S-VGGT: Structure-Aware Subscene Decomposition for Scalable 3D Foundation Models

Xinze Li[1], Pengxu Chen[2], Yiyuan Wang[3,1], Weifeng Su[1,4], Wentao Cheng[1,*]
[1]Beijing Normal-Hong Kong Baptist University [2]Jilin University [3]Hong Kong Baptist University
[4]Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science

***Abstract*—Feed-forward 3D foundation models face a key challenge: the quadratic computational cost introduced by global attention, which severely limits scalability as input length increases. Concurrent acceleration methods, such as token merging, operate at the token level. While they offer local savings, the required nearest-neighbor searches introduce undesirable overhead. Consequently, these techniques fail to tackle the fundamental issue of structural redundancy dominant in dense capture data. In this work, we introduce S-VGGT, a novel approach that addresses redundancy at the structural frame level, drastically shifting the optimization focus. We first leverage the initial features to build a dense scene graph, which characterizes structural scene redundancy and guides the subsequent scene partitioning. Using this graph, we softly assign frames to a small number of subscenes, guaranteeing balanced groups and smooth geometric transitions. The core innovation lies in designing the subscenes to share a common reference frame, establishing a parallel geometric bridge that enables independent and highly efficient processing without explicit geometric alignment. This structural reorganization provides strong intrinsic acceleration by cutting the global attention cost at its source. Crucially, S-VGGT is entirely orthogonal to token-level acceleration methods, allowing the two to be seamlessly combined for compounded speedups without compromising reconstruction fidelity. Code is available at https://github.com/Powertony102/S-VGGT.**

***Index Terms*—3D reconstruction, feed-forward 3D models, subscene partitioning, scene graph, attention acceleration.**

I. Introduction

The advent of 3D foundation models has marked a new phase in multi-view geometry, enabling efficient, optimization-free reconstructions with strong generalization. Early pointmap prediction architectures such as DUSt3R [1] established the foundations for robust and unconstrained feed-forward 3D reconstruction. This progress was further advanced by memory-augmented approaches with persistent or spatial state [2], [3], as well as large-scale reconstruction systems capable of handling thousands of views [4], [5]. End-to-end feed-forward Structure-from-Motion (SfM) pipelines [6], [7] further showed that effective pose and depth estimation is possible with minimal reliance on bundle adjustment.

Building on these advancements, global attention systems such as VGGT [8] and other large-scale models [9], [10] have enabled efficient 3D reconstruction by recovering camera poses, depth maps, and point clouds in a single forward pass. However, a critical bottleneck remains, as the quadratic computational cost incurred by global attention grows rapidly with the input sequence length. This problem is particularly pronounced in dense capture scenarios, where frames exhibit significant visual redundancy, and the information gain from additional frames diminishes. In such cases, the challenge shifts to how we can efficiently extract the geometric details from each frame, ensuring that crucial spatial context is preserved without incurring prohibitive computational costs.

Fig. 1: Comparison of VGGT (2.69 FPS) and S-VGGT (10.13 FPS) on a 500-frame scene. S-VGGT achieves a significant speedup by processing subscenes in parallel while maintaining reconstruction quality.

A prominent line of work focuses on reducing redundancy at the token level. Techniques initially proposed for visual encoders [11], [12] compress local representations by estimating token similarity and merging those judged redundant. Adaptations for 3D geometric models [13] follow the same principle and have demonstrated meaningful speedups. However, large-scale merging inherently requires computing similarity over a substantial token set, adding nontrivial overhead that partially offsets the acceleration. Furthermore, the fusion of multiple

* Corresponding author: Wentao Cheng. Email: wentaocheng@bnbu.edu.cn

tokens modifies the underlying feature distribution, leading to a representation shift that may degrade geometric precision under high compression ratios.

