HYBRIDDEPTH: ADAPTIVE COORDINATION OF LOCAL AND GLOBAL MEMORY FOR ONLINE VIDEO DEPTH ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Online video depth estimation must preserve useful temporal evidence while processing frames causally under a bounded memory budget. Existing streaming methods typically rely on one of two memory strategies with complementary limitations: a short causal window preserves recent fine-grained information but discards older observations, whereas a fixed-size recurrent state carries information over longer horizons but compresses history and can lose frame-specific details. The key challenge is therefore not simply to retain more history, but to coordinate recent explicit evidence with compressed longer-range context under bounded memory. We propose **HybridDepth**, a coordinated dual-memory framework for online video depth estimation. HybridDepth maintains a local token window and a recurrent global state as separate memory paths, allowing the current frame to access both recent detail and longer-range context without growing temporal storage with sequence length. To coordinate these memories, we introduce an attention-evidence gating mechanism that modulates each branch at the token level using the model's own query-key relevance. We further introduce an orthogonal learning objective that discourages the local and global branches from producing redundant corrections. Together, these designs preserve distinct temporal roles while allowing their contributions to adapt to the current observation. Across NuScenes, KITTI, Bonn, and TUM-Dynamic, HybridDepth achieves the best average depth accuracy among the compared online methods at 50, 100, and 200 frames. On 50-frame Bonn sequences, it also achieves the lowest temporal alignment error, providing direct evidence of improved adjacent-frame geometric consistency. At 500 frames, HybridDepth remains competitive, although it is not the best average method, while retaining bounded temporal memory. It operates at 24.94 FPS on a single H200. These results demonstrate a useful accuracy, temporal consistency, and memory trade-off for causal video depth estimation.

1 INTRODUCTION

Online video depth estimation predicts depth from a video stream as frames arrive, using only past and current observations. Unlike offline video depth methods, which can exploit future frames, bidirectional temporal modeling, or full-sequence optimization (Wang et al., 2023; Chen et al., 2025a; Luo et al., 2020; Shao et al., 2024; Hu et al., 2024; Kuang et al., 2024), an online model must make each prediction immediately while keeping its memory and computation practical as the stream grows. The central difficulty is therefore temporal: useful evidence may have appeared many frames earlier, yet retaining the complete history is incompatible with bounded-memory streaming.

Recent online methods (Li et al., 2023; Khan et al., 2023; Vaishakh et al., 2020; Chou et al., 2025; Feiden et al., 2025; Wang et al., 2025c; Wu et al., 2025; Zhuo et al., 2025; Chen et al., 2025b; Li et al., 2025) mainly address this problem through two memory paradigms. Window-based approaches (Feiden et al., 2025; Zhuo et al., 2025; Li et al., 2025; Lan et al., 2026) explicitly retain tokens from a small number of recent frames. This preserves fine-grained evidence such as object boundaries, newly visible structures, and recent motion, but information disappears once it leaves the window. Recurrent approaches (Wang et al., 2025c; Chou et al., 2025; Chen et al., 2025b) instead compress the accumulated history into a fixed-size state. This provides a constant-memory route to

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.

older observations, but repeated compression can dilute or overwrite frame-specific information. An object that was estimated correctly when clearly observed can therefore become inaccurate later as the stream evolves.

These two failure modes reveal a more fundamental problem in online depth estimation. Under a bounded memory budget, the model needs both *recent explicit evidence* and *compressed longer-range context*, but the two serve different purposes and are not equally useful for every current observation. Simply enlarging a local window increases memory and computation, while relying only on a recurrent state places all temporal information into the same compressed representation. Moreover, even if both sources are available, naively combining them does not ensure that they remain complementary: one branch may dominate, or both may learn similar corrections.

We address this problem with **HybridDepth**, which explicitly maintains a short local token memory and a recurrent global state, but keeps their contributions separate until the current-frame update. The resulting formulation is more than a concatenation of two memory types: the local and global histories are read through parallel branches, so their contributions remain explicitly controllable at each token. This provides a direct mechanism for preserving recent high-detail evidence while retaining access to information summarized from earlier observations, with temporal-state storage bounded by the fixed local window and global-state sizes.

The second question is how these two memory sources should be coordinated. Their usefulness changes with the current observation: recently observed details may be important around dynamic or newly revealed content, whereas information accumulated over a longer horizon may provide useful scene-level context. HybridDepth therefore derives a lightweight relevance signal from the attention evidence of each memory branch and uses it to modulate the corresponding residual update at the token level. We use this attention statistic as a relevance proxy rather than as calibrated confidence. Because two parallel branches can still converge to redundant corrections, we further introduce an orthogonal learning objective that discourages highly correlated local and global updates. Thus, the gate controls *how strongly* each memory contributes, while orthogonal learning encourages the two branches to contribute *different* information.

This coordinated design leads to consistent improvements in the regime where the method is empirically strongest. Across four real-world indoor and outdoor benchmarks, HybridDepth achieves the best average depth accuracy among the compared online methods at 50, 100, and 200 frames. Controlled memory ablations show that the full hybrid model outperforms both local-only and global-only variants, including at 200 frames. Beyond depth accuracy, HybridDepth obtains the lowest temporal alignment error on 50-frame Bonn sequences, providing direct evidence of improved adjacent-frame geometric consistency. At 500 frames, the method remains competitive but is not the best average method, exposing the limitation of fixed-capacity memory at very long horizons. HybridDepth nevertheless retains bounded temporal-state storage and operates at 24.94 FPS on a single H200.

In summary, our contributions are as follows:

- We introduce a **coordinated dual-memory formulation** for causal video depth estimation that keeps recent local evidence and recurrent longer-range context as separate read paths under bounded temporal memory. This addresses the complementary limitations of relying on either a short causal window or a compressed recurrent state alone.

- We propose **attention-evidence gating** to adapt the contribution of each memory branch at the token level, together with an **orthogonal learning objective** that discourages redundant local and global corrections. The two mechanisms jointly coordinate memories with different temporal roles rather than treating them as independent additions.

- Extensive evaluation shows that HybridDepth achieves the best average accuracy among the compared online methods at 50–200 frames, improves adjacent-frame geometric consistency on the Bonn temporal evaluation, and consistently outperforms local-only and global-only variants in controlled ablations. At 500 frames, it remains competitive while maintaining bounded temporal memory, making the empirical scope and limitations of the method explicit.

