RegVGGT: Sustainable Visual Geometry Grounding for Streaming via Regulated Memory

Hongbo Mao, Junjun Jiang*, Youyu Chen, Jiaxin Zhang, Zhemeng Dong, and Xianming Liu

Harbin Institute of Technology, Harbin 150001, China
jiangjunjun@hit.edu.cn
https://github.com/amao996/RegVGGT

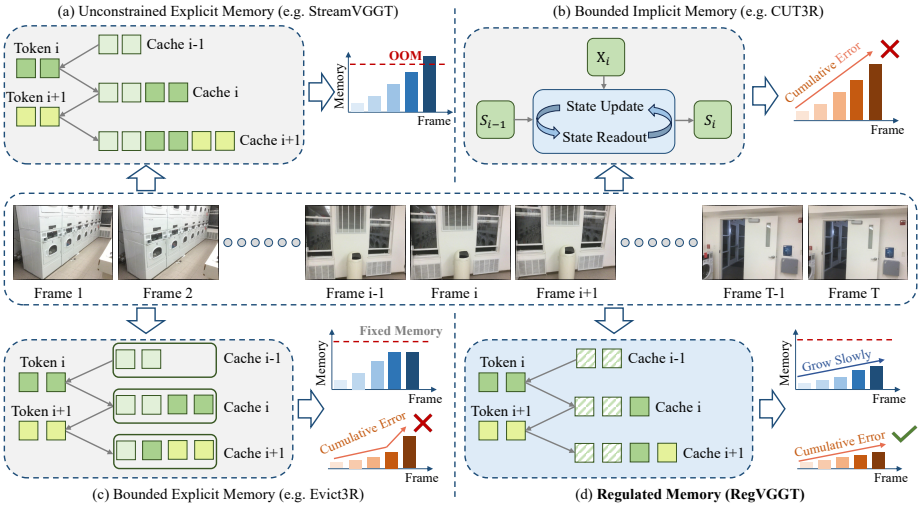

Fig. 1: Paradigm Comparison between previous streaming reconstruction paradigms and our RegVGGT. (a) Unconstrained explicit memory retains all historical KV pairs, causing memory usage to inflate rapidly. (b) Bounded implicit memory compresses historical information into compact hidden states, resulting in catastrophic forgetting. (c) Bounded explicit memory enforces a fixed memory budget, relying on the implicit prior of scene scales to properly measure the budget. (d) Our regulated memory admits at most 1% of tokens per frame to update the context memory, dramatically suppressing the memory inflation as the stream progresses.

Abstract. 3D reconstruction from a lengthy video stream input poses a dilemma for **f**eed-**f**orward **r**econstruction **m**odels (FFRMs), that a whole-stream inference context cannot be retained under limited GPU memory. Recent studies seek to resolve this problem via a trade-off between the integrity of inference context and GPU memory usage, which either suffer from a rapid memory inflation or degraded context integrity due to artificially capping memory usage. Driven by our key observation

* Corresponding anthor.

that the initial saliency of a token reliably dictates its long-term importance across the stream, we propose **RegVGGT**, a training-free token regulation method which aggressively regulates the tokens of incoming frames. By admitting at most 1% of tokens per frame to update the context memory, our method dramatically suppresses memory inflation as the stream progresses. Equipped with a FlashAttention-compatible token saliency estimation scheme, RegVGGT is capable of processing thousands of frames on a consumer-grade GPU with negligible compromise to reconstruction quality. Extensive experiments demonstrate that RegVGGT achieves state-of-the-art performance on long-horizon benchmarks across diverse FFRM prediction tasks, surpassing prior FFRM-based stream reconstruction baselines by a large margin.

Keywords: Feed-forward Reconstruction · Streaming 3D Reconstruction · Training-free Token Regulation

1 Introduction

Dense 3D geometry reconstruction from 2D image collections has long been a key topic of 3D computer vision, underpinning a variety of downstream applications such as augmented reality (AR) [8, 13, 36], embodied robotics [16, 30, 34], and autonomous driving [9, 10, 37]. Building upon decades of foundational research [7, 20, 32], recent advances have achieved an end-to-end reconstruction paradigm to predict dense 3D geometry directly from input images in a feed-forward manner [27, 29, 31]. These feed-forward reconstruction models (FFRMs) are built upon visual transformers [6], which embed input images into tokens, perform reconstruction with attention mechanisms [5, 25] across the tokens, and decode the tokens into dense 3D geometry such as point maps and camera poses.

Despite their remarkable achievements, these FFRMs operate in an offline inference regime that forces reprocessing all historical frames whenever a new frame arrives, making it infeasible to process lengthy video streams. By implementing the token attention operation using a temporal causal approach to replace the canonical global attention, recent studies [12, 38] adapt FFRMs to process incoming frames one-by-one, namely online inference. This largely alleviates the computational overhead for stream reconstruction, because the historical tokens can be directly accessed from memory eliminating the need to recompute them as each new frame arrives. To further extend the capacity of historical frame context embedded into limited memory, various memory bank strategies have been introduced to maintain historical frame tokens as context implicitly [2, 15, 28] or explicitly [17, 24, 35] for reconstructing the incoming frames, as shown in Fig. 1. While these methods succeed in capping memory usage, they suffer from forgetting crucial historical context as the processed frames increase, resulting in degraded reconstruction accuracy.

In this paper, we propose RegVGGT, a training-free token regulation method designed to adapt a prior FFRM [38] to lengthy stream input, which *merely retains at most 1% tokens for each incoming frame* as the context memory for

future frames. Our method stands upon a key observation, that the saliency of each FFRM token to all other tokens across the stream is temporally consistent. Thus, the initial saliency of a FFRM token reliably dictates its long-term importance across the stream. This allows us to aggressively prune redundant tokens online for each frame without discarding possibly salient tokens for the future frames. To further improve the efficiency for temporal causal attention of our method, we introduce a FlashAttention-compatible saliency estimation mechanism that recovers precise attention scores on-the-fly via Log-Sum-Exp (LSE) statistics without materializing the complete attention map. Applying our method to a pretrained FFRM using temporal causal attention [38] to strengthen its historical context memory efficiency, RegVGGT achieves state-of-the-art performance on long-horizon benchmarks [3, 19, 23] across diverse FFRM prediction tasks, surpassing prior FFRM-based stream reconstruction baselines [17, 24, 35] comprehensively by a large margin. The experimental results showcase the effectiveness of our method, providing valuable insights to the community for adapting FFRMs to stream reconstruction.

