S4GS: RESOLUTION-ASYMMETRIC 4K FEED-FORWARD GAUSSIAN RECONSTRUCTION FROM SPARSE VIEWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Feed-forward 3D Gaussian Splatting (3DGS) provides an efficient way to reconstruct renderable 3D scenes from sparse views, making it attractive for mobile immersive applications. However, immersive rendering demands 4K-level scene reconstruction, while directly processing such high-resolution inputs with large geometry backbones incurs prohibitive GPU memory costs. Existing high-resolution 3DGS methods mainly rely on per-scene optimization and are often enhanced by image or video super-resolution priors, making them difficult to support real-time immersive rendering. We present **S4GS**, a **resolution-asymmetric framework for 4K feed-forward Gaussian reconstruction from sparse views.** The key idea is to perform expensive scene-level reasoning at low resolution, while reserving high-resolution processing for dense geometry construction and local appearance refinement. For global reasoning, S4GS uses a low-resolution VGGT backbone to estimate camera parameters and build a coarse Gaussian scaffold. For high-resolution recovery, we anchor Gaussian positions with dense depth maps and use a lightweight RGB branch to extract high-resolution image features and predict residual Gaussian attributes for fine-detail recovery. By separating global 3D reasoning from high-resolution scene recovery, S4GS enables the full training and inference pipeline on a single 24 GB RTX 4090 GPU. Experiments on real-world mobile stereo data show that S4GS produces high-fidelity 4K novel views and supports variable-baseline stereo rendering for practical mobile-to-headset immersive applications.

1 INTRODUCTION

Feed-forward 3D Gaussian Splatting (3DGS) reconstructs a renderable scene directly from a few input images (Charatan et al., 2024; Chen et al., 2024; Jiang et al., 2025). This capability offers a practical route to mobile 3D content creation, allowing users to capture a scene with a smartphone and explore the reconstruction on a headset. Mobile stereo images provide practical input, but native-resolution processing places substantial demands on computational resources. Making high-resolution feed-forward reconstruction practical under limited GPU memory therefore remains an important challenge.

Recent large-scale visual geometry models provide a strong basis for feed-forward reconstruction by jointly reasoning about cameras and scene structure across input views (Wang et al., 2025a; Jiang et al., 2025). However, immersive rendering requires 4K-level reconstruction, while directly processing such high-resolution inputs with these backbones incurs prohibitive GPU memory costs. Existing high-resolution 3DGS methods mainly rely on per-scene optimization and are often enhanced by image or video super-resolution priors (Feng et al., 2024; Shen et al., 2024; Xie et al., 2025; Wan et al., 2025). Although these approaches improve visual detail, their scene-specific optimization and post-processing overhead make them difficult to use for real-time immersive rendering. This motivates a resolution-asymmetric design that retains the global reasoning capability of a large geometry backbone while recovering high-resolution geometry and appearance within a limited memory budget.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We present **S4GS**, a *resolution-asymmetric framework for 4K feed-forward Gaussian reconstruction from sparse views*. It uses a large pretrained 3D geometry reasoning model at low resolution to estimate cameras and a coarse Gaussian scaffold, while native-resolution stereo inputs provide dense depth anchors and RGB features for residual refinement. This division recovers a high-resolution Gaussian field without routing native-resolution images through the expensive global backbone.

This resolution split keeps the complete training and inference pipeline within the memory capacity of a single 24 GB RTX 4090 GPU. Evaluation on real-world mobile stereo captures demonstrates detailed 4K reconstruction for immersive viewing. We further examine the effect of global reasoning resolution, dense geometry anchoring, and high-resolution RGB refinement to assess how each component contributes to reconstruction quality and efficiency.

Our main contributions are summarized as follows:

- We introduce S4GS, a resolution-asymmetric framework for 4K feed-forward Gaussian reconstruction from sparse views, enabling the full training and inference pipeline on a single 24 GB RTX 4090 GPU by separating global reasoning from high-resolution scene recovery.
- We introduce a dense geometry-anchored lifting strategy that constructs fixed high-resolution Gaussian centers from aligned depth, providing explicit spatial support for the high-resolution representation.
- We introduce a lightweight RGB-guided residual refinement scheme that combines coarse Gaussian attributes with native-resolution image features to recover fine appearance details for high-fidelity novel-view rendering and immersive stereo output.