2 Related Work

Offline Video Depth Estimation methods predict per-frame depth maps given the entire video sequence, allowing the model to access both past and future frames. Existing approaches can be broadly categorized into three groups. The first group (Chen et al., 2025a; Wang et al., 2023) extends single-frame depth models (Yang et al., 2024b;c; Ke et al., 2024) by freezing the pretrained encoder or backbone and learning an additional temporal module to enforce cross-frame consistency. The second group (Yang et al., 2024a; Hu et al., 2024; Shao et al., 2024) leverages powerful video generation or diffusion models to produce temporally coherent depth predictions. The third group (Wang et al., 2024; 2025b; Zhang et al., 2025; Lu et al., 2024) predicts point maps across the sequence and reconstructs the scene as a global point cloud. Despite their strong performance, these methods typically involve high computational cost, batch-wise processing, or post-hoc optimization, which significantly limit their practicality in real-world streaming scenarios.

Online Video Depth Estimation aims to predict depth maps from streaming video frames in a causal manner, where future frames beyond the current timestep are unavailable. Early approaches (Vaishakh et al., 2020; Zhang et al., 2019; Khan et al., 2023; Li et al., 2023) were constrained by limited model capacity or training data, resulting in reduced accuracy and temporal instability. More recent methods can be broadly categorized into two groups. The first group (Feiden et al., 2025; Chou et al., 2025) builds upon single-image depth models by freezing the encoder and adapting the decoder for streaming prediction. The second group (Wang et al., 2025c; Chen et al., 2025b; Zhuo et al., 2025; Wu et al., 2025; Wang & Agapito, 2024) leverages modern 3D foundation models (Wang et al., 2025b; 2024) to reconstruct scene geometry through point map prediction from streaming inputs. While these methods improve practicality for real-world deployment, relatively little attention has been devoted to systematically studying the memory mechanisms required for robust and long-horizon online depth estimation.

Memory Mechanism is essential for maintaining temporal context under causality constraints. Existing approaches mainly adopt two types of memory designs: KV-cache-based causal attention and persistent latent states. KV-cache-based methods (Xiao et al., 2024; Dai et al., 2019; Lan et al., 2026; Zhuo et al., 2025; Li et al., 2025; Wang & Agapito, 2024) store previously computed tokens as key–value pairs, allowing future queries to attend to historical representations for streaming prediction. In contrast, another line of work (Hochreiter & Schmidhuber, 1997; Jaegle et al., 2022; Wang et al., 2025c; Chen et al., 2025b) follows the recurrent paradigm by compressing history into a fixed-size latent state, from which new inputs retrieve contextual cues.

3 Proposed Method

Task Definition Let $\{\mathbf{I}_t\}_{t=1}^{\infty}$ be a video stream whose frames arrive sequentially. Online depth estimation must produce $\mathbf{D}_t$ immediately after observing $\mathbf{I}_t$, without access to future frames:

$$\mathbf{D}_t = f_\theta(\mathbf{I}_1, \ldots, \mathbf{I}_t). \tag{1}$$

Retaining every frame or its features makes temporal storage grow with the stream. A bounded-memory model must therefore discard or compress history while preserving the evidence needed for the current prediction. HybridDepth addresses this problem by maintaining two bounded forms of history with different temporal roles and coordinating their contributions before decoding each depth map (Figure 1).

Window and Recurrent Memory A causal KV cache explicitly retains image tokens from recent frames. Given an incoming frame, a shared encoder produces $\mathbf{F}_t = \mathrm{Enc}(\mathbf{I}_t) \in \mathbb{R}^{K \times C}$. Keeping all past tokens requires $O(tK)$ temporal storage at time t, whereas truncating the cache to W frames bounds it at $O(WK)$. The resulting window preserves recent fine-grained evidence, but information is discarded once it leaves the cache. Recurrent state-token models instead compress accumulated history into a fixed number S of tokens, providing access to older context at $O(S)$ temporal storage. Repeated compression, however, can lose frame-specific information. These complementary information-loss mechanisms motivate keeping recent explicit evidence and compressed longer-range context as separate memory paths.

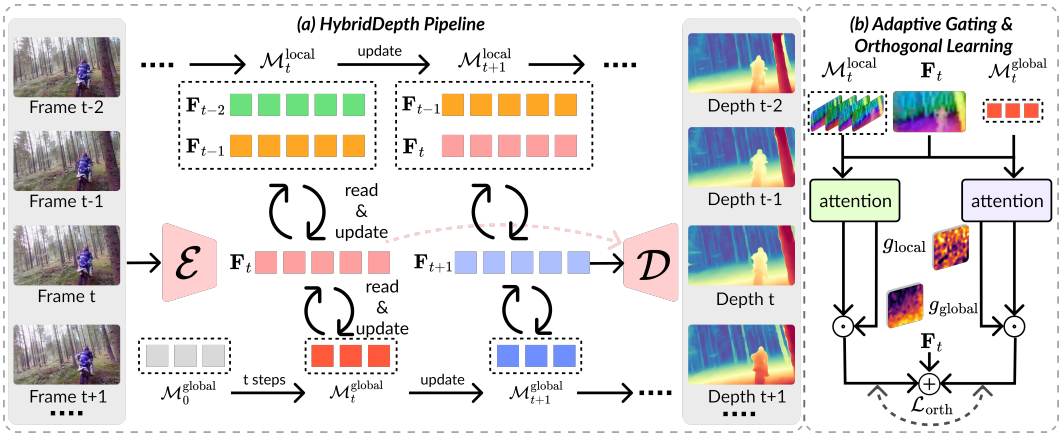


Figure 1: Overview of the HybridDepth framework. (a) Our model takes streaming videos as input, where a newly observed frame is input to the model at a new timestep. The feature of the frame t interacts with the local and global memories $\mathcal{M}_t^{\text{local}}$, $\mathcal{M}_t^{\text{global}}$ by reading and update operators. Then the updated feature is decoded to the predicted depth. (b) Our adaptive gating and orthogonal learning strategy. In the decoder block where the current image feature does cross attention with both memories, the attention score can be utilized as an adaptive gate for feature update, and these two branches are forced to learn different knowledge by $\mathcal{L}_{\text{orth}}$.

3.1 Coordinated Local and Global Memory

Our full pipeline is illustrated in Figure 1. At each timestep t, HybridDepth maintains (i) a causal local memory containing tokens from the most recent W frames and (ii) a compact recurrent global state summarizing information accumulated from earlier observations. We denote them by $\mathcal{M}_t^{\text{local}}$ and $\mathcal{M}_t^{\text{global}}$, respectively.

