VGGT-Based COLMAP-Free Initialization for Sparse-View 3D Gaussian Splatting

Shreeya Venkatraman*, Yuvan Raj Krishna*, and Jiji C V
*Joint first authors
Department of Computer Science and Engineering, Shiv Nadar University Chennai
shreezzz0704@gmail.com, yuvan22110466@snuchennai.edu.in, jijicv@snuchennai.edu.in

***Abstract*—Standard 3D Gaussian Splatting (3DGS) pipelines for Novel View Synthesis (NVS) are bottlenecked by Structure-from-Motion (SfM) initialization. In casual, sparse-view scenarios, feature matching breaks down, causing the entire reconstruction process to fail. We replace this brittle dependency with a COLMAP-free, feed-forward initializer powered by a Visual Geometry Grounded Transformer (VGGT). By leveraging VGGT, our pipeline jointly estimates camera parameters and dense scene geometry across all views in a single pass. A Bridge Module then robustly normalizes the scene scale and conditions initial Gaussian opacity on geometric confidence to discourage floater artifacts during densification. Our framework reduces the initialization phase from minutes (full-scene SfM) to seconds and achieves 100% initialization success from as few as three unposed images (a regime where COLMAP succeeds on only 1 of 7 Mip-NeRF 360 scenes). Project page:** `https://github.com/yuvanrajkrishna/VGGT-Sparse-3DGS`.

***Index Terms*—3DGS, VGGT, COLMAP-free initialization, Sparse-view capture, NVS**

I. Introduction

Generating photorealistic novel views from sparse, wide-baseline images remains a central challenge in neural rendering. While Neural Radiance Fields (NeRFs) [1] and fast hash-encoded variants [2] achieve high-quality results, their reliance on volumetric sampling limits real-time performance. Sparse-NeRF regularization [3] and joint pose optimization [4] add further computational overhead, making them ill-suited to interactive applications.

Conversely, 3D Gaussian Splatting (3DGS) [5] achieves real-time rendering ($>$ 100 FPS) but is strictly bounded by its initialization, relying on Structure-from-Motion (SfM) libraries like COLMAP [6]. In casual, sparse-view regimes, feature matching routinely breaks down, causing optimization to fail entirely. Recent efforts address this by applying feed-forward regression [7]–[9], joint pose optimization [10], [11], or depth and point-cloud priors [12], [13]. While 3D foundation models [14], [15] provide strong geometric anchors, deploying them directly often demands complex global alignment operations [16].

To resolve this bottleneck, we introduce a COLMAP-free initialization pipeline. While pairwise initializers like InstantSplat [16] utilize stereo priors to initialize 3DGS, their $\mathcal{O}(N^2)$ architecture requires post-hoc scale alignment across independently computed relative poses. In contrast, our pipeline leverages VGGT [17]. By integrating cross-view context via joint N-view global attention, we implicitly encourage scale consistency and directly output globally aligned geometry in a single pass. This architecture enables coherent scene inference even in extreme wide-baseline regimes where pairwise matching collapses, recovering coherent scene structure from as few as three unposed images. Our contributions are:

1) **A Single-Pass, Globally Consistent Initializer:** By exploiting VGGT's joint N-view processing, we recover globally consistent camera parameters and dense scene geometry in a single forward pass. This architectural distinction is the core enabler of robust initialization under wide-baseline, unposed conditions.
2) **The Bridge Module:** A mechanism bridging VGGT's unbounded neural geometry and 3DGS's bounded optimization space. Global Coordinate Normalization anchors the scene to prevent vanishing positional gradients. Uncertainty-Gated Seeding conditions initial Gaussian opacity on geometric confidence, designed to suppress the floater artifacts that unconstrained densification would otherwise amplify.

II. Related Work

We categorize sparse-view 3DGS methods by their geometric initialization strategies.

A. SfM-Free Radiance Fields

To bypass SfM fragility, early methods jointly optimized geometry and poses [4]; CF-3DGS [10] adapts this to 3DGS, but resolving unconstrained local minima requires expensive auxiliary constraints like tri-consistency supervision (TriGS [11]).

