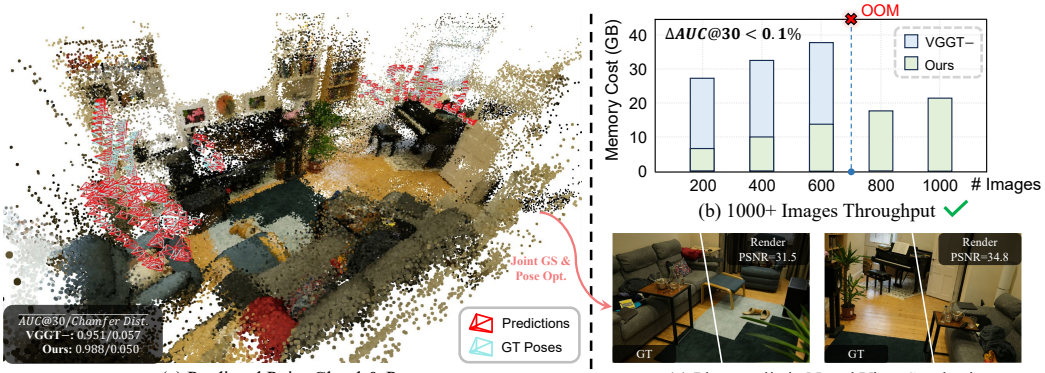
VGGT-X: WHEN VGGT MEETS DENSE NOVEL VIEW SYNTHESIS

Yang Liu[1,2], Chuanchen Luo[4], Zimo Tang[3], Junran Peng[5] ✉, & Zhaoxiang Zhang [1,2] ✉
[1] NLPR, MAIS, Institute of Automation, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] Huazhong University of Science and Technology
[4] Shandong University [5] University of Science and Technology Beijing
{liuyang2022, zhaoxiang.zhang}@ia.ac.cn, u202315173@hust.edu.cn
chuanchen.luo@sdu.edu.cn, jrpeng4ever@126.com



(a) Predicted Point Cloud & Poses (b) 1000+ Images Throughput (c) Photorealistic Novel View Synthesis

Figure 1: Reconstruction and Novel View Synthesis results. In part (a), we extend VGGT to handle dense multi-view inputs and incorporate an efficient global alignment, yielding highly accurate predictions. Part (b) demonstrates that eliminating redundant VRAM usage enables inference throughput over 1000 images without compromising performance. The VGGT− here denotes VGGT with the elimination of redundant intermediate features. Finally, part (c) illustrates that, with an appropriate joint pose and 3DGS optimization strategy, a photorealistic rendering can be realized.

ABSTRACT

We study the problem of applying 3D Foundation Models (3DFMs) to dense Novel View Synthesis (NVS). Despite significant progress in Novel View Synthesis powered by NeRF and 3DGS, current approaches remain reliant on accurate 3D attributes (e.g., camera poses and point clouds) acquired from Structure-from-Motion (SfM), which is often slow and fragile in low-texture or low-overlap captures. Recent 3DFMs showcase orders of magnitude speedup over the traditional pipeline and great potential for online NVS. But most of the validation and conclusions are confined to sparse-view settings. Our study reveal that naively scaling 3DFMs to dense views encounters two fundamental barriers: dramatically increasing VRAM burden and imperfect outputs that degrade initialization-sensitive 3D training. To address these barriers, we introduce **VGGT-X**, incorporating a memory-efficient VGGT implementation that scales to 1,000+ images, an adaptive global alignment for VGGT output enhancement, and robust 3DGS training practices. Extensive experiments show that these measures substantially close the fidelity gap with COLMAP-initialized pipelines, achieving state-of-the-art results in dense COLMAP-free NVS and pose estimation. Additionally, we analyze the causes of remaining gaps with COLMAP-initialized rendering, providing insights for the future development of 3D foundation models and dense NVS. Our project page is available at https://dekuliutesla.github.io/vggt-x.github.io/.

1 INTRODUCTION

Novel View Synthesis (NVS) reconstructs a 3D scene from multi-view images to render photorealistic novel views. Implicit representations like Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) set a new standard in rendering fidelity, while recent explicit 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) revolutionizes the area by enabling realistic rendering with real-time speed. Both families, however, typically require tens of minutes to train and depend on accurate initialization from external sensors or reconstruction pipelines (e.g., COLMAP (Schönberger & Frahm, 2016)), which incurs expensive hardware or additional minutes-to-hours overhead (Li et al., 2025).

Recent 3D Foundation Models (3DFMs) offer a promising alternative by dramatically accelerating components of the pipeline. For example, VGGT can infer camera poses and depth for 200 images in 10s (Wang et al., 2025a), and Anysplat produces 3DGS from 64 views in 5s (Jiang et al., 2025a), suggesting orders-of-magnitude speedups over classic pipelines. Yet these methods are largely demonstrated in sparse-view regimes (tens of images), leaving open the question: **what would happen if 3DFMs are applied to dense NVS?**

This work investigates applying 3DFMs to dense NVS and identifies two central obstacles. First, 3DFM computation and memory cost increase dramatically with the number of views (e.g., VRAM of VGGT rises from 5.6 GB to 40.6 GB when input rises from 20 to 200 views (Wang et al., 2025a)), making direct dense inference of 3DGS properties infeasible on commercial GPUs. Second, even when used as a drop-in replacement for traditional reconstruction pipelines like COLMAP, 3DFM outputs exhibit higher noise levels. Such noise undermines the learning of initialization-sensitive 3D primitives and leads to significant degradation in rendering quality.

To remove the obstacles and explore the answer to the question, we take VGGT (Wang et al., 2025a) as a representative 3DFM and pursue two directions. On the 3DFM side, we remove redundant feature caching, reduce numeric precision, and adopt batched frame-wise operations to losslessly scale VGGT inference to 1,000+ images (see part (b) of Fig. 1). On the 3DGS side, we study the effect of initializing 3DGS with VGGT outputs. Tab. 4 shows substantial degradation under naïve initialization. We further investigate whether the mitigation strategy exists. We propose an efficient adaptive global alignment under epipolar constraints to refine VGGT predictions. Besides, we adopt MCMC-3DGS (Kheradmand et al., 2024) and joint pose optimization to increase robustness to noisy initialization, along with a point-cloud initialization strategy through comparative analysis. Through these approaches, we largely mitigate the fidelity gap and obtain state-of-the-art rendering under COLMAP-free settings. We also analyze remaining discrepancies with COLMAP-initialized training, including overfitting and generalization problems, and offer concrete directions for stronger 3DFMs and more robust NVS training.

