

The Decoder Already Knows Where to Densify: Attention-Gated Capacity Allocation for Feed-Forward 3D Gaussian Splatting

Seunggyu Choi^{*}, Suwoong Yeom^{*}, and Suk-Ju Kang[†]

Sogang University

^{*} Indicates Equal Contribution [†] Corresponding Author

Abstract. Feed-forward 3D Gaussian Splatting (3DGS) reconstructs scenes in a single pass, yet distributes its primitive budget almost uniformly, since the photometric gradients that drive densification in optimization-based 3DGS are unavailable at inference time. Recent feed-forward methods recover some adaptivity via post-hoc voxelization and pruning or via auxiliary density predictors, but these either generate excess primitives only to filter them, or add learnable modules trained to estimate scene complexity. We instead ask whether the backbone already exposes such a cue. We find that two parameter-free statistics of the geometry decoder’s attention, Attention Sharpness and Cross-Head Disagreement, are spatially correlated with where reconstruction error concentrates, and we use their fusion as an inference-time proxy for capacity demand. Building on this descriptor, we propose Attention-Gated Capacity Allocation (AGCA), an end-to-end differentiable framework with two coupled branches: Continuous Attention-Gated Densification, which reallocates spatial capacity by soft-routing child Gaussians rather than hard thresholding, and Signal-Gated Representational Adaptation, which gates a zero-initialized LoRA so that a frozen decoder adapts only where structure demands it. AGCA attaches to frozen backbones as a plug-and-play module: trained once on RE10K, it improves PSNR by 0.4–1.8 dB across three architectures and transfers without fine-tuning to DTU, ScanNet, and ACID, while updating less than 0.1% of the total parameters and preserving single-pass inference.

Keywords: 3D Gaussian Splatting · Feed-Forward 3DGS · Adaptive Capacity Allocation · Attention-Guided Capacity Allocation

1 Introduction

Generalizable 3D scene reconstruction aims to recover complex spatial structures from limited observations and serves as a core component for applications such as AR/VR and world modeling. Recently, feed-forward 3D Gaussian Splatting (3DGS) [1, 3–5, 7, 10, 13, 16, 18–20, 26, 29, 34, 35, 38, 41] has enabled scene reconstruction without the time-consuming per-scene optimization process, shifting the paradigm of 3D vision toward single-pass inference. These models leverage

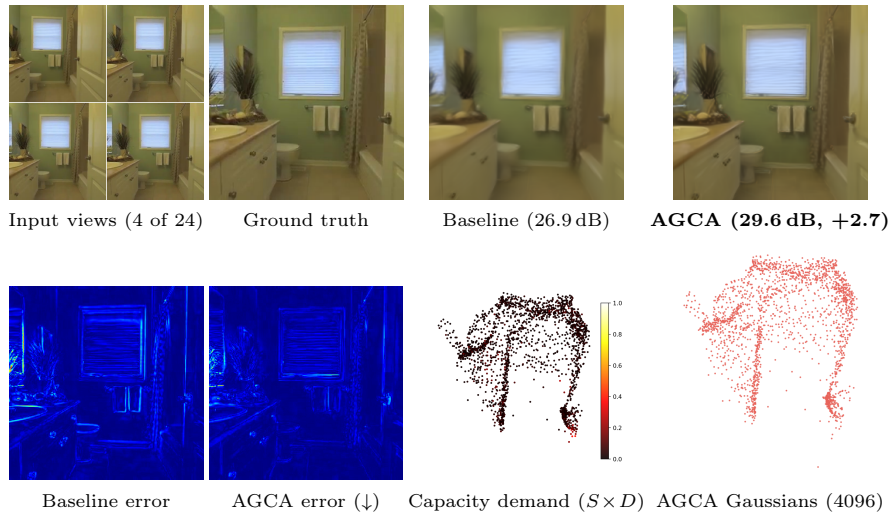


Fig. 1: Motivation for Attention-Gated Capacity Allocation. **Top:** A feed-forward 3DGS baseline allocates a fixed budget of 2048 Gaussians uniformly; AGCA improves PSNR by +2.7 dB. **Bottom-left:** Error maps show AGCA reduces reconstruction error in geometrically complex regions. **Bottom-right:** The capacity-demand signal $S \times D$, extracted from the decoder’s attention, spatially correlates with error; AGCA uses it to concentrate additional Gaussians (2048 $\rightarrow$ 4096) in high-demand areas.

large-scale foundation backbones [21, 28, 30, 32] to directly predict Gaussian primitives. However, this paradigm inherits a dilemma rooted in the nature of the 3D Gaussian representation itself.