Our work is motivated by the observation that redundancy in dense sequences manifests primarily at the frame level. In scenarios where neighboring images share significant geometric overlap, the marginal information gain provided by global attention diminishes rapidly, making the quadratic computational cost unjustifiable. This phenomenon mirrors concepts in classical SfM, where strategies such as view graph sparsification [14]–[17] and reconstruction via submaps [18], [19] utilize structural decomposition to tractable subproblems. Our approach capitalizes on this insight: by addressing redundancy at the frame level, we can reduce the effective input sequence length before attention is applied, thereby avoiding the quadratic cost at its source. Crucially, the successful mechanisms used in classical SfM, such as iterative pose refinement and explicit alignment, conflict with the forward only, single pass requirement of modern foundation models. As a result, we require a novel formulation that bridges the gap between structural efficiency and the single pass nature of geometric foundation models.

To realize this vision, we introduce S-VGGT, a framework designed for efficient inference by targeting systematic redundancy. Our core strategy is to decompose the input sequence into a small number of coherent partitions that preserve the original spatial structure while reducing the effective frame count. We leverage the model's intrinsic intermediate features to derive a density-aware affinity score, which captures the inter-frame correlations and guides the sequence decomposition. A key aspect of our approach is assigning a common reference frame to all subscenes. This mechanism allows for independent and parallel processing within a unified coordinate system, eliminating the need for explicit alignment after inference. As a result, we achieve significant reductions in global attention complexity, delivering substantial speedups without sacrificing reconstruction quality.

We evaluate S-VGGT on several datasets, including ScanNet [20], NRGBD [21], and 7Scenes [22], and demonstrate a significant speed advantage over VGGT. As shown in Fig. 1, S-VGGT achieves substantial acceleration without compromising reconstruction fidelity. Importantly, our sequence optimization is fully orthogonal to token techniques, such as those used in FastVGGT [13]. This orthogonality enables seamless integration with existing token-based acceleration methods. By combining S-VGGT with techniques like token merging, we achieve even greater speedups, leveraging the strengths of both approaches to reduce computational bottlenecks while preserving reconstruction fidelity.

II. Method

In this section, we introduce the key components of our method. First, we provide an overview of the foundational components of VGGT [8], upon which our approach is built. We then detail our strategy for addressing frame-level redundancy through soft partitioning, followed by an explanation of how we evaluate frame similarity and organize frames into coherent subscenes. Next, we describe how the subscenes are processed in parallel and how we ensure consistency across them using anchor frame sharing. Finally, we present the optimization techniques used to achieve efficient grouping and show how they enable significant computational savings without compromising reconstruction quality. The overall framework of our method is illustrated in Fig. 2.

A. Preliminaries

We briefly summarize the relevant components of VGGT [8] that our method builds upon. Given a sequence of N RGB images $\mathbf{I} = \{\mathbf{I}_s\}_{s=1}^{N}$ with each $\mathbf{I}_s \in \mathbb{R}^{3\times H\times W}$, VGGT first encodes each frame using a DINOv2 [23] backbone:

$$\mathbf{F}_s = f_{\mathrm{DINO}}(\mathbf{I}_s) \in \mathbb{R}^{P\times C},$$

producing P patch tokens per frame. Following the original architecture, one camera token and four register tokens are appended, so each frame contains $(P+5)$ tokens in total.

VGGT employs an alternating-attention backbone consisting of L layers. Each layer first performs *frame-wise* self-attention independently on all frames, and then applies a *global* cross-frame attention on the concatenation of all tokens:

$$\mathbf{F} = \mathbf{F}_1 \,\|\, \mathbf{F}_2 \,\|\, \cdots \,\|\, \mathbf{F}_N \in \mathbb{R}^{N(P+5)\times C}.$$

As a result, every global-attention operation jointly processes $N(P+5)$ tokens, leading to a quadratic computational cost with respect to the sequence length N. This global-attention step is the dominant bottleneck when processing dense or long multi-view sequences.

