \mathbf{S}_{t-1}$, where g_t is a decreasing function of δ_t , the SIG module acts as a non-expansive mapping that constrains the state trajectory relative to the initial anchor $\mathbf{S}_0$.*

Proof. Consider the distance to the anchor after the SIG update:

$$\|\tilde{\mathbf{S}}_t - \mathbf{S}_0\|_2 = \|g_t \mathbf{S}'_t + (1 - g_t) \mathbf{S}_{t-1} - \mathbf{S}_0\|_2 \quad (16)$$

By the triangle inequality:

$$\|\tilde{\mathbf{S}}_t - \mathbf{S}_0\|_2 \leq g_t \|\mathbf{S}'_t - \mathbf{S}_0\|_2 + (1 - g_t) \|\mathbf{S}_{t-1} - \mathbf{S}_0\|_2 \quad (17)$$

Substituting the innovation δ_t :

$$\|\tilde{\mathbf{S}}_t - \mathbf{S}_0\|_2 \leq g_t \delta_t + (1 - g_t) \|\mathbf{S}_{t-1} - \mathbf{S}_0\|_2 \quad (18)$$

Since the sigmoid gate $g_t \rightarrow \eta$ (lower bound) as $\delta_t \rightarrow \infty$, the update weight shifts predominantly toward the previous state $\mathbf{S}_{t-1}$ whenever the candidate $\mathbf{S}'_t$ exhibits anomalous drift. This ensures that the state evolution is not solely driven by potentially noisy current observations but is regularized by the historical trajectory and the geometric anchor $\mathbf{S}_0$. Thus, the latent representation is prevented from migrating into regions where the frozen decoder lacks reconstruction fidelity.

C.2 Geometric Manifold Consistency

A fundamental limitation of previous recurrent architectures like CUT3R and TTT3R is the unbounded accumulation of error variance over long sequences, leading to structural distortion. We formally prove that the BASU module strictly bounds the state uncertainty.

Theorem 2 (Uncertainty Boundedness). *Consider the recurrent state update in BASU where the process noise covariance $\mathbf{Q}_t$ is bounded by $\|\mathbf{Q}_t\|_2 \leq Q_{max}$ and the observation noise covariance $\mathbf{R}_t$ (derived from the confidence map C_t) is bounded by $\|\mathbf{R}_t\|_2 \leq R_{max}$. For any sequence length $t > 0$, the spectral norm of the posterior state covariance $\Sigma_{t|t}$ is strictly upper-bounded by the maximum observation noise.*

Proof. From the recursive covariance update, the prior covariance is $\Sigma_{t|t-1} = \Sigma_{t-1|t-1} + \mathbf{Q}_t$. The posterior covariance update can be reformulated using the matrix inversion lemma as:

$$\Sigma_{t|t} = (\Sigma_{t|t-1}^{-1} + \mathbf{R}_t^{-1})^{-1} \quad (19)$$

Given that $\Sigma_{t|t-1}$ is a positive-definite matrix resulting from the accumulation of initial variance $\Sigma_{0|0}$ and process noise $\mathbf{Q}_t$, its inverse $\Sigma_{t|t-1}^{-1}$ is also positive-definite with $\|\Sigma_{t|t-1}^{-1}\|_2 > 0$. Applying the properties of the spectral norm for positive-definite matrices:

$$\|\Sigma_{t|t}\|_2 = \left\| \left(\Sigma_{t|t-1}^{-1} + \mathbf{R}_t^{-1} \right)^{-1} \right\|_2 < \|(\mathbf{R}_t^{-1})^{-1}\|_2 = \|\mathbf{R}_t\|_2 \leq R_{max} \quad (20)$$

This inequality holds for all t as long as incoming observations are constrained by the confidence-driven noise model $\mathbf{R}_t$. Thus, the representational uncertainty of BAT3R does not grow with sequence length t , preventing the catastrophic drift observed in heuristic gating methods.

C.3 Bayesian Fusion Optimality

We further justify that the BASU update rule provides the optimal integration of novel scene geometry.