Unlike discrete pixels or point clouds, Gaussians possess spatial extent and naturally allow overlap. When feed-forward models regress these primitives in a single forward pass, overlapping primitives often fail to organize into geometrically favorable structures, resulting in spatial interference and redundant allocations. As illustrated in Figure 1, this limitation typically manifests as blurred structures and concentrated errors in a small number of geometrically complex regions, while a comparable share of the budget is spent on flat, easily represented areas. The design of the backbones that predict these primitives has advanced rapidly, but the distribution of a fixed primitive budget across a scene at inference time has received considerably less attention.

Conventional optimization-based 3DGS addresses this issue through iterative topological optimization, where photometric loss gradients serve as feedback signals for splitting or pruning Gaussians. Feed-forward models acquire global generalization ability through photometric supervision during large-scale training, but no scene-specific gradient feedback is available once training ends. Without such post-hoc optimization, feed-forward networks cannot adjust primitive density to the geometric complexity of a particular test scene. To mitigate this limitation, recent efforts [11, 17, 23, 27, 31, 33, 36] have explored post-hoc voxeliza-

tion and pruning of Gaussians, or introduced auxiliary networks that explicitly estimate spatial density scores for topology control. These approaches recover a degree of adaptivity, but they either generate excessive primitives and filter them afterwards, or add learnable modules that must themselves be trained to estimate scene complexity. Both routes assume that the capacity-demand signal has to be constructed externally. We take the opposite view and look for a cue that the backbone already exposes.

We find such a cue within the joint self-attention mechanism of the geometry decoder, and extract two parameter-free statistics from its attention distribution. The first is Attention Sharpness ($\mathcal{S}$), the smoothed inverse of head-averaged attention entropy, which is high when the decoder concentrates a Gaussian query on a narrow set of patches, as occurs at depth discontinuities and thin structures. The second is Cross-Head Disagreement ($\mathcal{D}$), the variance of attention weights across heads, which we interpret as a proxy for correspondence ambiguity in textureless or occluded regions. As shown in Figure 1, the fusion of these two statistics is spatially correlated with where reconstruction error concentrates, which makes it usable as an inference-time stand-in for the densification cues that photometric gradients supply in optimization-based 3DGS. The descriptor requires no ground-truth error maps, no auxiliary head, and no additional forward computation beyond reading attention that the backbone already produces.

Driven by this observation, we propose Attention-Gated Capacity Allocation (AGCA), an end-to-end differentiable framework for scene-adaptive capacity allocation, instantiated by two coupled components. Continuous Attention-Gated Densification (CAD) handles spatial capacity. Rather than relying on discrete, hard-thresholding heuristics that disrupt the computation graph, CAD formulates densification as a soft-routing mechanism [6, 14]: a lightweight continuous gate, driven by the intrinsic descriptor, modulates the opacities of child Gaussians so that redundant primitives fade out while latent children materialize at complex boundaries, keeping the primitive count topologically fixed and the rendering path differentiable.

Adding Gaussians is insufficient if the decoder cannot coherently represent them. To accommodate the newly synthesized primitives without catastrophic forgetting of pre-trained priors, Signal-Gated Representational Adaptation (SRA) injects a zero-initialized Low-Rank Adaptation (LoRA) module [12, 24, 39] into the frozen decoder. Gated per token by the same descriptor, the adapter stays dormant on low-demand tokens and activates only where structure demands it, so that spatial and representational capacity are reallocated under a single shared signal rather than by two independent policies.