B. Frame Similarity and Scene Density

As in classical incremental SfM pipelines, where a scene graph is built by comparing pairs of images, our framework also begins by evaluating pairwise frame similarity. Although dedicated image retrieval features could be applied here, we find that lightweight descriptors computed directly from the inherited per-frame feature maps in VGGT are sufficient for capturing viewpoint overlap and coarse structural cues. Given the token representation $\mathbf{F}_s$ of frame s, we obtain a single C-dimensional descriptor by averaging its patch tokens.

Pairwise similarity is computed using cosine similarity,

$$S_{ij} = \frac{\mathbf{d}_i^\top \mathbf{d}_j}{\|\mathbf{d}_i\|\,\|\mathbf{d}_j\|},$$

resulting in a similarity matrix $S \in \mathbb{R}^{N\times N}$ that summarizes how strongly each frame is supported by the others. High similarity corresponds to substantial overlap in observed content, whereas low similarity indicates wider viewpoint deviation.

To quantify how densely the frames cover the scene, we count for each frame how many other frames exhibit similarity above a fixed threshold. Averaging this quantity over all frames gives a single density value that reflects the overall redundancy of the input: large values indicate that many frames observe nearly identical content, whereas small values imply substantial viewpoint variation. This density value directly determines

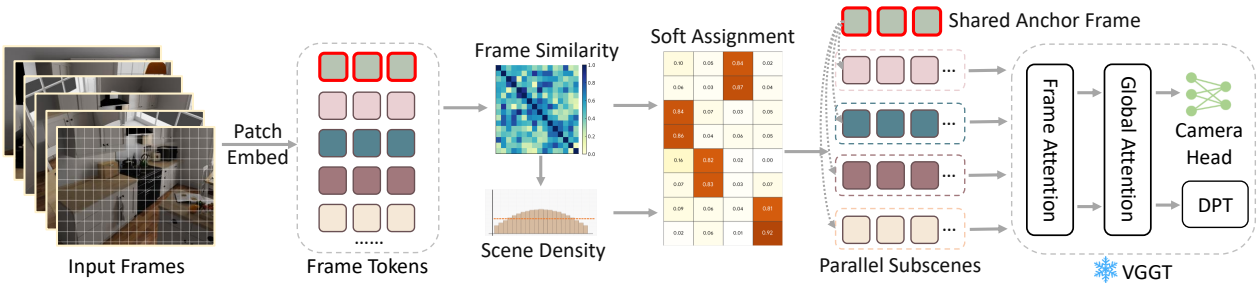

Fig. 2: The framework of S-VGGT. The input frames are first embedded into tokens, and frame similarity is calculated to assess redundancy. Frames are then grouped into subscenes via soft assignment, ensuring parallel processing. A shared reference frame across subscenes enables efficient global and frame attention operations, with the model architecture based on VGGT [8].

the number of subgroups. Instead of relying on manually chosen constants or complex heuristics, we simply clamp the density by a predefined maximum group count $K_{\max}$, using $\min(\text{density}, K_{\max})$ as the final number of groups. In dense inputs the density is large and thus produces fewer, larger groups, while more diverse inputs naturally yield a larger number of finer groups. This adaptive choice of the group count sets the stage for the grouping mechanism introduced next.

C. Grouping via Soft Assignment

The objective of this stage is to partition the input frames into K subscenes. Unlike traditional submap-based reconstruction methods, which require costly explicit geometric alignment during final fusion, our approach aims to produce a partitioning that is compatible with efficient feed-forward inference. Concretely, we construct subscenes that satisfy three properties: (1) strong internal connectivity, so that frames within a subscene provide sufficient mutual support for robust local reconstruction; (2) reduced internal redundancy, which lowers the computational overhead; and (3) fidelity to the overall scene geometry, which allows subscenes to share anchor frames and remain implicitly aligned in a common coordinate system.