Feed-forward regressors offer rapid inference but remain domain-constrained. For example, pixelSplat [7] is tailored for narrow-baseline stereo. Extensions like MVSplat [8], NoPoSplat [9], and AnySplat [18] broaden this scope but still struggle in the extreme wide-baseline, unposed regime our approach targets.

B. Prior-Guided Geometric Initialization

Foundation models like DUSt3R [14] and MASt3R [15] extract dense pointmaps from image pairs, and InstantSplat [16] uses these stereo priors to initialize 3DGS in a pose-free framework. Because it optimizes the scene and cameras jointly from unposed images, its learned coordinate space does not

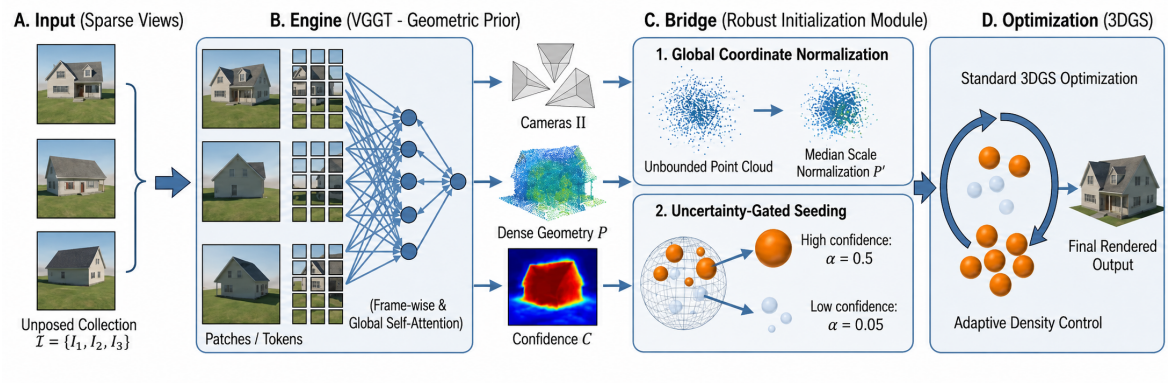


Fig. 1. **Method Overview. (A)** Sparse unposed images $\mathcal{I}$ where SfM fails. **(B)** VGGT regresses cameras Π, dense geometry $\mathbf{P}$, and confidence $\mathbf{C}$ via Alternating Attention. **(C)** The Bridge applies Global Coordinate Normalization and Uncertainty-Gated Seeding. **(D)** Standard 3DGS optimization.

align with the benchmark test views; evaluation therefore relies on a post-training refinement phase in which the fully trained model is frozen and each test camera's extrinsics are iteratively adjusted to minimize photometric error against the held-out ground-truth test images before metrics are computed.

In contrast, our method addresses geometric noise before optimization and adheres strictly to a train-only evaluation paradigm: VGGT [17] jointly processes all N views to natively enforce scale consistency in a single forward pass, and our Bridge Module stabilizes the result by down-weighting low-confidence points through their initial opacity prior to densification—using only the input training views and never exposing the pipeline to ground-truth test images. Moreover, while InstantSplat also consumes its prior's confidence maps, it uses them to rank views and to scale per-point learning rates, seeding all Gaussians at a uniform opacity; our Uncertainty-Gated Seeding instead injects confidence directly into the initial opacity state, gating densification before optimization begins.

While generative approaches like GSFixer [19] restore views via costly video diffusion, the alternating-attention backbone directly predicts globally aligned geometry—which our Bridge Module then bounds for 3DGS—avoiding both alignment and generative overhead entirely.

III. Methodology

3DGS parameterizes scenes as 3D Gaussians with position $\mu \in \mathbb{R}^3$, covariance Σ, opacity $\alpha \in [0, 1]$, and Spherical Harmonics (SH) color; novel views are rendered by differentiable α-blending, and optimization efficacy depends heavily on initial spatial quality.

