SPECGS: FREQUENCY–SPATIAL SUPPORT COUPLED GAUSSIAN SPLATTING FOR MEMORY-EFFICIENT LARGE-SCALE SCENE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

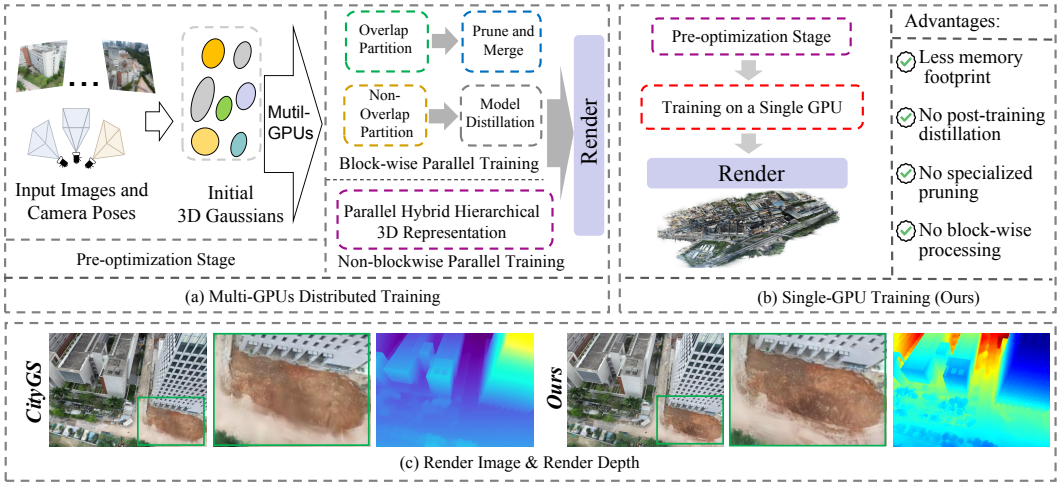


Figure 1: (a) Existing city-scale 3D modeling methods rely on multi-GPU parallelization, including overlapping partitioning (causing merging artifacts), distributed training with distillation (introducing extra cost), and parallel hierarchical representations with batch rendering (still multi-GPU dependent). All are limited by single-GPU memory. (b) In contrast, our method follows the standard 3DGS pipeline without additional components and runs on a single consumer-grade GPU for large-scale urban reconstruction. (c) Compared with CityGS (Liu et al., 2024a), our method produces sharper high-frequency details in both rendered images and depth maps.

ABSTRACT

Despite the success of 3D Gaussian Splatting (3DGS) in real-time rendering, reconstructing large-scale urban scenes remains challenging. Due to the inherent low-pass nature of Gaussian primitives, capturing high-frequency details (e.g., sharp edges and intricate textures) typically requires excessive primitive densification. Consequently, many existing methods rely on multi-GPU training or scene partitioning to scale reconstruction. We propose SpecGS, a 3DGS framework that improves scalability through frequency-guided Gaussian adaptation. SpecGS estimates Gaussian-level high-frequency evidence from image-space frequency responses and couples it with spatial-support regulation and densification prioritization, enabling representation capacity to be preferentially allocated to regions requiring fine-detail reconstruction. Moreover, to ensure stable optimization at scale, we introduce a three-stage optimization strategy with an Adaptive Convergence Trigger (ACT) that automatically schedules coarse-to-fine resolution progression based on optimization saturation. At the final stage, Difference-of-Gaussian (DoG) supervision further emphasizes high-frequency reconstruction while providing the frequency evidence used for Gaussian adaptation. Experiments on multiple urban datasets demonstrate that SpecGS achieves a favorable quality–efficiency trade-off with significantly fewer primitives and reduced memory consumption, enabling practical city-scale reconstruction on a single GPU.

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.

1 INTRODUCTION

Recent advances in neural rendering have been revolutionized by 3D Gaussian Splatting (3DGS) Kerbl et al. (2023), which employs anisotropic Gaussian primitives together with an efficient rasterization pipeline to achieve state-of-the-art novel view synthesis with fast training. Despite these successes, extending 3DGS to large-scale urban environments remains challenging. A fundamental limitation is that a standard Gaussian primitive exhibits a low-pass spectral response, whose smooth spatial support attenuates high-frequency appearance details (Hamdi et al., 2024). However, urban scenes contain abundant high-frequency structures, including sharp geometric edges, intricate textures, and repetitive structural patterns Lin et al. (2024). To capture these details, existing methods typically resort to brute-force strategies, such as increasing the density of primitives or using higher-order spherical harmonics (SH) (Liu et al., 2024b; Barron et al., 2022a; Huang et al., 2024). In practice, this often leads to tens of millions of Gaussians, resulting in excessive GPU memory consumption and unstable optimization. Recent works attempt to alleviate these issues through system-level solutions, such as spatial partitioning, multi-resolution tiling (Li et al., 2025), or multi-GPU parallelization (Gao et al., 2025; Liu et al., 2024a). While these approaches mitigate memory bottlenecks to some extent, they mainly improve scalability through partitioned or distributed computation, without explicitly adapting Gaussian capacity to the spatially varying frequency demand of large-scale scenes.

To address this limitation, we propose SpecGS, a 3D Gaussian Splatting framework that improves scalability by allocating representation capacity according to local high-frequency demand rather than uniformly increasing primitive density. Unlike existing city-scale approaches that rely on scene partitioning or multi-GPU parallelization (See Fig. 1), SpecGS retains the standard Gaussian splatting rasterization pipeline while introducing frequency-guided spatial adaptation, enabling high-fidelity urban reconstruction on a single consumer-grade GPU. Specifically, SpecGS estimates a Gaussian-level frequency score from image-space high-frequency evidence and couples it with a bounded spatial-support control variable. Rather than treating frequency as an optimizable primitive parameter, the estimated frequency score serves as a conditioning signal that regulates spatial support and prioritizes densification. Gaussians associated with stronger high-frequency evidence therefore receive stronger suppression against excessive support expansion and higher priority for additional representation capacity, while smoother regions avoid unnecessary densification. This design enables frequency-aware allocation of Gaussian capacity without modifying the standard rasterization process, substantially reducing redundant primitives in large-scale reconstruction (See Tab. 2).

Moreover, to ensure stable optimization under frequency-guided spatial adaptation, we introduce a three-stage optimization strategy with an Adaptive Convergence Trigger (ACT) that governs a coarse-to-fine training curriculum. As illustrated in Fig. 2, the first two stages perform coarse and intermediate-resolution structural optimization, while the third stage is dedicated to full-resolution frequency-guided refinement. Lower-resolution training naturally emphasizes coarse-to-medium-frequency structures and provides a stable geometric scaffold before fine-scale optimization. ACT monitors optimization saturation and advances the rendering resolution only after the current stage has sufficiently converged, effectively using resolution progression as an implicit low-pass constraint. In the third stage, SpecGS aggregates Difference-of-Gaussian (DoG) Lowe (2004) responses into Gaussian-level frequency evidence, which guides spatial-support regularization and densification priority while also providing frequency-aware reconstruction supervision. In this way, ACT determines when fine-scale refinement should begin, whereas frequency evidence determines where additional representation capacity should be allocated.

Our contributions are summarized as follows:

- We propose **SpecGS**, a frequency-guided Gaussian splatting framework that estimates Gaussian-level high-frequency evidence and uses it to jointly regulate spatial support and densification priority, enabling more efficient allocation of representation capacity in large-scale scenes.
- We introduce a three-stage optimization strategy with an **Adaptive Convergence Trigger (ACT)**, a data-driven coarse-to-fine training strategy that schedules resolution progression according to optimization saturation, establishing stable geometry before full-resolution frequency-guided refinement.

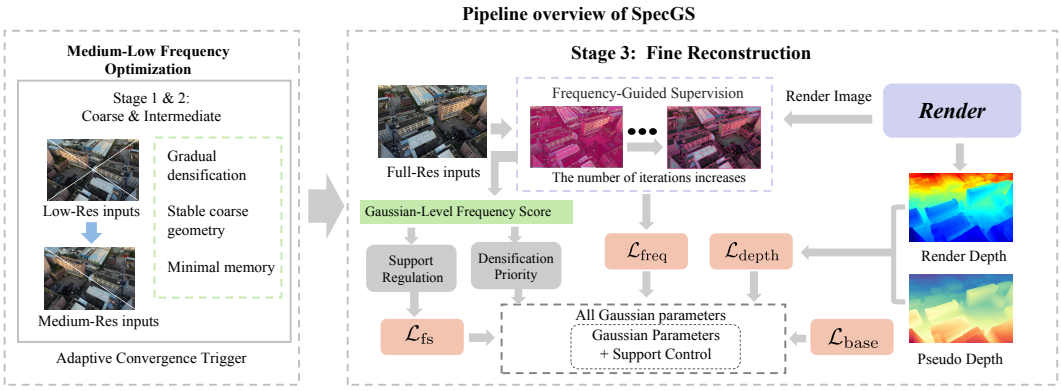


Figure 2: Overview of SpecGS. Stages I and II progressively establish a stable geometric scaffold under ACT. At Stage III, DoG-based high-frequency evidence is aggregated into a Gaussian-level frequency score f_k, which conditions spatial-support regularization and densification priority while also supporting frequency-aware reconstruction supervision. ACT determines when to refine, whereas f_k guides where representation capacity is preferentially allocated.

- We demonstrate on large-scale urban datasets that SpecGS achieves strong rendering quality with significantly fewer primitives and reduced memory consumption, enabling practical city-scale reconstruction on a single GPU.

2 Related Works

Novel View Synthesis. Novel view synthesis generates photorealistic images from unseen viewpoints by learning a coherent 3D scene representation from input images (Zhu et al., 2025; Barron et al., 2022b; Chen et al., 2024; Li et al., 2026; Mallick et al., 2024). Early volumetric methods such as Soft3D (Penner & Zhang, 2017) were computationally expensive. The advent of Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) transformed the field by proposing a continuous volumetric scene representation trained through differentiable rendering. Its follow-up works (Barron et al., 2022b; Müller et al., 2022; Fridovich-Keil et al., 2022) further enhanced both reconstruction quality and computational efficiency. Recently, a significant breakthrough was introduced by 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023), which achieves real-time rendering through the use of explicit 3D Gaussians. Recent studies have advanced 3DGS primarily along two main directions: rendering quality—e.g., alias-free multi-scale rendering with Mip-splatting (Yu et al., 2024) and memory-efficient hierarchical reconstruction with HRGS (Li et al., 2025); and computational efficiency—e.g., kernel-level acceleration and reduced memory usage in FlashGS (Feng et al., 2025). Together, these developments extend 3DGS to large-scale, high-resolution settings.

Large-Scale Scene Reconstruction. Reconstructing large-scale scenes presents challenges in efficiency, memory, and geometric consistency. Early NeRF-based methods such as Block-NeRF (Tancik et al., 2022) and Mega-NeRF (Turki et al., 2022a) partition scenes for parallel training, while 3DGS-based approaches further improve efficiency through explicit primitives (Kerbl et al., 2023). Recent methods mainly rely on system-level strategies, including block-wise training (Lin et al., 2024), divide-and-conquer with LoD rendering (Liu et al., 2024a), post-processing (Fan et al., 2024), distributed optimization (Chen & Lee, 2024), and parallel hybrid hierarchies (Gao et al., 2025). However, CityGS-v2 (Liu et al., 2024b) reduces computational overhead at the cost of reconstruction quality, while MixGS (Liu et al., 2025) supports single-GPU training but still requires more than 10 hours of optimization on large-scale scenes. Unlike these approaches, SpecGS addresses scalability through frequency-guided Gaussian adaptation, using local high-frequency evidence to regulate spatial support and prioritize densification. This enables single-GPU, partition-free reconstruction on mainstream urban datasets (Lin et al., 2022; Turki et al., 2022b; Xiong et al., 2024) without relying on block-wise decomposition or excessive primitive densification.

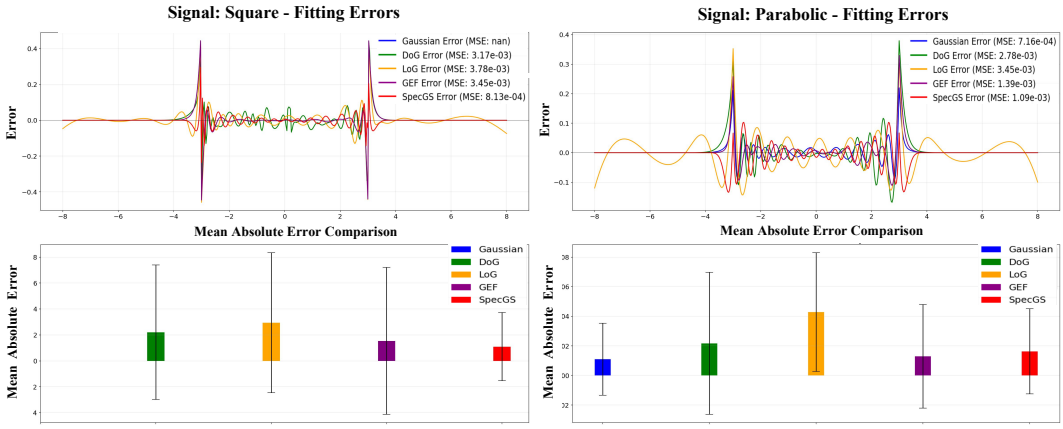


Figure 3: Numerical simulation results of different mixture-based representations are presented, demonstrating fitting performance on both high-frequency (Square) and low-frequency (Parabolic) signals via Mean Absolute Error (MAE) comparisons; NaN indicates fitting failure. Additional signal comparisons are provided in the supplementary material (See Fig. 14).

Table 1: Quantitative comparisons on the Mill19 (Turki et al., 2022b) and UrbanScene3D (Lin et al., 2022) datasets for novel view synthesis. ↑ indicates higher is better, while ↓ indicates lower is better. The top three results are highlighted with red, orange, and yellow backgrounds, respectively. † denotes results obtained without the decoupled appearance encoding.