To achieve the desired grouping behavior, we maintain a soft assignment matrix $\mathbf{A} \in \mathbb{R}^{N\times K}$, where each row encodes a distribution over subscenes and each entry $\mathbf{A}_{sk}$ reflects the degree to which frame s belongs to subscene k. This soft formulation remains fully differentiable and GPU-friendly, avoiding the instability and iterative overhead of hard clustering. We optimize $\mathbf{A}$ using three lightweight regularization terms. First, to ensure that each subscene retains meaningful connectivity and remains consistent with the overall scene structure, we favor partitions in which frames share consistent similarity relationships with the rest of the inputs. Each subscene is summarized by a soft group mean $\mathbf{h}_k = \frac{1}{m_k}\sum_{s=1}^{N}\mathbf{A}_{sk}\mathbf{S}_{s:}$ with size $m_k = \sum_{s=1}^{N}\mathbf{A}_{sk}$, and is compared against the global mean $\mathbf{h}_{\text{avg}} = \frac{1}{N}\sum_{s=1}^{N}\mathbf{S}_{s:}$. The coherence loss

$$\mathcal{L}_{\text{coh}} = \sum_{k=1}^{K}\|\mathbf{h}_k - \mathbf{h}_{\text{avg}}\|_2^2$$

encourages each subscene to preserve global consistency.

To prevent any subscene from becoming disproportionately large, we regularize the soft group sizes $m_k = \sum_{s=1}^{N}\mathbf{A}_{sk}$ to stay close to the ideal size N/K. This yields a simple balance loss,

$$\mathcal{L}_{\text{bal}} = \sum_{k=1}^{K}\left(m_k - \frac{N}{K}\right)^2,$$

which discourages severely unbalanced partitions and helps retain the computational benefits of grouping.

To obtain clear and GPU-friendly discretization, we encourage each row of $\mathbf{A}$ to form a confident assignment. A lightweight sharpness regularizer,

$$\mathcal{L}_{\text{sharp}} = \sum_{s=1}^{N}\sum_{k=1}^{K}\mathbf{A}_{sk}\,(1-\mathbf{A}_{sk}),$$

drives the soft assignments toward one-hot vectors without resorting to iterative hard clustering. The full grouping objective,

$$\mathcal{L}_{\text{group}} = \lambda_{\text{coh}}\mathcal{L}_{\text{coh}} + \lambda_{\text{bal}}\mathcal{L}_{\text{bal}} + \lambda_{\text{sharp}}\mathcal{L}_{\text{sharp}},$$

is optimized with a small number of gradient descent steps on $\mathbf{A}$ alone. After optimization, we determine hard assignments by selecting the subscene that maximizes the assignment matrix. Each frame is then associated with a subscene based on this assignment. Each subscene contains a nearly equal number of frames. To ensure they fit within a batch, we perform a lightweight correction by reassigning a few nearby frames based on their similarity. This process remains strictly feed-forward and does not involve any additional model evaluations.

D. Anchor Frame Sharing

We partition the sequence into independent subscenes to decouple the attention mechanism, effectively eliminating the quadratic complexity of global attention and allowing the model to ignore irrelevant long-range dependencies. This enables parallel processing of distinct subscenes, helping the system scale to long videos without memory bottlenecks. However, independent processing risks geometric misalignment, as VGGT anchors its 3D coordinate system to the first frame. We address this by introducing Anchor Frame Sharing: by