We evaluate AGCA as a plug-and-play module on three frozen feed-forward backbones, C3G [1], DepthSplat [33], and AnySplat [17]. Trained once on RE10K and applied without fine-tuning, it improves PSNR by 0.4 to 1.8 dB across these backbones and transfers to DTU, ScanNet, and ACID, while updating less than 0.1% of the total parameters and preserving single-pass inference.

Our contributions are summarized as follows:

- We show that two parameter-free statistics of a feed-forward geometry decoder’s self-attention, Attention Sharpness and Cross-Head Disagreement, are spatially correlated with regions where reconstruction error concentrates, and can therefore serve as an inference-time capacity-demand signal obtained without ground-truth error maps, auxiliary predictors, or scene-specific optimization.
- We propose Attention-Gated Capacity Allocation, a differentiable framework that redistributes Gaussian primitives according to this signal. Its continuous gating formulation avoids hard Top- K selection and post-hoc pruning, keeping topology control compatible with end-to-end training.
- We show that spatial reallocation and representational adaptation are complementary, and that driving both from one shared descriptor outperforms either branch alone or an ungated combination, while updating less than 0.1% of the total parameters and preserving single-pass inference.

2 Related Work

2.1 Feed-Forward 3D Gaussian Splatting

3DGS [20] achieves high-quality rendering through differentiable rasterization, but its per-scene optimization framework lacks the generalization ability for unseen environments. To enable generalizable 3D reconstruction, numerous feed-forward models [1, 3–5, 7, 10, 13, 16, 18, 19, 26, 29, 34, 35, 38, 41] based on large-scale VFMs [21, 28, 30, 32] have recently been proposed to generate 3D scenes from sparse images through a single forward pass.

However, the uniform regression paradigm of single-pass inference introduces a structural dilemma when coupled with the intrinsic properties of Gaussian representations. Unlike discrete points, Gaussians possess continuous spatial extent and naturally allow overlap. Consequently, feed-forward settings must regress primitive positions and scales at once from limited observations without post-hoc topological coordination, making optimization difficult. Because geometric complexity cannot be dynamically reflected, this often results in wasted representational resources or spatial interference.

2.2 Adaptive Densification in Feed-Forward Models

Conventional optimization-based methods regulate primitive density using photometric gradients after rendering, which are unavailable in feed-forward inference. To address this, recent works [11, 23, 33, 36] explore adaptive resource allocation within feed-forward pipelines. Some aggregate Gaussians into voxel grids for merging and refinement [17, 31], while others use patch entropy to vary primitive allocation [27]. These results suggest that topology control is feasible in feed-forward settings, but existing approaches either rely on heavy 3D post-processing or use statistical heuristics for allocation.

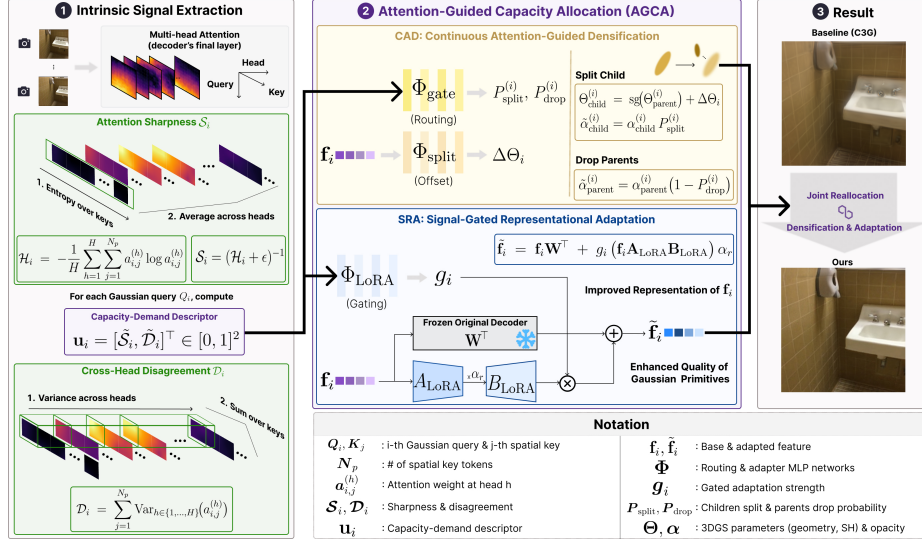


Fig. 2: Overview of the proposed AGCA framework. The method extracts an intrinsic capacity-demand signal from decoder attention, uses it to route Gaussian densification through CAD, and reuses the same signal to gate decoder adaptation via SRA.