Local and Global Memories Given $\mathbf{I}_t$, we first encode the current frame as $\mathbf{F}_t = \text{Enc}(\mathbf{I}_t) \in \mathbb{R}^{K\times C}$. The two memory sources are

$$\mathcal{M}_t^{\text{local}} := \mathbf{F}_{t-W} \oplus \mathbf{F}_{t-W+1} \oplus \cdots \oplus \mathbf{F}_{t-1} \in \mathbb{R}^{WK\times C}, \tag{2}$$

$$\mathcal{M}_t^{\text{global}} := \mathbf{S}_{t-1} \in \mathbb{R}^{S\times C}. \tag{3}$$

For early timesteps ($t \le W$), the local memory uses the available past frames and is padded with canonical tokens $\mathbf{F}_0$ when needed. Thus, the local memory explicitly retains recent frame-specific evidence, while the global state provides a compressed route to information accumulated over a longer temporal horizon.

Parallel Memory Read A key design choice is to keep the two memory paths separate when conditioning the current representation. Both memories independently interact with the same current tokens:

$$\Delta\mathbf{F}_t^{\text{local}} = \text{CrossAttn}^{\text{local}}\left(\mathbf{F}_t, \mathcal{M}_t^{\text{local}}\right), \tag{4}$$

$$\Delta\mathbf{F}_t^{\text{global}} = \text{CrossAttn}^{\text{global}}\left(\mathbf{F}_t, \mathcal{M}_t^{\text{global}}\right). \tag{5}$$

Before adaptive coordination, their residual contributions can be written as

$$\hat{\mathbf{F}}_t = \mathbf{F}_t + \Delta\mathbf{F}_t^{\text{local}} + \Delta\mathbf{F}_t^{\text{global}}. \tag{6}$$

Keeping these updates explicit is important: rather than first collapsing local and global history into one representation, HybridDepth can control how much each source contributes to every current token through the gate introduced in subsection 3.2. The resulting tokens $\hat{\mathbf{F}}_t$ are decoded into the depth prediction.

Memory Update After frame t is processed, the local memory slides forward by appending the current tokens and discarding the oldest ones:

$$\mathcal{M}_{t+1}^{\text{local}} := \mathbf{F}_{t-W+1} \oplus \cdots \oplus \mathbf{F}_t. \tag{7}$$

The global state recurrently incorporates the current observation:

$$\mathbf{S}_t = \mathbf{S}_{t-1} + \mathrm{CrossAttn}^{\text{update}}(\mathbf{S}_{t-1}, \mathbf{F}_t)\,, \tag{8}$$

and $\mathcal{M}_{t+1}^{\text{global}} := \mathbf{S}_t$. The retained temporal state therefore remains bounded by the fixed local window and global-state sizes rather than growing with sequence length.

3.2 Adaptive Gating from Attention Evidence

The two memories contain different forms of temporal evidence, and their usefulness can vary across the current frame. Recent observations may be more informative for newly appearing or rapidly changing regions, whereas the recurrent state may provide useful context accumulated from earlier observations. Applying both branches with fixed strength cannot account for this variation.

HybridDepth therefore adaptively modulates the two memory-conditioned updates using a lightweight relevance signal derived from their own attention evidence. This signal is used as a parameter-free proxy for query-memory compatibility, rather than as calibrated confidence. The mechanism is illustrated in Figure 1(b).

Attention Evidence For the local and global read branches in Equation 4 and Equation 5, let their pre-softmax attention scores be

$$S_{\text{local}} = \frac{Q_{\text{local}} K_{\text{local}}^{\top}}{\sqrt{d}}, \qquad S_{\text{global}} = \frac{Q_{\text{global}} K_{\text{global}}^{\top}}{\sqrt{d}}. \tag{9}$$

Each score tensor has shape $\mathbb{R}^{H\times K\times L}$, where H is the number of heads, K is the number of current-frame queries, and L is the number of memory keys, with $L_{\text{local}} = WK$ and $L_{\text{global}} = S$. We aggregate these scores as a lightweight measure of compatibility between each current token and the corresponding memory source.

Per-token Memory Coordination For each current token k, we average the attention evidence over heads and memory tokens and apply an activation σ:

$$g_{\text{local}}(k) = \sigma\left(\frac{1}{HL_{\text{local}}}\sum_{h=1}^{H}\sum_{l=1}^{L_{\text{local}}} S_{\text{local}}(h,k,l)\right), \tag{10}$$

$$g_{\text{global}}(k) = \sigma\left(\frac{1}{HL_{\text{global}}}\sum_{h=1}^{H}\sum_{l=1}^{L_{\text{global}}} S_{\text{global}}(h,k,l)\right). \tag{11}$$

Normalization by the number of heads and memory tokens makes the aggregation independent of the different memory lengths. The two gates independently modulate the residual updates:

$$\hat{\mathbf{F}}_t = \mathbf{F}_t + g_{\text{local}} \odot \Delta\mathbf{F}_t^{\text{local}} + g_{\text{global}} \odot \Delta\mathbf{F}_t^{\text{global}}. \tag{12}$$

The gates are not constrained to sum to one: both memories may contribute when both are relevant, or either contribution may be suppressed when its attention evidence is weak. Thus, the gate controls how strongly each temporal source influences the current prediction without introducing a separate reliability predictor.

3.3 Orthogonal Learning for Different Memories

Adaptive gating controls the strength of the two memory updates, but does not by itself prevent them from learning similar corrections. Because both branches condition the same current tokens, their updates can become redundant, weakening the intended benefit of maintaining two distinct temporal representations.

Orthogonality Regularization We therefore impose a soft orthogonality regularization between the token-wise local and global updates:

$$\mathcal{L}_{\text{orth}} = \frac{1}{K}\sum_{k=1}^{K}\left|\frac{\left\langle \Delta\mathbf{F}_{t,k}^{\text{local}}, \Delta\mathbf{F}_{t,k}^{\text{global}}\right\rangle}{\left\|\Delta\mathbf{F}_{t,k}^{\text{local}}\right\|_2 \left\|\Delta\mathbf{F}_{t,k}^{\text{global}}\right\|_2 + \epsilon}\right|, \tag{13}$$

where $\Delta\mathbf{F}_{t,k} \in \mathbb{R}^C$ denotes the update of token k and ϵ is a small constant.