TABLE I: Quantitative results of point cloud reconstruction on the Neural RGB-D [21] and 7-Scenes [22] datasets with 500-frame input sequences. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	NRGBD							7 Scenes						
	Acc ↓		Comp ↓		NC ↑		FPS ↑	Acc ↓		Comp ↓		NC ↑		FPS ↑
	Mean	Med.	Mean	Med.	Mean	Med.		Mean	Med.	Mean	Med.	Mean	Med.	
Fast3R	0.088	0.040	0.031	0.011	0.607	0.640	5.484	0.058	0.025	0.049	0.009	0.572	0.609	5.312
CUT3R	0.286	0.208	0.105	0.036	0.567	0.597	**15.342**	0.175	0.121	0.083	0.083	0.546	0.563	**15.435**
Spann3R	0.700	0.343	0.221	0.128	0.559	0.587	7.961	0.379	0.242	0.163	0.080	0.534	0.548	7.895
VGGT*	0.031	0.019	0.025	**0.010**	**0.642**	**0.767**	2.732	0.019	**0.009**	0.028	0.010	**0.632**	**0.716**	2.612
FastVGGT	**0.027**	**0.018**	0.022	**0.010**	0.638	0.764	8.092	**0.018**	**0.009**	0.029	0.010	0.625	0.702	8.011
Ours	0.031	0.022	**0.020**	**0.010**	0.622	0.717	9.934	0.022	0.011	**0.022**	**0.008**	0.622	0.697	9.425

TABLE II: Quantitative results of camera pose estimation and point cloud reconstruction on the ScanNet dataset with input sequences of 1000 images. *OOM* denotes out-of-memory.

Method	ATE ↓	ARE ↓	RPE-rot ↓	RPE-trans ↓	FPS ↑
Fast3R	1.065	42.024	28.461	0.456	2.673
CUT3R	1.235	56.756	0.968	0.048	**11.725**
Spann3R	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*
VGGT*	0.190	4.351	0.864	0.038	1.458
FastVGGT	0.162	3.805	**0.656**	**0.030**	5.200
Ours	**0.145**	**3.576**	0.665	0.053	5.699

prepending global Frame 0 to each subscene, we ensure that all groups share the same reference point. This simple approach is highly effective in practice. By sharing the anchor frame, all subscenes align in a unified global coordinate system, eliminating the need for complex geometric optimization or rigid alignment, while preserving the efficiency gains from partitioning.

E. Complexity Analysis

We analyze the computational complexity to demonstrate the intrinsic acceleration of S-VGGT. The bottleneck in the baseline VGGT is the global attention mechanism, which scales quadratically with the sequence length, i.e., $\mathcal{O}((NT)^2)$, where N is the number of frames and T is the number of tokens per frame. By decomposing the sequence into K independent subscenes, S-VGGT reduces the attention cost to $\sum_{k=1}^{K} \mathcal{O}((NT/K)^2) = \mathcal{O}((NT)^2/K)$. This yields a theoretical speedup factor of K. The overhead introduced by our method—computing frame similarity and soft assignments—scales as $\mathcal{O}(N^2)$ based on frame-level descriptors. Since the number of tokens per frame satisfies $T \gg 1$ (typically $T \approx 1000$), this overhead is negligible compared to the token-level attention cost $\mathcal{O}(N^2T^2)$. Consequently, S-VGGT achieves substantial reductions in both latency and memory usage.

III. Experiments

A. Experimental Setup

a) Datasets and Metrics: We evaluate our framework on three widely used benchmarks to assess performance across different scene types and trajectory lengths. ScanNet [20] serves as the primary benchmark for large-scale camera pose estimation. Following standard protocols [1], [8], we report Absolute Trajectory Error (ATE), Absolute Rotation Error (ARE), and Relative Pose Errors (RPE) on unseen trajectories. For dense reconstruction quality, we utilize Neural RGB-D [21] and 7-Scenes [22]. Here, we report standard geometric metrics: Accuracy (Acc), Completeness (Comp), and Normal Consistency (NC). To rigorously test scalability and efficiency, we construct long-sequence inputs (500–1000 frames) for all evaluations, challenging the models' ability to handle dense structural redundancy.

b) Baselines: We primarily compare S-VGGT against VGGT [8], the state-of-the-art global-attention foundation model. Specifically, we use the VRAM-efficient variant, denoted as VGGT*, to enable fair comparison on long sequences within memory limits. We also compare against FastVGGT [13], a representative token-level acceleration method. For broader context, we include recent feed-forward reconstruction baselines derived from DUSt3R [1], including Fast3R [5] (parallel multi-view), Spann3R [3] (spatial memory), and CUT3R [2] (recurrent state), to highlight the robustness of our approach in long-sequence settings.