Our framework replaces brittle SfM point clouds with a feed-forward neural prior. The pipeline (Fig. 1) comprises: (1) **Feed-Forward Inference** via VGGT, (2) a **Bridge Module**, and (3) **Standard 3DGS Optimization**.

A. Feed-Forward Geometric Inference

Standard 3DGS requires COLMAP [6] for camera extrinsics and sparse point clouds (Fig. 1A); in wide-baseline scenarios feature matching fails, leaving no viable starting point. We employ VGGT [17] as our geometric backbone (Fig. 1B).

Architecture: VGGT processes $\mathcal{I}$ via a pre-trained DINOv2 encoder to extract semantic tokens. The transformer applies **Alternating Attention** across 24 paired layers, interleaving *frame-wise* and *global* self-attention. This joint processing implicitly aggregates cross-view geometric evidence, synthesizing consistent predictions even in occluded or textureless regions without reliable local correspondences.

Scale & Robustness: Pre-trained on diverse 3D datasets (e.g., Co3Dv2, ScanNet, MegaDepth), VGGT's 1.2 billion parameters enable zero-shot generalization to in-the-wild sparse captures where classical optimization fails.

Output Heads: Tokens are decoded into camera parameters Π and a dense point cloud $\mathbf{P} \in \mathbb{R}^{N\times H\times W\times 3}$ with a per-point confidence map $\mathbf{C} \in \mathbb{R}^{N\times H\times W}$ reflecting the predicted aleatoric uncertainty; low confidence indicates unreliable feature-space triangulation. We utilize $\mathbf{P}$ and $\mathbf{C}$ to seed the initial Gaussian field.

Pose Inaccuracy Robustness: While VGGT's cameras lack bundle-adjusted sub-pixel precision, their cross-view consistency provides a coherent initialization, and Uncertainty-Gated Seeding is designed to further mitigate pose noise by down-weighting low-confidence regions before densification.

B. Robust Initialization Strategies

Neural models predict unbounded geometry, whereas 3DGS requires bounded spatial coordinates to stabilize positional gradients. Our Bridge Module (Fig. 1C) resolves this.

1) Global Coordinate Normalization: We normalize the scene scale by the median point-to-origin distance, a statistic robust to the extreme "fly-away" outliers common in transformer predictions. The resulting scale factor s is applied to both the point cloud and the camera translations, preserving relative geometry while normalizing the scene extent. The normalized geometry $\mathbf{P}'$ is:

$$\mathbf{P}' = s \cdot \mathbf{P}, \qquad s = 1 \,/\, \text{median}_i \|\mathbf{p}_i\|_2 \tag{1}$$

where $\mathbf{p}_i \in \mathbb{R}^3$ is the i-th point of $\mathbf{P}$ and the median is taken over all predicted points, i.e., s is the reciprocal of the median

point-to-origin distance. This transformation stably anchors the initial Gaussians for the rasterizer.

2) Uncertainty-Gated Seeding: Standard adaptive density control amplifies noisy initializations by misinterpreting errors as missing details. To prevent runaway densification, *Uncertainty-Gated Seeding* uses the per-point confidence score c_i emitted by VGGT's uncertainty head (monotonically decreasing in the predicted aleatoric uncertainty). This score modulates the initial opacity $\alpha_{\text{init}}^{(i)}$ via a bounded linear mapping:

$$\alpha_{\text{init}}^{(i)} = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min}) \frac{c_i - c_{\min}}{c_{\max} - c_{\min}} \tag{2}$$

where $c_{\min}$ and $c_{\max}$ are the minimum and maximum confidence scores over the predicted point cloud, computed per scene. By construction, $\alpha_{\text{init}}^{(i)} \in [\alpha_{\min}, \alpha_{\max}]$, so no explicit clamping is required.

The bounds $\alpha_{\min} = 0.05$ and $\alpha_{\max} = 0.5$ prevent low-confidence points from remaining permanently transparent, while the conservative upper cap stops over-confident initializations from forming gradient-blocking opaque walls.