Effect Minimizing $\mathcal{L}_{\text{orth}}$ discourages highly correlated local and global update directions, reducing redundant corrections in the additive fusion. The regularizer does not itself assign semantic roles to the branches; rather, together with their different memory contents, it encourages the two sources to contribute complementary information. In this way, adaptive gating determines how strongly each branch contributes, while orthogonal learning discourages both branches from making the same correction.

3.4 Sequential Learning Objective

We supervise HybridDepth at the **sequence level** using the scale-and-shift invariant (SSI) objective from Ranftl et al. (2022). Instead of normalizing every frame independently, one scale and shift are estimated for the complete training sequence, encouraging a shared affine depth relation across time.

Sequence-level SSI Supervision Given predicted depths $\{\hat{\mathbf{D}}_t\}_{t=1}^T$, ground-truth depths $\{\mathbf{D}_t^*\}_{t=1}^T$, and valid masks $\{\Omega_t\}_{t=1}^T$, we aggregate all valid sequence pixels and fit global parameters $\hat{s}_{\text{seq}}, \hat{t}_{\text{seq}}$ for the predictions and $s^*_{\text{seq}}, t^*_{\text{seq}}$ for the ground truth. This yields the sequence-level SSI loss $\mathcal{L}_{\text{ssi}}^{\text{seq}}$.

Relative Depth Supervision Sequence-level SSI enforces a shared affine relation across the clip, but does not explicitly normalize residuals by depth magnitude. We therefore complement it with a sequence-level absolute relative error:

$$\mathcal{L}_{\text{rel}}^{\text{seq}} = \frac{1}{\sum_t |\Omega_t|} \sum_{t=1}^{T} \sum_{\mathbf{p}\in\Omega_t} \frac{\left| \frac{s^*_{\text{seq}}}{\hat{s}_{\text{seq}}} \hat{\mathbf{D}}_t(\mathbf{p}) - \frac{s^*_{\text{seq}}}{\hat{s}_{\text{seq}}} \hat{t}_{\text{seq}} + t^*_{\text{seq}} - \mathbf{D}_t^*(\mathbf{p}) \right|}{|\mathbf{D}_t^*(\mathbf{p})| + \epsilon}. \tag{14}$$

This term balances supervision across different depth ranges and complements the sequence-level SSI objective.

Final Objective The final training objective is

$$\mathcal{L} = \lambda_{\text{ssi}}\mathcal{L}_{\text{ssi}}^{\text{seq}} + \lambda_{\text{rel}}\mathcal{L}_{\text{rel}}^{\text{seq}} + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}}. \tag{15}$$

Here, $\mathcal{L}_{\text{ssi}}^{\text{seq}}$ provides sequence-level depth supervision, $\mathcal{L}_{\text{rel}}^{\text{seq}}$ balances errors across depth ranges, and $\mathcal{L}_{\text{orth}}$ discourages redundant local and global memory updates.

4 Experiments

Experimental Settings Our model is initialized from pretrained CUT3R (Wang et al., 2025c) weights. We use a temporal-window KV cache with $W = 4$ and train on 8-frame clips. The network is optimized with Adam for 10 epochs, using learning rates of 5×10^{-5} for the decoder and DPT head (Ranftl et al., 2021), and 5×10^{-6} for the encoder. Training is performed on two NVIDIA H200 GPUs for about one day, and all evaluations are conducted on a single H200.

Datasets Starting from the pretrained CUT3R initialization, we fine-tune on Dynamic-Replica (Karaev et al., 2023), PointOdyssey (Zheng et al., 2023), ARKitScenes (Baruch et al., 2021), VirtualKITTI2 (Cabon et al., 2020), Waymo (Sun et al., 2020), and SpatialVID (Wang et al., 2025a). We evaluate on four real-world benchmarks: KITTI (Geiger et al., 2012), NuScenes (Caesar et al., 2020), Bonn (Palazzolo et al., 2019), and TUM-Dynamic (Sturm et al., 2012), using sequence lengths of 50, 100, 200, and 500 frames.

Baselines and Metrics We compare with CUT3R (Wang et al., 2025c), FlashDepth (Chou et al., 2025), TTT3R (Chen et al., 2025b), and StreamVGGT (Zhuo et al., 2025) under the same causal evaluation protocol. Following prior work (Chen et al., 2025a; Wang et al., 2025c; Chen et al., 2025b; Wang et al., 2023; Chou et al., 2025; Zhuo et al., 2025), predicted depths are aligned once per sequence to ground truth using a least-squares scale and shift. We report absolute relative error (Abs Rel) and threshold accuracy (δ_1). Cross-method results compare complete systems, while the matched ablations below isolate the contribution of our design.

Table 1: **Video Depth Quantitative Evaluation.** We report performance on NuScenes, KITTI, Bonn, and TUM-Dyn. The best and second-best results in each category are **bold** and underlined. At 500 frames, NuScenes is unavailable and the average therefore covers KITTI, Bonn, and TUM-Dyn only. − indicates unavailable results.

Method	NuScenes		KITTI		Bonn		TUM-Dyn		Average	
	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑
50 frames										
CUT3R	17.1	74.2	9.66	90.7	7.53	91.6	10.1	88.7	11.1	86.3
TTT3R	16.8	75.6	9.26	91.4	7.00	96.5	9.70	89.0	10.7	88.1
FlashDepth	17.7	80.1	11.9	88.6	6.79	96.9	13.6	83.6	12.5	87.3
StreamVGGT	17.6	74.7	11.1	86.6	**5.41**	97.3	**8.20**	**91.7**	10.6	87.6
Ours	**13.5**	**83.4**	**8.51**	**93.9**	5.47	**97.4**	10.3	89.3	**9.5**	**91.0**
100 frames										
CUT3R	17.7	72.6	11.2	87.7	6.95	94.1	10.5	87.9	11.6	85.6
TTT3R	17.4	72.9	10.1	91.0	6.10	96.5	9.79	88.5	10.9	87.2
FlashDepth	18.6	78.7	12.4	87.4	6.60	97.2	17.3	81.8	13.7	86.3
StreamVGGT	20.8	65.0	15.6	74.1	**5.36**	**97.5**	**8.64**	**91.7**	12.6	82.1
Ours	**15.9**	**81.6**	**9.04**	**93.1**	5.37	**97.5**	10.7	88.7	**10.3**	**90.2**
200 frames										
CUT3R	19.2	70.3	10.8	89.8	6.76	95.1	10.5	88.2	11.8	85.9
TTT3R	17.3	74.7	9.60	91.6	6.34	95.9	9.84	89.0	10.8	87.8
FlashDepth	18.3	79.8	12.6	87.3	7.81	95.8	16.3	82.8	13.8	86.4
StreamVGGT	20.2	67.1	23.6	56.0	6.12	**96.8**	**9.01**	**92.4**	14.7	78.1
Ours	**15.4**	**83.4**	**8.72**	**93.5**	**5.93**	96.6	12.1	87.5	**10.5**	**90.2**
500 frames										
CUT3R	–	–	11.9	87.6	8.30	94.0	13.2	82.8	11.1	88.1
TTT3R	–	–	10.7	89.7	**7.20**	**95.6**	**11.5**	**85.0**	**9.8**	**90.1**
FlashDepth	–	–	13.8	83.6	10.1	91.9	16.2	80.5	13.4	85.3
StreamVGGT	–	–	OOM	OOM	OOM	OOM	OOM	OOM	N/A	N/A
Ours	–	–	**9.60**	**91.3**	7.65	95.4	13.1	83.4	10.1	90.0