2 Related Work

2.1 Feed-Forward Visual Geometry and Gaussian Reconstruction

Feed-forward visual geometry models simplify multi-view reconstruction by predicting scene geometry directly from images. DUSt3R (Wang et al., 2024) predicts dense point maps from image pairs, and MASt3R (Leroy et al., 2024) strengthens this paradigm with geometry-aware matching. VGGT (Wang et al., 2025a) jointly estimates cameras, depth, point maps, and point tracks, while Fast3R (Yang et al., 2025) and CUT3R (Wang et al., 2025b) study scalable and persistent feed-forward perception. Together, these models show that pretrained visual geometry networks can recover globally consistent structure with little or no per-scene optimization.

This paradigm has also been combined with 3D Gaussian representations. pixelSplat (Charatan et al., 2024) and MVSplat (Chen et al., 2024) predict renderable Gaussians from sparse views, while Splatt3R (Smart et al., 2024) and NoPoSplat (Ye et al., 2025) reduce dependence on calibrated cameras. AnySplat (Jiang et al., 2025) jointly estimates cameras, geometry, and Gaussians from unconstrained views. Recent methods further emphasize geometric priors: DepthSplat (Xu et al., 2025) uses pretrained depth features, Depth Anything 3 (Lin et al., 2026) provides spatially consistent dense geometry, and AnchorSplat (Zhang et al., 2026) explicitly anchors Gaussian prediction in 3D. EcoSplat (Park et al., 2026) instead studies efficiency by reducing redundant primitives.

These methods establish an effective route to fast feed-forward 3D reconstruction, but their globally coupled reasoning is usually performed below native multi-megapixel mobile resolution. Increasing backbone resolution rapidly raises token and activation memory, whereas downsampling discards information needed for high-resolution rendering. Our work addresses this resolution gap by retaining low-resolution global reasoning and explicitly recovering high-resolution geometry and appearance from the original observations.

2.2 High-Resolution 3D Reconstruction and Super-Resolution

High-resolution scene reconstruction seeks to improve spatial fidelity without making computation prohibitive. Before the emergence of 3D Gaussian Splatting, GridNeRF (Xu et al., 2023) used compact multi-resolution feature grids for large scenes. 3D Gaussian Splatting (Kerbl et al., 2023) substantially improves optimization and rendering efficiency, but high-resolution reconstruction still requires dense spatial support and many primitives.

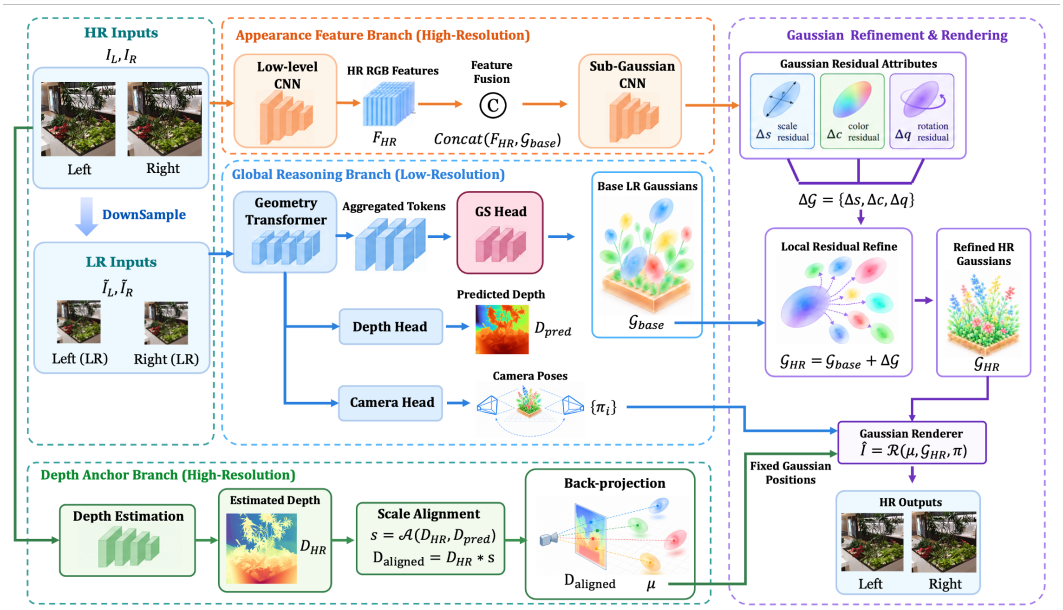

Figure 1: Overview of our resolution-asymmetric reconstruction framework.