C. Optimization and Rendering

Following initialization, we revert to standard 3DGS optimization (Fig. 1D), minimizing photometric error. Uncertainty-Gated Seeding is intended to let standard adaptive density control function robustly even with 3–6 views. We use an $\mathcal{L}_1$ and D-SSIM loss ($\lambda = 0.2$):

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_1 + \lambda \mathcal{L}_{\text{D-SSIM}} \tag{3}$$

Our approach leaves the downstream renderer untouched, ensuring modularity.

IV. Experiments

A. Experimental Setup

Dataset & Protocol. We evaluate our method on a standard subset of 7 scenes (bicycle, bonsai, counter, garden, kitchen, room, stump) from the Mip-NeRF 360 dataset [20]. To test zero-shot generalization, we also evaluate on Tanks & Temples (8 scenes: *Ballroom*, *Barn*, *Church*, *Family*, *Francis*, *Horse*, *Ignatius*, *Museum*) and MVImgNet (7 object-centric scenes: *bench*, *bicycle*, *car*, *chair*, *ladder*, *suv*, *table*), following the evaluation protocol of InstantSplat [16]. None of our evaluation datasets appears in VGGT's published training-data list. Images are downsampled by a factor of 8 (`-r 8`). Training view counts range from $n = 3$ to $n = 12$. For evaluation, every 8th image of each Mip-NeRF 360 scene is held out as the test set (206 test images across the 7 scenes), while Tanks & Temples and MVImgNet use the fixed InstantSplat split of 12 test images per scene (96 and 84 test images, respectively). All reported metrics are computed exclusively on these held-out test views, rendered at their reference camera poses; the test set is identical at every n. To evaluate our robust initialization under rapid convergence, we fix 3DGS optimization at 10,000 iterations (Spherical Harmonics degree of 3) instead of the standard 30,000 iterations. Training extended to 30K remains stable (no divergence or late-stage artifacts) but yields no test-side gain in this sparse regime (e.g., *counter*, n=12: 14.66 dB at 10K vs. 14.60 dB at 30K), validating the 10K budget.

TABLE I

Novel View Synthesis Results on Three Datasets. COLMAP vs. VGGT (Ours) under an identical train-only protocol (n views, 10K iterations). **Succ.**: scenes that COLMAP initialization succeeds on. **COLMAP (matched)** and **Ours (matched)** columns average both methods over that row's COLMAP-surviving scenes (bold: best); Δ is the matched PSNR difference (Ours − COLMAP). **Ours (all scenes)** columns are averages over all scenes, because our method succeeds on all scenes at every n.