4.1 Quantitative and Qualitative Results

Table 1 compares HybridDepth with recent online methods over sequence lengths from 50 to 500 frames.

Depth Accuracy HybridDepth achieves the best average Abs Rel and δ_1 among the compared methods at 50, 100, and 200 frames. Relative to the next-best average Abs Rel, the error is reduced by 10.4%, 5.5%, and 2.8%, respectively. The gains are not uniform across every dataset, most notably on TUM-Dyn, but the average improvement is consistent across these three horizons. At 500 frames, NuScenes is unavailable and the average covers only KITTI, Bonn, and TUM-Dyn. HybridDepth is not the best average method at this horizon: TTT3R obtains $9.8/90.1$ compared with our $10.1/90.0$. HybridDepth nevertheless achieves the best KITTI result and completes 500-frame inference, while StreamVGGT is OOM under the same evaluation setting. The common three-benchmark comparison in Table 5 further quantifies this long-horizon limitation.

Temporal Consistency Abs Rel and δ_1 measure depth accuracy but do not directly establish temporal consistency. We therefore additionally evaluate temporal alignment error (TAE) using a single alignment for each complete sequence. On 50-frame Bonn sequences, HybridDepth obtains the lowest TAE of 0.55, compared with 0.63 for StreamVGGT and 0.67 for TTT3R (Table 4). This provides direct evidence of improved adjacent-frame geometric consistency, while not implying a globally consistent 3D reconstruction.

Qualitative Results Figure 2 illustrates the same behavior on a dynamic indoor sequence. HybridDepth preserves sharper human boundaries and more stable balloon-human depth ordering than the compared methods. This example is complementary to the quantitative TAE evaluation rather than standalone evidence of general temporal consistency.

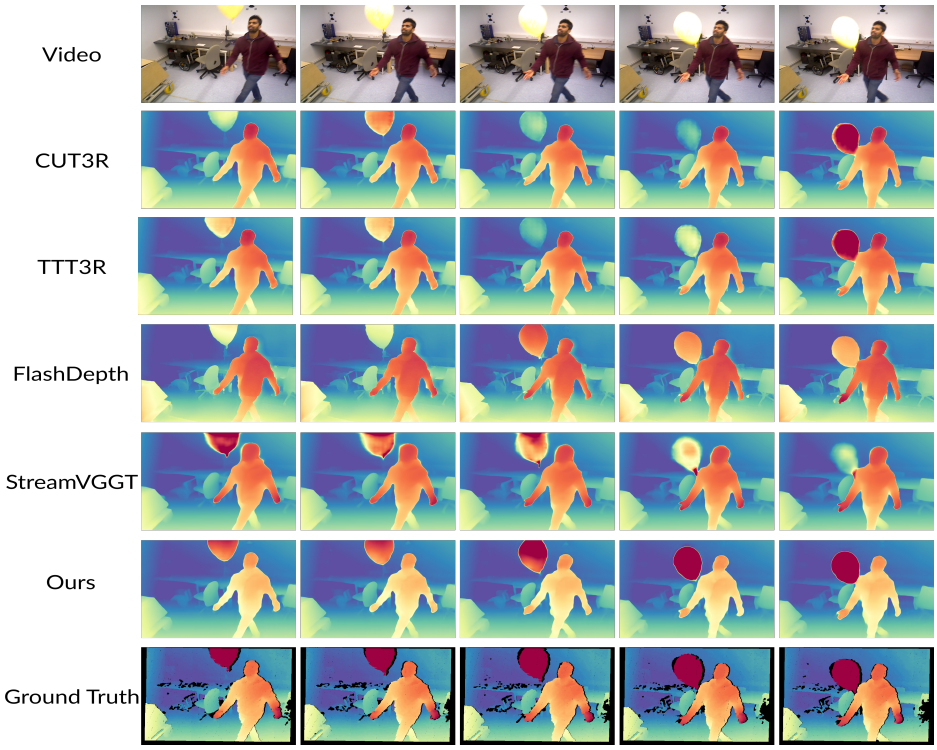


Figure 2: Qualitative comparison on a dynamic indoor sequence. HybridDepth preserves sharper depth discontinuities and more stable depth ordering around the moving human and balloon while maintaining coherent background structure.

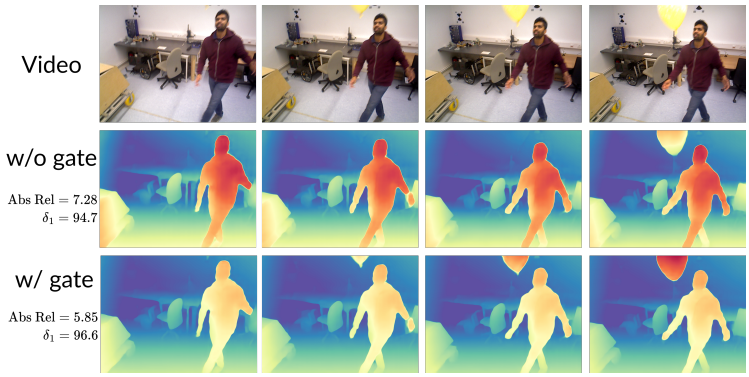


Figure 3: Ablation on adaptive gating. In this example, gating reduces the depth corruption around the newly appearing balloon.