c) Implementation Details: All experiments are conducted on a single NVIDIA A100 GPU using `bfloat16` precision to optimize memory efficiency. For S-VGGT, unless otherwise specified, we employ a maximum subscene count of $K_{\max} = 8$ for standard sequences, scaling proportionally for longer inputs. The grouping module performs a lightweight inference-time optimization (typically 10 iterations) to refine soft assignments. Crucially, our framework requires no fine-tuning of the pre-trained VGGT weights; all results are obtained in a strictly zero-shot manner.

B. 3D Reconstruction

We evaluate the 3D reconstruction capabilities of S-VGGT on long-sequence inputs from the Neural RGB-D and 7-Scenes datasets. As summarized in Table I, our framework demonstrates a dominant efficiency advantage while preserving high geometric fidelity. In terms of runtime performance, S-VGGT achieves an inference speed of approximately 10 FPS on Neural RGB-D, representing a nearly $3.6\times$ speedup over the VGGT* baseline (2.73 FPS) and consistently outperforming the token-level acceleration of FastVGGT (8.09 FPS).

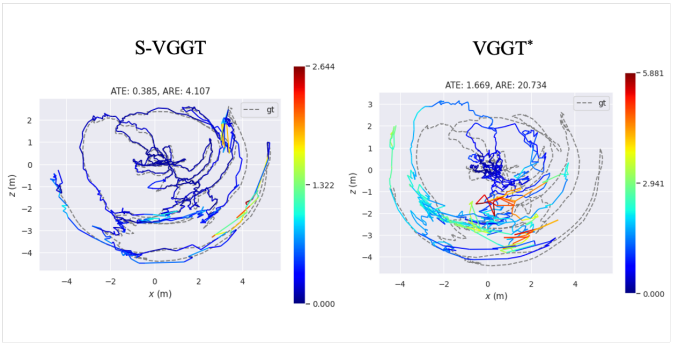

Fig. 3: Qualitative comparison of camera pose estimation performance between S-VGGT and VGGT*.

Similar gains are observed on 7-Scenes, confirming that our frame-level structural partitioning effectively circumvents the quadratic computational cost associated with dense inputs.

Crucially, this acceleration is achieved without compromising reconstruction quality. On both benchmarks, S-VGGT maintains accuracy and completeness metrics comparable to the full-attention baseline. For instance, on 7-Scenes, our method yields an accuracy of 0.022, closely tracking the baseline performance and significantly surpassing DUSt3R-based variants such as Fast3R and CUT3R. This suggests that our density-aware subscenes successfully capture the essential geometric context required for robust depth and point map prediction. Furthermore, unlike methods that struggle with global consistency over long trajectories, our approach maintains high normal consistency, validating the effectiveness of the anchor frame sharing mechanism in preserving a unified coordinate system across parallel subscenes.

C. Camera Pose Estimation

We evaluate the camera pose estimation performance on unseen trajectories drawn from a uniformly sampled subset of ScanNet [20]. All experiments are conducted on challenging input sequences of length 1000 images, which demonstrate the crucial efficiency and robustness advantages of our method under long-sequence configurations. As shown in Table II, our framework excels in both accuracy and runtime efficiency. Specifically, S-VGGT achieves the best absolute pose accuracy, with an ATE of 0.145, significantly outperforming the original VGGT* baseline (ATE 0.190) and even the token-accelerated FastVGGT. This result is particularly notable, as the structural partitioning in our approach, which intentionally limits the scope of attention, unexpectedly improves geometric accuracy.