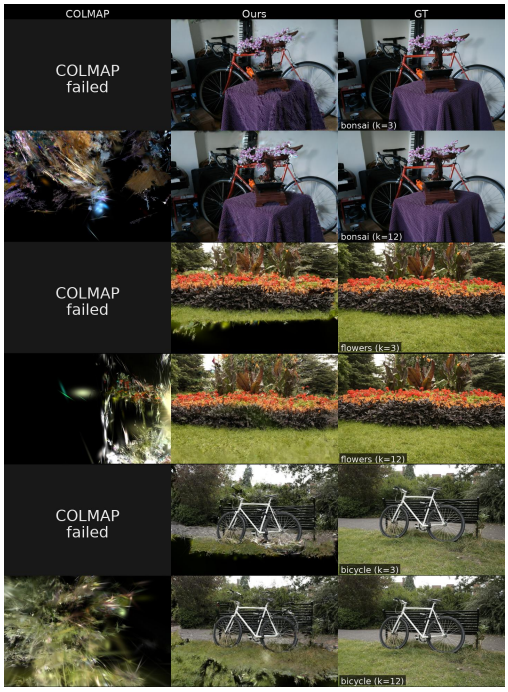
n	Succ.	COLMAP (matched)			Ours (matched)			Δ	Ours (all scenes)		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓		PSNR↑	SSIM↑	LPIPS↓
Mip-NeRF 360 (7 scenes)											
3	1/7	9.60	0.124	0.669	**10.80**	**0.176**	**0.622**	+1.20	12.34	0.167	0.607
6	2/7	10.33	0.144	0.647	**15.35**	**0.244**	**0.461**	+5.02	14.12	0.228	0.523
9	6/7	12.65	0.200	0.578	**14.61**	**0.275**	**0.495**	+1.96	14.88	0.257	0.502
12	6/7	14.36	0.292	0.487	**15.91**	**0.309**	**0.439**	+1.55	15.96	0.287	0.454
Tanks & Temples (8 scenes)											
3	6/8	12.45	0.287	0.550	**13.65**	**0.322**	**0.472**	+1.20	14.01	0.334	0.458
6	7/8	15.62	0.390	0.396	**16.30**	**0.454**	**0.352**	+0.68	16.50	0.457	0.346
9	7/8	**16.74**	0.435	**0.325**	16.59	**0.436**	0.336	−0.15	16.80	0.439	0.331
12	7/8	16.73	0.435	0.327	**16.80**	**0.456**	0.327	+0.07	16.98	0.457	0.323
MVImgNet (7 scenes)											
3	3/7	10.66	0.270	0.578	**15.21**	**0.391**	**0.474**	+4.55	14.77	0.352	0.481
6	3/7	15.54	0.500	0.415	**17.55**	**0.532**	**0.356**	+2.01	16.73	0.386	0.414
9	2/7	**20.18**	**0.664**	**0.287**	17.85	0.566	0.305	−2.33	16.98	0.389	0.401
12	2/7	**19.79**	**0.636**	**0.230**	17.22	0.545	0.317	−2.57	16.76	0.386	0.406

Hardware & Runtime. Experiments were conducted on a single NVIDIA GeForce RTX 4080 SUPER (16 GB). VGGT feed-forward inference (including model load and export) requires ≈10 s on average. The complete pipeline (from raw unposed images to a fully trained 3DGS model at 10K iterations) averages ≈137 s (101–166 s for n=3–12), reducing the initialization phase from minutes (full-scene SfM) to seconds. Peak GPU memory is dominated by the single VGGT forward pass (≈14 GB for 12 views); 3DGS optimization itself typically requires under 2 GB (up to ∼5 GB on heavily densified outdoor scenes), so the full pipeline runs on one consumer GPU.

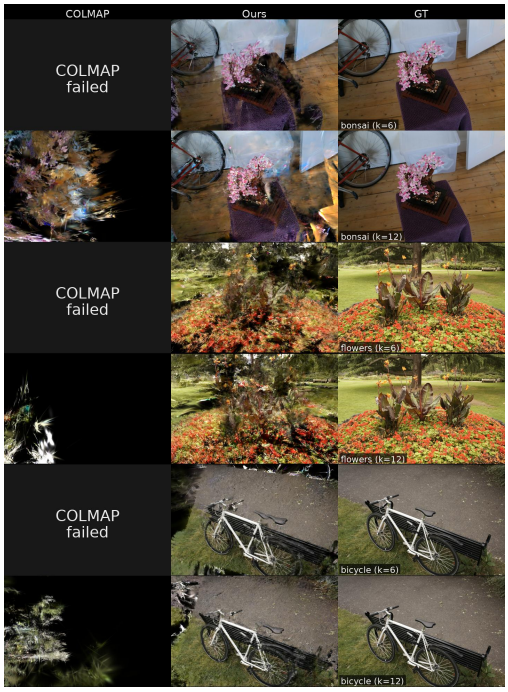
B. Robustness in the Sparse Regime

The fundamental premise of our work is that classical SfM is severely bottlenecked by wide-baseline sparsity. Table I validates this claim through a strict *train-only* comparison, where the initialization methods are completely denied access to test views. We define initialization success as producing a non-empty, structurally coherent point cloud and valid cameras sufficient to commence 3DGS optimization—distinct from final reconstruction quality.