4.2 Ablation Studies

Ablation on Memory Coordination Table 2 first examines the mechanisms used to coordinate the two memory paths. The gate alone (#2) improves NuScenes and TUM-Dyn but slightly degrades KITTI and Bonn, showing that attention-based modulation is useful but not uniformly sufficient by itself. Adding $\mathcal{L}_{\text{orth}}$ (#3) improves the result on all four datasets relative to the gated model, consistent with reducing redundant branch updates. Combining the gate with sequence-level relative supervision (#4) provides another complementary improvement. The complete coordinated model (#7) achieves the strongest result on every dataset.

Ablation on Memory Design The local-only and global-only variants use the same training setup but retain only one temporal source. Neither is consistently stronger across datasets, whereas the hybrid model achieves the best result on all four. This isolates the benefit of retaining both recent explicit evidence and compressed longer-range context. The matched 200-frame evaluation in Ta-

Table 2: **Ablation Studies.** All variants use the same initialization, training data, and optimization, and are evaluated at 50 frames. We evaluate Adaptive Gating (**Gate**), Orthogonal Regularization ($\mathcal{L}_{\textbf{orth}}$), Sequential Relative Depth Loss ($\mathcal{L}_{\textbf{rel}}^{\textbf{seq}}$), and different memory designs. **Bold** indicates the best performance.

#	Setting	Gate	$\mathcal{L}_{\text{orth}}$	$\mathcal{L}_{\text{rel}}^{\text{seq}}$	NuScenes		KITTI		Bonn		TUM-Dyn	
					Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑
Ablation on Different Modules												
1	Baseline				15.2	80.6	9.34	92.4	6.12	96.5	11.7	87.8
2	+ Gate	✓			14.6	81.9	9.41	92.2	6.20	96.2	11.0	88.6
3	+ Gate + $\mathcal{L}_{\text{orth}}$	✓	✓		14.3	82.5	8.78	93.3	5.74	97.1	10.9	88.7
4	+ Gate + $\mathcal{L}_{\text{rel}}^{\text{seq}}$	✓		✓	13.9	82.7	9.08	92.8	6.05	96.7	10.4	89.1
Ablation on Memory Design												
5	Local only	✓		✓	14.1	82.6	8.92	93.1	5.78	97.0	10.5	89.0
6	Global only	✓		✓	14.6	81.8	8.63	93.6	5.61	97.2	10.9	88.5
7	**Hybrid (Ours)**	✓	✓	✓	**13.5**	**83.4**	**8.51**	**93.9**	**5.47**	**97.4**	**10.3**	**89.3**

ble 6 shows the same pattern: the full model reduces Abs Rel by 1.28 relative to local-only memory and by 0.56 relative to global-only memory. Thus, the complementarity of the two memory paths extends beyond the original 50-frame ablation.

4.3 Limitations

HybridDepth remains limited in several respects. First, both memory paths depend on reliable observations, so severe blur, occlusion, abrupt illumination changes, or fast motion can corrupt the evidence available to both branches. Second, the direct temporal evaluation measures adjacent-frame geometric consistency on Bonn; it does not establish a globally consistent 3D map, which the model does not explicitly optimize. Third, performance degrades at very long horizons: HybridDepth is not the best average method at 500 frames, reflecting the remaining limitation of fixed-capacity global memory and the gap between 8-frame training clips and much longer inference sequences. Finally, the additional memory branches increase computation, and the reported real-time throughput applies to a single H200 rather than edge hardware.

5 Conclusion

We introduced HybridDepth, a bounded-memory framework for online video depth estimation that coordinates two complementary forms of temporal evidence: recent explicit information stored in a local window and longer-range context compressed into a recurrent global state. Rather than merging these memories into a single representation, HybridDepth keeps their read paths separate and adaptively modulates their contributions using attention evidence. An orthogonal learning objective further discourages redundant local and global updates, allowing the two memory sources to contribute complementary information under a fixed temporal-memory budget. Experiments across real-world indoor and outdoor benchmarks show that HybridDepth achieves the best average depth accuracy among the compared online methods at 50, 100, and 200 frames. Controlled ablations consistently outperform both local-only and global-only variants, including at 200 frames, supporting the benefit of coordinated local and global memory beyond either memory source alone. HybridDepth also achieves the lowest temporal alignment error on the Bonn evaluation, providing direct evidence of improved adjacent-frame geometric consistency. At 500 frames, HybridDepth remains competitive but is not the best average method, highlighting the remaining limitation of fixed-capacity memory at very long horizons. Nevertheless, it retains bounded temporal storage and practical streaming efficiency. Overall, these results show that online depth estimation benefits not simply from retaining more history, but from explicitly coordinating recent detailed evidence with compressed longer-range context. This provides a practical direction for more accurate and temporally consistent depth prediction under causal and bounded-memory constraints.

AI USE STATEMENT

We used LLM to assist with editing portions of the manuscript, including clarifying the method description, interpreting and refining the discussion of reported ablation results, and adjusting LaTeX formatting and pagination. The authors take responsibility for the final text, scientific claims, results, and citations, including material produced with AI assistance.

REFERENCES

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://machinelearning.apple.com/research/arkitscenes.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. URL https://arxiv.org/abs/2001.10773.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving, 2020. URL https://arxiv.org/abs/1903.11027.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. *arXiv:2501.12375*, 2025a.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025b.

Gene Chou, Wenqi Xian, Guandao Yang, Mohamed Abdelfattah, Bharath Hariharan, Noah Snavely, Ning Yu, and Paul Debevec. Flashdepth: Real-time streaming video depth estimation at 2k resolution. In *The IEEE International Conference on Computer Vision (ICCV)*, 2025.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1285. URL https://aclanthology.org/P19-1285/.

Johann-Friedrich Feiden, Tim Küchler, Denis Zavadski, Bogdan Savchynskyy, and Carsten Rother. Online video depth anything: Temporally-consistent depth prediction with low memory consumption, 2025. URL https://arxiv.org/abs/2510.09182.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. URL https://www.cvlibs.net/publications/Geiger2012CVPR.pdf.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 11 1997. doi: 10.1162/neco.1997.9.8.1735.