To summarize our contributions:

1. We report a heuristic behavior of FFRM inference on image stream input, that the initial saliency of a FFRM token reliably dictates its long-term importance across the stream.
2. We propose RegVGGT, which aggressively retains at most 1% of tokens for each incoming frame from the stream input with minimal compromise in the integrity of historical stream context.
3. Extensive experiments show that RegVGGT achieves state-of-the-art performance on long-horizon benchmarks, surpassing baseline methods comprehensively by a large margin.

2 Related Works

Feed-Forward 3D Reconstruction. Building upon decades of foundational research [7, 20, 32], recent advances have achieved an end-to-end reconstruction paradigm to predict dense 3D geometry directly from input images in a feed-forward manner. Pioneering feed-forward reconstruction models (FFRMs) such as DUSt3R [29] and MASt3R [14] formulate 3D reconstruction as pairwise pointmap regression, predicting dense 3D geometry by leveraging cross-attention between image pairs, but require costly global alignment post-processing. To address this, VGGT [27] employs global self-attention across all image tokens, predicting point maps and camera poses in a single forward pass, eliminating the need for such alignment. Subsequent works optimize VGGT across multiple dimensions, such as flexible inputs [11] and equivariant designs [31]. Despite their impressive reconstruction accuracy, these FFRMs operate in an offline inference regime that forces reprocessing all historical frames whenever a new frame arrives, making it infeasible to process lengthy video streams.

Feed-Forward Streaming 3D Reconstruction. Streaming 3D reconstruction targets the incremental processing of image streams. To adapt FFRMs to this paradigm, the critical first step is to avoid reprocessing all historical frames whenever a new frame arrives. Recent streaming reconstruction methods [12,38] replace canonical global attention with a temporal causal approach, explicitly retaining all historical Key–Value (KV) pairs as context memory. This largely alleviates the computational overhead for stream reconstruction, because historical tokens can be directly accessed from memory, eliminating the need to recompute them upon the arrival of each new frame. However, their unconstrained retention of full KV pairs causes memory usage to grow rapidly, rendering them unsustainable for long-horizon reconstruction.

Memory Bank Strategy. To further extend the capacity of historical frame context embedded into limited memory, various memory bank strategies have been introduced to implicitly or explicitly maintain historical frame tokens as context for reconstructing the incoming frames. Implicit strategies store latent features from historical frames [26, 33] or compress historical information into compact hidden states [2,15,28]. While memory-efficient, they suffer from catastrophic forgetting of crucial historical context as the processed frames increase. Conversely, recent explicit strategies [17, 24, 35] cap the memory usage by enforcing a fixed memory budget. However, they rely on the implicit prior of scene scales to properly measure the pre-defined budget. A fixed budget is inevitably redundant for small-scale scenes and insufficient for large ones, suffering from the same contextual forgetting as the implicit strategies. In contrast, we admit at most 1% of salient tokens per frame to update the context memory, dramatically suppressing the memory inflation as the stream progresses with negligible compromise to reconstruction quality.

3 Method

3.1 Preliminaries

Visual Geometry Grounded Transformer (VGGT) [27]. Our method is to adapt a VGGT-like FFRM to lengthy video stream input. A VGGT-like FFRM takes an image sequence of T frames as input,

$$\mathcal{I} = \{\mathbf{I}_t\}_{t=1}^{T}, \quad \mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}, \tag{1}$$

to directly predict view-wise camera poses, depth maps and point maps in a single forward pass. This process is accomplished with visual transformer [6], by encoding images into tokens via an encoder [18], performing reconstruction in the embedding space with alternated frame-attention and global-attention layers, and finally decoding the tokens with respective prediction heads to produce the dense reconstruction results. Though such an architectural design renders impressive reconstruction performance, inference with all input views in one

shot and the underlying quadratic growth of memory occupancy against view number prohibit VGGT-like FFRMs from processing lengthy video stream input with hundreds of frames.

Canonical Attention Operation. For a better understanding of the following discussions, we first introduce canonical attention operation in visual transformer [6], along with necessary notations. Given three token sequences of length $T \cdot N$ as queries, keys and values, which are denoted as $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{(T \cdot N) \times D}$ respectively, a canonical attention is defined to calculate,

$$\mathbf{O} = \bigoplus_{h=1}^{H} \text{Attn}(\mathbf{Q}_{(h)}, \mathbf{K}_{(h)}) \times \mathbf{V}_{(h)}, \text{Attn}\left(\mathbf{Q}_{(h)}, \mathbf{K}_{(h)}\right) = \text{Softmax}\left(\frac{\mathbf{Q}_{(h)} \times \mathbf{K}_{(h)}^{\top}}{\sqrt{D/H}}\right), \tag{2}$$

where N denotes the number of tokens of a frame, $\bigoplus$ denotes concatenation operation along feature dimension, H denotes the number of attention heads, D denotes the dimensional size of token features, and $\mathbf{Q}_{(h)}, \mathbf{K}_{(h)}, \mathbf{V}_{(h)} \in \mathbb{R}^{(T \cdot N) \times \frac{D}{H}}$ so that $\mathbf{Q} = \bigoplus_h \mathbf{Q}_{(h)}, \mathbf{K} = \bigoplus_h \mathbf{K}_{(h)}, \mathbf{V} = \bigoplus_h \mathbf{V}_{(h)}$. In the remainder of this paper, by default we take $H = 1$ for brevity, shortening the (h) in subscript unless otherwise specified.

Adaption to Stream Input via Temporal Causal Attention. To adapt VGGT-like FFRMs to stream input, the first step is to avoid reprocessing all historical frames each time when a new frame is passed. This is achieved by replacing the canonical global attention operation in VGGT-like architectures with temporal causal attention operation [12, 38]. Specifically, temporal causal attention breaks the queries into view-wise splits that $\mathbf{Q}_t \in \mathbb{R}^{N \times D}$ denotes the query tokens for the t-th frame, being the same for keys and values. Further notating $\mathbf{K}_{\leq t} = \bigcup_{m=1}^{m<=T} \mathbf{K}_m$ (the same for $\mathbf{Q}$ and $\mathbf{V}$), there is,

$$\mathbf{O}_t = \text{Attn}(\mathbf{Q}_t, \mathbf{K}_{\leq t}) \times \mathbf{V}_{\leq t}, \mathbf{O}_t \in \mathbb{R}^{N \times D}. \tag{3}$$

By storing $\mathbf{K}_{\leq t-1}$ and $\mathbf{V}_{\leq t-1}$ in memory, namely KV caches, one only need to focus on resolving tokens from the t-th frame to reconstruct this new incoming frame by efficiently looking up memory for historical context. However, retaining complete KV caches in memory lead to a rapid inflation of memory footprint.