Theorem 3 (Minimum MSE Optimality). *Under the assumption that the latent state innovation follows a Gaussian distribution, the Kalman gain $\mathbf{K}_t$ derived in BASU yields the Best Linear Unbiased Estimator (BLUE) for the scene state $\mathbf{S}_t^*$, minimizing the expected Mean Squared Error (MSE) $\mathbb{E}[\|\mathbf{S}_t^* - \boldsymbol{\mu}_{t|t}\|_2^2]$.*

Proof. Let the error of the posterior estimate at time t be $\mathbf{e}_t = \mathbf{S}_t^* - \boldsymbol{\mu}_{t|t}$. The objective is to minimize the trace of the posterior covariance matrix $J = \text{Tr}(\boldsymbol{\Sigma}_{t|t})$. Substituting the update rule $\boldsymbol{\mu}_{t|t} = \boldsymbol{\mu}_{t-1|t-1} + \mathbf{K}_t(\tilde{\mathbf{S}}_t - \boldsymbol{\mu}_{t-1|t-1})$ into the error term, we obtain the posterior covariance $\boldsymbol{\Sigma}_{t|t}$ as a function of $\mathbf{K}_t$:

$$\boldsymbol{\Sigma}_{t|t} = (\mathbf{I} - \mathbf{K}_t)\boldsymbol{\Sigma}_{t|t-1}(\mathbf{I} - \mathbf{K}_t)^\top + \mathbf{K}_t\mathbf{R}_t\mathbf{K}_t^\top \quad (21)$$

Taking the derivative of the trace $\frac{\partial \text{Tr}(\boldsymbol{\Sigma}_{t|t})}{\partial \mathbf{K}_t}$ and setting it to zero:

$$-2(\mathbf{I} - \mathbf{K}_t)\boldsymbol{\Sigma}_{t|t-1} + 2\mathbf{K}_t\mathbf{R}_t = 0 \quad (22)$$

Solving for $\mathbf{K}_t$ yields the optimal gain $\mathbf{K}_t = \boldsymbol{\Sigma}_{t|t-1}(\boldsymbol{\Sigma}_{t|t-1} + \mathbf{R}_t)^{-1}$. While BAT3R utilizes a square-root stabilized variant to handle the scale mismatches in high-dimensional transformer latent spaces, it remains anchored to this MSE-minimizing objective, ensuring that well-consolidated memory is protected while novel content is optimally assimilated.

C.4 Information-Theoretic Gain and Update

BASU utilizes a temporal reliability gate λ_t derived from the symmetric KL divergence (Information Gain, IG_t) between the prior and posterior distributions. We prove that this formulation yields a robust filter that naturally suppresses both redundant observations and anomalous outliers.

Theorem 4 (Update Informativeness). *The information gain $IG_t = D_{KL}(q_t||q_{t|t-1}) + D_{KL}(q_{t|t-1}||q_t)$ is minimized when the current observation $\tilde{\mathbf{S}}_t$ provides no novel information relative to the existing state uncertainty, and maximized when the observation signals a significant but reliable structural change.*

Proof. For Gaussian beliefs, the symmetric KL divergence is given by Eq. 13 in the main text. Consider two extreme cases:

- **Redundancy:** If $\tilde{\mathbf{S}}_t \approx \boldsymbol{\mu}_{t-1|t-1}$ and $\mathbf{R}_t \gg \boldsymbol{\Sigma}_{t|t-1}$, the posterior $\boldsymbol{\mu}_{t|t} \rightarrow \boldsymbol{\mu}_{t-1|t-1}$ and $\boldsymbol{\Sigma}_{t|t} \rightarrow \boldsymbol{\Sigma}_{t|t-1}$. In this limit, $IG_t \rightarrow 0$, leading to a reliability gate $\lambda_t \rightarrow 0$ via the sigmoid squashing function. This prevents redundant computations and preserves well-consolidated memory.
- **Anomaly:** If $\|\tilde{\mathbf{S}}_t - \boldsymbol{\mu}_{t-1|t-1}\|_2 \rightarrow \infty$ while $\mathbf{R}_t$ is small, IG_t grows quadratically with the innovation residual. By mapping IG_t through a sigmoid with a high-gain threshold, BAT3R penalizes these anomalously large updates as potential outliers (e.g., dynamic occlusions), effectively shielding the persistent state from corruption.

The dual-suppression mechanism ensures that the state plasticity is selectively preserved only for genuinely informative updates that reflect consistent scene geometry.