In contrast, our method avoids expensive refinement and handcrafted rules, instead performing proactive densification by leveraging signals of multi-view consistency collapse implicitly encoded during feature fusion, offering a complementary approach to adaptive topology control.

2.3 Attention Statistics as Capacity-Demand Signals

Interpreting attention in foundation models, particularly vision transformers (ViTs) [9], through information bottlenecks and uncertainty has gained traction in deep learning theory. Prior work shows that attention entropy reflects uncertainty in information flow and correlates with over-concentrated reliance on limited evidence, leading to narrow contextual reasoning [37]. Attention statistics also align with semantic boundaries and spatial ambiguity in images [2].

In multi-view transformers, we hypothesize that attention statistics indicate where additional primitives are required. Concentrated attention (high Attention Sharpness, S) may signal structures that are hard to represent under a fixed budget, while cross-head disagreement (Cross-Head Disagreement, $\mathcal{D}$) captures correspondence ambiguity and epistemic uncertainty. We use these signals as intrinsic descriptors to anticipate capacity-critical regions before rendering, yielding an inference-time analogue of the densification cues provided by photometric gradients in optimization-based 3DGS.

3 Method

We present Attention-Gated Capacity Allocation (AGCA), an end-to-end differentiable framework that enables feed-forward 3DGS models to dynamically reallocate both spatial (Gaussian density) and representational (decoder adaptation) capacity using shared attention-derived capacity-demand signals, while updating fewer than 0.1% of the total model parameters.

As illustrated in Figure 2, the pipeline consists of three stages that bridge the gap between static feed-forward inference and adaptive scene reconstruction. First, Signal Extraction derives intrinsic per-token capacity-demand descriptors from the geometry decoder’s attention, providing a parameter-free, inference-time alternative to densification cues used in optimization-based 3DGS (Sec. 3.1). Second, Continuous Attention-Gated Densification (CAD) maps this signal to a differentiable soft-routing policy, reallocating spatial capacity by synthesizing child Gaussians as bounded residuals (Sec. 3.2). Third, Signal-Gated Representational Adaptation (SRA) modulates decoder capacity via per-token gated LoRA, selectively enhancing representation where structurally necessary while preserving pre-trained priors (Sec. 3.3) [12, 24, 39]. The entire system is trained end-to-end using a photometric loss, enabling joint reallocation of spatial and representational resources without requiring auxiliary supervision.

3.1 Intrinsic Capacity-Demand Signal Extraction

Optimization-based per-scene 3DGS relies on iterative view-space positional gradients to identify regions requiring densification, a strategy that is fundamentally incompatible with the single-pass inference of feed-forward models. A naive alternative, such as predicting a 2D error map via an auxiliary head, incurs additional parameters, demands per-pixel ground-truth supervision, and suffers from multi-view inconsistencies (e.g., specular highlights). We hypothesize instead that *the geometry decoder naturally encodes per-token capacity demand and epistemic uncertainty within its attention distribution*, allowing us to extract this latent signal at negligible computational overhead.

Architecture-Agnostic Attention Extraction. While feed-forward 3DGS architectures differ in their specific decoder designs, they fundamentally rely on an attention mechanism to map 2D visual evidence to 3D representations. Let $\mathbf{X}^{\text{patch}}$ denote the multi-view visual patch tokens and $\mathbf{X}^{\text{gs}}$ denote the latent Gaussian tokens. Whether the decoder employs joint self-attention over concatenated tokens (e.g., C3G) or explicit cross-attention blocks (e.g., NoPoSplat), the final decoder layer typically produces an attention distribution indicating how each Gaussian queries the visual patches.