In summary, our contributions are fourfold:

- We identify and analyze the key problems that prevent current 3DFMs from scaling to dense NVS.
- We explore and reveal how the key problems can be alleviated by introducing VGGT-X, a memory-efficient VGGT implementation combined with an adaptive global alignment and 3DGS training practices tailored to imperfect initialization.
- We analyze the residual gap to COLMAP-initialized pipelines and provide insights to strengthen future 3DFMs and NVS training.
- Extensive experiments confirm our state-of-the-art performance in both pose estimation and COLMAP-free NVS.

2 RELATED WORKS

2.1 NOVEL VIEW SYNTHESIS

Novel view synthesis (NVS) seeks to generate photorealistic images from novel viewpoints given a set of input images captured from different perspectives of a 3D scene. This task fundamentally relies on reconstructing a faithful 3D representation of the scene. A landmark in this field is **Neural Radiance Fields (NeRF)** (Mildenhall et al., 2021), which employs multi-layer perceptrons (MLPs)

to implicitly encode scene geometry and appearance. Subsequent works have advanced NeRF along multiple directions, including improved reflectance modeling (Verbin et al., 2022; Attal et al., 2023), anti-aliasing techniques (Barron et al., 2021; 2022), and acceleration of both training and inference (Zhang et al., 2023; Müller et al., 2022; Yu et al., 2021). More recently, **3D Gaussian Splatting (3DGS)** (Kerbl et al., 2023) has emerged as a powerful alternative, offering substantial efficiency gains while preserving high rendering quality. Building on this foundation, recent research has extended 3DGS to large-scale scene reconstruction (Lin et al., 2024; Liu et al., 2024; 2025), compact storage and transmission (Fan et al., 2024a; Lee et al., 2024), and artifact mitigation (Yu et al., 2024; Ye et al., 2024; Radl et al., 2024). Despite these advances, NVS methods still require accurate camera parameters, and 3DGS in particular remains highly sensitive to the quality of the initial point cloud. Inaccurate poses or noisy geometry often result in visual artifacts and geometric misalignments in the synthesized views.

2.2 3D Foundation Models

3D foundation models aim to infer fundamental 3D attributes—such as camera parameters, point clouds, depth maps, point tracks, or even neural radiance fields—directly from image collections. Current approaches are broadly instantiated through two architectural paradigms: diffusion-based models (Ho et al., 2020) and feed-forward ViT-based models (Dosovitskiy et al., 2021). Based on input types, the 3D foundation models can be categorized into **4 types** (Cong et al., 2025). **(i)** For **uncalibrated image pairs**, DUSt3R (Wang et al., 2024b) and its successors (Leroy et al., 2024; Fan et al., 2024b; Zhang et al., 2025a; Ye et al., 2025; Lu et al., 2025; Smart et al., 2024; Chen et al., 2025) predict point clouds (with auxiliary properties such as confidence) within the coordinate frame of the first camera. Through additional correspondence matching and reprojection loss optimization, these local geometries can be aligned into a consistent global frame (Duisterhof et al., 2025). **(ii)** For **unordered multi-view image collections**, models such as (Yang et al., 2025; Wang et al., 2025a;c; Fang et al., 2025) employ inter- and intra-view cross-attention to directly produce globally consistent poses and geometry. **(iii)** For **image streams**, models like Spann3R (Wang & Agapito, 2025) and CUT3R (Wang et al., 2025b) predict next-frame geometry by leveraging current features and temporal memory, while diffusion-based approaches (Team et al., 2025; Jiang et al., 2025b; Xu et al., 2025) cast geometry estimation as a conditional generative process. **(iv)** For **uncalibrated sparse views**, FLARE (Zhang et al., 2025b) adopts a cascaded, feed-forward pipeline that first regresses camera poses and then conditions global geometry and appearance estimation. Despite rapid progress, most existing models incur substantial computational overhead and exhibit degraded performance when scaled to hundreds or thousands of images. Our study aims to address this gap and provide new insights into the development of scalable 3D foundation models.

2.3 3D Radiance Field Learning with Pose Optimization

To mitigate the dependence on accurate camera poses, recent NVS approaches have explored a variety of strategies. A widely adopted solution is the joint optimization of camera parameters alongside the neural radiance field, often complemented by multi-view correspondence losses (Wang et al., 2021; Jeong et al., 2021). Methods such as NoPe-NeRF (Bian et al., 2023) and SPARF (Truong et al., 2023) incorporate depth supervision, whereas (Bian et al., 2024; Huang et al., 2025b) employ MLPs to regress pose updates, enhancing robustness and exploiting global scene context. NeRF-based techniques further investigate strategies to mitigate sub-optimal convergence caused by high-frequency positional embeddings (Lin et al., 2021; Chng et al., 2022; Xia et al., 2022). In the context of 3DGS, MCMC-3DGS (Kheradmand et al., 2024) enhances robustness to initialization by reformulating the Gaussian Splatting update mechanism, while (Fu et al., 2024; Chen et al., 2024; Ji & Yao, 2025) perform incremental local geometry reconstruction and pose refinement for unposed image sequences. More recently, approaches leveraging 3D foundation models or tracking models (Huang et al., 2025a;b; Shi et al., 2025) have been proposed to efficiently obtain high-quality initializations of poses and geometry. Despite these advances, a notable performance gap remains compared to COLMAP-initialized optimization, and scaling these methods to large image collections remains largely unexplored. Our work aims to advance this frontier, providing insights into training photorealistic neural radiance fields from imperfectly registered poses and point clouds.