As shown in Fig. 2, at extreme sparsity ($n = 3$ and $n = 6$), classical feature matchers fail to establish valid epipolar geometry. This results in fragmented or completely empty point clouds. On Mip-NeRF 360, COLMAP successfully initializes only 1 of 7 scenes at $n = 3$, yielding a very low 9.60 dB single-scene average. By replacing this brittle local feature matching with VGGT's global attention prior, which implicitly aggregates cross-view geometric evidence, our method achieves a 100% initialization success rate across



(a) Extreme sparsity ($n = 3$). Left: COLMAP output. Middle: VGGT (Ours). Right: Ground Truth.



(b) Sparse regime ($n = 6$). Left: COLMAP output. Middle: VGGT (Ours). Right: Ground Truth.

Fig. 2. **Qualitative Comparison of Initialization Robustness.** The grids compare **COLMAP** (left), **VGGT (Ours)** (middle), and **Ground Truth** (right). For extremely sparse inputs ($n = 3$ and $n = 6$), COLMAP frequently returns null, fragmented, or degenerate reconstructions, while our neural initializer consistently recovers coherent global structure and appearance. The *flowers* panels are qualitative only (outside the 7-scene quantitative subset).

all three datasets and all view counts. Notably, on Mip-NeRF 360 at n=6, our method delivers a **+5.02 dB improvement** on the matched subset of scenes COLMAP manages to initialize, while additionally reconstructing the five scenes for which COLMAP returns nothing.

Table I also shows the initialization remains stable on unseen domains: on Tanks & Temples and MVImgNet our pipeline again initializes every scene at every view count and trains stably, even at the extreme $n = 3$ boundary. The only configurations where COLMAP overtakes our pipeline by a meaningful margin on the matched subset are MVImgNet at $n \geq 9$, where its surviving set collapses to the two most SfM-friendly object-centric scenes (on Tanks & Temples at n=9 the two are effectively tied, -0.15 dB); on such near-ideal captures a fully converged SfM solution is more precise than our feed-forward prior. This is consistent with our positioning: the proposed initializer guarantees a usable reconstruction where SfM fails outright, rather than surpassing SfM where it already succeeds.

Pose accuracy. To quantify the initialization directly, we measure the Sim(3)-aligned absolute trajectory error of VGGT's training cameras against Mip-NeRF 360's full-scene reference COLMAP poses (available for all 7 scenes at every n): only 1.7–5.8% of the RMS trajectory radius, with relative rotation errors nearly flat at 1.2–1.5° regardless of n—evidence that joint N-view inference yields globally coherent camera geometry up to a single similarity transform. The sole outlier (*bicycle*, n=9: 10.7° relative translation direction error) is also the weakest rendering cell at that view count, corroborating that residual failures stem from pose drift (Sec. V).

TABLE II

STATE-OF-THE-ART COMPARISON (MIP-NERF 360, $n = 12$).

FULL-DATASET METHODS EXTRACT POSES USING ALL AVAILABLE IMAGES. UNDER STRICT TRAIN-ONLY CONSTRAINTS, OUR METHOD SUBSTANTIALLY OUTPERFORMS THE CLASSICAL BASELINE IN PSNR AND LPIPS WITH COMPARABLE SSIM. **BEST** AND SECOND BEST FULL-DATASET SCORES HIGHLIGHTED. SPARSE-VIEW ROWS ARE 7-SCENE MEANS UNLESS MARKED; †6 TRAIN-ONLY COLMAP-SURVIVING SCENES.

Method	Pose Prior	PSNR ↑	SSIM ↑	LPIPS ↓
Requires Full-Dataset Images				
COLMAP (Sparse) [6]	Full-Dataset Poses	17.76	0.496	0.383
COLMAP (Dense) [21]	Full-Dataset Poses	18.05	0.520	**0.358**
3DGS Baseline [5]	Full-Dataset Poses	18.52	0.523	0.415
FSGS [22]	Full-Dataset Poses	18.80	0.531	0.418
CoR-GS [23]	Full-Dataset Poses	19.52	0.558	0.418
DropGaussian [24]	Full-Dataset Poses	**19.74**	**0.577**	0.364
Strict Sparse-View Constraints				
COLMAP Baseline [6]†	Train-Only Poses	14.36	**0.292**	0.487
VGGT (Ours)	**Train-Only Poses**	**15.96**	0.287	**0.454**