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver io: A general architecture for structured inputs & outputs, 2022. URL https://arxiv.org/abs/2107.14795.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. *CVPR*, 2023.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9492–9500, 2024. doi: 10.1109/CVPR2024.01234. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Ke_Repurposing_Diffusion-Based_Image_Generators_for_Monocular_Depth_Estimation_CVPR_2024_paper.pdf.

Numair Khan, Eric Penner, Douglas Lanman, and Lei Xiao. Temporally consistent online depth estimation using point-based fusion. *Computer Vision and Pattern Recognition (CVPR)*, 2023.

Zhengfei Kuang, Tianyuan Zhang, Kai Zhang, Hao Tan, Sai Bi, Yiwei Hu, Zexiang Xu, Milos Hasan, Gordon Wetzstein, and Fujun Luan. Buffer anytime: Zero-shot video depth and normal from image priors. *arXiv preprint arXiv:2411.17249*, 2024. URL https://arxiv.org/abs/2411.17249.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. STream3R: Scalable sequential 3D reconstruction with causal transformer. In *ICLR*, 2026.

Zhaoshuo Li, Wei Ye, Dilin Wang, Francis X Creighton, Russell H Taylor, Ganeshnbsp Venkatesh, and Mathias Unberath. Temporally consistent online depth estimation in dynamic scenes. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

Zizun Li, Jianjun Zhou, Yifan Wang, Haoyu Guo, Wenzheng Chang, Yang Zhou, Haoyi Zhu, Junyi Chen, Chunhua Shen, and Tong He. Wint3r: Window-based streaming reconstruction with camera token pool, 2025. URL https://arxiv.org/abs/2509.05296.

Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3r: Aligned monocular depth estimation for dynamic videos. *arXiv preprint arXiv:2412.03079*, 2024.

Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. In *ACM Transactions on Graphics (TOG)*, 2020. URL https://dl.acm.org/doi/10.1145/3386569.3392377.

E. Palazzolo, J. Behley, P. Lottes, P. Giguère, and C. Stachniss. ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals. *arXiv*, 2019. URL https://arxiv.org/abs/1905.02082.

Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. URL https://arxiv.org/abs/2103.13413.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022.

Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Vitor Guizilini, Yue Wang, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors, 2024. URL https://arxiv.org/abs/2406.01493.

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*, pp. 573–580, 2012. doi: 10.1109/IROS.2012.6385773.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2446–2454, June 2020.

Patil Vaishakh, Wouter Van Gansbeke, Dengxin Dai, and Luc Van Gool. Dont forget the past: Recurrent depth estimation from monocular video. *IEEE Robotics and Automation Letters*, 5(4): 6813–6820, 2020.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. Spatialvid: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024.

Yiran Wang, Min Shi, Jiaqi Li, Chaoyi Hong, Zihao Huang, Juewen Peng, Zhiguo Cao, Jianming Zhang, Ke Xian, and Guosheng Lin. Neural video depth stabilizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://github.com/RaymondWang987/NVDS.

Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. *arXiv preprint arXiv:2507.02863*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5e5fd18f863cbe6d8ae392a93fd271c9-Paper-Conference.pdf.

Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin, Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. *arXiv preprint arXiv:2410.10815*, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Yang_Depth_Anything_Unleashing_the_Power_of_Large-Scale_Unlabeled_Data_CVPR_2024_paper.pdf.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024c. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/26cfdcd8fe6fd75cc53e92963a656c58-Abstract-Conference.html.

Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, and Youliang Yan. Exploiting temporal consistency for real-time video depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1725–1734, October 2019.

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. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=lJpqxFgWCM.

Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *ICCV*, 2023.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025.

A Implementation Details

Our method is initialized from the pretrained CUT3R (Wang et al., 2025c) weights (specifically, cut3r-512-dpt-4-64.pth). To model short-term temporal context, we maintain a temporal-window KV cache that stores tokens from the most recent $\mathrm{W} = 4$ frames, which models the lossless history information in the latest observations. The hybrid memory block is inserted into all layers inside CUT3R's decoder blocks, in total 12 blocks.

During training, each sample is formed by randomly selecting the start and interval to be an 8-frame clip from a video sequence. At each timestep, the cache is updated by appending the current-frame tokens and removing the oldest ones. We train the model with Adam, using a maximum learning rate of 5×10^{-5} for the decoder and DPT head (Ranftl et al., 2021), and 5×10^{-6} for the encoder, with the min-lr of 1×10^{-7} for a total of 10 epochs. Training is performed on two NVIDIA H200 GPUs for approximately one day, and all evaluations are conducted on a single H200 GPU.

B Details about the Datasets

B.1 Training Datasets

We train our model on a mixture of indoor and outdoor datasets: Dynamic-Replica (Karaev et al., 2023), PointOdyssey (Zheng et al., 2023) with 4500 samples per epoch, ARKitScenes (Baruch et al., 2021) with 9000 samples per batch for the low-resolution part, and 4500 samples per batch for the high-resolution part, VirtualKITTI2 (Cabon et al., 2020) with 4500 samples per batch, Waymo (Sun et al., 2020) with 4500 samples per batch, and SpatialVID (Wang et al., 2025a) with 9000 samples per batch.

This choice ensures our model achieves robust zero-shot generalization and maintains strict spatial-temporal consistency across highly diverse environments. Our training corpus is meticulously balanced with a mixture of indoor and outdoor, synthetic and real-world datasets, totaling 40,500 sampled sequences per epoch.

While recent foundational frameworks like CUT3R (Wang et al., 2025c) rely on massive-scale pre-training utilizing over 400K clips across 32 datasets, we demonstrate that an extensively curated, compute-efficient mixture of approximately 40K sequences is sufficient to achieve highly competitive generalization. By meticulously balancing indoor, outdoor, and dynamic priors, our compact training subset distills the essential geometric and temporal constraints without demanding prohibitive computational resources.

B.2 Evaluation Benchmarks

To guarantee that all evaluations are in real-world scenarios and keep a balance between both indoor and outdoor data, we follow Chen et al. (2025b) to preprocess NuScenes (Caesar et al., 2020) with 20 videos, KITTI (Geiger et al., 2012) with 13 videos, Bonn (Palazzolo et al., 2019) with 5 videos, and TUM-Dynamic (Sturm et al., 2012) with 8 videos, into evaluation video clips of different lengths, including 50 frames, 100 frames, 200 frames, and 500 frames. This makes the evaluation not only reflect the model's generalization ability across different types of scenes, but also assess if the model can work robustly from a short horizon to a very long horizon.