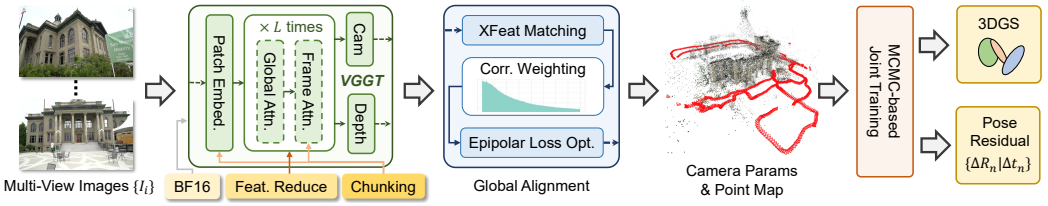


Figure 2: Overall pipeline of our model.

3 METHOD

3.1 PRELIMINARY

3D Gaussian Splatting (Kerbl et al., 2023) models a 3D scene using a collection of ellipsoids parameterized by 3D Gaussian distributions, i.e., $\mathbb{G} = \{\mathcal{G}_i \mid i = 1, \ldots, N_G\}$. Each Gaussian is associated with learnable attributes, including its center $\boldsymbol{\mu_i} \in \mathbb{R}^{3\times1}$, covariance matrix $\boldsymbol{\Sigma_i} \in \mathbb{R}^{3\times3}$, opacity $\sigma_i \in [0, 1]$, and spherical harmonics (SH) features $\boldsymbol{f_i} \in \mathbb{R}^{3\times16}$ for view-dependent appearance modeling. The covariance matrix is further decomposed into a scaling matrix $\mathbf{S_i}$ and a rotation matrix $\mathbf{R_i}$, such that $\boldsymbol{\Sigma_i} = \mathbf{R_i}\mathbf{S_i}\mathbf{S_i}^T\mathbf{R_i}^T$. For a given pixel p, the color $\boldsymbol{c}_p$ is obtained via alpha blending. Given a ground-truth image $\mathbf{I}$, the optimization of 3DGS is driven by the total loss $\mathcal{L}_{\text{total}}$, defined as the weighted combination of the L1 loss $\mathcal{L}_1$ and the D-SSIM loss $\mathcal{L}_{\text{SSIM}}$. To mitigate under- or over-reconstruction, 3DGS employs a heuristic adaptive density control strategy guided by the view-space position gradient $\nabla_{\text{densify}} = \partial\mathcal{L}/\partial\boldsymbol{\mu}_i$. Gaussians with gradients exceeding a predefined threshold are either cloned or split. We refer readers to the original paper (Kerbl et al., 2023) for additional details.

3DGS-MCMC (Kheradmand et al., 2024) improves 3DGS in both rendering fidelity and robustness to noisy initialization. The key insight is that the optimization of 3DGS can be reformulated as a Stochastic Gradient Langevin Dynamics (SGLD) update:

$$\mathcal{G} \leftarrow \mathcal{G} - \lambda_{\text{lr}} \cdot \nabla_{\mathcal{G}} \mathbb{E}_{\mathbf{I}\sim\mathcal{I}}[\mathcal{L}_{\text{total}}(\mathcal{G}; \mathbf{I})] + \lambda_{\text{noise}} \cdot \boldsymbol{\epsilon}, \tag{1}$$

where λ_{lr} and λ_{noise} denote hyperparameters that control the learning rate and the magnitude of stochastic exploration, respectively, and $\boldsymbol{\epsilon}$ represents noise sampled for exploration. To mitigate the dependency on precise initialization, we adopt 3DGS-MCMC as our baseline for NVS.

3.2 MEMORY-EFFICIENT VGGT IMPLEMENTATION

As illustrated in Fig. 2, the network structure of VGGT comprises three main components: per-frame DINO-based patch embedding extractor, stacked transformer layers alternating between global and frame-wise attention (i.e., AA layers), and a decoder for camera parameter regression and dense predictions (Wang et al., 2025a). Although VGGT contains 24 AA layers, only the output features from layers 4, 11, 17, and 23 are utilized for dense prediction. To eliminate redundancy, we discard intermediate outputs from other layers, thereby reducing VRAM consumption. This modification increases image throughput from 150 to 600 images, and we refer to this variant as VGGT−.

Another source of redundancy lies in data precision. While automatic mixed precision is enabled, the majority of operations and tensor storage still default to Float32. We observe that switching to BFloat16, except for MLP in heads, introduces no noticeable degradation in performance. In contrast, it reduces the peak GPU memory usage by up to 74%, leading to a substantial improvement in inference throughput. Moreover, since both DINO feature extraction and frame-wise attention involve only intra-frame computation, frames can be processed asynchronously. Consequently, N input images can be divided into $\lceil N/S \rceil$ chunks, which are sequentially processed. By selecting an appropriate chunk size S, peak memory usage in these modules can be effectively controlled. For convenience, this version is named as VGGT−−.

3.3 CAMERA PARAMETERS GLOBAL ALIGNMENT (GA)

After the feedforward inference of VGGT, we obtain estimated camera parameters $\{\mathcal{K}_n, \mathcal{R}_n, t_n\}_N$, where for the n-th camera, $\mathcal{K}_n$ denotes the intrinsic matrix, while $\mathcal{R}_n$ and t_n represent the rotation matrix and translation vector of the extrinsic matrix, respectively. These parameters can be refined using image correspondences by minimizing the epipolar distance loss:

$$\mathcal{L}_{EG} = \sum_m \sum_k w_k \, e_{m,k} \,/\, \sum_m \sum_k w_k, \quad e_{m,k} = x'_k F_m x_k, \tag{2}$$

where $e_{m,k}$ is the epipolar distance for the k-th correspondence in the m-th image pair, x'_k and x_k are the corresponding keypoints, and F_m is the fundamental matrix derived from the paired cameras. The weights w_k reflect the reliability of each correspondence, making their estimation crucial for effective optimization.

Not all C_N^2 image pairs have overlapping fields of view. Following (Jeong et al., 2021), we restrict candidate pairs to those with view angles below a certain threshold. For these pairs, VGGT's tracking head can provide correspondences and confidence scores. However, as shown in Tab. 3, these predictions are insufficiently reliable for camera refinement. We therefore adopt `XFeat` (Potje et al., 2024), a recent neural feature matcher known for its efficiency. While `XFeat` provides accurate matches, it does not supply correspondence weights w_k. Using VGGT's depth confidence as a proxy also proves suboptimal (cf. Tab. 3).