A common solution is to reconstruct a low-resolution Gaussian scene and enhance it with pretrained 2D or video priors. SRGS (Feng et al., 2024) uses super-resolved images to guide Gaussian densification, while SuperGS (Xie et al., 2025) adopts a structured coarse-to-fine representation (Lu et al., 2024) with additional consistency constraints. S2Gaussian (Wan et al., 2025) combines depth regularization, controlled densification, and super-resolved source and pseudo views for sparse-view reconstruction. SuperGaussian (Shen et al., 2024) instead repurposes a pretrained video upsampling model and consolidates its predictions into a 3D-consistent representation. These methods demonstrate the value of image and video priors for enhancing details missing from a low-resolution scene.

Most existing approaches follow a post-hoc coarse-to-fine pipeline: a low-resolution scene is reconstructed first and then densified or re-optimized using enhanced observations. This becomes expensive as the target resolution and Gaussian count grow. Our setting is different because native high-resolution stereo observations are already available; the bottleneck is the memory required to process them with a large global geometry backbone. We therefore formulate the task as *resolution-asymmetric high-resolution reconstruction*. Global reasoning is performed at reduced resolution, while the original observations directly provide dense geometric anchors and high-frequency appearance cues, avoiding native-resolution backbone processing and a separate iterative super-resolution stage.

3 Method

We reconstruct a high-resolution 3D Gaussian field from sparse input views without applying a large visual geometry backbone directly to multi-megapixel images. Our key observation is that global scene reasoning and high-resolution scene recovery have different resolution requirements. We therefore use downsampled observations for global geometry, while reusing the original images for dense geometry and high-frequency appearance. The resulting pipeline has three stages: resolution-decoupled global reasoning, geometry-anchored Gaussian lifting, and RGB-guided appearance refinement.

3.1 Overview

Let $\mathcal{V} = \{1, \ldots, M\}$ index the sparse input views, and let $\mathcal{I}^H = \{I_i^H\}_{i \in \mathcal{V}}$ denote their native-resolution observations. A downsampling operator produces the low-resolution inputs $\mathcal{I}^L$. Our

target is a high-resolution Gaussian field $\mathcal{G}^H = \{(\boldsymbol{\mu}_j^H, \boldsymbol{\theta}_j^H)\}_{j=1}^{N_H}$, where $\boldsymbol{\mu}_j^H$ and $\boldsymbol{\theta}_j^H$ are the center and rendering attributes of Gaussian j. The field combines centers recovered from dense geometry with attributes refined from high-resolution appearance cues. Figure 1 summarizes the proposed pipeline.

3.2 Resolution-Decoupled Global Geometry Reasoning

Applying a large pretrained 3D geometry reasoning model at native resolution rapidly increases token and activation memory. We instead apply a generic global geometry model $\mathcal{F}_{\text{global}}$ to the downsampled views to recover the global scene prior:

$$\left(\mathcal{C}^H, \mathcal{D}_{\text{g}}^L, \mathcal{Z}^L\right) = \mathcal{F}_{\text{global}}\left(\mathcal{I}^L\right). \tag{1}$$

Here $\mathcal{C}^H = \{C_i^H\}_{i\in\mathcal{V}}$ contains the reference cameras, $\mathcal{D}_{\text{g}}^L$ is the coarse depth prediction, and $\mathcal{Z}^L$ contains aggregated multi-view features. The aggregated features are subsequently mapped by the Gaussian head to a coarse scaffold:

$$\mathcal{G}^L = \mathcal{H}_{\text{gauss}}\left(\mathcal{Z}^L\right). \tag{2}$$

The scaffold $\mathcal{G}^L$ provides a globally coherent prior rather than the final high-resolution representation. This separation lets the global model focus on scene-level structure while the subsequent branches recover dense local information.

3.3 Dense Geometry-Anchored Gaussian Lifting

The coarse scaffold does not provide sufficient spatial support for a native-resolution Gaussian field. A generic pretrained depth model $\mathcal{F}_{\text{depth}}$ therefore estimates dense depth from each original observation:

$$D_{\text{d},i}^H = \mathcal{F}_{\text{depth}}(I_i^H), \qquad i \in \mathcal{V}. \tag{3}$$

The depth predicted from the native-resolution image and the coarse depth produced by the global model may use different scales. For each view, we resize the dense depth to the coarse depth resolution and estimate a scale factor from their median values:

$$\gamma_i = \frac{\text{med}(D_{\text{g},i}^L)}{\text{med}(\mathcal{S}_D(D_{\text{d},i}^H))}. \tag{4}$$