	Building			Rubble			Residence			Sci-Art		
	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓
VastGaussian †	0.728	21.80	0.225	0.742	25.20	0.264	0.699	21.01	0.261	0.761	22.64	0.261
3DGS	0.738	22.53	0.214	0.725	25.51	0.316	0.745	22.36	0.247	0.791	24.13	0.262
DoGaussian	0.759	22.73	0.204	0.765	25.78	0.257	0.740	21.94	0.244	0.804	24.42	0.219
CityGaussian	0.778	21.55	0.246	0.813	25.77	0.228	0.813	22.00	0.211	0.837	21.39	0.230
CityGS-X	0.781	22.01	0.251	0.814	25.69	0.236	0.789	21.98	0.205	0.842	22.21	0.215
PGSR	0.480	16.12	0.573	0.728	23.09	0.334	0.746	20.57	0.289	0.799	19.72	0.275
CityGaussianV2	0.650	19.07	0.397	0.720	23.75	0.322	0.769	21.15	0.234	0.810	20.66	0.266
Ours	0.804	22.84	0.210	0.801	25.85	0.210	0.825	22.51	0.205	0.815	22.62	0.208

3 METHODS

SpecGS follows a progressive three-stage optimization pipeline (Fig. 2) for memory-efficient large-scale scene reconstruction. Stages I and II optimize standard 3D Gaussians at progressively higher resolutions to establish a stable geometric scaffold, while an Adaptive Convergence Trigger (ACT) determines when the optimization should advance. At Stage III, SpecGS activates frequency-guided refinement. Specifically, image-space high-frequency evidence is aggregated onto visible Gaussian primitives to obtain a Gaussian-level frequency score f_k. This score conditions spatial-support regularization and densification priority, allowing representation capacity to be preferentially allocated to regions requiring fine-detail reconstruction.

3.1 PRELIMINARY

We briefly review 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023). A scene is represented by a set of Gaussian primitives $\mathcal{G} = \{G_k\}_{k=1}^{K}$, where K denotes the number of Gaussians. Each primitive G_k is parameterized by a 3D center $\mathbf{p}_k \in \mathbb{R}^3$, opacity $\alpha_k \in [0, 1]$, scaling vector $\boldsymbol{\sigma}_k \in \mathbb{R}^3$, rotation $\mathbf{R}_k \in \mathrm{SO}(3)$, and spherical-harmonic (SH) coefficients $\mathbf{c}_k^{\mathrm{SH}}$. The SH coefficients produce a view-dependent RGB color $\mathbf{c}_k$ under the current viewing direction. The covariance of G_k is

$$\Sigma_k = \mathbf{R}_k \mathbf{S}_k \mathbf{S}_k^\top \mathbf{R}_k^\top, \qquad \mathbf{S}_k = \mathrm{diag}(\boldsymbol{\sigma}_k). \tag{1}$$

During rendering, projected Gaussians are composited in front-to-back order. For pixel $\mathbf{u}$, let $a_k(\mathbf{u}) = \alpha_k \mathcal{G}_k^{2D}(\mathbf{u})$ denote the effective pixel-wise opacity contribution of Gaussian G_k, where

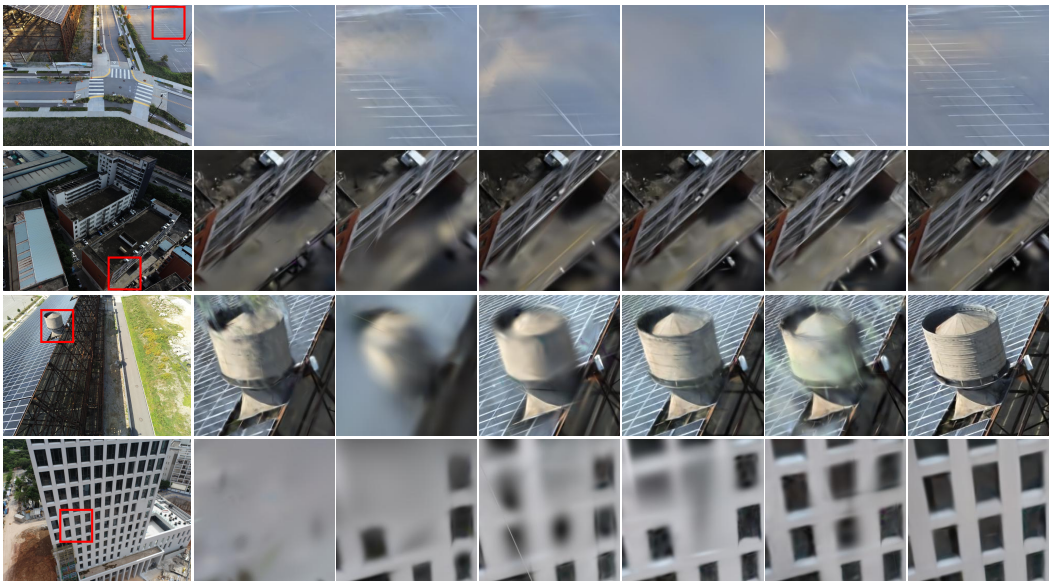


Figure 4: Qualitative comparison with existing methods on Mill-19 (Turki et al., 2022b) and UrbanScene3D (Lin et al., 2022).

$\mathcal{G}_k^{2D}(\mathbf{u})$ is its projected 2D Gaussian response. The rendered color is

$$\hat{\mathbf{C}}(\mathbf{u}) = \sum_{k \in \mathcal{D}(\mathbf{u})} \mathbf{c}_k \, a_k(\mathbf{u}) \prod_{j<k} (1 - a_j(\mathbf{u})) \,, \tag{2}$$

where $\mathcal{D}(\mathbf{u})$ denotes the depth-ordered set of Gaussians contributing to pixel $\mathbf{u}$.

3.2 Frequency–Spatial Support Adaptation

Gaussian-Level Frequency Evidence. For training iteration t and current view v, let u index image pixels. We first extract image-space high-frequency evidence from the ground-truth image $\mathbf{I}_v$ using a Difference-of-Gaussians (DoG) response:

$$\mathcal{M}_v(u) = \frac{|\mathcal{K}_{\kappa_1} * \mathbf{I}_v(u) - \mathcal{K}_{\kappa_2} * \mathbf{I}_v(u)|}{\max_{u'} |\mathcal{K}_{\kappa_1} * \mathbf{I}_v(u') - \mathcal{K}_{\kappa_2} * \mathbf{I}_v(u')| + \epsilon}, \tag{3}$$

where $\mathcal{K}_\kappa$ denotes a Gaussian filtering kernel with bandwidth κ, and $\epsilon > 0$ is a small constant for numerical stability. The resulting map $\mathcal{M}_v$ emphasizes edges and fine-scale textures. During Stage III, the DoG bandwidth follows a triangular schedule controlled by the normalized Stage-III progress t_{prog}: the bandwidth gradually increases during the first half of refinement and decreases during the second half, allowing the supervision to cover different detail scales during full-resolution refinement. Here, t_{prog} is independent of the Gaussian-level frequency score f_k.

For each visible Gaussian G_k, we aggregate this image-space evidence according to its rendering contribution:

$$\hat{f}_{k,v}^{(t)} = \frac{\sum_u \text{sg}\left(w_{k,u,v}^{(t)}\right) \mathcal{M}_v(u)}{\sum_u \text{sg}\left(w_{k,u,v}^{(t)}\right) + \epsilon}, \tag{4}$$

where $w_{k,u,v}^{(t)} \geq 0$ denotes the alpha-compositing contribution of Gaussian G_k to pixel u in view v, and $\text{sg}(\cdot)$ denotes the stop-gradient operator. The view-conditioned estimate is accumulated across training observations using an exponential moving average (EMA):

$$f_k^{(t)} = (1 - \rho_f) f_k^{(t-1)} + \rho_f \hat{f}_{k,v}^{(t)}, \tag{5}$$

where $\rho_f \in (0, 1]$ is the EMA update rate.

Importantly, f_k is a dynamically maintained training statistic rather than an optimizable model parameter. It summarizes the high-frequency evidence associated with each Gaussian and serves as a conditioning signal for spatial-support regulation and primitive allocation.

Frequency–Spatial Support Coupling. To explicitly regulate the effective spatial extent of each Gaussian, we introduce a bounded support-control variable s_k:

$$\tilde{\boldsymbol{\sigma}}_k = s_k \boldsymbol{\sigma}_k, \qquad s_k = 2\ \mathrm{sigmoid}(\gamma \beta_k), \tag{6}$$

where β_k is an optimizable latent variable and $\gamma > 0$ controls the activation slope. The corresponding effective covariance is

$$\tilde{\Sigma}_k = \mathbf{R}_k \tilde{\mathbf{S}}_k \tilde{\mathbf{S}}_k^\top \mathbf{R}_k^\top, \qquad \tilde{\mathbf{S}}_k = \mathrm{diag}(\tilde{\boldsymbol{\sigma}}_k). \tag{7}$$

Thus, s_k provides an explicit bounded variable for controlling spatial support while preserving the anisotropy and orientation of the original Gaussian.

We couple the frequency evidence with spatial-support control through

$$\mathcal{L}_{\mathrm{fs}} = \frac{1}{|\mathcal{V}_v|} \sum_{k \in \mathcal{V}_v} f_k\ \max(0, s_k - 1), \tag{8}$$

where $\mathcal{V}_v$ denotes the set of Gaussians visible in the current training view. For Gaussians with $s_k > 1$, higher frequency evidence therefore imposes a stronger penalty against excessive support expansion. Note that f_k itself is not differentiated; instead, it modulates the regularization applied to the optimizable support variable s_k.

Frequency-Guided Densification. The same Gaussian-level frequency score is also used to prioritize primitive allocation during densification. Let g_k denote the standard 3DGS densification signal and τ_g its base threshold. We define a Gaussian-specific effective threshold

$$\tau_k = \frac{\tau_g}{1 + \eta f_k}, \tag{9}$$

where $\eta \geq 0$ controls the strength of frequency guidance. A Gaussian is selected for densification when

$$g_k > \tau_k. \tag{10}$$

Since τ_k decreases monotonically with f_k, Gaussians associated with stronger high-frequency evidence are preferentially allocated additional representation capacity. The subsequent densification operation follows the standard 3DGS procedure; f_k affects the allocation priority rather than the geometric displacement of newly generated primitives.

3.3 Stage-wise Optimization with Adaptive Scheduling

Adaptive Convergence Trigger. Fixed iteration-based schedules may either increase the rendering resolution prematurely or waste computation after the current stage has saturated. ACT instead monitors two consecutive windows of training losses, $\mathcal{W}_{\mathrm{prior}}$ and $\mathcal{W}_{\mathrm{recent}}$. Their relative improvement is

$$\delta_t = \frac{\mu(\mathcal{W}_{\mathrm{prior}}) - \mu(\mathcal{W}_{\mathrm{recent}})}{\mu(\mathcal{W}_{\mathrm{prior}}) + \epsilon}, \tag{11}$$

where $\mu(\cdot)$ denotes the mean loss within a window. When $0 \leq \delta_t < \tau_{\mathrm{act}}$, with τ_{act} denoting the convergence threshold, ACT regards the current stage as saturated and increases the rendering resolution. This produces a data-driven coarse-to-fine curriculum in which fine-scale frequency-guided refinement is activated only after a stable geometric foundation has been established.

Progressive Objective Formulation. Stages I and II optimize geometric structure at $0.5\times$ and $0.75\times$ of the full rendering resolution, respectively. Both stages use the standard 3DGS photometric objective $\mathcal{L}_{\mathrm{base}}$ together with monocular depth supervision $\mathcal{L}_{\mathrm{depth}}$:

$$\mathcal{L}_{\mathrm{I,II}} = \mathcal{L}_{\mathrm{base}} + \mathcal{L}_{\mathrm{depth}}. \tag{12}$$

Table 2: Comparison of Large-Scale Scene Modeling Methods. The best result for each metric is highlighted in **bold**. * indicates multi-GPU training. **GS(M)**: Gaussians (millions); **Size(G)**: model size (GB); **Mem(G)**: GPU memory (GB).

Method	Building			Rubble			Residence			Sci-Art		
	GS(M)	Size(G)	Mem(G)	GS(M)	Size(G)	Mem(G)	GS(M)	Size(G)	Mem(G)	GS(M)	Size(G)	Mem(G)
CityGaussian*	9.65	3.18	>24	9.91	2.29	>24	9.66	2.23	>24	3.79	0.89	>24
3DGS	12.94	2.99	>24	8.58	2.08	>24	8.68	2.01	>24	3.55	0.84	>24
CityGS-X*	9.97	3.02	14.35	10.21	2.93	14.12	11.23	3.28	15.23	4.56	1.23	12.46
Ours	**6.39**	**1.28**	**12.09**	**5.89**	**1.27**	**11.45**	**5.26**	**1.29**	**9.31**	**1.98**	**0.45**	**10.82**



Figure 5: Rendering results across the three ACT-controlled optimization stages.

After ACT advances the optimization to Stage III, training proceeds at full resolution and activates Gaussian-level frequency estimation, frequency-conditioned support regularization, and frequency-guided densification. The same DoG evidence map is additionally used to emphasize high-frequency reconstruction residuals:

$$\mathcal{L}_{\text{freq}} = \left\| \mathcal{M}_v \odot (\mathbf{I}_{\text{ren},v} - \mathbf{I}_v) \right\|_1, \tag{13}$$

where $\mathbf{I}_{\text{ren},v}$ denotes the image rendered from view v. The Stage-III objective is

$$\mathcal{L}_{\text{III}} = \mathcal{L}_{\text{base}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{fs}} + \lambda_{\text{dog}} \mathcal{L}_{\text{freq}}, \tag{14}$$

where λ_{dog} controls the contribution of the frequency-guided reconstruction term.

Overall, ACT determines *when* the optimization should progress toward fine-scale reconstruction, while the Gaussian-level frequency score f_k determines *where* representation capacity should be preferentially allocated.

4 Experiments

4.1 Experimental Setup

Our experiments cover four representative scenes drawn from two datasets: Building and Rubble from Mill-19 (Turki et al., 2022b), and Residence and Sci-Art from UrbanScene3D (Lin et al., 2022). Unless otherwise specified, all experiments are conducted on a single NVIDIA A5000 GPU. Additional comparisons on other datasets and further implementation details are provided in Sec. B and Sec. A.

Table 4: Core design module ablation study on the Sci-Art dataset. Each variant removes one core design module of SpecGS.

Variant	PSNR↑	Mem(G)↓
w/o Spatial Support	21.85	>24
w/o Freq-Guided Densification	21.94	20.33
w/o Freq-Spatial Coupling	21.90	11.48
w/o Freq-Guided Supervision	22.15	10.92
Ours	**22.62**	**10.82**

Following the original 3DGS preprocessing protocol, images wider than 1600 pixels are automatically resized to a width of 1600 pixels when the resolution parameter is set to -1; we refer to this setting as *Full Resolution*. ACT starts from further downsampled versions of this resolution and progressively increases the rendering scale, providing a consistent coarse-to-fine optimization schedule. We further evaluate SpecGS on additional large- and small-scale scenes in the supplementary material.

4.2 Toy Experiments

We conduct a controlled 1D signal-fitting study using several mixture-based representations, including Gaussian Mixture Models (GMM), Difference of Gaussians (DoG), Generalized Exponential

Table 3: Comparison with recent large-scale Gaussian Splatting methods. The upper block reports the average performance and training time over Building, Rubble, Residence, and Sci-Art under each method's reported single-GPU setting. The lower block provides a controlled comparison on Rubble using the same CityGS pipeline and 6×A5000 GPUs.