To address this issue, we propose an adaptive weighting strategy. Intuitively, when both the 3D foundation model and the matching model provide reliable estimates, most $e_{m,k}$ values should cluster near zero, and such correspondences should be assigned higher weights. Conversely, correspondences with large $e_{m,k}$ are more likely to be outliers and should be down-weighted. The "Global Alignment" panel in Fig. 2 illustrates a typical histogram of $e_{m,k}$ with the x-axis limited to $[0, 20]$. As observed, $e_{m,k}$ exhibits a long-tail distribution, which aligns naturally with this intuition. Accordingly, we first compute $e_{m,k}$ using VGGT-predicted camera parameters as defined in Eq. (2), and then estimate the adaptive weights as:

$$w_k = \left(\frac{f(e_{m,k})}{\text{Avg}(f(e_{m,k}))} \right)^{\alpha}, \tag{3}$$

where f is the probability density function approximated via histogram, $\text{Avg}(f(e_{m,k}))$ denotes the average density over all $e_{m,k}$, and α is empirically set to 0.5. As validated in Tab. 3, this weighting scheme enables more efficient convergence during camera optimization.

Finally, we adapt the learning rate to different convergence regimes. When VGGT's initialization is accurate, a small learning rate suffices for fine alignment. However, in challenging cases, such a setting fails to provide adequate updates. To adaptively control learning, we use the median epipolar distance as an indicator and adjust the learning rate according to the following empirical rule:

$$\text{lr} = \begin{cases} \text{lr}_0, & \text{if Median}(e_{m,k}) < b_1, \\ \text{lr}_1, & \text{if } b_1 < \text{Median}(e_{m,k}) < b_2, \\ \text{lr}_2, & \text{if Median}(e_{m,k}) > b_2, \end{cases} \tag{4}$$

where lr_0, lr_1, lr_2 and the bounds b_1, b_2 are specified in Sec. 4.1. As shown in Tab. 3, this adaptive strategy is critical for ensuring robust convergence in camera parameter optimization.

3.4 3DGS TRAINING WITH IMPERFECT POSES

The global alignment procedure in Sec. 3.3 substantially improves the accuracy of estimated camera parameters, thereby facilitating convergence of 3DGS training. Nonetheless, the performance gap relative to COLMAP remains, which is detrimental for initialization-sensitive models such as vanilla 3DGS (cf. Tab. 4). To mitigate this issue, we adopt MCMC-3DGS, which offers improved robustness under noisy or imperfect poses.

Figure 3: Qualitative comparison of rendering results. 3DGS† here means 3DGS trained with COLMAP initialization, and is mainly for reference. Here, *Apple* is from *CO3Dv2* dataset, *Garden* and *Stump* are from *MipNeRF360* dataset, *Ignatius* and *Caterpillar* are from *TnT* dataset.

In addition, we adopt a joint optimization scheme in which residual camera poses are optimized alongside Gaussian parameters under photometric supervision. Concretely, we estimate the residual translation $\Delta t_n \in \mathbb{R}^3$ and residual rotation $\Delta r_n \in \mathbb{R}^6$. Following (Zhou et al., 2019), the 6D rotation representation Δr_n is converted into a residual rotation matrix $\Delta \mathcal{R}_n \in \mathbb{R}^{3\times3}$, which is then applied to refine $\mathcal{R}_n$. In addition, we leverage the correspondence weights introduced in Sec. 3.3 to select reliable initialization points, providing a stronger starting configuration for 3DGS training. As shown in Tab. 4, this strategy leads to consistently improved performance.

4 Experiments

4.1 Experimental Setup

Datasets & Metrics. We evaluate our model on widely used multi-view reconstruction benchmarks, including *MipNeRF360* (Barron et al., 2022), *Tanks and Temple (TnT)* (Knapitsch et al., 2017), and *CO3Dv2* (Reizenstein et al., 2021), with maximum image sequence lengths of 311, 1106, 202 and scene numbers of 9, 5, 5, respectively. *MipNeRF360* is employed for our ablation studies. We follow (Wang et al., 2025a) for pose and point map estimation. **Pose accuracy** is evaluated using the standard AUC@30 metric, which integrates Relative Translation Error (RTE) and Relative Rotation Error (RRE). RTE and RRE compute the relative angular errors in translation and rotation for each image pair. We note that *AUC@30 is not order-invariant* and introduce a minor modification to address this; further details are provided in the Appendix B. **Point map quality** is measured using Chamfer Distance, alongside accuracy and completeness metrics. For multi-view reconstruction, we adhere to the dataset splits and training view resolutions reported in prior works (Kerbl et al., 2023; Fu et al., 2024). **Rendering quality** is assessed via PSNR, SSIM, and LPIPS. For computational efficiency, we report both runtime and VRAM usage measured on a 40G A100 GPU.

Implementation Details. In our experiments, the chunk size S for the frame-wise operation described in Sec. 3.2 is set to 128. For the global alignment procedure in Sec. 3.3, the angle-of-view threshold is set to 30 degrees. The learning rates lr_0, lr_1, and lr_2 are configured to 5×10^{-4}, 1×10^{-3}, and 1×10^{-2}, respectively, while the parameters b_1 and b_2 are set to 2.5 and 7.5. The maximum number of correspondences per image pair is limited to 4096, and the optimization is run for 300 iterations. When extracting COLMAP results, only matched points with weights exceeding 0.3 are retained. During MCMC-3DGS training, the maximum number of Gaussians per scene is matched to that of the vanilla 3DGS to ensure fairness. Pose embeddings are initialized with a learning rate of

Table 1: Comparison with SOTA methods on rendering quality. † means initialized with COLMAP. Note that for fairness, 3RGS is also trained on predictions from our VGGT−− with GA. The best performance of each part is in **bold**.