Here $\mathcal{S}_D$ resizes a depth map to the coarse resolution. The aligned native-resolution depth is then obtained as $\widetilde{D}_{\text{d},i}^H = \gamma_i D_{\text{d},i}^H$, transferring the global scale estimated by the geometry model to the dense depth. Using the reference camera C_i^H, every valid native-resolution depth pixel is back-projected into the common reconstruction coordinate system:

$$\mathbf{x}_{i,\mathbf{p}} = \Pi^{-1}\left(\mathbf{p}, \widetilde{D}_{\text{d},i}^H(\mathbf{p}), C_i^H\right). \tag{5}$$

The back-projected point $\mathbf{x}_{i,\mathbf{p}}$ is directly assigned as the corresponding high-resolution Gaussian center $\boldsymbol{\mu}_j^H$ and remains fixed during appearance refinement. This operation is the key geometric lifting step: native-resolution depth supplies dense spatial support, while the reference cameras and scale alignment preserve the global coordinate frame. Because the centers are fixed, the residual branch only needs to refine Gaussian attributes and cannot explain high-resolution structure by drifting the 3D support.

3.4 RGB-Guided High-Frequency Gaussian Residual Refinement

The coarse scaffold provides a globally consistent Gaussian parameterization, but downsampling removes local color and appearance cues. We therefore use a lightweight low-level RGB CNN to extract native-resolution features:

$$F_i^H = \mathcal{E}_{\text{rgb}}(I_i^H), \qquad i \in \mathcal{V}. \tag{6}$$

In parallel, the coarse Gaussian attributes are rasterized into view-aligned maps Θ_i^L and upsampled to the native resolution:

$$\Theta_i^{0,H} = \mathcal{U}(\Theta_i^L). \tag{7}$$

The upsampled base attributes and the high-resolution RGB features are fused by a lightweight convolutional refinement network $\mathcal{R}_\phi$. It predicts a residual map rather than a new Gaussian field:

$$\Delta\Theta_i^H = \mathcal{R}_\phi\Big(\text{Concat}\left[\Theta_i^{0,H}, F_i^H\right]\Big). \tag{8}$$

The refined attribute map is then obtained by residual addition, $\Theta_i^H = \Theta_i^{0,H} + \Delta\Theta_i^H$. For a fixed anchor $\boldsymbol{\mu}_j^H$, sampling these maps gives $\boldsymbol{\theta}_j^{0,H} = \Theta_i^{0,H}(\mathbf{p})$, $\Delta\boldsymbol{\theta}_j^H = \Delta\Theta_i^H(\mathbf{p})$, and $\boldsymbol{\theta}_j^H = \boldsymbol{\theta}_j^{0,H} + \Delta\boldsymbol{\theta}_j^H$. The principal residual components are color, rotation, and scale:

$$\Delta\boldsymbol{\theta}_j^H = \big(\Delta\mathbf{c}_j, \Delta\mathbf{q}_j, \Delta\mathbf{s}_j\big), \tag{9}$$

where $\Delta\mathbf{c}_j$, $\Delta\mathbf{q}_j$, and $\Delta\mathbf{s}_j$ denote corrections to color, rotation, and scale, respectively. The center $\boldsymbol{\mu}_j^H$ is excluded from this residual, so the RGB branch restores high-frequency appearance while the geometry branch remains responsible for 3D localization.

3.5 Training Objective

The final field is rendered from the reference cameras, $\widehat{I}_i^H = \mathcal{R}_{\text{gs}}(\mathcal{G}^H, C_i^H)$. We optimize a reconstruction objective that combines pixel fidelity with perceptual similarity:

$$\mathcal{L} = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\left[\left\|\widehat{I}_i^H - I_i^H\right\|_2^2 + \lambda_{\text{LPIPS}}\,\text{LPIPS}(\widehat{I}_i^H, I_i^H)\right]. \tag{10}$$

We set $\lambda_{\text{LPIPS}} = 0.2$ in all experiments. Native-resolution supervision encourages the refined field to recover high-frequency appearance while remaining anchored to explicit dense geometry. During training, rendering and supervision use random crops from the native-resolution frames to limit memory, whereas inference evaluates the complete frame.

4 Experiments

We evaluate S4GS in terms of rendering quality, geometric consistency, and computational cost for ultra-high-resolution reconstruction under a constrained GPU memory budget. We first describe the experimental setup in Sec. 4.1 and compare the complete framework with alternative reconstruction strategies in Sec. 4.2. We then analyze memory scaling and conduct ablations in Sec. 4.3 to examine global reasoning resolution, dense geometry anchoring, and high-frequency RGB refinement.