C Efficiency Analysis

Table 3 compares the computational efficiency of different streaming 3D perception methods, including CUT3R (Wang et al., 2025c), TTT3R (Chen et al., 2025b), StreamVGGT (Zhuo et al., 2025), and our HybridDepth. We report the number of parameters, FLOPs, peak GPU memory consumption during inference, as well as the runtime throughput (FPS) and per-frame latency. All methods are evaluated under the same input resolution and hardware setup for fair comparison.

CUT3R and TTT3R achieve the highest throughput (around 30 FPS) with identical parameter counts and computational costs. However, both methods require storing intermediate states for sequential updates, resulting in a peak memory usage of 6.31 GB. StreamVGGT, on the other hand, has substan-

Table 3: **Efficiency analysis.** We compare the computational efficiency of CUT3R (Wang et al., 2025c), TTT3R (Chen et al., 2025b), StreamVGGT (Zhuo et al., 2025), and our HybridDepth. For fair comparison, all methods are evaluated under the same input resolution and hardware setup. We report the number of parameters, FLOPs, peak GPU memory during inference, throughput (FPS), and per-frame latency. Our HybridDepth only uses about 4% more parameters, but can keep a good balance between local details and long-horizon consistency.

Method	Params (M) ↓	FLOPs (G) ↓	Peak Memory (GB) ↓	FPS ↑	Latency (ms) ↓
CUT3R	793.31	24336.78	6.31	30.32	32.98
TTT3R	793.31	24336.78	6.31	29.55	33.84
StreamVGGT	1256.54	36033.19	13.43	10.64	93.99
Ours	821.66	25434.10	5.82	24.94	40.31

tially larger computational cost and memory consumption, with 1.26B parameters and 36T FLOPs per frame, leading to significantly slower inference (10.64 FPS) and higher latency (93.99 ms).

Our HybridDepth introduces a hybrid memory design that explicitly separates short-term local information from long-range global context while keeping the computational overhead modest. As shown in Table 3, our method maintains a comparable model size (821.66M parameters) and computational cost (25.4T FLOPs) to CUT3R and TTT3R, while achieving the lowest peak GPU memory consumption (5.82 GB) among all compared methods. In terms of runtime, HybridDepth operates at 24.94 FPS with a latency of 40.31 ms, which remains suitable for real-time streaming applications.

Overall, these results show that the proposed hybrid memory formulation introduces only minor computational overhead while providing improved memory efficiency and maintaining practical real-time performance.

D Additional Temporal and Long-Horizon Evaluation

D.1 Temporal Geometric Consistency

We evaluate temporal alignment error (TAE) on 50-frame Bonn sequences. Predicted depths are aligned once per complete sequence with a single scale and shift. For each adjacent frame pair, source depths are back-projected using the ground-truth camera intrinsics, transformed using the ground-truth relative pose, and reprojected into the target view. We compute depth-reprojection error over valid, non-occluded pixels in both temporal directions and average over adjacent pairs. All methods use the same poses, intrinsics, masks, and alignment protocol. Lower TAE indicates better adjacent-frame geometric consistency; it does not imply a globally consistent 3D map.

Table 4: Temporal alignment error on 50-frame Bonn sequences. Lower is better.

Method	TAE ↓
CUT3R	0.78
TTT3R	0.67
FlashDepth	0.82
StreamVGGT	0.63
HybridDepth	**0.55**

HybridDepth attains the lowest TAE among the compared streaming methods (Table 4), complementing the per-frame depth-accuracy results in the main paper.

D.2 Accuracy on a Common Benchmark Subset

NuScenes has no 500-frame entry in Table 1. To compare the same benchmarks across horizons, Table 5 reports equal-weight averages over KITTI, Bonn, and TUM-Dyn at 50, 200, and 500 frames.

These values are computed from the rounded per-dataset entries in the main table; they must not be compared directly with its four-dataset averages at 50 and 200 frames.

Table 5: Common three-benchmark average across sequence lengths. Each entry is Abs Rel ↓ / δ_1 ↑. Values are recomputed from the displayed KITTI, Bonn, and TUM-Dyn results in Table 1; OOM denotes out of memory.

Method	50 frames	200 frames	500 frames
CUT3R	9.10 / 90.3	9.35 / 91.0	11.13 / 88.1
TTT3R	8.65 / 92.3	8.59 / 92.2	**9.80 / 90.1**
FlashDepth	10.76 / 89.7	12.24 / 88.6	13.37 / 85.3
StreamVGGT	8.24 / 91.9	12.91 / 81.7	OOM
HybridDepth	**8.09 / 93.5**	**8.92 / 92.5**	10.12 / 90.0

The three-benchmark comparison makes the long-horizon limitation explicit: HybridDepth improves on its 50-frame result at neither 200 nor 500 frames, and TTT3R has the better average at 500 frames. HybridDepth nevertheless has the best 500-frame KITTI result in Table 1. Its Abs Rel increase from 50 to 500 frames is 2.02 on this common subset, versus 1.15 for TTT3R.

D.3 Component Ablation Beyond 50 Frames

Table 6 extends the component and memory-design analysis from 50 to 200 frames using equal-weight averages over the same KITTI, Bonn, and TUM-Dyn benchmarks. The 50-frame entries are derived from the per-dataset results in Table 2; the 200-frame entries come from the additional matched-dataset ablation. The full model is the hybrid-memory variant.

Table 6: Component and memory-design ablations on the common three-benchmark subset. Each entry is Abs Rel ↓ / δ_1 ↑.

Setting	50 frames	200 frames
Baseline	9.05 / 92.2	10.15 / 90.6
+ Gate	8.87 / 92.3	9.81 / 91.0
+ Gate + $\mathcal{L}_{\text{orth}}$	8.47 / 93.0	9.34 / 91.8
+ Gate + $\mathcal{L}_{\text{rel}}^{\text{seq}}$	8.51 / 92.9	9.43 / 91.7
Local only	8.40 / 93.0	10.20 / 90.8
Global only	8.38 / 93.1	9.48 / 91.7
Hybrid (full model)	**8.09 / 93.5**	**8.92 / 92.5**

At 200 frames, hybrid memory lowers Abs Rel by 1.28 relative to local-only memory and by 0.56 relative to global-only memory, while also improving δ_1. This supports complementary roles for recent detail and recurrent context beyond the original 50-frame ablation.