Method	Hardware	Setting	PSNR↑	Training Time↓
MixGS	1× RTX 3090	Avg. 4 scenes	24.32	14 h 41 min
SpecGS	1× A5000	Avg. 4 scenes	23.46	**45 min**
SSAGS+CityGS	6× A5000	Rubble	27.33	84 min
SpecGS+CityGS	6× A5000	Rubble	**27.43**	**21 min**

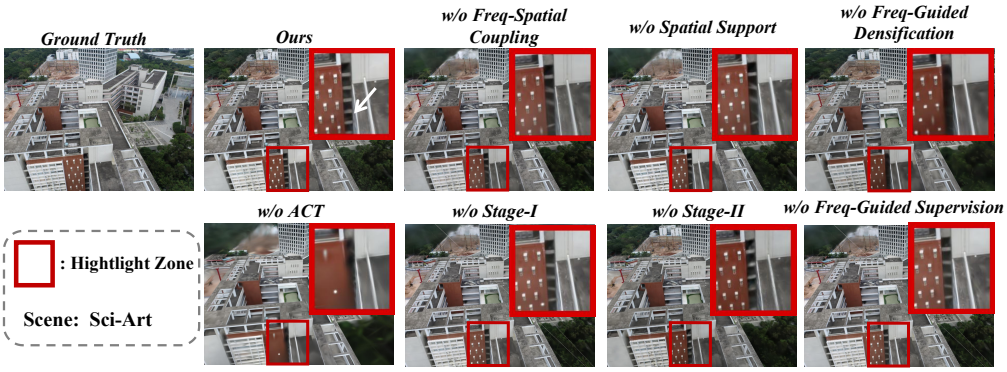


Figure 6: Visual comparison of ablation variants on Sci-Art (Lin et al., 2022). The full model better preserves both structural integrity and fine-scale texture.

Family (GEF), Laplacian of Gaussian (LoG), and the SpecGS-inspired formulation. The target signals span smooth low-frequency profiles and increasingly high-frequency structures, allowing us to examine how localized basis adaptation affects fitting efficiency under a limited component budget. Performance is measured by the fitting error between the reconstructed and ground-truth signals. As shown in Fig. 3, adaptive localized representations are particularly beneficial for signals containing sharp transitions and high-frequency variations. This controlled study serves as empirical motivation for allocating representation capacity according to local frequency demand, rather than as a reduced implementation of the full 3D SpecGS pipeline. Additional results are provided in Supplementary Sec. F.

4.3 Main Experiments

Novel View Synthesis. As shown in Tab. 1 and Fig. 4, SpecGS achieves competitive or superior rendering quality across the four benchmark scenes compared with representative 3DGS-based city-scale methods. In particular, SpecGS achieves the best PSNR on Building and Rubble and competitive perceptual quality across the remaining scenes. Qualitative comparisons further show that SpecGS preserves fine structural details, including building edges, window frames, and surface textures, while avoiding excessive smoothing. The progressive improvements shown in Fig. 5 also illustrate how coarse geometry is first established before full-resolution fine-detail refinement.

Memory and Computational Efficiency. As reported in Tab. 2, SpecGS performs large-scale reconstruction on a single GPU, whereas several city-scale baselines either require multi-GPU training or exceed 24 GB of GPU memory. Across the evaluated scenes, SpecGS uses substantially fewer Gaussian primitives and smaller model sizes without introducing explicit pruning or post-training distilla-

Table 5: Training strategy ablation study on the Sci-Art dataset (Lin et al., 2022). Each variant removes one training component of SpecGS.

Variant	PSNR↑	SSIM↑	LPIPS↓
w/o Stage-I	21.80	0.791	0.240
w/o Stage-II	22.00	0.804	0.225
w/o ACT	20.20	0.782	0.220
Ours	**22.62**	**0.815**	**0.208**

tion. These results support the effectiveness of frequency-guided support regulation and densification prioritization in reducing unnecessary primitive allocation while maintaining reconstruction quality. A more detailed analysis is provided in Sec. G.

Table 6: Sensitivity to the transition point of the Stage-III DoG bandwidth schedule.

t_{prog} (%)	Sci-Art		Building	
	PSNR↑	GS(M)↓	PSNR↑	GS(M)↓
45	22.61	1.98	22.83	6.39
50	**22.62**	1.98	22.84	6.39
55	22.59	1.98	**22.86**	6.39

Extension to Multi-GPU Training. SpecGS is designed as a standalone single-GPU framework, completing training in 45 minutes on a single NVIDIA A5000. For comparison, the recent single-GPU method MixGS (Liu et al., 2025) reports an average training time of 14 h 41 min on a single RTX 3090 across the four urban scenes, while achieving higher average PSNR. Beyond its default single-GPU setting, SpecGS can also be integrated into multi-GPU reconstruction pipelines. Specifically, we combine SpecGS with the CityGS (Liu et al., 2024a) partitioning framework and compare it with SSAGS (Xue et al., 2026), which is likewise built upon the multi-GPU CityGS pipeline. Under the same CityGS pipeline and $6\times$A5000 configuration on Rubble, SpecGS+CityGS improves PSNR from 27.33 to 27.43 while reducing training time from 84 to 21 minutes. These results show that SpecGS retains its efficiency advantage when extended from standalone single-GPU training to multi-GPU large-scale reconstruction.

4.4 Ablation Studies

We conduct ablation studies on the Sci-Art dataset (Lin et al., 2022) using a single NVIDIA A5000 GPU. Visual comparisons are provided in Fig. 6. We evaluate two aspects of SpecGS: the frequency-guided adaptation modules (Tab. 4) and the progressive optimization strategy (Tab. 5).

Core Design Modules. We ablate four components. (i) *w/o Spatial Support* removes the bounded spatial-support parameterization, resulting in memory overflow and indicating its importance for controlling Gaussian spatial extent. (ii) *w/o Freq-Guided Densification* removes the frequency-conditioned densification priority and restores the standard densification criterion, leading to substantially higher memory consumption and lower reconstruction quality. (iii) *w/o Freq-Spatial Coupling* removes the frequency-conditioned support regularization $\mathcal{L}_{\text{fs}}$, degrading both quality and memory efficiency. (iv) *w/o Freq-Guided Supervision* removes the DoG-weighted high-frequency reconstruction term, reducing the recovery of fine-scale details. Together, these results show that Gaussian-level frequency evidence is useful both for regulating spatial support and for directing additional capacity toward detail-rich regions.

DoG Schedule Sensitivity. We further examine the transition point of the triangular DoG bandwidth schedule within Stage III. As shown in Tab. 6, varying t_{prog} from 45% to 55% results in only minor PSNR changes while leaving the Gaussian count unchanged on both scenes. This indicates that SpecGS is insensitive to the exact transition point, and we therefore use 50% as a fixed default across all experiments.

Training Strategies. As shown in Tab. 5, removing either Stage I or Stage II degrades reconstruction quality, indicating that coarse- and intermediate-resolution optimization provides an important geometric foundation before full-resolution refinement. Removing ACT leads to the largest performance degradation, further demonstrating the importance of adapting resolution progression to optimization saturation.

5 Conclusion

To address the scalability challenges of city-scale 3D Gaussian Splatting, we propose SpecGS, a frequency-guided framework that allocates Gaussian representation capacity according to local high-frequency demand. SpecGS aggregates image-space frequency evidence into Gaussian-level scores, which condition spatial-support regularization and densification priority, thereby reducing unnecessary primitive allocation while preserving fine-scale details. Together with the ACT-controlled coarse-to-fine optimization strategy, SpecGS enables memory-efficient large-scale reconstruction on a single GPU. Experiments across multiple benchmarks demonstrate a favorable quality–efficiency trade-off with substantially fewer primitives and reduced memory consumption.

AI USE STATEMENT

Generative AI tools were used only for language polishing, grammar correction, and improving the clarity of the manuscript. They were not used to generate experimental results, perform data analysis, or make scientific claims. All technical content, experiments, and conclusions were produced and verified by the authors.

ETHICS STATEMENT

This work focuses on efficient 3D scene reconstruction and does not involve human subjects, personal data, or sensitive attributes. All experiments are conducted on publicly available research datasets following their intended academic use. We are not aware of any direct ethical concerns specific to the proposed method.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, optimization objectives, training schedule, hyperparameters, datasets, and evaluation protocols in the main paper and supplementary material. All experiments use consistent settings unless otherwise specified. We plan to release the implementation and configuration files to facilitate reproduction of the reported results.

REFERENCES

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, 2022a.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022b.

Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. PGSR: planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. *CoRR*, abs/2406.06521, 2024.

Yu Chen and Gim Hee Lee. DOGS: distributed-oriented gaussian splatting for large-scale 3d reconstruction via gaussian consensus. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejia Xu, and Zhangyang Wang. Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ FPS. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Guofeng Feng, Siyan Chen, Rong Fu, Zimu Liao, Yi Wang, Tao Liu, Boni Hu, Linning Xu, Zhilin Pei, Hengjie Li, Xiuhong Li, Ninghui Sun, Xingcheng Zhang, and Bo Dai. Flashgs: Efficient 3d gaussian splatting for large-scale and high-resolution rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 2025.

Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. 2022.

Yuanyuan Gao, Hao Li, Jiaqi Chen, Zhengyu Zou, Zhihang Zhong, Dingwen Zhang, Xiao Sun, and Junwei Han. Citygs-x: A scalable architecture for efficient and geometrically accurate large-scale scene reconstruction. *CoRR*, abs/2503.23044, 2025.

Abdullah Hamdi, Luke Melas-Kyriazi, Jinjie Mai, Guocheng Qian, Ruoshi Liu, Carl Vondrick, Bernard Ghanem, and Andrea Vedaldi. GES: generalized exponential splatting for efficient radiance field rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 2024.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH*, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 2023.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: benchmarking large-scale scene reconstruction. *ACM Trans. Graph.*, 36(4):78:1–78:13, 2017.

Changbai Li, Haodong Zhu, Hanlin Chen, Juan Zhang, Tongfei Chen, Shuo Yang, Shuwei Shao, Wenhao Dong, and Baochang Zhang. HRGS: hierarchical gaussian splatting for memory-efficient high-resolution 3d reconstruction. *CoRR*, abs/2506.14229, 2025.

Changbai Li, Haodong Zhu, Hanlin Chen, Xiuping Liang, Tongfei Chen, Shuwei Shao, Linlin Yang, Huobin Tan, and Baochang Zhang. Urbangs: A scalable and efficient architecture for geometrically accurate large-scene reconstruction. *CoRR*, abs/2602.02089, 2026.

Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, and Wenming Yang. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *CVPR*, 2024.

Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. Capturing, reconstructing, and simulating: The urbanscene3d dataset. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13668 of *Lecture Notes in Computer Science*, pp. 93–109. Springer, 2022.

Chuandong Liu, Huijiao Wang, Lei Yu, and Gui-Song Xia. Holistic large-scale scene reconstruction via mixed gaussian splatting. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

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 Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XVI*, volume 15074 of *Lecture Notes in Computer Science*, pp. 265–282. Springer, 2024a.

Yang Liu, Chuanchen Luo, Zhongkai Mao, Junran Peng, and Zhaoxiang Zhang. Citygaussianv2: Efficient and geometrically accurate reconstruction for large-scale scenes. *CoRR*, abs/2411.00771, 2024b.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60, 2004.

Saswat Subhajyoti Mallick, Rahul Goel, Bernhard Kerbl, Markus Steinberger, Francisco Vicente Carrasco, and Fernando De la Torre. Taming 3dgs: High-quality radiance fields with limited resources. In Takeo Igarashi, Ariel Shamir, and Hao (Richard) Zhang (eds.), *SIGGRAPH Asia 2024 Conference Papers, SA 2024, Tokyo, Japan, December 3-6, 2024*, pp. 2:1–2:11. ACM, 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*, 2021.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022.

Eric Penner and Li Zhang. Soft 3d reconstruction for view synthesis. *ACM Transactions on Graphics*, 2017.

Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben P. Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 8238–8248. IEEE, 2022.

Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly- throughs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 12912–12921. IEEE, 2022a.

Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly- throughs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 12912–12921. IEEE, 2022b.

Butian Xiong, Nanjun Zheng, Junhua Liu, and Zhen Li. Gauu-scene V2: assessing the reliability of image-based metrics with expansive lidar image dataset using 3dgs and nerf. *CoRR*, abs/2404.04880, 2024.

Weiyi Xue, Fan Lu, Chi Zhang, Tianhang Wang, Sanqing Qu, Zehan Zheng, Boyuan Zheng, Junqiao Zhao, and Guang Chen. Signal structure-aware gaussian splatting for large-scale scene reconstruction. *CoRR*, abs/2607.01698, 2026.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *CVPR*, 2024.

Haodong Zhu, Changbai Li, Yangyang Ren, Zichao Feng, Xuhui Liu, Hanlin Chen, Xiantong Zhen, and Baochang Zhang. Surf3r: Rapid surface reconstruction from sparse RGB views in seconds. *CoRR*, abs/2508.04508, 2025.

A Implementation Details of SpecGS

All experiments are conducted using a unified training protocol across datasets and scenes. Training follows a three-stage coarse-to-fine curriculum with a total budget of $100{,}000$ iterations. Following the standard 3DGS preprocessing protocol, the downsampled training images are treated as the effective full resolution. Stages I and II operate at $0.5\times$ and $0.75\times$ of this resolution, respectively, while Stage III performs full-resolution frequency-guided refinement.

Stage transitions are primarily governed by the Adaptive Convergence Trigger (ACT), which monitors optimization saturation using consecutive loss windows and advances the rendering resolution once the relative improvement falls below the predefined threshold. A maximum iteration limit is retained for each stage as a fallback to avoid excessively delayed transitions.

The total iteration budget is mainly used to support stable optimization across multiple resolution stages rather than to obtain gains from prolonged training. As shown in Tab. 7, increasing vanilla 3DGS training from 30k to 100k iterations improves PSNR by only 0.03 dB, while extending SpecGS from 100k to 200k iterations changes PSNR by only 0.01 dB. This indicates that the adopted training budget primarily supports the progressive optimization procedure.

For the bounded spatial-support parameterization in Eq. 6, we use $\gamma = 0.1$. The ACT saturation threshold is set to 1×10^{-2}. During Stage III, the Gaussian-level frequency score f_k is maintained according to the EMA formulation in the main paper and is used to condition spatial-support regularization and densification priority. The DoG kernels maintain the bandwidth relation $\kappa_2 = 2\kappa_1$, and the frequency-guided reconstruction term is weighted by $\lambda_{\text{dog}} = 0.5$. The DoG bandwidth follows the Stage-III triangular schedule described in Sec. E, with its transition fixed at $t_{\text{prog}} = 50\%$. All remaining densification, pruning, and photometric optimization settings follow the standard 3DGS pipeline unless otherwise specified.