4.1 Experimental Setup

Datasets. We evaluate S4GS on two datasets: a real-world self-collected mobile dataset and a synthetic virtual dataset. We introduce them below. **Self-collected Mobile Dataset** is captured with a smartphone supporting synchronized dual-camera acquisition and contains 2,056 stereo pairs at 3024×3024 resolution, including 1,852 pairs for training and 204 pairs reserved exclusively for final testing. The test set covers textured indoor and outdoor scenes, near-to-far subject distances, normal and low-light illumination, as well as difficult cases involving motion, occlusion, fine structures, and repetitive textures, providing a practical benchmark for multi-megapixel mobile reconstruction. We additionally evaluate on **StereoGenBench** (Cui et al., 2026), a synthetic Unreal Engine benchmark with a calibrated six-camera array and per-view RGB images, metric depth, and camera parameters. Its ground-truth target views enable reference-based evaluation of novel-view synthesis.

Baselines. We compare S4GS with three representative baselines using the same input views. AnySplat-LR+Bicubic reconstructs from 756×756 inputs and upsamples the rendered views to the target resolution with bicubic interpolation. AnySplat-LR+2DSR also reconstructs from 756×756 inputs, but applies EDSR (Lim et al., 2017) to the rendered views. AnySplat-HR directly processes the native 3024×3024 inputs whenever memory permits, providing a native-resolution reference for the memory comparison.

Metrics. For novel-view synthesis on StereoGenBench, we report PSNR, SSIM, and LPIPS against target-view ground truth over valid image regions. On the mobile dataset, novel-view quality is evaluated without ground truth using BRISQUE, MUSIQ, MANIQA, and NIQE. Geometric consistency is assessed using mean absolute relative depth error after cross-view reprojection (Depth AbsRel; lower is better). Efficiency is measured by peak allocated GPU memory and inference time per stereo pair; *OOM* denotes insufficient memory. For no-reference image quality, BRISQUE (Mittal et al., 2012) and NIQE (Mittal et al., 2013) measure deviations from natural-scene statistics (lower is better), while MUSIQ (Ke et al., 2021) and MANIQA (Yang et al., 2022) predict perceptual quality (higher is better).

Implementation details. All experiments use a single 24 GB RTX 4090. Starting from the public AnySplat checkpoint, we freeze VGGT-1B and train only the Gaussian head, RGB feature extractor, and SubGaussian CNN. We use AdamW with a peak learning rate of 10^{-4}, batch size one, and 50 epochs, optimizing MSE plus LPIPS with weight 0.2. Training renders random 756×756 crops; evaluation uses full 3024×3024 frames. Mixed precision and gradient checkpointing reduce memory use.

4.2 End-to-End Reconstruction on a Single 24 GB GPU

Capability for 4K Novel View Synthesis. For the self-collected mobile dataset, we use the synchronized stereo pair as context views and synthesize novel views by changing the camera parameters to produce a three-times stereo-baseline displacement. Because these views have no ground-truth images, all methods are evaluated with the four no-reference metrics BRISQUE, MUSIQ, MANIQA, and NIQE under the same rendering protocol. For StereoGenBench, cameras 0 and 3 from the same frame serve as context views, while cameras 1 and 2 provide the target views; all methods follow this identical protocol and are evaluated against the target-view ground truth using PSNR, SSIM, and LPIPS.

Table 1: Novel-view synthesis on the self-collected mobile dataset under a 24 GB GPU budget.

Method	Peak Mem. (GB)	Inf. Time (s)	BRISQUE ↓	MUSIQ ↑	MANIQA ↑	NIQE ↓
AnySplat-LR+Bicubic	7.15	0.241	70.8432	18.9714	0.2116	10.0886
AnySplat-LR+2DSR*	7.32	0.863	67.1629	19.9776	0.2017	9.4849
AnySplat-HR	OOM	-	-	-	-	-
Ours	16.36	0.812	**41.865**	**22.406**	**0.1948**	**5.383**

* EDSR (Lim et al., 2017) is used for 2D super-resolution.

Table 2: Novel-view synthesis on StereoGenBench under a 24 GB GPU budget.

Method	Peak Mem. (GB)	Inf. Time (s)	PSNR ↑	SSIM ↑	LPIPS ↓
AnySplat-LR+Bicubic	7.33	0.285	21.36	0.790	0.405
AnySplat-LR+2DSR*	7.56	0.907	21.33	0.789	0.405
AnySplat-HR	OOM	-	-	-	-
Ours	16.43	0.870	**22.51**	**0.836**	**0.281**

* EDSR (Lim et al., 2017) is used for 2D super-resolution.