In terms of efficiency, S-VGGT achieves a remarkable 3.9× speedup over VGGT* . While methods like CUT3R achieve high raw FPS, their absolute pose estimation performance suffers from accumulated error over long sequences. Similarly, Spann3R, which relies on spatial memory, fails entirely due to memory limitations. The key to our success lies in addressing the accumulated error introduced by global attention in previous VGGT variants, which arises from noisy long-range correlations as the frame count increases. By partitioning the sequence into subscenes, we effectively constrain attention to density-consistent regions, filtering out irrelevant correlations. This design allows S-VGGT to achieve superior prediction quality while maintaining scalability for large, dense datasets. A qualitative comparison of camera pose estimation performance between S-VGGT and VGGT* is shown in Fig. 3.

D. Complementarity with Token-Level Acceleration

A core claim of our work is that structural redundancy reduction (frame-level) is fully orthogonal to feature redundancy reduction (token-level). To empirically validate this, we evaluate a hybrid configuration, denoted as "Ours+Fast," by combining our frame partitioning module with the token-merging strategy of FastVGGT. As shown in Fig. 4, the hybrid approach significantly improves efficiency compared to FastVGGT alone. For example, "Ours+Fast" achieves a speedup of 3.4× for a 300-frame sequence, compared to 2.2× for FastVGGT alone. This trend continues across longer sequences, with "Ours+Fast" reaching up to 5.8× speedup for a 700-frame sequence, providing a substantial advantage over FastVGGT.

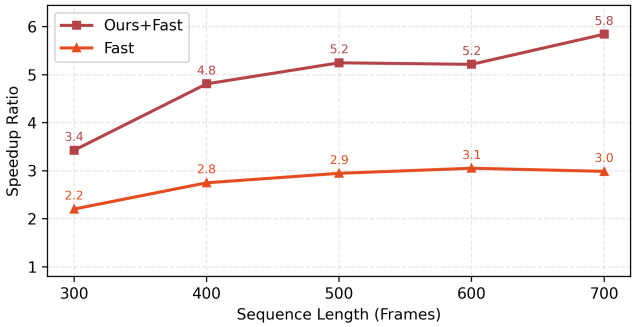

Fig. 4: Compounded Speedup (vs. VGGT*) on NRGBD [21] Results validate the complementarity of frame-level S-VGGT ("Ours") and token-level FastVGGT ("Fast"), showing enhanced acceleration across varying sequence lengths.

E. Analysis

While the consistency loss ($\mathcal{L}_{\text{cons}}$) ensures accuracy, the balance ($\mathcal{L}_{\text{bal}}$) and sharpness ($\mathcal{L}_{\text{sharp}}$) terms are critical for computational viability. Their removal leads to unbalanced, degenerate partitions that eliminate the parallel speedup. Furthermore, the Anchor Frame Sharing protocol is essential. Its removal either causes catastrophic geometric misalignment or introduces costly post-hoc optimization, fundamentally compromising our efficient inference goal.

The time breakdown comparison between VGGT* and S-VGGT (Fig. 5) reveals that both methods allocate the majority of their time to global attention. However, S-VGGT introduces two additional modules: similarity graph and soft assignment. While these modules add complexity, they are efficiently integrated, with minimal impact on the overall time cost. This demonstrates that S-VGGT effectively mitigates the global

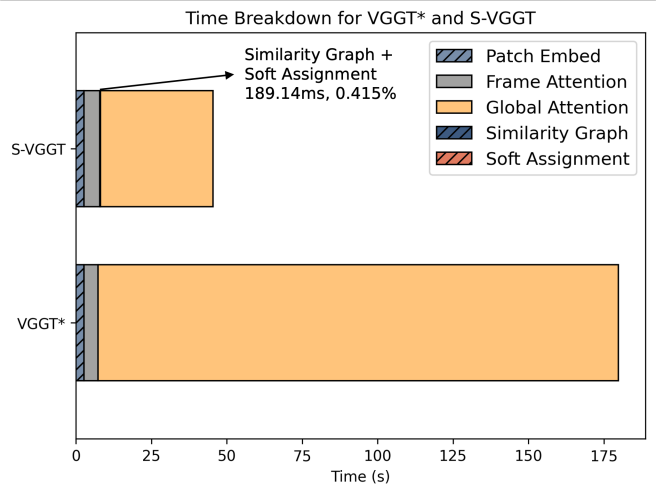

Fig. 5: Time breakdown of VGGT* vs. S-VGGT.

attention bottleneck while maintaining competitive efficiency. The integration of these modules enhances S-VGGT's ability to handle long sequences, making it a scalable solution for 3D perception tasks.