We generalize this by isolating the *Gaussian-to-Patch* cross-attention matrix, denoted as $\mathbf{A}^{\text{gs} \rightarrow \text{patch}} \in [0, 1]^{H \times N_g \times N_p}$, where H , N_g , and N_p are the number of attention heads, Gaussian tokens, and patch tokens, respectively. This sub-matrix serves as a general, naturally available byproduct of the forward pass, mapping structural dependencies without requiring auxiliary operations

or architecture-specific modifications. From this matrix, we derive two distinct spatial indicators.

Signal 1: Multi-Head Attention Sharpness. For each Gaussian token $i \in \{1, \dots, N_g\}$, we compute the head-averaged attention entropy to quantify the dispersion of its visual queries. Let $a_{i,j}^{(h)}$ denote the (h, i, j) -th element of $\mathbf{A}^{\text{gs} \rightarrow \text{patch}}$, representing the attention probability from the i -th Gaussian to the j -th patch at the h -th head. The entropy $\mathcal{H}_i$ is given by:

$$\mathcal{H}_i = -\frac{1}{H} \sum_{h=1}^H \sum_{j=1}^{N_p} a_{i,j}^{(h)} \log a_{i,j}^{(h)} \quad (1)$$

We define the corresponding *Sharpness* as its smoothed inverse, $\mathcal{S}_i = (\mathcal{H}_i + \epsilon)^{-1}$, where ϵ is a small numerical stabilizer. A high $\mathcal{S}_i$ indicates that the decoder focuses intensely on a sharp structural cue (e.g., a depth edge) with limited primitives. This suggests insufficient primitive support for representing localized structure.

Signal 2: Cross-Head Disagreement. While $\mathcal{S}_i$ measures intra-head focus, it is insensitive to inter-head conflict. To capture epistemic uncertainty, we introduce the *Cross-Head Disagreement* $\mathcal{D}_i$, defined as the cumulative variance of the attention weights across all H heads:

$$\mathcal{D}_i = \sum_{j=1}^{N_p} \text{Var}_{h \in \{1, \dots, H\}}(a_{i,j}^{(h)}) \quad (2)$$

A high $\mathcal{D}_i$ suggests that different attention heads fundamentally disagree on which spatial patch provides the most plausible visual support. This pattern often arises in textureless or heavily occluded regions, effectively serving as an epistemic-uncertainty proxy.

To ensure scale-invariant properties across diverse scenes, both signals are min-max standardized across the N_g Gaussian token dimension. The final per-token capacity-demand descriptor is defined as $\mathbf{u}_i = [\tilde{\mathcal{S}}_i, \tilde{\mathcal{D}}_i]^\top \in [0, 1]^2$. Conceptually, $\tilde{\mathcal{S}}_i$ acts as a densification-promoting signal (targeting sharp, under-resourced structures), while $\tilde{\mathcal{D}}_i$ acts as a densification-suppressing signal, helping suppress floater generation in ambiguous regions.

3.2 Continuous Attention-Gated Densification

Standard 3DGS densification relies on hard thresholding and discrete cloning/splitting operations, which inherently break the computation graph. To render densification fully end-to-end learnable purely from rendering losses, CAD reformulates the process as a differentiable soft-routing mechanism.

Continuous gate network. A lightweight multi-layer perceptron (MLP), Φ_{gate} , maps the descriptor $\mathbf{u}_i$ to per-token continuous split and drop probabilities:

$$[P_{\text{split}}^{(i)}, P_{\text{drop}}^{(i)}] = \sigma(\Phi_{\text{gate}}(\mathbf{u}_i)) \in [0, 1]^2, \quad (3)$$

where $\sigma(\cdot)$ is the element-wise sigmoid activation. To prevent catastrophic layout disruption during early training phases, the final layer of Φ_{gate} is initialized with a conservative negative bias. This inductive bias yields a near-identity behavior at initialization ($t = 0$), allowing the routing policy to emerge gradually under photometric supervision rather than from arbitrary initialization noise.