3.2 Insights to FFRMs with Temporal Causal Attention

To mitigate the rapid inflation of memory consumption in temporal causal attention based FFRM, recent and concurrent works [17, 24, 35] artificially adopt a fixed token budget to bound the memory footprint. However, as the number of processed frames increases, the budget must decide which memory to abandon for incoming frames, resulting in a distortion of historical frame context. Furthermore, it remains unclear *how to ensure that the tokens currently abandoned are not salient for the future frames*. We resolve these two questions with three empirical observations, which are discussed below.

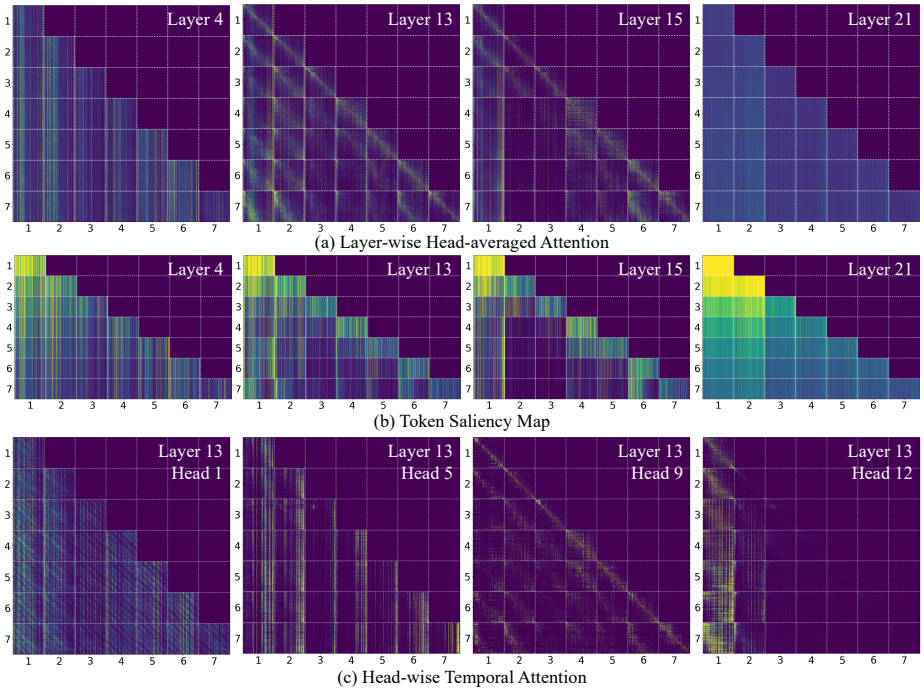

Fig. 2: Visualization of temporal causal attention. We visualize attention maps to reveal the interaction between queries (rows) and keys (columns). Frames are sampled in a stride of 100 from ScanNet [3] dataset to depict the long-term behaviour of causal temporal attention. a) We depict layer-wise head-averaged attention map, visualizing $\mathrm{Attn}(\mathrm{Q}_t, \mathrm{K}_{t\leq 1})$ in Eq. (3) across frame index t. b) We sum up each column, *i.e.* the attention of different queries for the same key, showing a temporal non-increasing response of key token saliency. c) The head-wise attention maps within Layer 13.

Observation 1: Temporal Consistency of Token Saliency. As illustrated in Fig. 2, the attention scores of each key token are highly consistent with respect to queries from subsequent frames. Key tokens assigned high attention scores in their source frame consistently remain salient for the future frames, whereas those with initially low attention scores rarely become salient later.
Implication. The temporal consistency behaviour of token saliencies indicates that they could be confidently estimated online using attention scores of the current frame, with negligible risk of discarding tokens which are possibly salient for future frames.

Observation 2: Differentiated Attention Behaviour Across Layers. We profile the difference in the behaviour of attention maps across different transformer layers, visualizing the head-averaged temporal causal attention maps in Fig. 2 (a). In most early and late layers (*e.g.*, Layer 4 and 21, denoted as $\mathcal{L}_{\mathrm{qa}}$), the

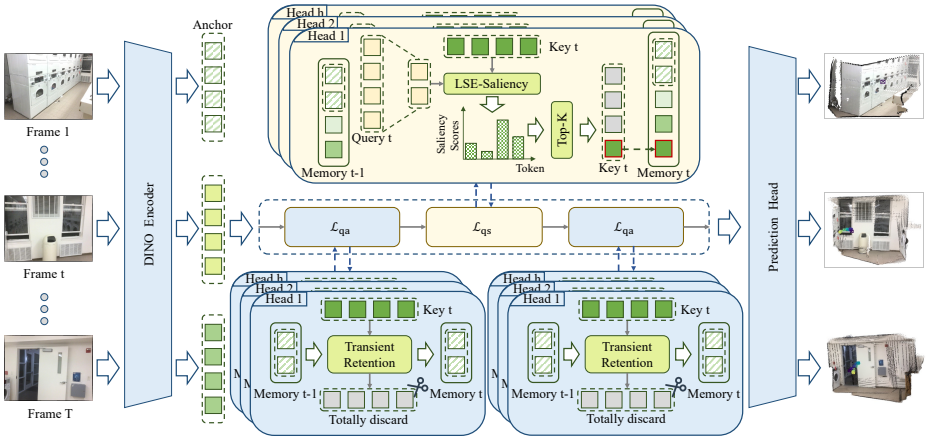

Fig. 3: Overview of RegVGGT. Instead of enforcing a fixed memory budget, RegVGGT regulates the per-frame information admitted into memory through a training-free token regulation strategy. We decouple layers into two types. $\mathcal{L}_{\text{qa}}$ entirely discard intermediate history, while $\mathcal{L}_{\text{qs}}$ strictly preserve the top 1% of salient tokens per frame. Coupled with LSE-Saliency, RegVGGT achieves long-term geometric consistency and true scene-agnostic scalability.

attention maps exhibit query-agnostic intensity along the column. In contrast, the attention maps in the intermediate layers (*e.g.*, Layers 13 and 15, denoted as $\mathcal{L}_{\text{qs}}$) display query-specific, sparse and diagonally structured attention patterns.
Implication. For layers with query-agnostic attention maps, we empirically find that we can only preserve the key/value tokens of the first frame without hindering reconstruction, which is to say zero update in memory is demanded beyond memorizing the first frame. It suggests that these high response key tokens might be consistent anchors in physical world to align the subsequent frames with the first frame, thus a preservation of first frame tokens is sufficient to make these layers function well. For other layers, their sparse attention intensity suggests a possibility of applying aggressive token pruning on them to suppress the update to historical token memory, minimizing the memory footprint.

Observation 3: Distinct Function of Attention Heads. Zooming into layers with query-specific attentions, we investigate attention maps at the level of attention head, as shown in Fig. 2(c). The attention maps show head-distinct behaviour, indicating their difference between each other.
Implication. The saliency of tokens should be evaluated independently for each attention head for best quantification.