IV. Conclusion

We address the scalability bottleneck in global attention-based 3D foundation models by introducing S-VGGT, a novel framework that reduces structural redundancy at the frame level. By leveraging density-aware partitioning and anchor frame sharing, our method enables efficient parallel inference. S-VGGT offers significant intrinsic acceleration over the baseline while improving geometric accuracy by mitigating long-range noisy correlations. Additionally, the framework is fully orthogonal to existing token-level techniques, achieving compounded acceleration when combined. This work provides a scalable and efficient solution for high-fidelity 3D perception in future applications.

V. Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. 62306154) and BNBU Research Grant (Grant No. R0200028-26).

References

[1] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud, "Dust3r: Geometric 3d vision made easy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20697–20709.

[2] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa, "Continuous 3d perception model with persistent state," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 10510–10522.

[3] Hengyi Wang and Lourdes Agapito, "3d reconstruction with spatial memory," in *2025 International Conference on 3D Vision (3DV)*. IEEE, 2025, pp. 78–89.

[4] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy, "Must3r: Multi-view network for stereo 3d reconstruction," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1050–1060.

[5] Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli, "Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 21924–21935.

[6] Sven Elflein, Qunjie Zhou, and Laura Leal-Taixé, "Light3r-sfm: Towards feed-forward structure-from-motion," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 16774–16784.

[7] Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud, "Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion," in *2025 International Conference on 3D Vision (3DV)*. IEEE, 2025, pp. 1–10.

[8] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny, "Vggt: Visual geometry grounded transformer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5294–5306.

[9] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al., "Mapanything: Universal feed-forward metric 3d reconstruction," *arXiv preprint arXiv:2509.13414*, 2025.

[10] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He, "π^3: Permutation-equivariant visual geometry learning," *arXiv preprint arXiv:2507.13347*, 2025.

[11] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman, "Token merging: Your vit but faster," in *ICLR*, 2023.

[12] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang, "An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models," in *European Conference on Computer Vision*. Springer, 2024, pp. 19–35.

[13] You Shen, Zhipeng Zhang, Yansong Qu, and Liujuan Cao, "Fastvggt: Training-free acceleration of visual geometry transformer," *arXiv preprint arXiv:2509.02560*, 2025.

[14] Noah Snavely, Steven M Seitz, and Richard Szeliski, "Skeletal graphs for efficient structure from motion," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[15] Kai Ni, Drew Steedly, and Frank Dellaert, "Out-of-core bundle adjustment for large-scale 3d reconstruction," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[16] Siyu Zhu, Runze Zhang, Lei Zhou, Tianwei Shen, Tian Fang, Ping Tan, and Long Quan, "Very large-scale global sfm by distributed motion averaging," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4568–4577.

[17] Rajvi Shah, Visesh Chari, and PJ Narayanan, "View-graph selection framework for sfm," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 535–550.

[18] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L Schönberger, "Global structure-from-motion revisited," in *European Conference on Computer Vision*. Springer, 2024, pp. 58–77.

[19] Johannes L Schonberger and Jan-Michael Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[20] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[21] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies, "Neural rgb-d surface reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6290–6301.

[22] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon, "Scene coordinate regression forests for camera relocalization in rgb-d images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2930–2937.

[23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al., "Dinov2: Learning robust visual features without supervision," *Transactions on Machine Learning Research Journal*, 2024.