B Additional Experiments and Analysis

B.1 More Experimental Validation on Generalization and Scalability

Evaluation on large-scale urban scenes. To further evaluate the scalability of SpecGS beyond the main benchmarks, we conduct experiments on the city-scale GauU-Scene dataset (Xiong et al., 2024). As shown in Table 12, SpecGS achieves the highest PSNR on both the "Russian" and "Village" scenes while remaining feasible on a single GPU. On "Village", SpecGS achieves 25.76 dB PSNR using only 3.21M Gaussians, with a model size of 0.76 GB, 14.93 GB peak GPU memory, and 67 minutes of training. Compared with the strongest baseline in PSNR, this corresponds to a 0.13 dB improvement while substantially reducing the Gaussian count. On "Russian", SpecGS similarly achieves the highest PSNR and the shortest training time among the compared methods. These results demonstrate that the proposed frequency-guided allocation strategy remains effective when applied to additional large-scale scenes.

Evaluation on standard benchmarks. We further evaluate SpecGS on the widely used Mip-NeRF360 (Barron et al., 2022a) and Tanks & Temples (Knapitsch et al., 2017) benchmarks. As reported in Table 14, SpecGS achieves the highest PSNR and SSIM on both datasets while using the fewest Gaussian primitives and the lowest GPU memory among the evaluated methods. Its LPIPS remains competitive with the strongest perceptual baselines. These results indicate that the benefit of frequency-guided support regulation and primitive allocation is not restricted to large-scale urban scenes. Qualitative comparisons in Fig. 11 further show that SpecGS preserves fine structures and object boundaries across these general-scene benchmarks.

Comparison with Mip-Splatting. We additionally compare SpecGS with Mip-Splatting, which primarily addresses aliasing by integrating Gaussian responses over pixel footprints. While this strategy improves multi-scale rendering, it does not explicitly adapt primitive allocation according to local high-frequency demand.

As shown in Table 8, Mip-Splatting produces substantially more Gaussians and requires considerably higher GPU memory across all four scenes. In contrast, SpecGS uses Gaussian-level frequency

Table 7: Impact of training iterations on performance in Sci-Art scene (Lin et al., 2022).

Method	Iters	SSIM ↑	PSNR ↑	LPIPS ↓
3DGS	30k	0.791	24.13	0.262
3DGS	100k	0.793	24.16	0.258
Ours	100k	0.815	22.62	0.208
Ours	200k	0.815	22.63	0.207

Table 8: Comparison with Mip-Splatting across different scenes.

Scene	Method	GS (M) ↓	Size (GB) ↓	Mem (GB) ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Building	Mip-Splatting	19.67	4.62	>60	20.65	0.684	0.345
	Ours	**6.39**	**1.28**	**12.09**	**22.84**	**0.804**	**0.210**
Rubble	Mip-Splatting	17.21	4.04	>60	24.04	0.724	0.323
	Ours	**5.89**	**1.27**	**11.45**	**25.85**	**0.801**	**0.210**
Residence	Mip-Splatting	20.03	4.70	>60	20.98	0.747	0.272
	Ours	**5.26**	**1.29**	**9.31**	**22.51**	**0.825**	**0.205**
Sci-Art	Mip-Splatting	6.11	1.43	>40	21.74	0.806	0.257
	Ours	**1.98**	**0.45**	**10.82**	**22.62**	**0.815**	**0.208**

evidence to regulate spatial support and prioritize densification in detail-rich regions, thereby reducing unnecessary primitive allocation while achieving higher reconstruction quality in these experiments.

Comparison with FDS-GS. We also compare SpecGS with FDS-GS on the Rubble scene. As shown in Table 13, SpecGS improves PSNR from 25.12 to 25.85 and SSIM from 0.729 to 0.801, while reducing LPIPS from 0.301 to 0.210. In the corresponding reproduced experiments, FDS-GS requires more than 30 GB of peak GPU memory, whereas SpecGS uses 12.45 GB. This comparison provides additional evidence that frequency-guided capacity allocation can improve reconstruction quality without relying on excessive memory consumption.

Qualitative results. Visual comparisons in Fig. 7 show that SpecGS reconstructs fine-scale structures such as architectural edges, window frames, and surface textures while avoiding excessive smoothing. Additional cross-method and cross-scene comparisons are provided in Fig. 10. Together with the quantitative results above, these comparisons demonstrate that SpecGS maintains a favorable quality–efficiency trade-off across different scene scales and datasets.

B.2 Extension to Multi-GPU Training

SpecGS is designed primarily as a standalone single-GPU framework. We nevertheless investigate whether its frequency-guided allocation mechanism remains compatible with existing distributed reconstruction pipelines. For this purpose, we integrate SpecGS into the CityGS (Liu et al., 2024a) multi-GPU framework and compare it with SSAGS, which is also built upon CityGS.

Table 9 reports a controlled comparison across the four urban scenes using two densification thresholds, denoted by "L" (0.00012) and "S" (0.00015). The results show complementary behavior across scenes. SpecGS+CityGS performs particularly well on Building and Rubble, achieving higher PSNR under both threshold settings and generally competitive SSIM and LPIPS. On Residence, the two methods exhibit closely matched performance, whereas SSAGS+CityGS performs better on Sci-Art. These results indicate that SpecGS remains competitive when incorporated into the same system-level partitioning framework, rather than relying on an advantage from a different distributed training pipeline.

We further perform a controlled training-efficiency comparison on Rubble using the same CityGS pipeline and 6× NVIDIA A5000 GPUs. As shown in Table 10, SpecGS+CityGS achieves 27.43 dB PSNR, 0.841 SSIM, and 0.185 LPIPS, compared with 27.33 dB, 0.839, and 0.191 for

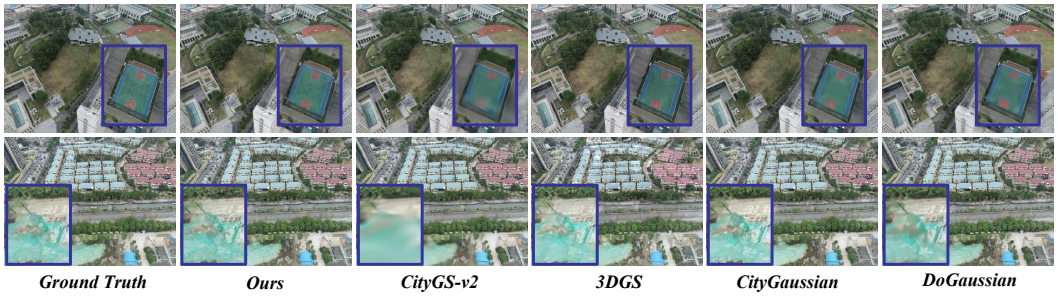


Figure 7: Qualitative results of ours and other methods in image rendering on GauU-Scene (Xiong et al., 2024).

SSAGS+CityGS. More importantly, the training time is reduced from 84 minutes to 21 minutes under the same hardware configuration. Thus, although SpecGS is designed for single-GPU reconstruction, its frequency-guided allocation mechanism can also be integrated into multi-GPU reconstruction pipelines without sacrificing its computational efficiency.

B.3 Efficiency of Frequency-Guided Supervision

We analyze the additional computational cost introduced by the DoG-based frequency-guided supervision used during Stage III. Following the standard 3DGS preprocessing protocol, the downsampled training images are treated as the effective full-resolution inputs. The DoG response is therefore computed at the resolution actually used for Stage-III optimization rather than at the original raw dataset resolution.

The DoG computation requires only two Gaussian filtering operations followed by a point-wise difference and normalization. As shown in Fig. 8(b), the additional GPU memory remains below 80 MB throughout Stage III, which is small relative to the overall memory footprint of the reconstruction model.

We further quantify its practical effect in Table 11. Removing frequency-guided supervision reduces PSNR from 22.62 to 22.15, while the measured GPU memory changes only from 10.82 GB to 10.92 GB. The measured rendering throughput also remains comparable (111.17 versus 107.23 FPS). These results indicate that the DoG-based supervision provides a clear reconstruction benefit while introducing negligible memory and runtime overhead in the evaluated setting.

B.4 Frequency-Domain Analysis of Reconstruction Fidelity

We further analyze the reconstructed images in the frequency domain to examine whether the proposed frequency-guided allocation strategy improves the recovery of fine-scale scene content. Importantly, this analysis concerns the spectral characteristics of the reconstructed outputs rather than the intrinsic expressiveness of individual Gaussian primitives. SpecGS does not introduce an explicit spectral carrier or learnable frequency parameter into the 3D Gaussian representation.

All analyses in this subsection are conducted on the Rubble scene (Turki et al., 2022b). For each rendered image, we remove the DC component and compute its two-dimensional Fourier spectrum. We then radially aggregate the spectral responses to compare how different methods distribute reconstruction energy across spatial-frequency ranges.

As shown in Fig. 9(a), SpecGS preserves more high-frequency content than 3DGS and CityGS. This observation is consistent with the qualitative results, where SpecGS better retains fine structures, boundaries, and local texture details.

Higher high-frequency energy alone, however, does not necessarily imply more accurate reconstruction, since spurious artifacts may also increase spectral energy. We therefore additionally analyze the frequency distribution of the reconstruction error between the rendered image and the ground truth. As shown in Fig. 9(b), SpecGS exhibits lower reconstruction error across the analyzed frequency ranges, with the most noticeable improvement in the high-frequency regime. Together with

Table 9: Controlled comparison with SSAGS (Xue et al., 2026) under the same CityGS (Liu et al., 2024a) multi-GPU training pipeline. "L" and "S" denote densification thresholds of 0.00012 and 0.00015, respectively.

Method	Building			Rubble			Residence			Sci-Art		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SSAGS+CityGS-L	22.86	0.778	0.225	27.35	0.843	0.189	23.46	0.840	0.194	25.94	0.890	0.170
SpecGS+CityGS-L	23.11	0.815	0.204	27.43	0.841	0.185	23.35	0.844	0.192	24.14	0.871	0.182
SSAGS+CityGS-S	22.98	0.780	0.225	27.21	0.830	0.208	23.50	0.839	0.196	25.12	0.863	0.200
SpecGS+CityGS-S	23.08	0.810	0.208	27.36	0.833	0.188	23.29	0.840	0.197	24.10	0.866	0.195

Table 10: Controlled training-efficiency comparison on Rubble using the same CityGS pipeline and $6\times$ NVIDIA A5000 GPUs.

Method	PSNR↑	SSIM↑	LPIPS↓	Time↓
SSAGS+CityGS	27.33	0.839	0.191	84 min
SpecGS+CityGS	**27.43**	**0.841**	**0.185**	**21 min**

Fig. 9(a), this indicates that the additional high-frequency content preserved by SpecGS corresponds to more accurate detail reconstruction rather than simple spectral amplification.

We further examine the spatial-support distribution of the optimized Gaussians. As shown in Fig. 9(c), SpecGS allocates a larger fraction of its representation to compact-support Gaussians than standard 3DGS. This trend is consistent with the proposed frequency-guided adaptation mechanism: stronger high-frequency evidence imposes greater resistance to excessive support expansion while simultaneously increasing densification priority in detail-rich regions.

Overall, the frequency-domain analysis provides complementary evidence that SpecGS recovers fine-scale scene content more accurately while maintaining a compact representation. We emphasize that these observations should not be interpreted as evidence that individual SpecGS Gaussians possess an explicit frequency modulation or intrinsically larger spectral bandwidth. Instead, the improvement arises from frequency-guided spatial-support regulation and more selective Gaussian capacity allocation.

C Frequency-Spatial Support Coupled Gaussian Representation

C.1 Bounded Spatial-Support Control

The spatial-support mechanism in SpecGS provides a stable and bounded multiplicative control over the effective extent of each Gaussian. For Gaussian G_k, let $\boldsymbol{\sigma}_k \in \mathbb{R}^3_+$ denote its original anisotropic scaling vector. We introduce an optimizable latent variable $\beta_k \in \mathbb{R}$ and define the support-control factor as

$$s_k = 2\ \mathrm{sigm}(\gamma\beta_k), \qquad \mathrm{sigm}(z) = \frac{1}{1+\exp(-z)}, \tag{15}$$

where $\gamma > 0$ controls the slope of the mapping. The effective scaling vector is

$$\tilde{\boldsymbol{\sigma}}_k = s_k \boldsymbol{\sigma}_k. \tag{16}$$

Since $\mathrm{sigm}(z) \in (0,1)$ for any finite z,

$$0 < s_k < 2. \tag{17}$$

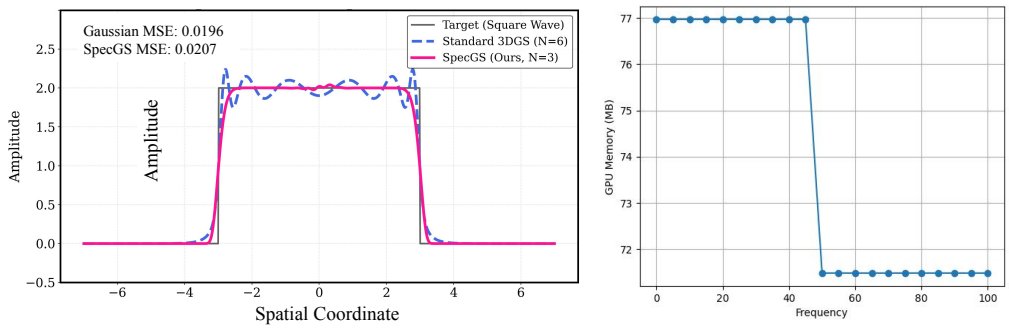
When $\beta_k = 0$,

$$s_k = 2\ \mathrm{sigm}(0) = 1, \tag{18}$$

so the additional support modulation is neutral and the effective scale coincides with the original Gaussian scale.

The corresponding effective covariance is

$$\widetilde{\boldsymbol{\Sigma}}_k = \mathbf{R}_k \widetilde{\mathbf{S}}_k \widetilde{\mathbf{S}}_k^\top \mathbf{R}_k^\top, \qquad \widetilde{\mathbf{S}}_k = \mathrm{diag}(\tilde{\boldsymbol{\sigma}}_k). \tag{19}$$



(a) Controlled 1D Square-Wave Fitting Comparison (b) GPU Memory Overhead of DoG Supervision

Figure 8: Controlled analysis of representation behavior and computational overhead. (a) In the 1D square-wave fitting experiment, the SpecGS-inspired formulation uses three components while achieving comparable fitting error to a six-component standard Gaussian mixture. (b) The DoG-based supervision introduces less than 80 MB of additional GPU memory throughout Stage III, indicating negligible overhead relative to the overall reconstruction pipeline.