	MipNeRF360			Tanks and Temple			CO3Dv2		
Model	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓
3DGS†	0.8148	27.39	0.1849	0.8509	24.85	0.1550	0.9379	32.58	0.0954
MCMC†	0.8357	27.91	0.1536	0.8674	25.76	0.1391	0.9407	33.21	0.0968
MCMC	0.5484	22.19	0.2822	0.6789	21.42	0.2778	0.7121	25.71	0.2008
CF-3DGS	0.2344	12.38	0.7186	0.3914	12.19	0.6082	0.6110	20.18	0.4354
HT-3DGS	0.3796	14.79	0.6691	0.4508	13.83	0.5850	0.8326	28.28	0.2298
3RGS	0.7128	25.39	0.2158	0.7497	21.47	0.3002	0.8781	31.07	0.1283
Ours	**0.7821**	**26.40**	**0.1774**	**0.8419**	**24.77**	**0.1676**	**0.9105**	**31.85**	**0.1128**

Table 2: Comparison with SOTA methods on pose estimation. The units for RRE and RTE are degrees. Note that for fairness, 3RGS is also trained on predictions from our VGGT−− with GA. The best performance of each part is in **bold**. "OOM" here means fail to run on all scenes due to Out-of-Memory error.

	MipNeRF360			Tanks and Temple			CO3Dv2		
Model	RRE↓	RTE↓	AUC@30↑	RRE↓	RTE ↓	AUC@30↑	RRE↓	RTE↓	AUC@30↑
MASt3R-Sfm	17.18	10.25	0.718	21.02	14.10	0.687	11.72	15.32	0.618
π^3	3.244	3.470	0.889	OOM	OOM	OOM	**0.924**	**1.719**	**0.956**
VGGT−−	1.094	1.759	0.951	2.034	1.891	0.953	3.035	4.659	0.841
VGGT−−, +GA	0.678	0.686	0.986	1.783	1.479	0.967	2.002	2.811	0.906
CF-3DGS	104.0	56.45	0.001	110.9	55.20	0.006	15.2	21.5	0.336
HT-3DGS	93.69	56.55	0.003	100.0	51.87	0.010	12.30	12.25	0.501
3RGS	0.605	0.484	0.991	4.855	6.762	0.846	1.972	2.583	0.911
Ours	**0.601**	**0.484**	**0.992**	**1.738**	**1.259**	**0.971**	1.984	2.687	0.909

1×10^{-4} and decayed exponentially by a factor of 0.1, while the learning schedule for other 3DGS attributes follows (Kheradmand et al., 2024). During rendering quality assessment, we would freeze trained Gaussians and tune the pose embedding for the test view and minimize the photometric loss, following practice of (Huang et al., 2025b). The tuning iteration is 10,000 for *TnT* and 5,000 for other datasets, while the other setting of learning schedule aligns with the training progress.

Baselines. For 3D key attributes prediction, we compare the performance of MASt3R-SfM (Duisterhof et al., 2025), π^3 (Wang et al., 2025c), and VGGT (Wang et al., 2025a). For MASt3R-SfM, we employ the retrieval mode in scene graph construction to achieve a balance between accuracy and efficiency. For COLMAP-free 3DGS training, we consider CF-3DGS (Fu et al., 2024), HT-3DGS (Ji & Yao, 2025), 3RGS (Huang et al., 2025b), and MCMC-3DGS (Kheradmand et al., 2024). To ensure a fair comparison, we replace the initial poses and point cloud in 3RGS with our globally aligned, higher-accuracy results.

4.2 Comparison with SOTA Methods

In Tab. 1, we compare rendering performance against recent advances and include results with COLMAP initialization as an upper-bound reference. Our model achieves state-of-the-art performance, as further confirmed by the qualitative results in Fig. 3, which show that our method more effectively suppresses blurry artifacts and floaters while preserving fine-grained textures. It is worth noting that the rendering quality of CF-3DGS on *CO3Dv2* is noticeably worse than reported in its original paper, likely due to reproducibility issues documented in its repository[1].

In Tab. 2, we compare pose estimation accuracy. The results demonstrate that both our global alignment and joint optimization strategies consistently improve performance, surpassing all previous

[1] https://github.com/NVlabs/CF-3DGS/issues/7

approaches that jointly optimize poses and 3DGS. We also evaluate the pose accuracy of 3D foundation models and provide trajectory visualizations in Fig. 4. Our model exhibits closer alignment with ground-truth trajectories, achieving the highest accuracy on *MipNeRF360* and *TnT*, and ranking second on *CO3Dv2*.

4.3 Ablation

Table 3: Ablation on model components in pose and point map estimation. The experiments are conducted on *MipNeRF360* (Barron et al., 2022). "-XFeat" here means replacing XFeat with tracking predicted by VGGT itself. "- PDF Weight" means using confidence predicted by VGGT to replace adaptive weight proposed in Sec. 3.3. Computation costs are evaluated on 40G A100.

	Pose Estimation			Point Map Estimation			Cost	
Model	RRE(°)↓	RTE(°)↓	AUC@30↑	Acc.↓	Comp.↓	Overall↓	T(min)	Mem.(GB)
VGGT	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
VGGT−	1.090	1.740	0.951	0.064	0.051	0.058	**0.98**	28.87
VGGT−−	1.094	1.759	0.951	**0.063**	0.050	0.057	1.29	**9.66**
VGGT−−, +BA	**0.640**	**0.392**	**0.994**	0.064	**0.037**	**0.050**	157	24.26
VGGT−−, +GA	**0.652**	**0.643**	**0.988**	0.069	0.039	0.054	**1.78**	**11.12**
- XFeat	2.096	2.250	0.920	0.220	0.184	0.202	4.46	13.49
- Adaptive LR	0.732	0.751	0.984	**0.064**	**0.037**	**0.051**	1.78	11.12
- PDF Weight	2.705	2.970	0.892	0.108	0.058	0.083	1.88	11.12
- Rand Order	0.681	0.691	0.986	0.068	0.040	0.054	1.78	11.12