As shown in Table 1, S4GS achieves a clear improvement in the no-reference quality measures on real mobile captures, indicating more natural and faithful synthesized views. It also completes 4K reconstruction within the 24 GB memory budget and retains sub-second inference, whereas native-resolution AnySplat runs out of memory and the 2D super-resolution baseline adds substantial cost without creating a high-resolution 3D Gaussian field. Table 2 shows the same trend on StereoGenBench: S4GS provides the strongest overall novel-view fidelity across the reference-based metrics while remaining within the memory limit. These results show that the proposed resolution-asymmetric design recovers high-resolution geometry and appearance without native-resolution global inference.

Table 3: Novel-view image quality assessment of S4GS under diverse real-world conditions.

Category	Group	# Samples	BRISQUE ↓	MUSIQ ↑	MANIQA ↑	NIQE ↓
Scene distance	Near	42	41.826	22.123	0.1926	5.680
	Middle	113	41.408	22.653	0.1945	5.276
	Far	49	42.954	22.084	0.1973	5.369
Scene type	Indoor	61	43.890	21.929	0.2027	5.736
	Outdoor	143	41.027	22.603	0.1915	5.237
Lighting	Day	136	38.953	22.830	0.1957	5.134
	Night	68	47.733	21.552	0.1929	5.885

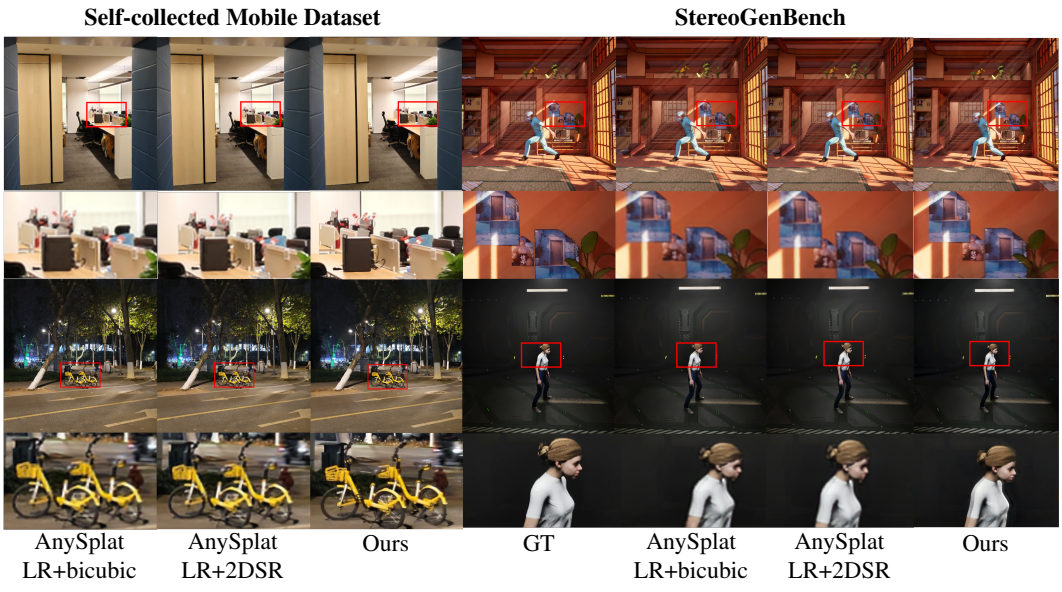

Figure 2: Qualitative comparison of synthesized novel views.

Figure 2 provides a visual comparison of the synthesized views. Compared with the baselines, S4GS produces sharper edges, richer complex textures, and more faithful details in hair and thin structures, demonstrating that the improvement comes from high-resolution 3D reconstruction rather than simple image-space sharpening.

Visual Quality Across Real-World Conditions. Table 3 further groups the self-collected test scenes by subject distance, scene type, and lighting condition. The consistently strong no-reference scores across these groups show that S4GS maintains stable high-fidelity novel-view synthesis in indoor and outdoor environments, across different distances and illumination conditions, supporting its robustness and practical value in real-world mobile capture scenarios.

4.3 Ablation and Analysis

We combine memory analysis and ablation experiments to examine the main design choices in S4GS. We first analyze how input resolution affects the memory requirements of the AnySplat baseline, then study dense geometry anchoring and high-resolution RGB refinement.

Backbone Memory Scaling with Input Resolution. We profile AnySplat on one stereo scene at four square input resolutions: 448, 756, 1512, and 3024 pixels per side. Table 4 reports the peak allocated memory of the global backbone, dense prediction heads, and full pipeline on a 24 GB GPU.

Table 4: AnySplat memory scaling on one stereo scene with two input views.