Table 11: Impact of frequency-guided supervision (DoG) on performance and efficiency.

Variant	PSNR ↑	Mem (GB) ↓	FPS ↑
w/o Freq-Guided Supervision	22.15	10.92	107.23
Ours	**22.62**	**10.82**	**111.17**

where $\mathbf{R}_k$ and $\mathbf{\Sigma}_k$ denote the rotation matrix and covariance matrix of the original Gaussian, respectively. Because s_k is a positive scalar,

$$\widetilde{\mathbf{\Sigma}}_k = s_k^2 \mathbf{\Sigma}_k. \tag{20}$$

Thus, the support-control factor preserves the anisotropy and orientation of the original Gaussian while modulating its overall spatial extent.

Smoothness and bounded sensitivity. The derivative of s_k with respect to β_k is

$$\frac{\partial s_k}{\partial \beta_k} = 2\gamma \,\mathrm{sigm}(\gamma\beta_k)\left[1 - \mathrm{sigm}(\gamma\beta_k)\right]. \tag{21}$$

Since

$$\mathrm{sigm}(z)\left[1 - \mathrm{sigm}(z)\right] \le \frac{1}{4}, \tag{22}$$

we obtain

$$0 < \frac{\partial s_k}{\partial \beta_k} \le \frac{\gamma}{2}. \tag{23}$$

Therefore, the mapping from β_k to s_k is smooth and globally Lipschitz with constant $\gamma/2$. This bounded derivative limits the sensitivity of the additional multiplicative support-control factor to updates of β_k and avoids abrupt support modulation during optimization.

C.2 Frequency-Conditioned Support Regularization

The Gaussian-level frequency score f_k is a dynamically maintained training statistic rather than an optimizable parameter. SpecGS uses this statistic to condition the regularization applied to the support-control factor:

$$\mathcal{L}_{\mathrm{fs}} = \frac{1}{|\mathcal{V}_v|} \sum_{k \in \mathcal{V}_v} f_k \,\max(0, s_k - 1), \tag{24}$$

where $\mathcal{V}_v$ denotes the set of Gaussians visible in the current training view.

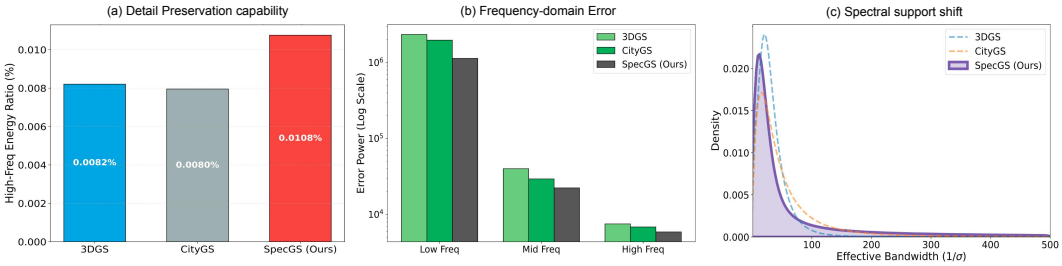


Figure 9: **Frequency-domain analysis of reconstruction fidelity.** (a) Radially averaged spectral statistics show that SpecGS preserves more high-frequency content in the reconstructed images. (b) Frequency-domain reconstruction-error analysis indicates lower error across frequency bands, particularly in the high-frequency regime. (c) The distribution of effective Gaussian support shows that SpecGS allocates a larger fraction of compact primitives to regions requiring fine-scale reconstruction.

Gradients are not propagated through f_k. For $s_k > 1$,

$$\frac{\partial \mathcal{L}_{\text{fs}}}{\partial \beta_k} = \frac{f_k}{|\mathcal{V}_v|} \frac{\partial s_k}{\partial \beta_k} = \frac{2\gamma f_k}{|\mathcal{V}_v|} \operatorname{sigm}(\gamma \beta_k) \left[1 - \operatorname{sigm}(\gamma \beta_k)\right], \tag{25}$$

whereas the contribution is zero for $s_k < 1$. At $s_k = 1$, the standard subgradient convention applies.

Combining Eqs. 23 and 25 yields

$$\left| \frac{\partial \mathcal{L}_{\text{fs}}}{\partial \beta_k} \right| \leq \frac{\gamma f_k}{2|\mathcal{V}_v|}. \tag{26}$$

Thus, stronger high-frequency evidence increases the penalty against excessive support expansion while maintaining a bounded response with respect to the additional support-control variable. Importantly, Eq. 24 does not force high-frequency Gaussians to shrink unconditionally. The regularizer becomes active only when $s_k > 1$. The final effective support remains jointly determined by the standard Gaussian scale, the reconstruction objective, and the additional support-control factor.

D Frequency-Guided Capacity Allocation

D.1 Monotonic Densification Priority

In addition to regulating spatial support, the Gaussian-level frequency score f_k is used to prioritize primitive allocation during densification. Let g_k denote the standard 3DGS densification signal and $\tau_g > 0$ its base threshold. SpecGS defines the Gaussian-specific threshold

$$\tau_k = \frac{\tau_g}{1 + \eta f_k}, \tag{27}$$

where $\eta \geq 0$ controls the strength of frequency guidance.

Its dependence on f_k is monotonic:

$$\frac{\partial \tau_k}{\partial f_k} = -\frac{\eta \tau_g}{(1 + \eta f_k)^2} \leq 0. \tag{28}$$

Therefore, increasing high-frequency evidence never increases the densification threshold. For $\eta > 0$, Gaussians associated with larger f_k receive lower thresholds and are consequently more likely to satisfy

$$g_k > \tau_k. \tag{29}$$

Equivalently,

$$g_k(1 + \eta f_k) > \tau_g. \tag{30}$$

This expression makes clear that f_k reweights the allocation priority of the standard densification signal rather than replacing it.

When f_k is small, τ_k remains close to the standard threshold τ_g. When f_k is larger, the threshold decreases and the corresponding Gaussian receives higher densification priority. Thus, the adaptive threshold biases additional representation capacity toward regions associated with stronger high-frequency evidence.

Importantly, Eq. 27 does not directly suppress standard densification in smooth regions. The reduction in the total Gaussian count observed in our experiments is therefore an empirical outcome of the complete coupled optimization, including coarse-to-fine training, support regulation, and frequency-guided densification, rather than a direct mathematical consequence of the adaptive threshold alone.

D.2 Separation of Allocation and Primitive Generation

The frequency-guided rule affects only the selection priority of Gaussian primitives. Once a Gaussian is selected, SpecGS follows the standard 3DGS densification procedure for primitive generation. In particular, f_k does not determine the displacement, orientation, scale, or other geometric parameters of newly generated Gaussians.

The overall information flow is

$$\mathcal{M}_v \longrightarrow \hat{f}_{k,v}^{(t)} \longrightarrow f_k^{(t)} \longrightarrow \{\mathcal{L}_{\text{fs}}, \tau_k\}\,. \tag{31}$$

Here, $\mathcal{M}_v$ is the image-space DoG evidence map, $\hat{f}_{k,v}^{(t)}$ is the current-view Gaussian-level estimate, and $f_k^{(t)}$ is its temporally accumulated statistic.

This formulation separates the roles of frequency evidence and Gaussian geometry. Image-space DoG responses provide local high-frequency evidence; f_k aggregates this evidence at the Gaussian level; and the resulting statistic conditions support regulation and densification priority. The underlying Gaussian rendering and primitive-generation mechanisms remain unchanged from standard 3DGS.

Consequently, SpecGS does not improve reconstruction by introducing an additional oscillatory rendering parameter into each Gaussian. Instead, it improves capacity allocation by directing support regulation and additional primitives toward regions associated with stronger high-frequency evidence.

E DoG-Based Frequency Evidence and Supervision

During Stage III, SpecGS uses a Difference-of-Gaussians (DoG) response to extract image-space evidence associated with edges, boundaries, and fine-scale textures. The same evidence map serves two purposes: (i) it provides the image-space observations from which the Gaussian-level frequency statistic f_k is estimated, and (ii) it weights the reconstruction residual in the frequency-guided supervision term. Let $\mathbf{I}_v$ denote the ground-truth image of the current training view v. We compute the DoG response using two Gaussian filters with bandwidths κ_1 and κ_2, where $\kappa_2 = 2\kappa_1$. The difference response is converted into a scalar image-space response and normalized by its maximum value to obtain the evidence map $\mathcal{M}_v$. Thus, $\mathcal{M}_v(\mathbf{u}) \in [0, 1]$, where $\mathbf{u}$ indexes image pixels, and larger values indicate stronger local variation under the current DoG bandwidth.

Stage-III bandwidth schedule. The DoG bandwidth is varied only during Stage III. Let $t_{\text{prog}} \in [0, 1]$ denote the normalized progress within Stage III. We employ a triangular schedule: the bandwidth gradually increases during the first half of Stage III and decreases during the second half. This schedule provides a smooth transition across different detail scales during full-resolution refinement. Importantly, t_{prog} is a global Stage-III progress variable and is completely independent of the Gaussian-level statistic f_k.

The transition is fixed at the midpoint of Stage III. As reported in the main paper, varying this transition point between 45% and 55% produces only minor changes in reconstruction quality and Gaussian count, indicating that the method is not sensitive to the exact transition location.

Gaussian-level evidence aggregation. For each visible Gaussian G_k, the DoG response is aggregated using its image-space rendering contribution:

$$\hat{f}_{k,v}^{(t)} = \frac{\sum_{\mathbf{u}} \mathrm{sg}\left(w_{k,\mathbf{u},v}^{(t)}\right) \mathcal{M}_v(\mathbf{u})}{\sum_{\mathbf{u}} \mathrm{sg}\left(w_{k,\mathbf{u},v}^{(t)}\right) + \epsilon}, \tag{32}$$

where $w_{k,\mathbf{u},v}^{(t)} \geq 0$ denotes the contribution of Gaussian G_k to pixel $\mathbf{u}$ in the current view and $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. The estimate is accumulated over training observations using the EMA defined in the main paper:

$$f_k^{(t)} = (1-\rho_f) f_k^{(t-1)} + \rho_f \hat{f}_{k,v}^{(t)}. \tag{33}$$

where $\rho_f \in (0,1]$ denotes the EMA update rate. With a non-negative initialization $f_k^{(0)} \geq 0$, the non-negative DoG evidence and rendering contributions ensure that $f_k^{(t)} \geq 0$ throughout training. Therefore, f_k is not an optimizable frequency parameter. It is a Gaussian-level statistic derived from image-space DoG evidence and is subsequently used to condition spatial-support regularization and densification priority.

Frequency-guided reconstruction. The same image-space evidence map is also used to emphasize reconstruction errors in regions containing stronger local detail:

$$\mathcal{L}_{\mathrm{freq}} = \left\| \mathcal{M}_v \odot (\mathbf{I}_{\mathrm{ren},v} - \mathbf{I}_v) \right\|_1, \tag{34}$$

where $\mathbf{I}_{\mathrm{ren},v}$ denotes the rendered image and $\odot$ denotes element-wise multiplication. Thus, the DoG response does not introduce a new rendering parameter; instead, it provides a shared image-space signal for both Gaussian-level capacity allocation and high-frequency reconstruction supervision.

Table 12: Performance comparison on GauU-Scene. GS(M): number of Gaussians in millions; Size: model size in GB; Mem: GPU memory consumption in GB. Methods marked with * use multi-GPU training. Best results for each metric are highlighted in bold.

Scene	Method	PSNR (dB)	GS (M)	Size (GB)	Mem (GB)	Time (min)	FPS
Russian	CityGaussian*	24.21	14.10	2.78	>24	223	**52.2**
	3DGS	24.08	12.94	2.55	>24	185	47.7
	CityGS-v2*	23.89	**6.97**	**0.38**	**14.35**	184	33.3
	Ours	**24.34**	9.41	1.17	17.09	**173**	49.5
Village	CityGaussian*	25.57	13.40	3.05	>24	140	**48.6**
	3DGS	25.63	12.30	2.85	>24	74	46.5
	CityGS-v2*	25.25	4.64	**0.63**	15.89	78	30.7
	Ours	**25.76**	**3.21**	0.76	**14.93**	**67**	47.9

F Controlled 1D Signal-Fitting Study

F.1 Purpose and Scope

We conduct a controlled one-dimensional signal-fitting study to examine how different localized basis families behave under limited component budgets. The goal of this experiment is to provide empirical motivation for adaptive representation allocation in signals with spatially varying complexity, rather than to reproduce the complete 3D SpecGS pipeline.

In particular, the SpecGS-inspired basis used in this study is a diagnostic 1D representation designed to expose the potential benefit of localized frequency adaptation. It should not be interpreted as the rendering primitive used by SpecGS in 3D. The full 3D method retains standard Gaussian primitives and instead uses image-space high-frequency evidence to regulate spatial support and densification priority.

We evaluate signals ranging from smooth profiles to sharp transitions and localized oscillations. This controlled setting allows us to investigate whether adaptive localized representations can achieve

Table 13: Comparison with FDS-GS on the Rubble scene. GPU memory denotes the peak memory measured in the corresponding experiment.

Method	PSNR↑	SSIM↑	LPIPS↓	Mem (GB)↓
FDS-GS	25.12	0.729	0.301	>30
SpecGS	**25.85**	**0.801**	**0.210**	**12.45**

Table 14: Quantitative comparisons on the Mip-NeRF360 (Barron et al., 2022a) and Tanks and Temples datasets (Knapitsch et al., 2017). For baselines marked with † we enable only those contributions that do not affect quality.

Method	Mip-NeRF360					Tanks and Temples				
	PSNR ↑	SSIM ↑	LPIPS ↓	VRAM ↓	#Gs ↓	PSNR ↑	SSIM ↑	LPIPS ↓	VRAM ↓	#Gs ↓
3DGS	27.53	0.815	0.256	8.8GiB	2.74M	23.77	0.852	0.204	4.7GiB	1.57M
2DGS	27.41	0.804	0.267	9.2GiB	2.84M	23.62	0.845	0.217	5.4GiB	1.61M
Taming-3DGS †	27.53	0.815	0.256	8.9GiB	2.73M	23.78	0.852	0.203	4.9GiB	1.57M
GES	26.91	0.794	**0.250**	6.2GiB	2.34M	23.14	0.841	**0.183**	3.4GiB	1.43M
Ours	**27.62**	**0.819**	0.258	**5.3GiB**	**1.91M**	**24.92**	**0.897**	0.185	**3.1GiB**	**0.98M**

lower fitting error or comparable accuracy with fewer components when the target contains substantial high-frequency structure.

F.2 Experimental Setup

All experiments are implemented in PyTorch using gradient-based optimization. The signal domain is $x \in [-8, 8]$ and is discretized into $M = 500$ uniformly spaced samples. Each representation is evaluated using component budgets $N \in \{5, 10, 15, 20\}$. Training proceeds for up to 2,000 iterations, with early stopping when the Mean Squared Error (MSE) falls below 10^{-5}.