Disentangled child generation. For tokens routed to densify, a non-linear split generator Φ_{split} predicts a bounded residual offset from the high-dimensional decoded feature $\mathbf{f}_i \in \mathbb{R}^D$ of token i :

$$\Delta\theta_i = \text{Clamp}(\Phi_{\text{split}}(\mathbf{f}_i), -\delta, +\delta), \quad \theta_{\text{child}}^{(i)} = \text{sg}(\theta_{\text{parent}}^{(i)}) + \Delta\theta_i, \quad (4)$$

where θ encapsulates the 3DGS parameters (3D mean, opacity, scale, rotation, and spherical harmonics), δ is a predefined clipping bound, and $\text{sg}(\cdot)$ represents the stop-gradient operator. The stop-gradient safely prevents the child branch from leaking unconstrained error signals into the parent pathway, thereby protecting the pre-trained latent space from collapse. Concurrently, the clamping mechanism anchors the child to the parent’s spatial topology, ensuring that child Gaussians act as *local structural refinements* rather than free-floating artifacts.

Differentiable soft routing. Naively concatenating a dynamic number of Gaussians disrupts differentiable alpha-compositing. Instead, we maintain a fixed topological cardinality (e.g., $2N_g$, accommodating all potential parents and children) and continuously modulate their opacities via the gating probabilities:

$$\tilde{\alpha}_{\text{parent}}^{(i)} = \alpha_{\text{parent}}^{(i)}(1 - P_{\text{drop}}^{(i)}), \quad \tilde{\alpha}_{\text{child}}^{(i)} = \alpha_{\text{child}}^{(i)} P_{\text{split}}^{(i)}. \quad (5)$$

Under this formulation, redundant Gaussians in flat regions gracefully fade out ($P_{\text{drop}} \rightarrow 1$), while latent child Gaussians materialize proportionally at complex boundaries ($P_{\text{split}} \rightarrow 1$). As both pathways are fully differentiable, rendering gradients flow through the gating mechanism without requiring non-differentiable approximations or auxiliary load-balancing losses.

3.3 Signal-Gated Representational Adaptation

Redistributing spatial capacity (Gaussians) only addresses half of the reconstruction bottleneck; the frozen decoder must also adapt its *representational capacity* to coherently decode the newly synthesized primitives. Full fine-tuning invariably risks degradation of multi-view priors established during pre-training. Conversely, applying a uniform LoRA updates the model in a generic, scene-agnostic manner, which can dilute the targeted effect of the CAD router. We introduce **Signal-Gated Representational Adaptation (SRA)**, modulated by the identical attention descriptor and implemented with zero-initialized Low-Rank Adaptation (LoRA).

Zero-initialized low-rank adaptation. For each frozen projection matrix $\mathbf{W}$ in the decoder’s attention sub-blocks, we inject trainable low-rank factors $\mathbf{A}_{\text{LoRA}}$ and $\mathbf{B}_{\text{LoRA}}$. The adapted forward pass for an input token $\mathbf{f}_i$ is gated by a token-specific scalar g_i :

$$\tilde{\mathbf{f}}_i = \mathbf{f}_i \mathbf{W}^\top + g_i (\mathbf{f}_i \mathbf{A}_{\text{LoRA}} \mathbf{B}_{\text{LoRA}}) \alpha_r, \quad (6)$$

where α_r is the LoRA scaling factor. Crucially, we strictly initialize $\mathbf{B}_{\text{LoRA}} = \mathbf{0}$. At step $t = 0$, the adapter contributes exactly zero, making the rendering quality functionally equivalent to the frozen baseline at initialization. This identity-preserving initialization stabilizes early training and allows the router to learn densification dynamics before the adapter becomes influential.