3.3 Proposed Method: RegVGGT

As established in Observation 1, token saliency can be reliably estimated on-the-fly using current-frame attention scores, ensuring the most salient tokens

retained at the current frame contain the most important context for the future frames. Instead of rigidly capping memory footprint, we strictly regulate the information admitted into memory at a constant rate per frame, where the rate can be extremely small according to Observation 2. A head-specific token retention strategy is further applied according to Observation 3. All proposals above, integrated with a FlashAttention-compatible saliency estimation strategy, together compose our RegVGGT, as illustrated in Fig. 3, which retains at most 1% tokens per frame to update the historical context memory without compromising reconstruction accuracy.

Global Anchor Preservation. In VGGT-like FFRMs [12, 27, 38], the first frame establishes the global coordinate system to which all subsequent frames are aligned. Validating this structural reliance, our empirical analysis reveals that tokens from the first frame receive sustained high attention over time, especially within the intermediate layers (*e.g.*, Layers 13 and 15 in Fig. 2). Pruning these tokens would severely damage long-term geometric consistency. Therefore, we designate the complete key–value pairs of the first frame as global anchor tokens, which are permanently preserved in the memory across all L layers:

$$\mathcal{M}_1^{(l)} = (\mathbf{K}_1^{(l)}, \mathbf{V}_1^{(l)}), \quad \forall l \in \{1, \dots, L\}, \tag{4}$$

where $\mathbf{K}_1^{(l)}$ and $\mathbf{V}_1^{(l)}$ denote the key/value tokens in the l-th layer of the 1-st frame, and $\mathcal{M}_1^{(l)}$ represents the corresponding anchor memory.

Layer-Specific Memory Allocation. Motivated by the varying attention behaviour across layers identified in Observation 2, we design a differentiated layer-specific memory allocation strategy for $\mathcal{L}_{\text{qa}}$ and $\mathcal{L}_{\text{qs}}$ (see Observation 2 for these two notations). For $\mathcal{L}_{\text{qa}}$, we only preserve the global anchor tokens in the first frame, while the KV caches from all subsequent frames are prohibited from memory,

$$\mathcal{M}_t^{(l)} = \mathcal{M}_{\text{anchor}}^{(l)}, \quad \forall l \in \mathcal{L}_{\text{qa}}, \tag{5}$$

where $\mathcal{M}_t^{(l)}$ denotes the historical memory maintained at temporal index t.

In contrast, capitalizing on the diagonally structured attention patterns, we regulate the per-frame information admitted into memory for $\mathcal{L}_{\text{qs}}$. In addition to preserving the global anchor tokens, we retain only the top k most salient tokens per incoming frame, where $k = \lfloor \gamma N \rfloor$,

$$\mathcal{M}_t^{(l)} = \mathcal{M}_{t-1}^{(l)} \bigcup \text{Top-}k(\mathbf{K}_t^{(l)}, \mathbf{V}_t^{(l)}), \quad \forall l \in \mathcal{L}_{\text{qs}}, \tag{6}$$

where $\gamma \in (0, 1]$ represents the per-frame token retention rate. Unlike fixed-budget baselines that enforce a rigid capacity constraint ($|\mathcal{M}_t| \leq B$), our memory footprint increases linearly at an extremely low rate, that $|\mathcal{M}_t| = |\mathcal{M}_1| + (t-1) \cdot \lfloor \gamma N \rfloor$. By setting the retention rate γ to merely 0.01, we ensure our memory footprint growth remains highly sustainable.

Head-Specific Token Retention. Motivated by Observation 3, we introduce a head-specific token retention strategy. First, to prevent the destructive interference caused by naively averaging attention scores across heads, we evaluate token saliency and maintain memory independently across attention heads within $\mathcal{L}_{\text{qs}}$. Second, to minimize computational overhead introduced by saliency estimation, we exploit the inherent spatial redundancy among neighboring queries. Specifically, we estimate token saliency using a spatially downsampled subset of queries, denoted as $\tilde{\mathbf{Q}}_t$. This subset is obtained by partitioning the current frame's queries into 2×2 non-overlapping windows and sampling the top-left query from each window. The saliency score S_j for the j-th token $\mathbf{k}_{t,j} \in \mathbb{R}^{1\times\frac{D}{H}}$ of $\mathbf{K}_t^{(l)}$ in head h is then defined as the cumulative attention it receives from this sampled query subset, which is the sum of the j-th column in $\text{Attn}(\tilde{\mathbf{Q}}_t, \mathbf{K}_t) \in \mathbb{R}^{\frac{N}{4}\times N}$, as depicted in the diagonal view patches of Fig. 2(b). Finally, the top-k tokens with the highest saliency scores are selected and retained independently for each attention head, effectively preserving sparse, fine-grained geometric cues.

FlashAttention-Compatible Saliency Estimation. Attention-based token retention strategies, including ours, rely on exact attention scores, which are not explicitly materialized by FlashAttention [4,5]. Constructing full attention matrices would incur $\mathcal{O}(N|\mathcal{M}_{\leq t}|)$ memory overhead, defeating the purpose of memory efficiency and low latency. To address this problem, we introduce LSE-Saliency, a FlashAttention-compatible saliency estimation mechanism that recovers exact attention scores on-the-fly using Log-Sum-Exp (LSE) statistics without materializing full attention maps. FlashAttention computes and caches a normalization factor (the log-partition function) for each query during its forward pass:

$$\text{LSE}_i = \log \sum_{m=1}^{t} \sum_{j=1}^{N} \exp(\mathbf{q}_{t,i}\mathbf{k}_{m,j}^{\top}/\sqrt{\frac{D}{H}}). \tag{7}$$

Leveraging this cached statistic, we can efficiently compute the exact attention scores of any token with respect to the entire historical context. Given the scaled attention logits between queries and current-frame keys, defined as $Z_{ij} = \mathbf{q}_i\mathbf{k}_j^{\top}/\sqrt{\frac{D}{H}}$, the exact attention scores can be recovered as:

$$A_{ij} = \exp(Z_{ij} - \text{LSE}_i), \tag{8}$$

where $A_{i,j}$ denotes the value in the i-th row and j-th column of $\text{Attn}(\tilde{\mathbf{Q}}_t, \mathbf{K}_t)$. This allows us to perform fine-grained token retention while fully preserving the hardware efficiency of FlashAttention.

4 Experiment

4.1 Experiment Setup

We evaluate RegVGGT on multiple 3D tasks, including camera pose estimation, video depth estimation, and 3D reconstruction. To assess scalability under long input streams, we focus on test scenes containing a large number of frames.

Table 1: Camera Pose Estimation Results on TUM Dynamics [23] and ScanNet [3].