Given a target signal $y(x)$ and its approximation $\hat{y}(x)$, the fitting objective is

$$\mathcal{L}_{\text{fit}} = \frac{1}{M} \sum_{m=1}^{M} \left(\hat{y}(x_m) - y(x_m)\right)^2 . \tag{35}$$

All compared representations follow the same sampling domain, optimization budget, and convergence criterion. Therefore, the study focuses on their relative fitting behavior under controlled component budgets. In addition to the above component-budget sweep, we conduct a separate illustrative square-wave experiment comparing three SpecGS-inspired components with six standard Gaussian components, as shown in Fig. 8(a). This auxiliary experiment is used only to illustrate fitting behavior under different component counts.

F.3 Target Signals

We consider six representative signals spanning different levels of smoothness and spectral complexity. To avoid confusion with the Gaussian scale $\boldsymbol{\sigma}_k$ used in the 3D formulation, we denote the characteristic signal width by W, with $W = 6$.

Square signal.

$$y_{\text{sq}}(x) = \begin{cases} 1, & |x| < W/2, \\ 0, & \text{otherwise.} \end{cases} \tag{36}$$

This signal contains two sharp discontinuities and therefore exhibits slow spectral decay.

Triangle signal.

$$y_{\text{tri}}(x) = \begin{cases} W/2 - |x|, & |x| < W/2, \\ 0, & \text{otherwise.} \end{cases} \tag{37}$$

The triangle signal is piecewise linear and is smoother than the square signal while retaining localized boundary transitions.

Parabolic signal.

$$y_{\mathrm{par}}(x) = \begin{cases} (W/2)^2 - x^2, & |x| < W/2, \\ 0, & \text{otherwise.} \end{cases} \tag{38}$$

Half-sinusoid signal.

$$y_{\mathrm{hs}}(x) = \begin{cases} \sin\left(\pi \frac{x+W/2}{W}\right), & |x| < W/2, \\ 0, & \text{otherwise.} \end{cases} \tag{39}$$

Gaussian signal.

$$y_{\mathrm{gau}}(x) = \exp\left(-\frac{x^2}{2(W/3)^2}\right). \tag{40}$$

Oscillatory signal.

$$y_{\mathrm{osc}}(x) = \begin{cases} \sin\left(\frac{4\pi x}{W}\right) \exp\left(-\frac{x^2}{8}\right), & |x| < W/2, \\ 0, & \text{otherwise.} \end{cases} \tag{41}$$

Together, these targets cover smooth, piecewise-smooth, discontinuous, and oscillatory signal structures.

F.4 Compared 1D Representations

We compare several localized mixture representations with different spatial and spectral behaviors. For clarity, the notation in this section is independent of the Gaussian-level frequency statistic f_k used by the full 3D SpecGS method.

Gaussian Mixture Model. A Gaussian mixture with N components is written as

$$g_{\mathrm{GMM}}(x) = \sum_{i=1}^{N} w_i \exp\left(-\frac{(x-\mu_i)^2}{2r_i^2}\right), \tag{42}$$

where μ_i, r_i, and w_i denote the component center, width, and weight, respectively.

Difference of Gaussians. The DoG representation is

$$g_{\mathrm{DoG}}(x) = \sum_{i=1}^{N} w_i \left[\exp\left(-\frac{(x-\mu_i)^2}{2r_{i,1}^2}\right) - \exp\left(-\frac{(x-\mu_i)^2}{2r_{i,2}^2}\right)\right]. \tag{43}$$

Generalized Exponential Family. The generalized exponential representation introduces a shape exponent p_i:

$$g_{\mathrm{GEF}}(x) = \sum_{i=1}^{N} w_i \exp\left(-\frac{|x-\mu_i|^{p_i}}{2r_i^2}\right), \tag{44}$$

where p_i controls the decay profile of each localized basis.

Laplacian of Gaussian. The LoG mixture is written as

$$g_{\mathrm{LoG}}(x) = \sum_{i=1}^{N} w_i \left(1 - \frac{(x-\mu_i)^2}{r_i^2}\right) \exp\left(-\frac{(x-\mu_i)^2}{2r_i^2}\right). \tag{45}$$

SpecGS-inspired localized basis. For the controlled diagnostic study, we additionally consider a localized Gaussian envelope equipped with an adaptive oscillatory component:

$$g_{\mathrm{SI}}(x) = \sum_{i=1}^{N} w_i \exp\left(-\frac{(x-\mu_i)^2}{2\tilde{r}_i^2}\right) A_i \cos(2\pi\nu_i d_i(x-\mu_i) + \varphi_i), \tag{46}$$

where ν_i denotes the oscillation frequency used only in this one-dimensional experiment, A_i and φ_i denote amplitude and phase, and d_i is the 1D directional factor used in the original simulation.

The adaptive spatial width is

$$\tilde{r}_i = \frac{r_i}{q_i}, \qquad q_i = 1 + \mathrm{softplus}(b_i), \tag{47}$$

where b_i is a toy-specific shape variable.

We emphasize that the symbols ν_i, q_i, and b_i are used only for this diagnostic 1D representation. They are intentionally distinct from the 3D quantities f_k, s_k, and β_k in the full SpecGS formulation. Moreover, Eq. 46 is not the rendering primitive of the proposed 3D method. Its role is solely to provide a controlled reference for studying how localized adaptive bases behave on signals with different frequency content.

F.5 Evaluation Protocol

We evaluate fitting accuracy primarily using MSE. Optimization is performed using MSE throughout the controlled study, while Fig. 14 additionally reports MAE as a complementary visualization metric. For each target signal and component budget, the final reconstruction error is computed using Eq. 35. The experiments therefore reveal how efficiently different localized basis families approximate signals of varying complexity under the same component budget.

Of particular interest are the square and oscillatory signals, where sharp transitions or localized oscillations create strongly non-uniform representation demands. Smooth signals such as the Gaussian and parabolic targets serve as complementary cases in which additional spectral flexibility is less critical.

Importantly, these experiments do not establish that the full 3D SpecGS method uses a more expressive rendering primitive. Instead, they provide empirical evidence for the broader design motivation that representation capacity should adapt to local signal complexity rather than being allocated uniformly.

F.6 Results and Discussion

The complete numerical results are shown in Figs. 14, 15, 16, 17, 18, 19, and 20.

Across the evaluated targets, different basis families exhibit distinct behavior depending on signal regularity. For smooth signals, standard localized bases already provide strong approximations, and the benefit of additional adaptive structure is limited. In contrast, signals containing discontinuities or oscillatory patterns place substantially greater demand on a fixed component budget.

The square-wave example in Fig. 8(a) illustrates this behavior particularly clearly. The SpecGS-inspired 1D representation uses three components while achieving fitting error comparable to that of a six-component standard Gaussian mixture. Similar qualitative trends can be observed in several signals containing sharp transitions or localized oscillations.

These observations motivate the capacity-allocation perspective adopted by the full 3D SpecGS framework. The 1D study shows that signal complexity is highly non-uniform and that adaptive localized representations can use a limited component budget more effectively in difficult regions. In the actual 3D method, however, this principle is realized differently: SpecGS retains standard Gaussian rendering primitives and uses image-space DoG evidence to derive the Gaussian-level statistic f_k, which regulates spatial support and prioritizes densification.

Therefore, the 1D experiments should be interpreted as controlled empirical motivation for frequency-guided capacity allocation, rather than as a reduced implementation or theoretical verification of the complete 3D SpecGS pipeline.

G Result Analysis and Discussion

We further analyze the mechanisms behind the observed improvements in reconstruction quality, memory efficiency, and Gaussian count. The key distinction of SpecGS is not an increase in the

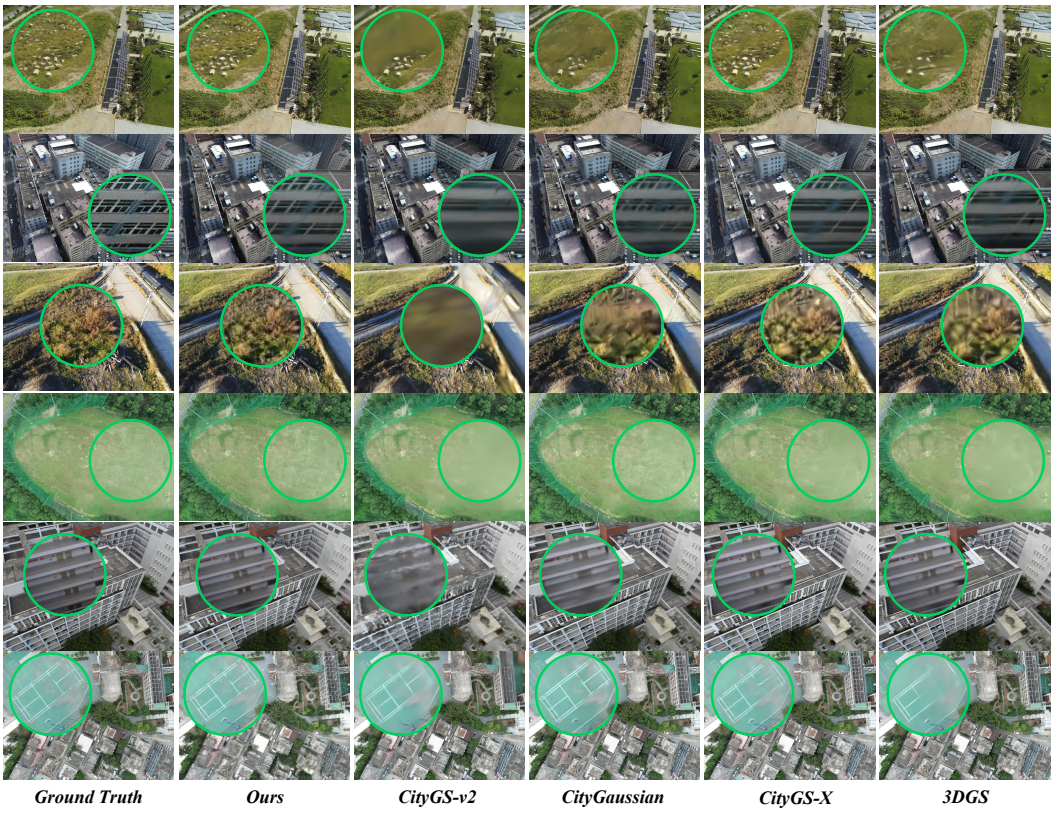


Figure 10: Qualitative results of ours and other methods in image rendering on GauU-Scene dataset (Xiong et al., 2024), mill19 dataset (Turki et al., 2022b) and the UrbanScene3D dataset (Lin et al., 2022).

intrinsic expressiveness of individual Gaussian primitives, but a more selective allocation of representation capacity according to local high-frequency demand.

Frequency-guided capacity allocation. Standard 3D Gaussian Splatting represents a scene using smooth, spatially localized Gaussian primitives. Regions containing sharp edges, fine textures, or repetitive structures typically require denser primitive allocation to achieve accurate reconstruction. However, applying the same densification behavior across regions with different levels of local complexity can introduce substantial redundancy.

SpecGS addresses this problem by estimating a Gaussian-level high-frequency statistic f_k from image-space DoG evidence. Importantly, f_k is not an optimizable frequency parameter and does not alter the rendering primitive itself. Instead, it conditions two complementary mechanisms. First, the frequency-spatial regularization $\mathcal{L}_{\text{fs}}$ assigns stronger penalties to excessive support expansion for Gaussians associated with larger f_k. Second, the Gaussian-specific densification threshold $\tau_k = \tau_g/(1+\eta f_k)$ decreases with f_k, giving such Gaussians higher priority for additional representation capacity.

These two mechanisms operate on different aspects of the representation: spatial-support regulation controls how existing Gaussians occupy space, whereas frequency-guided densification controls where additional Gaussians are introduced. The subsequent primitive-generation procedure remains unchanged from standard 3DGS; in particular, f_k does not determine the displacement or geometry of newly generated Gaussians.

Origin of the quality–efficiency trade-off. The reduction in Gaussian count should therefore be interpreted as a consequence of improved capacity allocation rather than enhanced per-primitive expressiveness. In smooth or weakly varying regions, frequency evidence remains relatively small,



Figure 11: **Comparative Visualization Across Methods.** Displayed are side-by-side comparisons between our proposed method and established techniques alongside their respective ground truth imagery. Specifically, we present comparative analyses of various methods on two benchmark datasets, namely Mip-NeRF360 Barron et al. (2022a) and Tanks & Temples (Knapitsch et al., 2017).

so SpecGS avoids assigning unnecessarily high densification priority. In regions with stronger local detail, larger frequency evidence both discourages excessive support expansion and increases the likelihood of densification. As a result, representation capacity is preferentially directed toward regions where additional Gaussians are more useful.

This interpretation is consistent with the ablation results in Tab. 4. Removing frequency-guided densification substantially increases memory consumption while degrading reconstruction quality, indicating that frequency-conditioned allocation is important for controlling primitive growth. Removing frequency-spatial coupling also reduces reconstruction quality, showing that support regulation provides a complementary role. Furthermore, removing the bounded spatial-support mechanism causes memory overflow, demonstrating the importance of maintaining stable control over Gaussian spatial extent.

The bounded support parameterization also contributes to optimization stability. Because $s_k = 2\ \mathrm{sigm}(\gamma\beta_k)$ remains strictly positive and bounded, the additional multiplicative support modulation introduced by β_k cannot grow without limit. At the same time, the hinge structure of $\mathcal{L}_{\mathrm{fs}}$ only penalizes expansion when $s_k > 1$. Thus, stronger frequency evidence does not force a Gaussian to shrink unconditionally; rather, it imposes stronger resistance to excessive support expansion while leaving the final scale jointly determined by the reconstruction objective and the standard Gaussian parameters.