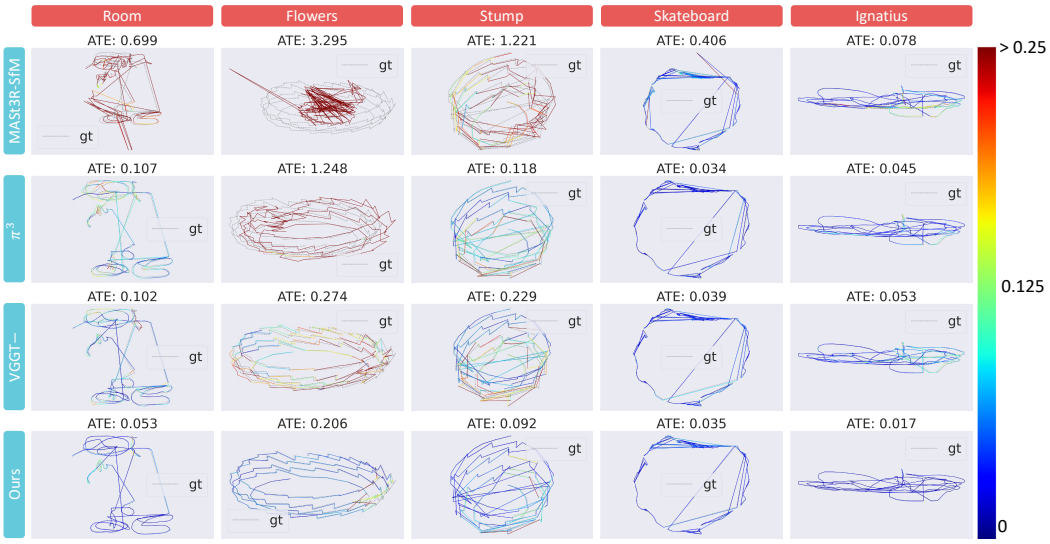


Figure 4: Qualitative comparison of estimated trajectories. Here we also report the Root Mean Square Error (RMSE) of the Absolute Trajectory Error (ATE) (in meters) (Matsuki et al., 2024). The color bar indicates trajectory distance. We recommend zooming in for better details.

First, we ablate the effect of different modules on 3D key attribute estimation. The primary reduction in computational overhead comes from redundant feature elimination and precision adjustment, which together lower VRAM usage by 83% on *MipNeRF360*. Batched attention further reduces memory by over 1 GB when scaling to more than 800 images. Combined these modifications together, the inference throughput is pushed to 1000+ images, as shown in Fig. 1. Noticeably, these optimizations have only a negligible impact on prediction accuracy, as indicated in Tab. 3.

Second, we examine strategies to enhance VGGT output quality. Replacing `XFeat` with the VGGT tracking head decreases AUC@30 by 6.8 points and increases Chamfer Distance by nearly fourfold. Similarly, leveraging VGGT-derived depth confidence to reweight `XFeat` correspondences results

Table 4: Ablation on model components in multi-view reconstruction. The experiments are conducted on *MipNeRF360* (Barron et al., 2022). The best performance of each metric is in **bold**.

	Initialization		Train Set			Test Set		
Model	Pose	Point Map	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓
3DGS	COLMAP	COLMAP	0.8869	29.58	0.1427	0.7194	27.39	0.1849
MCMC	COLMAP	COLMAP	0.9041	30.17	0.1231	0.8357	27.91	0.1536
+Pose Opt.	COLMAP	COLMAP	0.9042	30.19	0.1228	0.8359	27.95	0.1537
3DGS	VGGT−−	Rand. 500K	0.7284	23.95	0.2830	0.5321	21.10	0.3466
3DGS	VGGT−−, +GA	Rand. 500K	0.7538	25.00	0.2471	0.5675	22.23	0.3058
MCMC	VGGT−−, +GA	Rand. 500K	0.8178	26.41	0.1974	0.5563	22.37	0.2795
+Pose Opt.	VGGT−−, +GA	Rand. 500K	0.8965	29.25	**0.1229**	0.7731	26.28	0.1823
+Pose Opt.	VGGT−−, +GA	Rand. 500K	0.8965	29.25	**0.1229**	0.7731	26.28	0.1823
+Pose Opt.	VGGT−−, +GA	Filtered. 500K	0.8794	28.85	0.1473	0.7620	25.88	0.2005
+Pose Opt.	VGGT−−, +GA	Matched Points	**0.8966**	**29.59**	0.1314	**0.7821**	**26.40**	**0.1774**
+Pose Opt.	VGGT−, +BA	VGGT−, +BA	0.8948	29.23	0.1301	0.7765	26.33	0.1786

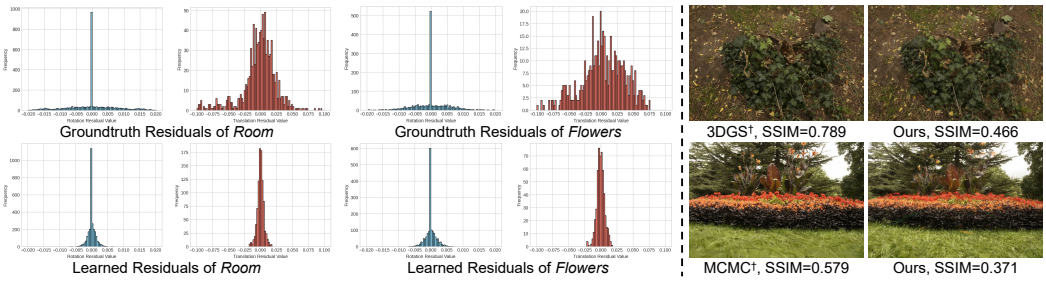


Figure 5: Bad case analysis. The blue and red histograms respectively correspond to rotation and translation residual distribution. The right part shows blurry artifacts caused by inaccurate poses.

in substantial performance degradation. In contrast, incorporating our adaptive learning rate yields consistently higher accuracy. Moreover, we observe that with permutation-equivariant AUC@30, a random input order still yields a slight performance gain, consistent with the findings of (Wang et al., 2025c). Besides, we also scale the official Bundle Adjustment (BA) strategy to hundreds of images by applying our architectural optimizations to VGG-SfM (Wang et al., 2024a). While this achieves higher accuracy, it requires over two hours to complete, and as shown in Tab. 4, its initialization does not improve NVS quality, confirming the superior efficiency of our strategy.

Finally, we ablate design choices for training high-quality 3DGS. As shown in Tab. 4, MCMC is more effective than vanilla 3DGS under imperfect initialization, and pose optimization proves essential for stable convergence and high rendering quality. Among initialization strategies, point clouds derived from high-confidence correspondences achieve the best performance. Limited by pages, we put additional ablations in Tab. 5 in the Appendix.