Model	Frames	TUM Dynamics			ScanNet		
		ATE (m)↓	RPE trans↓	RPE rot↓	ATE (m)↓	RPE trans↓	RPE rot↓
Evict3R* [17]	800	0.162	0.028	1.572	1.045	0.100	14.837
InfiniteVGGT* [35]		0.130	0.045	1.646	1.007	0.138	12.405
XStreamVGGT* [24]		0.121	0.020	1.354	1.000	0.089	10.225
RegVGGT* (ours)		**0.109**	**0.016**	**1.288**	**0.949**	**0.071**	**6.662**
Evict3R* [17]	900	0.160	0.027	1.632	1.073	0.100	15.159
InfiniteVGGT* [35]		0.132	0.043	1.709	1.036	0.146	11.861
XStreamVGGT* [24]		0.147	0.024	1.579	1.029	0.090	10.865
RegVGGT* (ours)		**0.125**	**0.016**	**1.373**	**0.976**	**0.072**	**6.779**
Evict3R* [17]	1000	0.174	0.027	2.004	1.114	0.101	15.333
InfiniteVGGT* [35]		0.141	0.042	1.865	1.061	0.141	11.622
XStreamVGGT* [24]		0.174	0.027	1.866	1.054	0.090	11.182
RegVGGT* (ours)		**0.135**	**0.016**	**1.431**	**0.997**	**0.072**	**6.656**

Implementation Details. RegVGGT is designed as a training-free token regulation strategy. We implement it on top of StreamVGGT [38], keeping the original network architecture and pre-trained weights frozen; our modifications are confined solely to the online token regulation mechanism. Furthermore, we adopt the VRAM-efficient implementation from FastVGGT [21], which discards intermediate features from 20 layers that are not required by subsequent modules. Models utilizing this baseline optimization are denoted with an asterisk (*e.g.*, StreamVGGT*). Unless otherwise specified, all experiments are conducted under this setting. All evaluations are performed on a single NVIDIA RTX 3090 GPU (24GB VRAM).

Configuration of RegVGGT. Guided by the empirical analyses in Sec. 3.2, we partition the 24 Transformer layers into two distinct functional groups. We designate Layers 1–10 and 18–24 as $\mathcal{L}_{\text{qa}}$, and Layers 11–17 as $\mathcal{L}_{\text{qs}}$. For $\mathcal{L}_{\text{qs}}$, we apply our token regulation strategy by performing a 2×2 spatial downsampling on queries and setting the retention rate to $\gamma = 0.01$ (retaining only 1% of salient tokens per incoming frame). To demonstrate the robustness of our method, these hyperparameter settings remain strictly fixed across all experiments and datasets.

Fair Comparison Protocol. Crucially, to ensure a rigorously fair comparison against baselines, we dynamically align the memory footprint. For any given input sequence length and dataset, we calculate the total number of tokens retained by RegVGGT, then enforce identical token budget for all baselines accordingly. This protocol ensures that all performance comparisons are conducted under identical memory footprints, thereby isolating the effectiveness of our token regulation mechanism.

Table 2: 3D Reconstruction Results on 7-Scenes [22] and NRGBD [1].

Model	Frames	7-Scenes						NRGBD					
		Acc↓		Comp↓		NC↑		Acc↓		Comp↓		NC↑	
		Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
Evict3R* [17]	200	0.106	0.047	0.059	0.017	0.578	0.621	0.107	0.061	0.035	0.010	0.638	0.741
InfiniteVGGT* [35]		0.067	0.036	0.054	0.005	0.579	0.623	0.062	0.042	0.017	**0.004**	0.631	0.725
XStreamVGGT* [24]		0.075	0.038	0.029	0.005	0.581	0.627	0.076	0.051	0.023	0.005	0.633	0.733
RegVGGT* (ours)		**0.036**	**0.020**	**0.024**	**0.004**	**0.591**	**0.643**	**0.043**	**0.031**	**0.016**	**0.004**	**0.641**	**0.745**
Evict3R* [17]	250	0.127	0.062	0.062	0.017	0.567	0.602	0.117	0.072	0.050	0.013	0.633	0.729
InfiniteVGGT* [35]		0.096	0.065	0.063	0.008	0.570	0.607	0.073	0.048	0.025	**0.004**	0.628	0.720
XStreamVGGT* [24]		0.102	0.054	0.030	0.007	0.571	0.610	0.080	0.051	0.025	**0.004**	0.629	0.724
RegVGGT* (ours)		**0.037**	**0.020**	**0.022**	**0.004**	**0.583**	**0.629**	**0.050**	**0.030**	**0.018**	**0.004**	**0.640**	**0.741**
Evict3R* [17]	300	0.163	0.091	0.080	0.032	0.552	0.579	0.168	0.112	0.062	0.019	0.611	0.683
InfiniteVGGT* [35]		0.122	0.091	0.061	0.009	0.563	0.596	0.081	0.058	**0.026**	**0.004**	0.622	0.707
XStreamVGGT* [24]		0.129	0.075	0.036	0.008	0.563	0.596	0.102	0.068	0.035	0.005	0.618	0.702
RegVGGT* (ours)		**0.038**	**0.020**	**0.022**	**0.004**	**0.576**	**0.617**	**0.068**	**0.041**	**0.026**	0.006	**0.630**	**0.724**

4.2 Camera Pose Estimation

Following CUT3R [28] and TTT3R [2], we evaluate camera pose accuracy on the TUM Dynamics [23] and ScanNet [3] datasets, adopting Absolute Trajectory Error (ATE), Relative Translation Error (RPE trans), and Relative Rotation Error (RPE rot) as our evaluation metrics. To systematically assess tracking robustness over long horizons, we follow the protocol established in TTT3R [2]. For TUM Dynamics [23], we extract the initial frames of each sequence. For ScanNet [3], we apply a temporal stride of 3. Both datasets yield evaluation sequences ranging from 50 to 1,000 frames in length. Quantitative results for extreme sequence lengths of 800, 900, and 1000 frames are presented in Tab. 1. As demonstrated, RegVGGT consistently outperforms all competing methods across all metrics and sequence lengths. Notably, as the input length increases, fixed-budget methods experience severe performance degradation, whereas RegVGGT sustains robust tracking accuracy.

4.3 3D Reconstruction

Following the standard protocol of CUT3R [28], we evaluate 3D reconstruction quality on the 7-Scenes [22] and NRGBD [1] datasets. To systematically assess long-horizon performance, we follow TTT3R [2] and sample each sequence with a temporal stride of 2, utilizing the first 50 to 300 frames as input. We report performance across three metrics, namely Accuracy (Acc), Completion (Comp), and Normal Consistency (NC). Quantitative results for extreme sequence lengths of 200 and 300 frames are presented in Tab. 2. As demonstrated, RegVGGT establishes new state-of-the-art performance across the vast majority of metrics. Crucially, as the sequence length scales from 200 to 300 frames, fixed-budget baselines suffer drastic degradation in accuracy and consistency (*e.g.*, Evict3R's mean accuracy error surges by over 50% on 7-Scenes). In contrast, RegVGGT remains highly stable with minimal performance degradation, validating the effectiveness of our method.