Why SpecGS is frequency-guided. The frequency-aware characterization of SpecGS originates from the source and use of the conditioning signal. During Stage III, a normalized DoG response extracts image-space evidence around edges, boundaries, and fine-scale textures. This evidence is aggregated according to the rendering contribution of each visible Gaussian and accumulated through an EMA to form f_k. The resulting statistic then conditions spatial-support regularization and

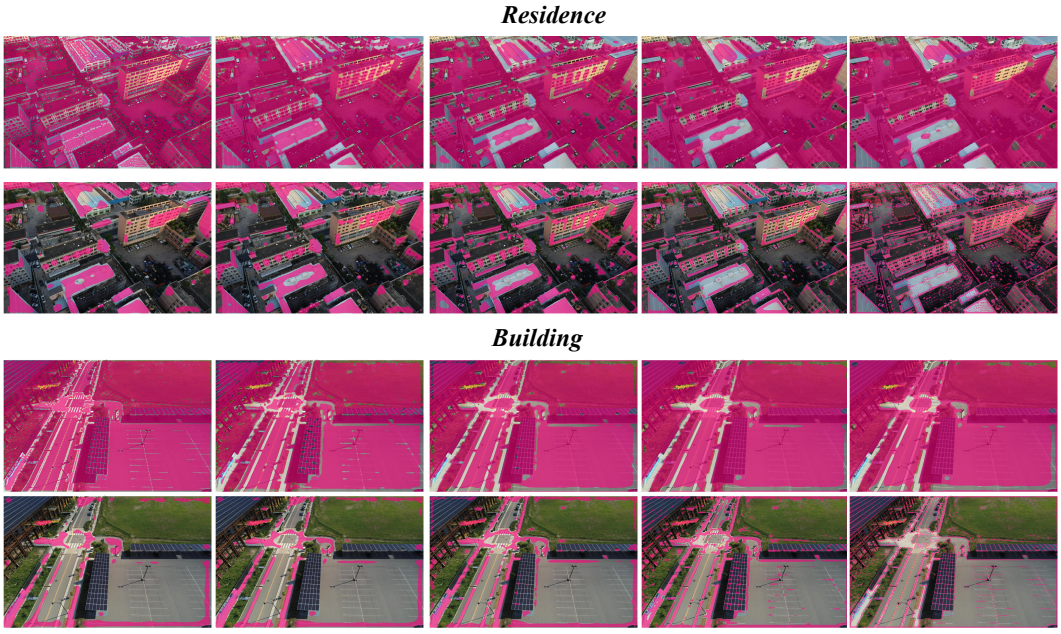


Figure 12: Visualization of the DoG-based frequency-guided supervision during Stage III. Highlighted regions indicate image locations receiving stronger reconstruction weighting under the current DoG response, including edges, boundaries, and fine-scale textures.

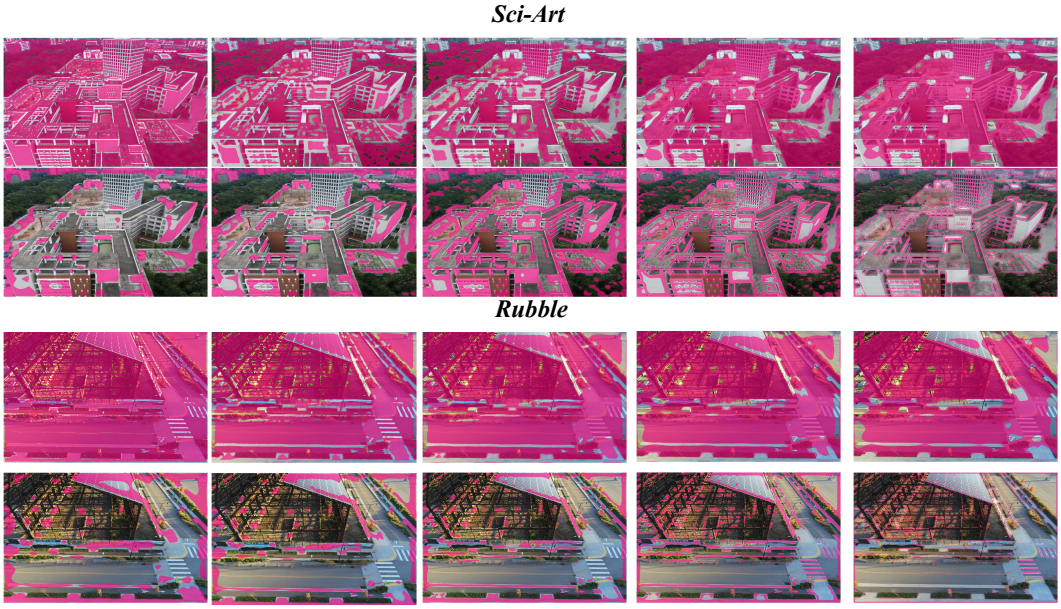


Figure 13: Visualization of the DoG-based frequency-guided supervision during Stage III. Highlighted regions indicate image locations receiving stronger reconstruction weighting under the current DoG response, including edges, boundaries, and fine-scale textures.

densification priority, while the same DoG map is also used to weight high-frequency reconstruction residuals.

Therefore, frequency information in SpecGS serves as a guidance signal for optimization and capacity allocation rather than as an additional spectral parameter of the Gaussian primitive. The frequency-domain analysis in Fig. 9 provides complementary evidence at the reconstruction level:

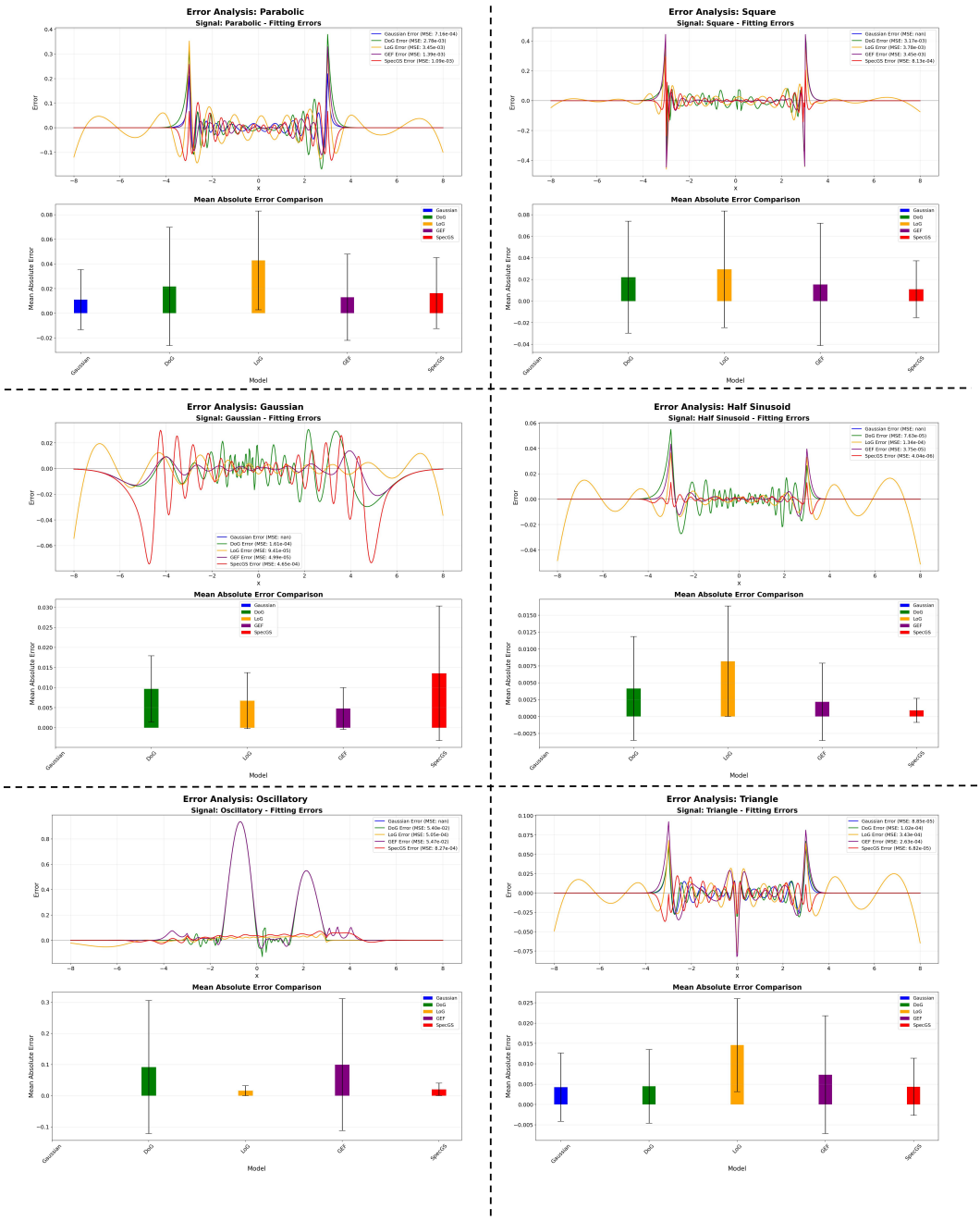


Figure 14: Numerical simulation results of different mixture-based representations are presented, demonstrating fitting performance on both high-frequency and low-frequency signals via Mean Absolute Error (MAE) comparisons; NaN indicates fitting failure.

SpecGS preserves more high-frequency scene content while reducing reconstruction error in the corresponding frequency ranges. These observations indicate more accurate recovery of fine-scale content, but should not be interpreted as evidence of explicit spectral modulation inside individual Gaussian primitives.

Interpretation of the 1D controlled study. The supplementary 1D experiments in Figs. 14, 15, 16, 17, 18, 19, and 20 provide controlled empirical motivation for the same broader principle. They compare localized basis families under limited component budgets and show that signals contain-

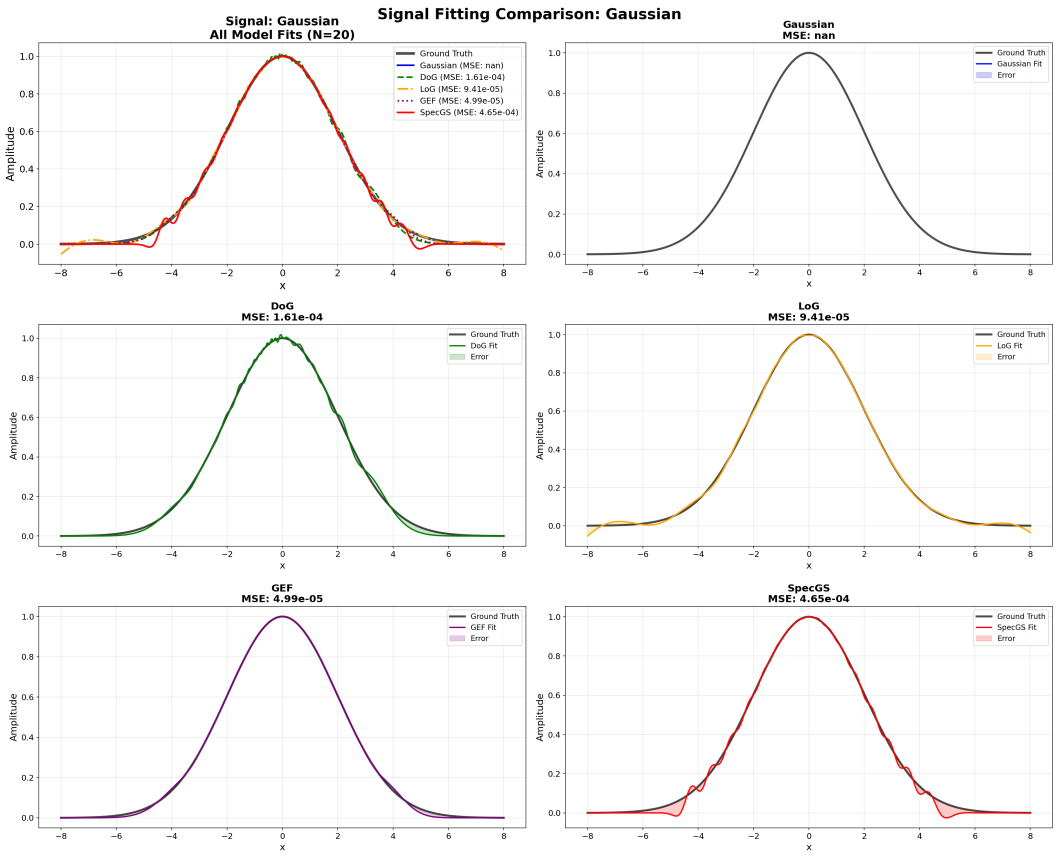


Figure 15: Numerical simulation results of different mixture-based representations are presented, demonstrating fitting performance on Gaussian signals via MSE comparisons; NaN indicates fitting failure.

ing sharp transitions or oscillatory structure place substantially different demands on representation capacity than smooth signals.

The SpecGS-inspired basis used in these experiments is only a diagnostic 1D representation and is not the rendering primitive used by the full 3D method. Accordingly, the 1D results should not be interpreted as demonstrating that SpecGS obtains its gains through a more expressive frequency-modulated Gaussian. Instead, they motivate the central design choice of the 3D framework: representation capacity should be adapted according to local signal complexity rather than allocated uniformly.

In summary, SpecGS improves large-scale reconstruction by shifting the focus from indiscriminate primitive growth to frequency-guided capacity allocation. Image-space high-frequency evidence determines where stronger support regulation and higher densification priority are needed, while the underlying Gaussian rendering formulation remains unchanged. This allocation strategy enables a favorable balance between reconstruction quality, Gaussian count, and memory consumption for city-scale reconstruction on a single GPU.

H LIMITATION

The Difference-of-Gaussian (DoG) mask used for frequency-guided supervision relies on spatial gradient statistics to emphasize high-frequency regions. In unstructured natural scenes including dynamic foliage, water ripples, or strong shadows, these high-frequency responses may be temporally inconsistent or spatially over-activated. This could potentially introduce misleading signals into the optimization. Since the resulting frequency statistic conditions support regulation and densifi-

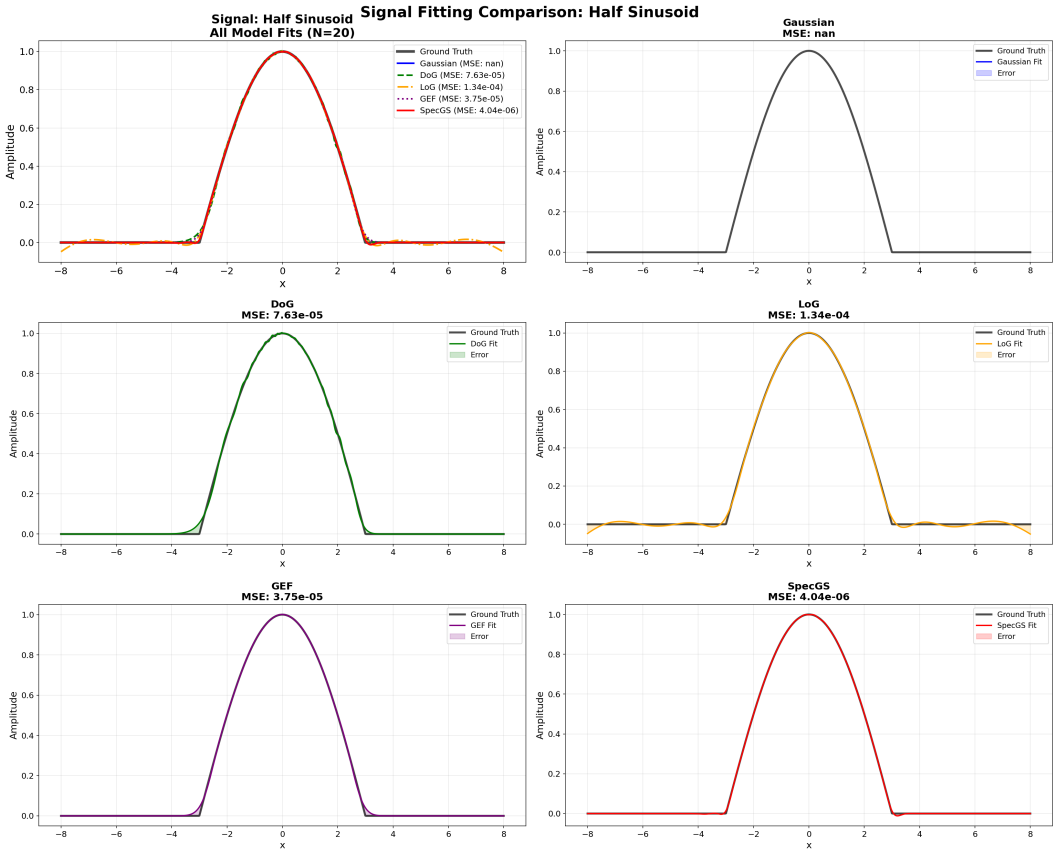


Figure 16: Numerical simulation results of different mixture-based representations are presented, demonstrating fitting performance on Half Sinusoid signals via MSE comparisons; NaN indicates fitting failure.

cation priority during Stage III, such non-structural high-frequency responses may lead to unnecessary support regulation or additional densification. This effect is limited in our targeted large-scale, predominantly static urban scenes where high-frequency content remains structure-aligned. Our coarse-to-fine scheduling with late-stage activation also partially mitigates instability. Extending SpecGS to dynamic or highly unstructured scenarios would nevertheless require additional regularization such as temporal consistency constraints, semantic-aware masking, or confidence-adaptive weighting. We leave this direction for future work.