4.4 Discussion

Although our model achieves state-of-the-art performance, a noticeable gap remains compared to 3DGS trained with COLMAP initialization, as shown in Tab. 1. Interestingly, Tab. 4 reveals that on the training set, our model even surpasses COLMAP-initialized 3DGS in rendering quality, yet its performance on the test set lags behind, suggesting a clear overfitting issue. This highlights the inherently ill-posed nature of the problem. And without reliable initialization, the optimization process is prone to getting trapped in local minima of the highly non-convex loss landscape. We also experimented with adding depth supervision (cf. Tab. 5), but found little improvement.

Besides, as illustrated in Tab. 2, even after joint optimization, pose accuracy still falls short of COLMAP. We further compare the learned camera pose residuals with ground-truth. The visualization is included in Fig. 5. We observe that while most residuals cluster near zero—indicating accurately predicted poses—the model struggles to sufficiently correct poses with large deviations.

Another noteworthy finding in Tab. 2 is that although VGGT substantially outperforms π^3 on *Mip-NeRF360*, it is surpassed on *CO3Dv2* by a considerable margin. This discrepancy suggests that the generalization ability of 3D foundation models remains an open challenge.

5 CONCLUSIONS

In this paper, we investigated the potential of applying 3D Foundation Models to dense novel view synthesis and identified two key challenges: the poor scalability in computational overhead and insufficient prediction accurarcy for subsequent radiance field fitting. To address these obstacles, we introduced VGGT-X, which integrates a memory-efficient VGGT implementation, adaptive global alignment, and robust 3DGS training strategies. Our approach substantially narrows the performance gap with COLMAP-initialized counterparts. Beyond these improvements, our analysis also sheds light on the remaining limitations and outlines promising directions for advancing both 3DFMs and NVS frameworks. We hope our findings provide valuable insights toward building fast, reliable, and fully COLMAP-free dense NVS systems.

REFERENCES

Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O'Toole, and Changil Kim. Hyperreel: High-fidelity 6-dof video with ray-conditioned sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16610–16620, 2023.

Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5855–5864, 2021.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5470–5479, 2022.

Jia-Wang Bian, Wenjing Bian, Victor Adrian Prisacariu, and Philip Torr. Porf: Pose residual field for accurate neural surface reconstruction. In *ICLR*, 2024.

Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4160–4169, 2023.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Easi3r: Estimating disentangled motion from dust3r without training. *arXiv preprint arXiv:2503.24391*, 2025.

Yu Chen, Rolandos Alexandros Potamias, Evangelos Ververas, Jifei Song, Jiankang Deng, and Gim Hee Lee. Zerogs: Training 3d gaussian splatting from unposed images. *arXiv preprint arXiv:2411.15779*, 2024.

Shin-Fang Chng, Sameera Ramasinghe, Jamie Sherrah, and Simon Lucey. Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation. In *European Conference on Computer Vision*, pp. 264–280. Springer, 2022.

Wenyan Cong, Yiqing Liang, Yancheng Zhang, Ziyi Yang, Yan Wang, Boris Ivanovic, Marco Pavone, Chen Chen, Zhangyang Wang, and Zhiwen Fan. E3d-bench: An end-to-end benchmark for 3d geometric foundation models. *ICCV*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

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 *International Conference on 3D Vision 2025*, 2025. URL https://openreview.net/forum?id=5uw1GRBFoT.

Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejia Xu, Zhangyang Wang, et al. Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps. *Advances in neural information processing systems*, 37:140138–140158, 2024a.

Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, et al. Large spatial model: End-to-end unposed images to semantic 3d. *Advances in neural information processing systems*, 37:40212–40229, 2024b.

Xianze Fang, Jingnan Gao, Zhe Wang, Zhuo Chen, Xingyu Ren, Jiangjing Lyu, Qiaomu Ren, Zhonglei Yang, Xiaokang Yang, Yichao Yan, et al. Dens3r: A foundation model for 3d geometry prediction. *arXiv preprint arXiv:2507.16290*, 2025.

Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20796–20805, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Guichen Huang, Ruoyu Wang, Xiangjun Gao, Che Sun, Yuwei Wu, Shenghua Gao, and Yunde Jia. Longsplat: Online generalizable 3d gaussian splatting from long sequence images. *arXiv preprint arXiv:2507.16144*, 2025a.

Zhisheng Huang, Peng Wang, Jingdong Zhang, Yuan Liu, Xin Li, and Wenping Wang. 3r-gs: Best practice in optimizing camera poses along with 3dgs. *arXiv preprint arXiv:2504.04294*, 2025b.

Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5846–5854, 2021.

Bo Ji and Angela Yao. Sfm-free 3d gaussian splatting via hierarchical training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21654–21663, 2025.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *arXiv preprint arXiv:2505.23716*, 2025a.

Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4d: Leveraging video generators for geometric 4d scene reconstruction, 2025b.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Shakiba Kheradmand, Daniel Rebain, Gopal Sharma, Weiwei Sun, Yang-Che Tseng, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 3d gaussian splatting as markov chain monte carlo. *Advances in Neural Information Processing Systems*, 37:80965–80986, 2024.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, and Eunbyung Park. Compact 3d gaussian representation for radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21719–21728, 2024.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024.

Jiahao Li, Haochen Wang, Muhammad Zubair Irshad, Igor Vasiljevic, Matthew R. Walter, Vitor Campagnolo Guizilini, and Greg Shakhnarovich. Fastmap: Revisiting structure from motion through first-order optimization. *https://arxiv.org/abs/2505.04612*, 2025.

Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5741–5751, 2021.

Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, et al. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5166–5175, 2024.

Yang Liu, Chuanchen Luo, Lue Fan, Naiyan Wang, Junran Peng, and Zhaoxiang Zhang. Citygaussian: Real-time high-quality large-scale scene rendering with gaussians. In *European Conference on Computer Vision*, pp. 265–282. Springer, 2024.

Yang Liu, Chuanchen Luo, Zhongkai Mao, Junran Peng, and Zhaoxiang Zhang. Citygaussianv2: Efficient and geometrically accurate reconstruction for large-scale scenes. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=a3ptUbuzbW.

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. *CVPR*, 2025.

Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18039–18048, 2024.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022.

Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, and Erickson R. Nascimento. Xfeat: Accelerated features for lightweight image matching. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2682–2691, 2024. doi: 10.1109/CVPR52733.2024.00259.

Lukas Radl, Michael Steiner, Mathias Parger, Alexander Weinrauch, Bernhard Kerbl, and Markus Steinberger. Stopthepop: Sorted gaussian splatting for view-consistent real-time rendering. *ACM Transactions on Graphics (TOG)*, 43(4):1–17, 2024.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021. doi: 10.1109/iccv48922.2021.01072. URL http://dx.doi.org/10.1109/iccv48922.2021.01072.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Dongbo Shi, Shen Cao, Lubin Fan, Bojian Wu, Jinhui Guo, Renjie Chen, Ligang Liu, and Jieping Ye. Trackgs: Optimizing colmap-free 3d gaussian splatting with global track constraints. *arXiv preprint arXiv:2502.19800*, 2025.

Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. 2024. URL https://arxiv.org/abs/2408.13912.

Aether Team, Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. *ICCV*, 2025.

Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4190–4200, 2023.

Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5481–5490. IEEE, 2022.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *3DV*, 2025.

Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21686–21697, 2024a.

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*, pp. 5294–5306, 2025a.

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*, pp. 10510–10522, 2025b.

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

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

Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF$--$: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021.

Yitong Xia, Hao Tang, Radu Timofte, and Luc Van Gool. Sinerf: Sinusoidal neural radiance fields for joint pose estimation and scene reconstruction. *arXiv preprint arXiv:2210.04553*, 2022.

Tian-Xing Xu, Xiangjun Gao, Wenbo Hu, Xiaoyu Li, Song-Hai Zhang, and Ying Shan. Geometrycrafter: Consistent geometry estimation for open-world videos with diffusion priors. *ICCV*, 2025.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

Botao Ye, Sifei Liu, Haofei Xu, Li Xueting, Marc Pollefeys, Ming-Hsuan Yang, and Peng Songyou. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *ICLR*, 2025.

Zongxin Ye, Wenyu Li, Sidun Liu, Peng Qiao, and Yong Dou. Absgs: Recovering fine details in 3d gaussian splatting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1053–1061, 2024.

Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5752–5761, 2021.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19447–19456, June 2024.

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. *ICLR*, 2025a.

Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views, 2025b. URL https://arxiv.org/abs/2502.12138.

Yuqi Zhang, Guanying Chen, and Shuguang Cui. Efficient large-scale scene representation with a hybrid of high-resolution grid and plane features. *arXiv preprint arXiv:2303.03003*, 2023.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5745–5753, 2019.

A Additional Ablations

Table 5: Additional ablation on model components in multi-view reconstruction. The experiments are conducted on *MipNeRF360* (Barron et al., 2022). This table showcases the aborted model designs. The best performance of each metric is in **bold**. The "Baseline" denotes MCMC-3DGS equipped with pose embedding. The modification of each following row is independent of the others.

	Initialization		Train Set			Test Set		
Model	Pose	Point Map	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓
Baseline	Ours	Matched Points	0.8966	29.59	0.1314	**0.7821**	**26.40**	0.1774
w MLP	Ours	Matched Points	0.8753	28.60	0.1436	0.7492	25.83	0.1934
w depth	Ours	Matched Points	0.8851	28.85	0.1415	0.7628	25.94	0.1954
w 2× pose lr	Ours	Matched Points	**0.9044**	**30.02**	**0.1243**	0.7740	26.16	**0.1737**
w Epi. Loss	Ours	Matched Points	0.8964	29.55	0.1318	0.7795	26.31	0.1780
w Epi. Loss	VGGT*	VGGT*	0.8499	27.21	0.1806	0.6440	23.22	0.2572

Here we provide additional ablation studies in Tab. 5. We experimented with design choices like MLP-based pose embedding learning (Huang et al., 2025b), epipolar loss during 3DGS training, and depth supervision. But none of them bring clear benefits to the rendering quality. We also tried to double learning rate and encourage to learn a broader distribution, but it turns out to aggregate the overfitting phenomenon. Moreover, the last row of Tab. 5 shows that integrating global alignment into GS training, rather than treating it as a separate process, leads to suboptimal results. Therefore, we adopt global alignment as an independent component.

B Permutation-Equivariant AUC@30

In this section, we analyze why the conventional AUC@30 metric is sensitive to the input image order and propose a simple yet effective modification to address this issue. AUC@30 first computes relative poses for all C_N^2 image pairs. Comparing the relative poses from ground truth and predictions, the relative rotation and translation errors can be derived for AUC@30 calculation. Specifically, for two images indexed by i and j (with $i < j$) and their corresponding extrinsics, the relative pose is computed as:

$$\Delta E_{ij} = {E_i}^{-1} E_j = \begin{pmatrix} {R_i}^T & -{R_i}^T t_i \\ 0 & 1 \end{pmatrix} \begin{pmatrix} R_j & t_j \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} {R_i}^T R_j & {R_i}^T (t_j - t_i) \\ 0 & 1 \end{pmatrix}. \tag{5}$$

However, if the image order is permuted and j precedes i, the relative pose becomes:

$$\Delta E_{ji} = \begin{pmatrix} {R_j}^T R_i & {R_j}^T (t_i - t_j) \\ 0 & 1 \end{pmatrix}. \tag{6}$$

While the orthogonality of R_i and R_j ensures that ${R_j}^T R_i = {R_i}^T R_j$, it is clear that ${R_i}^T (t_j - t_i) \neq {R_j}^T (t_i - t_j)$. Consequently, the relative translation angle—and hence AUC@30—is sensitive to the ordering of input images, which can lead to differences exceeding five points. To mitigate this, we include both E_{ij} and E_{ji} in the relative pose sequence instead of only E_{ij}. This modification preserves the relative rotation error while introducing permutation equivariance to relative translation error and AUC@30, resulting in a more robust and fair evaluation of pose estimation accuracy.

C Large Language Model Usage

We used LLMs solely as a writing assistant to improve grammar, clarity, and conciseness of the manuscript. The research ideas, technical contributions, experiments, and analyses were entirely conceived and conducted by the authors. No content was generated by LLMs beyond language refinement, and all scientific claims and results are the sole responsibility of the authors.