Input Resolution	VGGT Encoder	Depth Head	Gaussian Head	Full Pipeline
448×448	2.88	3.35	3.72	3.72
756×756	3.13	4.46	5.51	5.51
1512×1512	4.25	9.55	13.78	13.78
3024×3024	8.71	19.78*	–	OOM

* Peak observed before OOM in the depth head, not the memory required for a completed run. The Gaussian head was not reached.

Memory rises sharply with input resolution, and native-resolution inference fails before Gaussian prediction is completed. The trend identifies globally coupled reasoning as the bottleneck and motivates our asymmetric design: global reasoning remains at reduced resolution, while dedicated branches recover high-resolution geometry and appearance (Figure 3).

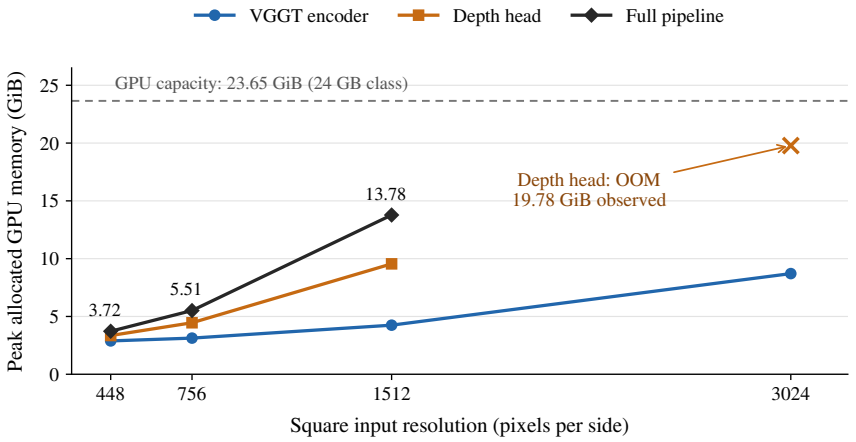

Figure 3: Peak allocated memory versus input resolution for AnySplat.

Effect of DA3-Based Geometry Anchoring. We examine whether dense depth from DA3 provides reliable spatial support for high-resolution Gaussians. The ablation removes explicit pre-anchoring and lets the residual network predict Gaussian centers, whereas the full model fixes centers to 3D points back-projected from aligned DA3 depth and refines the remaining attributes. Table 5 summarizes the resulting geometric consistency and rendering quality.

Table 5: Effect of DA3-based anchoring on geometric consistency and novel-view image quality.

Variant	Depth AbsRel ↓	BRISQUE ↓	MUSIQ ↑	MANIQA ↑	NIQE ↓
w/o DA3 Anchoring	0.094	75.0288	20.0409	0.1809	9.1915
Full Model (Ours)	**0.063**	**41.865**	**22.406**	**0.1948**	**5.383**

The lower Depth AbsRel indicates that explicit depth anchoring reduces cross-view reprojection error and improves geometric consistency. Explicit anchors also improve novel-view quality: without them, center errors accumulate as cross-view drift and ghosting, especially around sharp boundaries and thin structures. Fixing centers to aligned depth points preserves cleaner depth and RGB edges, as illustrated in Figure 4.

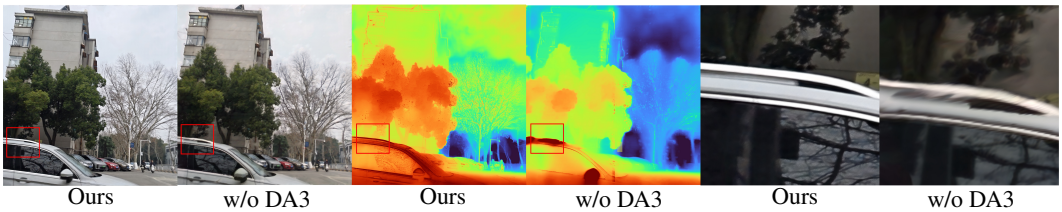

Figure 4: Effect of DA3 anchoring on rendered RGB, depth, and enlarged details (left to right).

Effect of High-Frequency RGB Refinement. After fixing the high-resolution Gaussian centers, we compare RGB-guided refinement with nearest-neighbor (block) upsampling of low-resolution Gaussian parameters. The ablation removes RGB guidance and residual refinement, whereas the full model extracts high-resolution RGB features and predicts residual Gaussian attributes. Table 6 reports the quantitative comparison.

Table 6: Effect of high-resolution RGB-guided refinement on novel-view image quality.