I Broader Social Impacts

This paper proposes a more memory-efficient and scalable representation for large-scale 3D scene reconstruction, with the primary goal of advancing the efficiency and expressiveness of machine learning–based visual representations. By reducing computational and memory requirements, this work may help lower the hardware barriers for large-scale 3D reconstruction and neural rendering, potentially making such techniques more accessible to a broader research and practitioner community. We do not anticipate any significant negative societal or ethical impacts arising directly from this work. The methods introduced are general-purpose and focus on improving computational efficiency and representation quality, without targeting sensitive application domains. As with many advances in machine learning, downstream applications should be deployed responsibly in accordance with existing ethical guidelines.

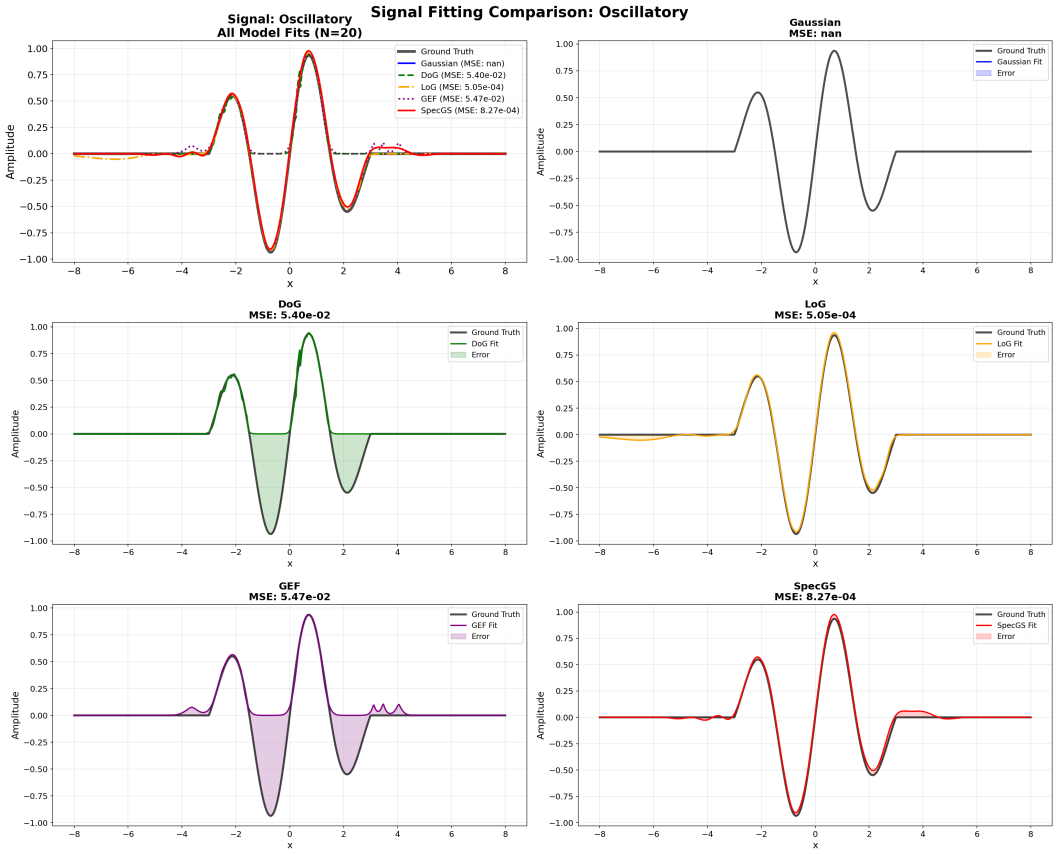


Figure 17: Numerical simulation results of different mixture-based representations are presented, demonstrating fitting performance on Oscillatory signals via MSE comparisons; NaN indicates fitting failure.

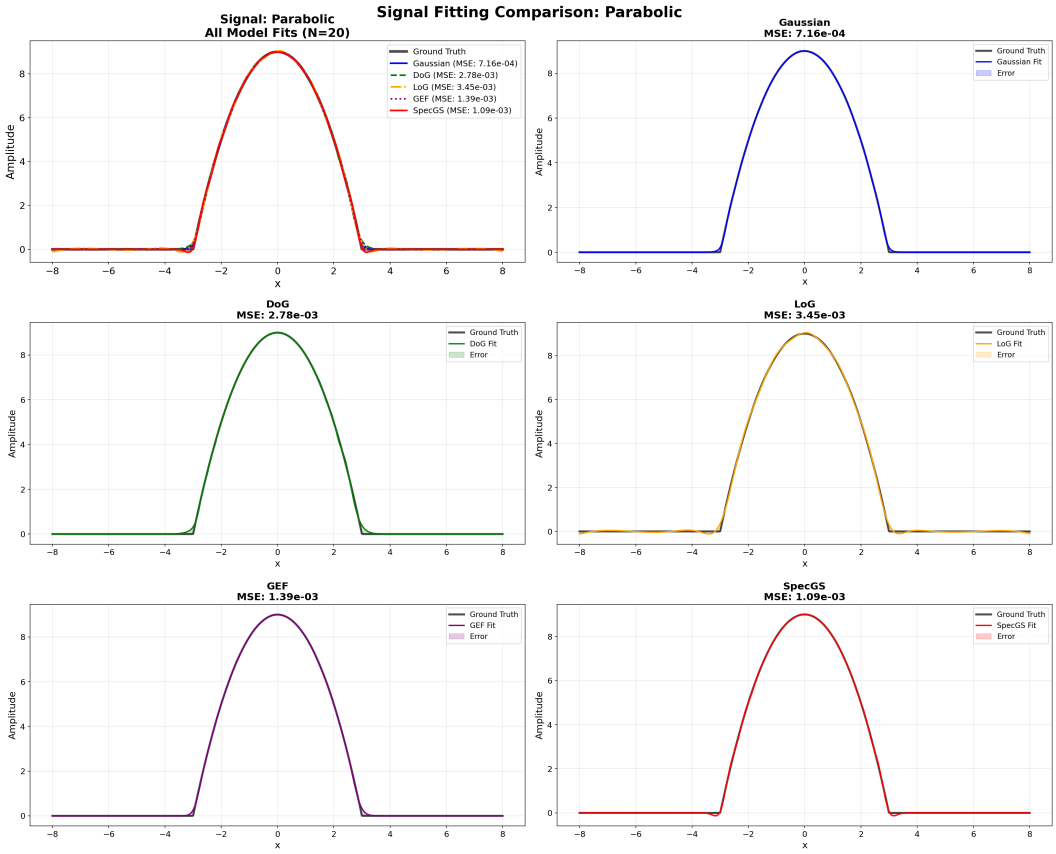


Figure 18: Numerical simulation results of different mixture-based representations are presented, demonstrating fitting performance on Parabolic signals via MSE comparisons; NaN indicates fitting failure.

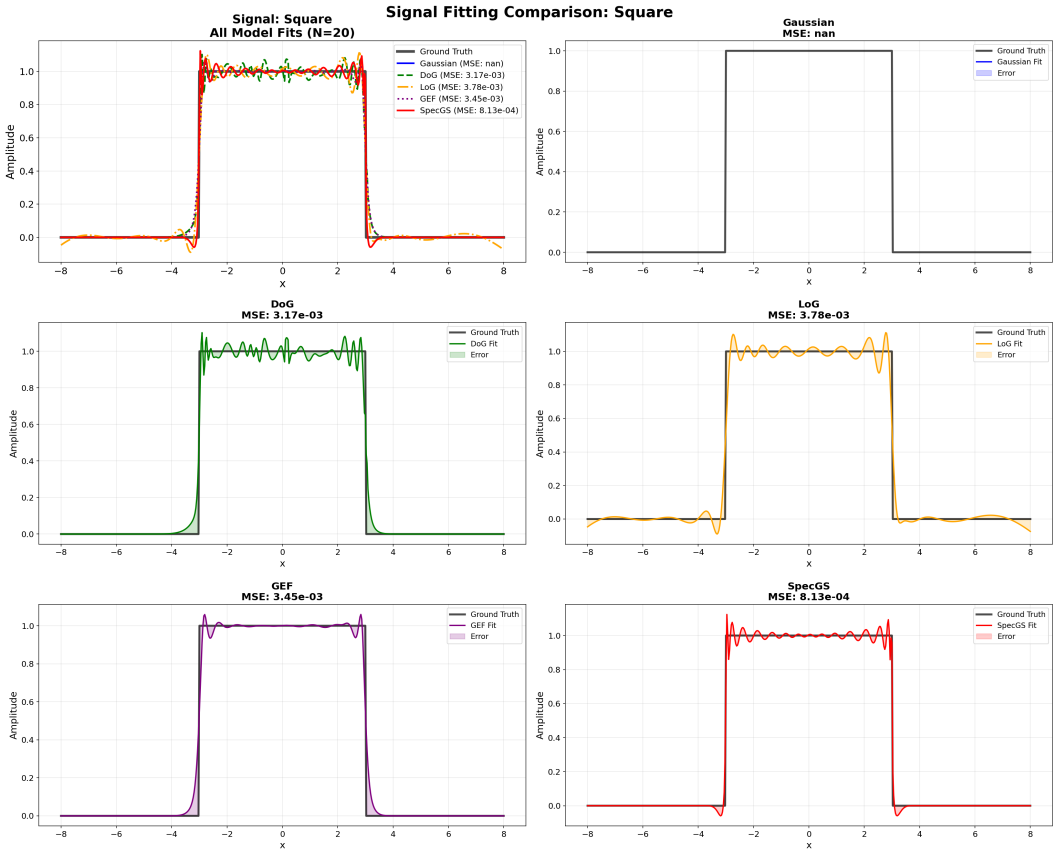


Figure 19: Numerical simulation results of different mixture-based representations are presented, demonstrating fitting performance on Square signals via MSE comparisons; NaN indicates fitting failure.

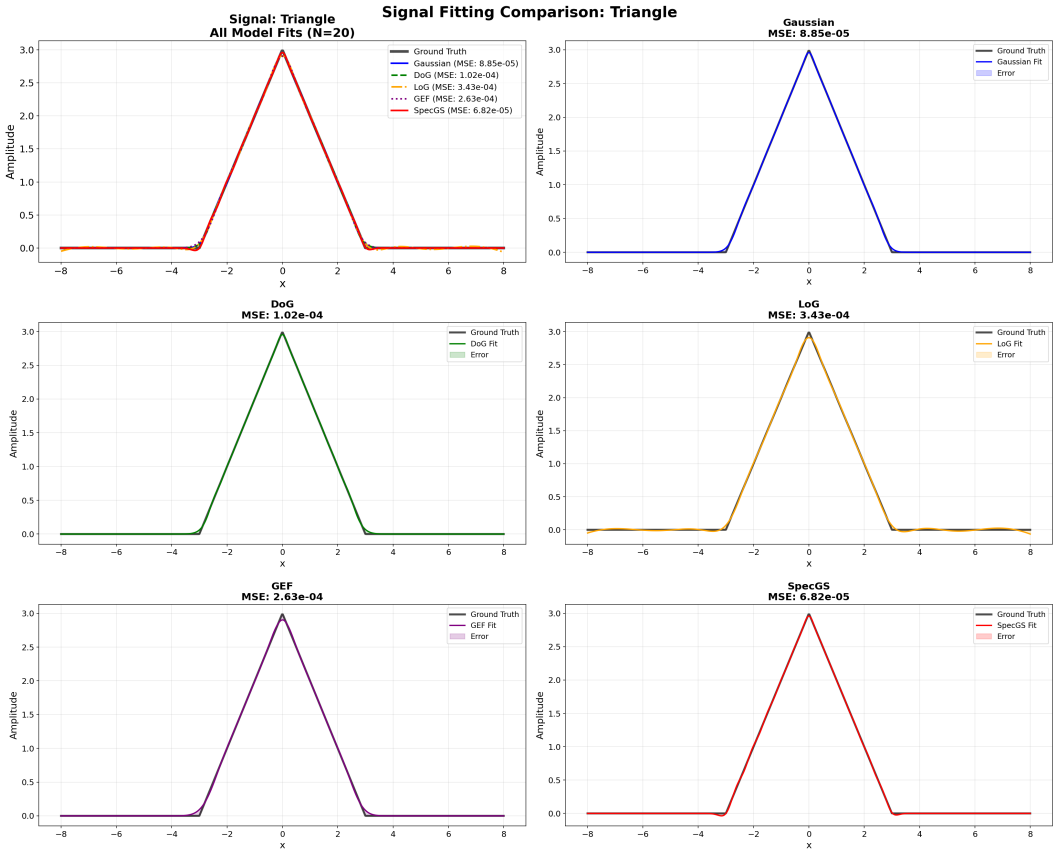


Figure 20: Numerical simulation results of different mixture-based representations are presented, demonstrating fitting performance on Triangle signals via MSE comparisons; NaN indicates fitting failure.