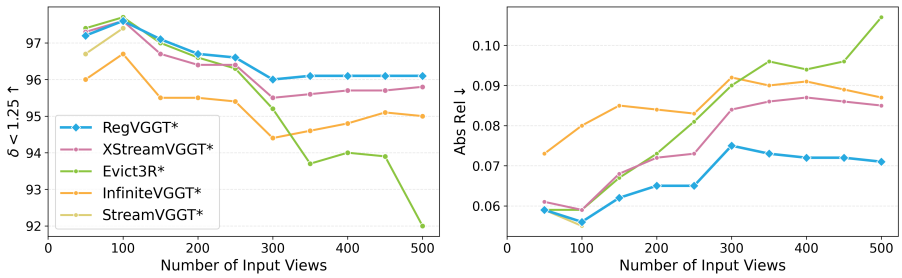

Fig. 4: Video Depth Estimation Results on Bonn [19].

Table 3: Ablation Study on Global Anchor Preservation.

Method	ATE (m)↓	RPE trans↓	RPE rot↓
w/o Global Anchor	0.175	0.063	4.005
w/ Global Anchor	**0.165**	**0.061**	**3.930**

Table 4: Ablation Study on Head-Specific Token Retention.

Method	ATE (m)↓	RPE trans↓	RPE rot↓	Latency (ms)↓
w/o Head-specific Retention	0.165	0.061	**3.840**	17.66
w/ Head-specific Retention	0.165	0.061	3.930	**8.29**

Table 5: Ablation Study on Layer-Specific Memory Allocation.

Method	ATE (m)↓	RPE trans↓	RPE rot↓
w/o Layer-Specific Retention	0.176	0.066	4.970
w/ Layer-Specific Retention	**0.165**	**0.061**	**3.930**

Table 6: Sensitivity Analysis of Query Downsampling Factor.

Downsample Factor	ATE (m)↓	RPE trans↓	RPE rot↓	Latency (ms)↓
1 × 1	**0.165**	**0.060**	**3.735**	17.64
2 × 2 (default)	**0.165**	0.061	3.930	8.29
3 × 3	0.167	0.062	4.224	7.69
4 × 4	0.168	0.062	4.260	**7.45**

4.4 Video Depth Estimation

Since most standard benchmarks contain a highly limited number of consecutive frames, we follow the protocol of TTT3R [2] to evaluate long-horizon performance on the Bonn [19] dataset. Specifically, after discarding the initial 30 frames of each sequence, we sample continuous subsequences ranging from 50 to 500 frames. Quantitative results, measured by Absolute Relative error (Abs Rel) and threshold accuracy ($\delta < 1.25$, denoting the percentage of predicted depths within a factor of 1.25 of the ground truth), are presented in Fig. 4. While Evict3R [17] performs competitively with RegVGGT on ultra-short sequences (50 and 100 frames), our model rapidly establishes state-of-the-art dominance as the temporal horizon extends. Crucially, as the input length increases, the performance gap between our method and fixed-budget baselines widens substantially. This diverging trend directly highlights the superior scalability and robustness of our regulated memory.

4.5 Ablation Study

We conduct ablation studies on the ScanNet [3] dataset to systematically evaluate the contribution of each key component in our architecture. Beyond validating the effectiveness of individual architectural components, we provide an in-depth analysis of critical parametric choices.

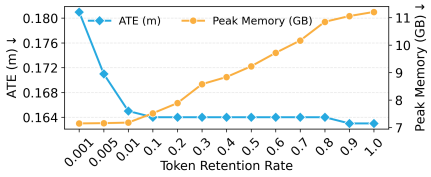

Fig. 5: Sensitivity Analysis of Token Retention Rate.

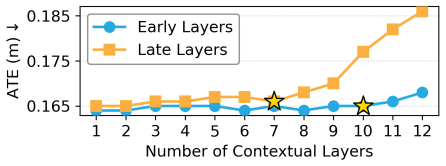

Fig. 6: Sensitivity Analysis of Layer Decoupling.

Effectiveness of Core Components. We ablate the global anchor preservation, the layer-specific memory allocation and the head-specific token retention under an identical total token budget to evaluate their individual contributions. As demonstrated in Tables 3 and 5, independently removing either the global anchor or the layer-specific memory allocation leads to a substantial degradation in tracking accuracy. Furthermore, while these two mechanisms secure tracking robustness, Tab. 4 reveals the critical efficiency contribution of our head-specific token retention strategy. Our strategy slashes the inference latency by more than half with a negligible impact on primary accuracy. Ultimately, these results confirm that all three components are indispensable to our architecture.

Sensitivity to Token Retention Rate. We analyze the effect of the token retention rate γ on tracking accuracy and peak memory footprint. As shown in Fig. 5, aggressively reducing γ below 0.01 leads to severe performance deterioration while yielding negligible memory savings, indicating that critical geometric cues are destructively over-pruned. However, performance rapidly stabilizes and remains highly robust once $\gamma \geq 0.01$. Furthermore, pushing γ beyond 0.1 provides negligible geometric benefits while continuously inflating the peak memory footprint. This trade-off curve demonstrates that retaining a mere 1% of salient tokens per frame is entirely sufficient for RegVGGT to preserve strict geometric consistency. Consequently, we adopt $\gamma = 0.01$ as the default configuration across all experiments.

Sensitivity to Layer Decoupling. Our method decouples Transformer layers into $\mathcal{L}_{\text{qa}}$ and $\mathcal{L}_{\text{qs}}$ based on the differentiated attention behaviour across layers revealed in Sec. 3.2. To translate this qualitative observation into an optimal architectural configuration, we evaluate the sensitivity to layer decoupling. Specifically, we progressively designate varying proportions of early and late layers as $\mathcal{L}_{\text{qa}}$, while treating the remaining intermediate layers as $\mathcal{L}_{\text{qs}}$. As shown in Fig. 6, we empirically determine the optimal decoupling to be the first 10 and last 7 layers (Layers 1–10 and 18–24) acting as $\mathcal{L}_{\text{qa}}$, and the intermediate 7 layers (Layers 11–17) acting as $\mathcal{L}_{\text{qs}}$. By discarding the entire intermediate history across 75% of the layers, this highly asymmetric design strikes a remarkably favorable balance between state-of-the-art accuracy and extreme memory efficiency.