Variant	BRISQUE ↓	MUSIQ ↑	MANIQA ↑	NIQE ↓
w/o RGB Guidance	72.2523	18.0432	0.1622	8.5019
Full Model (Ours)	**41.865**	**22.406**	**0.1948**	**5.383**

Nearest-neighbor expansion adds no image evidence, leaving local shapes, colors, and boundaries blurred or distorted. High-resolution RGB features provide local cues for residual prediction while the coarse scaffold preserves global structure, yielding sharper boundaries and more faithful colors (Figure 5).



Figure 5: Effect of RGB-guided refinement on the full image and enlarged details.

5 Conclusion

We present S4GS, a resolution-asymmetric framework for 4K feed-forward Gaussian reconstruction from sparse views. It performs global reasoning at low resolution, then uses native-resolution observations for DA3-based geometric anchoring and RGB-guided residual refinement, producing a globally consistent high-resolution Gaussian field within a single 24 GB GPU. Experiments and ablations show that S4GS enables high-fidelity 4K immersive reconstruction with real-time inference under a limited GPU memory budget.

AI USE STATEMENT

We used generative AI tools to assist with writing and polish the language of this manuscript. We reviewed all AI-assisted text and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

To support reproducibility, the Implementation details paragraph in Section 4.1 describes the hardware, model initialization, trainable components, and training settings used in our experiments.

REFERENCES

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3D gaussian splats from image pairs for scalable generalizable 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19457–19467, 2024.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3D gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pp. 370–386, 2024.

Yangzhi Cui, Feng Qiao, and Nathan Jacobs. StereoGenBench: A synthetic multi-camera benchmark for stereo generation under controlled baseline regimes, 2026. URL `https://arxiv.org/abs/2605.23237`.

Xiang Feng, Yongbo He, Yubo Wang, Yan Yang, Wen Li, Yifei Chen, Zhenzhong Kuang, Jiajun Ding, Jianping Fan, and Yu Jun. Srgs: Super-resolution 3D gaussian splatting. *arXiv preprint arXiv:2404.10318*, 2024.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, Dahua Lin, and Bo Dai. Anysplat: Feed-forward 3D gaussian splatting from unconstrained views. *ACM Transactions on Graphics*, 44(6):1–16, 2025. doi: 10.1145/3763326.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021. doi: 10.1109/ICCV48922.2021.00510.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. doi: 10.1145/3592433.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision*, pp. 71–91, 2024.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 136–144, 2017. doi: 10.1109/CVPRW.2017.151.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, 2026.

Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3D gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20654–20664, 2024.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. doi: 10.1109/TIP.2012.2214050.

Anish Mittal, Rajiv Soundararajan, and Alan Conrad Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726.

Jongmin Park, Minh-Quan Viet Bui, Juan Luis Gonzalez, Jaeho Moon, Jihyong Oh, and Munchurl Kim. Ecosplat: Efficiency-controllable feed-forward 3D gaussian splatting from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26010–26020, 2026.

Yuan Shen, Duygu Ceylan, Paul Guerrero, Zexiang Xu, Niloy J. Mitra, Shenlong Wang, and Anna Fruehstueck. Supergaussian: Repurposing video models for 3D super resolution. In *European Conference on Computer Vision*, pp. 215–233, 2024.

Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. *arXiv preprint arXiv:2408.13912*, 2024.

Yecong Wan, Mingwen Shao, Yuanshuo Cheng, and Wangmeng Zuo. S2gaussian: Sparse-view super-resolution 3D gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 711–721, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10510–10522, 2025b.

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*, pp. 20697–20709, 2024.

Shiyun Xie, Zhiru Wang, Yinghao Zhu, Xu Wang, Chengwei Pan, and Xiwang Dong. Supergs: Consistent and detailed 3D super-resolution scene reconstruction via gaussian splatting. *arXiv preprint arXiv:2505.18649*, 2025.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16453–16463, 2025.

Linning Xu, Yuanbo Xiangli, Sida Peng, Xingang Pan, Nanxuan Zhao, Christian Theobalt, Bo Dai, and Dahua Lin. Grid-guided neural radiance fields for large urban scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8296–8306, 2023.

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*, pp. 21924–21935, 2025.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1191–1200, 2022.

Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3D gaussian splats from sparse unposed images. In *International Conference on Learning Representations*, 2025.

Xiaoxue Zhang, Xiaoxu Zheng, Yixuan Yin, Tiao Zhao, Kaihua Tang, Michael Bi Mi, Zhan Xu, and Dave Zhenyu Chen. Anchorsplat: Feed-forward 3D gaussian splatting with 3D geometric priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18924–18933, 2026.