C. State-of-the-Art Comparisons

Table II evaluates our method against canonical baselines at $n = 12$. The baselines—DropGaussian [24], CoR-GS [23], FSGS [22], and the COLMAP MVS variants [21]—were chosen because they share our rigorous fixed-pose evaluation paradigm: no post-training test-view parameter optimization or test-image fitting, strictly preventing test-view leakage.

All full-dataset baselines are quoted from the DropGaussian benchmark [24], whose experimental framework (`-r 8`, 12 training views) fully aligns with ours.

While top performers like DropGaussian (19.74 dB) and CoR-GS (19.52 dB) rely on *Full-Dataset Poses*—pre-computed from all dataset images, including test views—our pipeline adheres to a strict *Train-Only* paradigm, inferring unobserved geometry from solely the 12 training views. Under the same train-only constraints, the standard COLMAP baseline collapses to 14.36 dB (averaged over its six surviving scenes). Despite operating without the dense geometric priors derived from full-dataset poses, our feed-forward approach successfully recovers highly coherent structures at 15.96 dB. The apparent SSIM deficit relative to the baseline row is an artifact of mismatched scene sets: on the identical six-scene subset, our method outperforms the COLMAP baseline on all three metrics (15.91 vs. 14.36 dB PSNR, 0.309 vs. 0.292 SSIM, 0.439 vs. 0.487 LPIPS). Our pipeline provides a robust initialization pathway for in-the-wild captures where dense multi-view data is unavailable and traditional SfM fails. We omit NoPoSplat [9] from Table II as its published evaluations utilize substantially different data and training configurations, precluding a direct 1-to-1 comparison under our standardized protocol.

V. Conclusion

We presented a COLMAP-free initialization framework for sparse-view 3DGS, replacing brittle feature matching with VGGT's joint N-view inference: 100% initialization success at $n = 3$, completing initialization in $\sim$10 s versus minutes for full-scene SfM. Its modular design leaves the downstream renderer untouched, serving as a drop-in for standard 3DGS pipelines.

Limitations & Future Work. Reconstruction quality is fundamentally bounded by VGGT's initialization accuracy; residual camera-pose drift on wide-baseline outdoor captures (e.g., *bicycle*) directly caps rendering quality. Primary failure modes include: (1) uniformly featureless regions (e.g., blank walls, overcast skies), where even token-based inference cannot recover depth; (2) highly specular objects violating photometric consistency; and (3) extreme foreground–background scale disparities, where the single global median-distance scale factor may under-represent near-field geometry. While Uncertainty-Gated Seeding down-weights low-confidence regions, the optimizer may still converge to suboptimal minima, and the pipeline is restricted to static scenes. Future work will incorporate train-view pose refinement (jointly optimizing training-camera extrinsics with the Gaussian parameters) and will extend the framework to dynamic, in-the-wild captures.

References

[1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 405–421.

[2] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM transactions on graphics (TOG)*, vol. 41, no. 4, pp. 102:1–102:15, 2022.

[3] M. Niemeyer, J. T. Barron, B. Mildenhall, M. S. Sajjadi, A. Geiger, and N. Radwan, "RegNeRF: Regularizing neural radiance fields for view synthesis from sparse inputs," in *Proc. IEEE/CVF CVPR*, 2022, pp. 5480–5490.

[4] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, "BARF: Bundle-adjusting neural radiance fields," in *Proc. IEEE/CVF ICCV*, 2021, pp. 5741–5751.

[5] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–14, 2023.

[6] J. L. Schonberger and J.-M. Frahm, "Structure-from-Motion revisited," in *Proc. IEEE CVPR*, 2016, pp. 4104–4113.