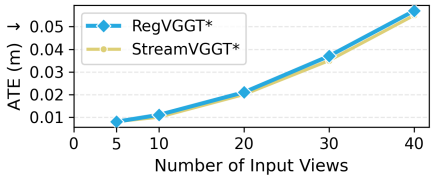

Fig. 7: Analysis of Context Integrity.

Table 7: Generalizability evaluation.

Method	Frames	Abs Rel↓	$\delta < 1.25$ ↑
Evict3R* [17]	400	0.073	95.8
Regulated Evict3R* [17]		**0.071**	**96**
Evict3R* [17]	500	0.076	95.8
Regulated Evict3R* [17]		**0.069**	**96.1**

Sensitivity to Query Downsampling. We evaluate the sensitivity of the downsampling factor on tracking accuracy and the extra inference latency introduced by saliency estimation. As presented in Tab. 6, applying a 2×2 downsampling factor drastically reduces the inference latency by more than 50% while achieving near-lossless primary trajectory accuracy. However, adopting more aggressive downsampling (3×3 and beyond) yields severely diminishing returns in computational speedups while introducing noticeable accuracy degradation. Therefore, we adopt the 2×2 downsampling factor as our default configuration, securing an optimal trade-off between efficiency and tracking accuracy.

Analysis of Context Integrity. To validate our claim that RegVGGT achieves extreme memory efficiency with negligible compromise to model performance, we conduct a comparison against the uncompressed FFRM backbone. Because the canonical backbone suffers from rapid memory inflation on lengthy streams, we strictly limit the evaluation lengths of this experiment to short sequences ranging from 5 to 40 frames. As illustrated in Fig. 7, despite retaining at most 1% of the most salient tokens per frame, RegVGGT exhibits negligible compromise to performance across all evaluation metrics. This compelling result confirms that our method successfully isolates the geometrically crucial information, thoroughly demonstrating that the omitted historical tokens are indeed highly redundant for dense 3D reconstruction.

Generalizability of Token Regulation Strategy. To further validate the effectiveness and generalizability of our token regulation strategy, which retains a mere 1% of salient tokens per frame, we conduct a plug-and-play experiment. Existing streaming methods, such as Evict3R [17], typically employ a fixed global token budget. However, such a rigid constraint is inevitably redundant for small-scale scenes and insufficient for large ones, inevitably leading to catastrophic forgetting as the input sequence grows. To address this, we replace the fixed budget module in Evict3R [17] with our proposed per-frame regulation strategy, ensuring that the evaluation is conducted under identical token budget. The video depth estimation results on the Bonn dataset [19] are presented in Tab. 7. As illustrated, our regulated Evict3R achieves consistent improvements across all metrics. This finding not only demonstrates that a dynamic, sequence-length-

Table 8: Robustness Evaluation in Challenging Scenarios.

Model	γ	Camera Pose			Point Map	
		ATE ↓	RPE-t ↓	RPE-rot ↓	CD ↓	NC ↑
StreamVGGT*	1.00	0.159	**0.668**	**48.970**	**0.054**	**0.750**
RegVGGT*	0.10	**0.154**	0.673	49.010	0.066	0.745
RegVGGT*	0.01	0.155	0.677	49.026	0.099	0.731

Table 9: Quantitative Metrics for Observation 1.

Attention Head	Time Offset (Δt)				
	+10	+30	+50	+70	+90
Head 1	0.997	0.990	0.980	0.979	0.976
Head 5	0.997	0.983	0.967	0.954	0.943
Head 9	0.820	0.603	0.527	0.507	0.439
Head 12	0.926	0.739	0.541	0.479	0.299

aware capacity is intrinsically superior to a fixed budget, but also highlights the strong transferability of our design across different baseline architectures.

Robustness in Challenging Scenarios. We simulate highly dynamic environments and large motions via aggressive 1/100 frame sampling. Specifically, we evaluate camera pose estimation on the TUM-D dataset and point map reconstruction on 7-Scenes (Tab. 8). RegVGGT robustly maintains the backbone's performance in highly dynamic settings. As 7-Scenes results show, large motion sharply reduces temporal redundancy by introducing abundant unique per-frame geometry, greatly alleviated by adopting a milder retention rate of $\gamma = 0.1$.

Quantitative Metrics for Observation 1. Using Layer 13 (Fig. 2) as an example, we quantify temporal consistency on ScanNet by computing the Spearman rank correlation of per-token saliency relative to the 10th frame (Tab. 9). Heads 1 and 5 maintain robust temporal consistency, while Heads 9 and 12 naturally decay as they prioritize global consistency and the reference frame.

5 Conclusion

In this paper, we presented RegVGGT, a training-free online token regulation method that resolves the trade-off between rapid memory inflation and degraded context integrity in FFRMs. Driven by the key observation that the initial saliency of a FFRM token reliably dictates its long-term importance across the stream, our method aggressively regulates the context by retaining at most 1% of the most salient tokens per frame. Coupled with a FlashAttention-compatible saliency estimation mechanism, RegVGGT enables the processing of thousands of frames on consumer-grade GPUs with negligible compromise to reconstruction quality. Extensive experiments demonstrate that our method comprehensively surpasses existing baselines on long-horizon benchmarks, providing valuable insights to the community for adapting FFRMs to stream reconstruction.

Acknowledgements

The research was supported by the National Natural Science Foundation of China (U23B2009, 62471158).