Signal-gated, per-token activation. A lightweight gating network, Φ_{lor_a} , maps the shared descriptor $\mathbf{u}_i$ to the activation score $g_i = \sigma(\Phi_{\text{lor}_a}(\mathbf{u}_i)) \in [0, 1]$. For low-demand tokens ($g_i \rightarrow 0$), the adapter remains dormant, preserving the frozen decoder’s stability. Conversely, for structurally complex tokens ($g_i \rightarrow 1$), the LoRA adaptation is fully activated. Visual patch tokens, serving as constant evidence, are assigned $g_i = 1$.

Signal-Conditioned Adapter Gating. Attention statistics extracted from the decoder’s initial evidence-gathering stage compute per-token adapter gates, g_i . These gates conditionally modulate the representational capacity of subsequent processing steps, which ultimately emit the final spatial densification decisions ($P_{\text{split}}, P_{\text{drop}}$). The specific execution of this signal flow seamlessly adapts to the underlying backbone design. In architectures featuring stacked decoder layers (e.g., the GMAE decoder in C3G [1]), this mechanism naturally forms a cross-layer cascade within a single forward pass. Conversely, for models employing monolithic decoders (e.g., AnySplat [17], DepthSplat [33]), an equivalent routing effect is achieved through an efficient two-pass scheme: the first pass extracts the intrinsic uncertainty signals, and the second pass applies the signal-conditioned gated adapters.

3.4 Training Objective and Parameter Efficiency

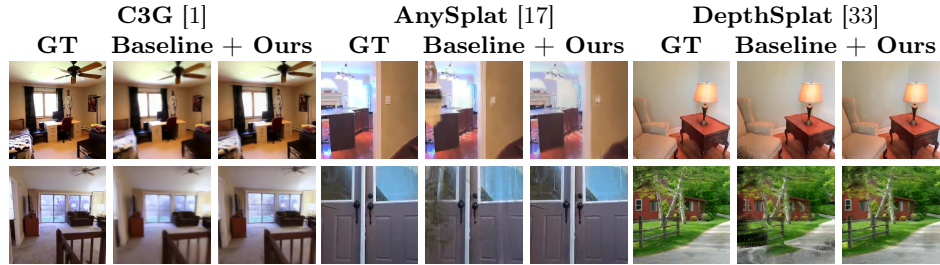
The entire framework is optimized end-to-end via a single photometric objective on novel views:

$$\mathcal{L} = \left\| \mathcal{R}(\{\Theta^{(i)}, \tilde{\alpha}^{(i)}\}_{i=1}^{2N_g}) - \mathbf{I}_{\text{gt}} \right\|_2^2, \quad (7)$$

where $\mathcal{R}(\cdot)$ denotes the differentiable Gaussian rasterizer. The total footprint of all trainable components, including the routing gates ($\Phi_{\text{gate}}, \Phi_{\text{lor}_a}$), the child residual generator (Φ_{split}), and the LoRA factors, remains less than 0.1% of the frozen baseline model. This parameter efficiency supports our core claim: the observed rendering improvements stem from *intelligent capacity reallocation* guided by intrinsic signals, rather than a brute-force increase in model capacity.

Table 1: Quantitative comparison on RE10K under 6-, 12-, and 24-view settings. Higher is better for PSNR and SSIM, while lower is better for LPIPS.

Methods	6 view			12 view			24 view		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
AnySplat [17]	19.208	0.665	0.312	18.936	0.627	0.342	18.652	0.612	0.355
AnySplat + Ours	19.659	0.675	0.303	19.226	0.632	0.336	19.110	0.628	0.330
DepthSplat [33]	20.450	0.692	0.285	20.780	0.705	0.312	21.110	0.722	0.340
DepthSplat + Ours	21.020	0.715	0.250	21.440	0.730	0.304	21.850	0.748	0.305
C3G [1]	18.425	0.540	0.391	18.798	0.553	0.320	18.995	0.561	0.351
C3G + Ours	19.624	0.590	0.337	20.221	0.607	0.301	20.442	0.614	0.313

**Fig. 3:** Qualitative comparison on RE10K. [40] For each backbone, we show the ground truth, the original feed-forward baseline, and the corresponding plug-and-play integration with our method.