[7] D. Charatan, S. L. Li, A. Tagliasacchi, and V. Sitzmann, "pixelSplat: 3D Gaussian splats from image pairs for scalable generalizable 3D reconstruction," in *Proc. IEEE/CVF CVPR*, 2024, pp. 19 457–19 467.

[8] Y. Chen, H. Xu, C. Zheng, B. Zhuang, M. Pollefeys, A. Geiger, T.-J. Cham, and J. Cai, "MVSplat: Efficient 3D Gaussian splatting from sparse multi-view images," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 370–386.

[9] B. Ye, S. Liu, H. Xu, X. Li, M. Pollefeys, M.-H. Yang, and S. Peng, "No pose, no problem: Surprisingly simple 3D Gaussian splats from sparse unposed images," in *The Thirteenth International Conference on Learning Representations*, 2025.

[10] Y. Fu, S. Liu, A. Kulkarni, J. Kautz, A. A. Efros, and X. Wang, "COLMAP-free 3D Gaussian splatting," in *Proc. IEEE/CVF CVPR*, 2024, pp. 20 796–20 805.

[11] C. Huang, Q. Zhang, Q. Zhang, N. Li, Y. Gong, X. Wang, and W. Feng, "TriGS: Tri-consistency 3D Gaussian splatting from sparse and unposed views," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 699–708.

[12] J. Li, J. Zhang, X. Bai, J. Zheng, X. Ning, J. Zhou, and L. Gu, "DNGaussian: Optimizing sparse-view 3D Gaussian radiance fields with global-local depth normalization," in *Proc. IEEE/CVF CVPR*, 2024, pp. 20 775–20 785.

[13] Y. Jang and E. Pérez-Pellitero, "CoMapGS: Covisibility map-based Gaussian splatting for sparse novel view synthesis," in *Proc. IEEE/CVF CVPR*, 2025, pp. 26 779–26 788.

[14] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "DUSt3R: Geometric 3D vision made easy," in *Proc. IEEE/CVF CVPR*, 2024, pp. 20 697–20 709.

[15] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3D with MASt3R," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 71–91.

[16] Z. Fan, W. Cong, K. Wen, K. Wang, J. Zhang, X. Ding, D. Xu, B. Ivanovic, M. Pavone, G. Pavlakos *et al.*, "InstantSplat: Sparse-view Gaussian splatting in seconds," *arXiv preprint arXiv:2403.20309*, 2024.

[17] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "VGGT: Visual geometry grounded transformer," in *Proc. IEEE/CVF CVPR*, 2025, pp. 5294–5306.

[18] L. Jiang, Y. Mao, L. Xu, T. Lu, K. Ren, Y. Jin, X. Xu, M. Yu, J. Pang, F. Zhao *et al.*, "AnySplat: Feed-forward 3D Gaussian splatting from unconstrained views," *ACM Transactions on Graphics (TOG)*, vol. 44, no. 6, pp. 1–16, 2025.

[19] X. Yin, Q. Zhang, J. Chang, Y. Feng, Q. Fan, X. Yang, C.-M. Pun, H. Zhang, and X. Cun, "GSFixer: Improving 3D Gaussian splatting with reference-guided video diffusion priors," *arXiv preprint arXiv:2508.09667*, 2025.

[20] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-NeRF 360: Unbounded anti-aliased neural radiance fields," in *Proc. IEEE/CVF CVPR*, 2022, pp. 5470–5479.

[21] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixel-wise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 501–518.

[22] Z. Zhu, Z. Fan, Y. Jiang, and Z. Wang, "FSGS: Real-time few-shot view synthesis using Gaussian splatting," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 145–163.

[23] J. Zhang, J. Li, X. Yu, L. Huang, L. Gu, J. Zheng, and X. Bai, "CoR-GS: sparse-view 3D Gaussian splatting via co-regularization," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 335–352.

[24] H. Park, G. Ryu, and W. Kim, "DropGaussian: Structural regularization for sparse-view Gaussian splatting," in *Proc. IEEE/CVF CVPR*, 2025, pp. 21 600–21 609.