References

1. Azinović, D., Martin-Brualla, R., Goldman, D.B., Nießner, M., Thies, J.: Neural rgb-d surface reconstruction. In: CVPR. pp. 6290–6301 (2022). https://doi.org/10.1109/CVPR52688.2022.00619
2. Chen, X., Chen, Y., Xiu, Y., Geiger, A., Chen, A.: Ttt3r: 3d reconstruction as test-time training. arXiv preprint arXiv:2509.26645 (2025). https://doi.org/10.48550/arXiv.2509.26645
3. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR. pp. 5828–5839 (2017). https://doi.org/10.48550/arXiv.1702.04405
4. Dao, T.: Flashattention-2: Faster attention with better parallelism and work partitioning. arXiv preprint arXiv:2307.08691 (2023). https://doi.org/10.48550/arXiv.2307.08691
5. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. NeurIPS **35**, 16344–16359 (2022). https://doi.org/10.48550/arXiv.2205.14135
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020). https://doi.org/10.48550/arXiv.2010.11929
7. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. IEEE TPAMI **32**(8), 1362–1376 (2009). https://doi.org/10.1109/TPAMI.2009.161
8. Hong, J.G., Noh, S.Y., Lee, H.K., Cheong, W.S., Chang, J.Y.: 3d clothed human reconstruction from sparse multi-view images. In: CVPR. pp. 677–687 (2024). https://doi.org/10.1109/CVPRW63382.2024.00072
9. Huang, N., Wei, X., Zheng, W., An, P., Lu, M., Zhan, W., Tomizuka, M., Keutzer, K., Zhang, S.: $s^3gaussian$: Self-supervised street gaussians for autonomous driving. CoRR (2024). https://doi.org/10.48550/arXiv.2405.20323
10. Huang, Y., Zheng, W., Zhang, Y., Zhou, J., Lu, J.: Tri-perspective view for vision-based 3d semantic occupancy prediction. In: CVPR. pp. 9223–9232 (2023). https://doi.org/10.1109/CVPR52729.2023.00890
11. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025). https://doi.org/10.48550/arXiv.2509.13414
12. Lan, Y., Luo, Y., Hong, F., Zhou, S., Chen, H., Lyu, Z., Yang, S., Dai, B., Loy, C.C., Pan, X.: Stream3r: Scalable sequential 3d reconstruction with causal transformer. arXiv preprint arXiv:2508.10893 (2025). https://doi.org/10.48550/arXiv.2508.10893
13. Lei, J., Weng, Y., Harley, A.W., Guibas, L., Daniilidis, K.: Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In: CVPR. pp. 6165–6177 (2025). https://doi.org/10.1109/CVPR52734.2025.00578
14. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: ECCV. pp. 71–91. Springer (2024). https://doi.org/10.1007/978-3-031-73220-1_5
15. Li, Z., Zhou, J., Wang, Y., Guo, H., Chang, W., Zhou, Y., Zhu, H., Chen, J., Shen, C., He, T.: Wint3r: Window-based streaming reconstruction with camera token pool. arXiv preprint arXiv:2509.05296 (2025). https://doi.org/10.48550/arXiv.2509.05296

16. Liu, S., Gao, Y., Zhang, T., Pautrat, R., Schönberger, J.L., Larsson, V., Pollefeys, M.: Robust incremental Structure-from-Motion with hybrid features. In: ECCV. pp. 249–269. Springer (2024). https://doi.org/10.48550/arXiv.2409.19811
17. Mahdi, S., Ayar, F., Javanmardi, E., Tsukada, M., Javanmardi, M.: Evict3r: Training-free token eviction for memory-bounded streaming visual geometry transformers. arXiv preprint arXiv:2509.17650 (2025). https://doi.org/10.48550/arXiv.2509.17650
18. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023). https://doi.org/10.48550/arXiv.2304.07193
19. Palazzolo, E., Behley, J., Lottes, P., Giguere, P., Stachniss, C.: Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7855–7862. IEEE (2019). https://doi.org/10.1109/IROS40897.2019.8967590
20. Schönberger, J.L., Frahm, J.M.: Structure-from-Motion revisited. In: CVPR. pp. 4104–4113 (2016). https://doi.org/10.1109/CVPR.2016.445
21. Shen, Y., Zhang, Z., Qu, Y., Zheng, X., Ji, J., Zhang, S., Cao, L.: Fastvggt: Training-free acceleration of visual geometry transformer. arXiv preprint arXiv:2509.02560 (2025). https://doi.org/10.48550/arXiv.2509.02560
22. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.: Scene coordinate regression forests for camera relocalization in rgb-d images. In: CVPR. pp. 2930–2937 (2013). https://doi.org/10.1109/CVPR.2013.377
23. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: 2012 IEEE/RSJ international conference on intelligent robots and systems. pp. 573–580. IEEE (2012). https://doi.org/10.1109/IROS.2012.6385773
24. Su, Z., Ye, W., Feng, H., Fan, K., Zhang, J., Yu, D., Liu, Z., Wong, N.: Xstreamvggt: Extremely memory-efficient streaming vision geometry grounded transformer with kv cache compression. arXiv preprint arXiv:2601.01204 (2026). https://doi.org/10.48550/arXiv.2601.01204
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017). https://doi.org/10.48550/arXiv.1706.03762
26. Wang, H., Agapito, L.: 3d reconstruction with spatial memory. arXiv preprint arXiv:2408.16061 (2024). https://doi.org/10.1109/3DV66043.2025.00013
27. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR. pp. 5294–5306 (2025). https://doi.org/10.1109/CVPR52734.2025.00499
28. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: CVPR. pp. 10510–10522 (2025). https://doi.org/10.1109/CVPR52734.2025.00983
29. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR. pp. 20697–20709 (2024). https://doi.org/10.1109/CVPR52733.2024.01956
30. Wang, T., Mao, X., Zhu, C., Xu, R., Lyu, R., Li, P., Chen, X., Zhang, W., Chen, K., Xue, T., et al.: EmbodiedScan: A holistic multi-modal 3D perception suite towards embodied AI. In: CVPR. pp. 19757–19767 (2024). https://doi.org/10.1109/CVPR52733.2024.01868

31. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Scalable permutation-equivariant visual geometry learning. arXiv e-prints pp. arXiv–2507 (2025). https://doi.org/10.48550/arXiv.2507.13347
32. Wu, C.: Towards linear-time incremental Structure from Motion. In: 2013 International Conference on 3D Vision-3DV 2013. pp. 127–134. IEEE (2013). https://doi.org/10.1109/3DV.2013.25
33. Wu, Y., Zheng, W., Zhou, J., Lu, J.: Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. arXiv preprint arXiv:2507.02863 (2025). https://doi.org/10.48550/arXiv.2507.02863
34. Wu, Y., Zheng, W., Zuo, S., Huang, Y., Zhou, J., Lu, J.: EmbodiedOcc: Embodied 3D occupancy prediction for vision-based online scene understanding. In: ICCV. pp. 26360–26370 (2025). https://doi.org/10.48550/arXiv.2412.04380
35. Yuan, S., Yang, Y., Yang, X., Zhang, X., Zhao, Z., Zhang, L., Zhang, Z.: Infinitevggt: Visual geometry grounded transformer for endless streams. arXiv preprint arXiv:2601.02281 (2026). https://doi.org/10.48550/arXiv.2601.02281
36. Zheng, S., Zhou, B., Shao, R., Liu, B., Zhang, S., Nie, L., Liu, Y.: Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In: CVPR. pp. 19680–19690 (2024). https://doi.org/10.1109/CVPR52733.2024.01861
37. Zhou, X., Lin, Z., Shan, X., Wang, Y., Sun, D., Yang, M.H.: Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In: CVPR. pp. 21634–21643 (2024). https://doi.org/10.1109/CVPR52733.2024.02044
38. Zhuo, D., Zheng, W., Guo, J., Wu, Y., Zhou, J., Lu, J.: Streaming 4d visual geometry transformer. arXiv preprint arXiv:2507.11539 (2025). https://doi.org/10.48550/arXiv.2507.11539