4 Experiments

Datasets. We evaluate on four diverse benchmarks: **RealEstate10K (RE10K)** [40] (indoor/outdoor walkthroughs), using the standard test split (7,286 scenes, 360×640) [5, 7]; **ScanNet** [8] (indoor RGB-D sequences), adopting the official 100-scene test split (1296×968); **ACID** [22] (outdoor aerial imagery), using the 1,848-scene split in the pixelSplat format; and **DTU** [15] (controlled indoor multi-view stereo), where evaluation images are center-cropped to 252×252 for fair cross-model comparison.

Baselines. To demonstrate AGCA’s generality, we integrate it into three diverse feed-forward Gaussian splatting architectures: **(1) C3G** [1] (VGGT [30] ViT-Large backbone), **(2) DepthSplat** [33] (DINOv2 [28] backbone with explicit depth priors), and **(3) AnySplat** (VGGT architecture supporting flexible input views). For all baselines, we keep the pretrained backbones entirely frozen and only train the lightweight AGCA module.

Implementation details. We train AGCA on RE10K for 10K steps using MSE loss and AdamW [25] (LR 5×10^{-5} , WD 10^{-4} , OneCycleLR with 10% warm-up), and evaluate on all four benchmarks without fine-tuning. AGCA introduces two

Table 2: Quantitative comparison on DTU [15], ScanNet [8], and ACID [22]. Higher is better for PSNR and SSIM, while lower is better for LPIPS.

Methods	DTU			ScanNet			ACID		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
AnySplat [17]	25.420	0.812	0.185	20.150	0.654	0.312	18.250	0.585	0.365
AnySplat + Ours	25.850	0.825	0.170	20.620	0.668	0.295	18.780	0.602	0.345
DepthSplat [33]	25.610	0.820	0.178	22.480	0.712	0.280	19.920	0.642	0.330
DepthSplat + Ours	26.040	0.833	0.162	23.150	0.735	0.255	20.650	0.665	0.305
C3G [1]	24.315	0.765	0.210	19.240	0.582	0.355	17.510	0.510	0.410
C3G + Ours	26.120	0.805	0.165	21.050	0.635	0.302	19.120	0.575	0.325

lightweight modules on top of frozen backbones: Continuous Attention-Gated Denisification (CAD) and Signal-Gated Representational Adaptation (SRA).

CAD is implemented as a 3-layer MLP that predicts additional Gaussians conditioned on the attention-derived capacity-demand descriptor. SRA is realized via *Signal-Gated LoRA*, which inserts rank-4 adapters into the final four attention layers, modulated by an 82-parameter MLP conditioned on the same descriptor. The descriptor itself is computed directly from self-attention maps using Attention Sharpness and Cross-Head Disagreement, without introducing additional learnable predictors.

Total trainable parameters range from 634K (AnySplat) to 1,135K (DepthSplat), adding $< 0.1\%$ to the backbones. Experiments run on a single NVIDIA H100 GPU, using bfloat16 for C3G and AnySplat, and float32 for DepthSplat.

4.1 Experimental Results and Analysis

Quantitative Comparisons. Table 1 reports quantitative results on RE10K [40] under 6-, 12-, and 24-view settings. We apply AGCA to several feed-forward 3DGS baselines, including AnySplat [17], DepthSplat [33], and C3G [1], to evaluate its plug-and-play behavior across architectures.

AGCA consistently improves reconstruction quality across models and view settings. For example, DepthSplat + Ours achieves the best PSNR in all three RE10K settings, reaching 21.020, 21.440, and 21.850 dB for 6, 12, and 24 views, respectively. Similar gains are also observed across other backbones such as AnySplat and C3G, indicating that the proposed routing strategy improves fidelity independently of the underlying architecture.

Table 2 further shows that these gains transfer to unseen datasets, including DTU [15], ScanNet [8], and ACID [22]. The consistent improvements across all three benchmarks suggest that our attention-derived capacity-demand signals capture general geometric cues rather than dataset-specific biases.

Qualitative Comparisons. Figure 3 visualizes the novel view synthesis quality of our method compared to the baselines. Vanilla feed-forward architectures (e.g.,

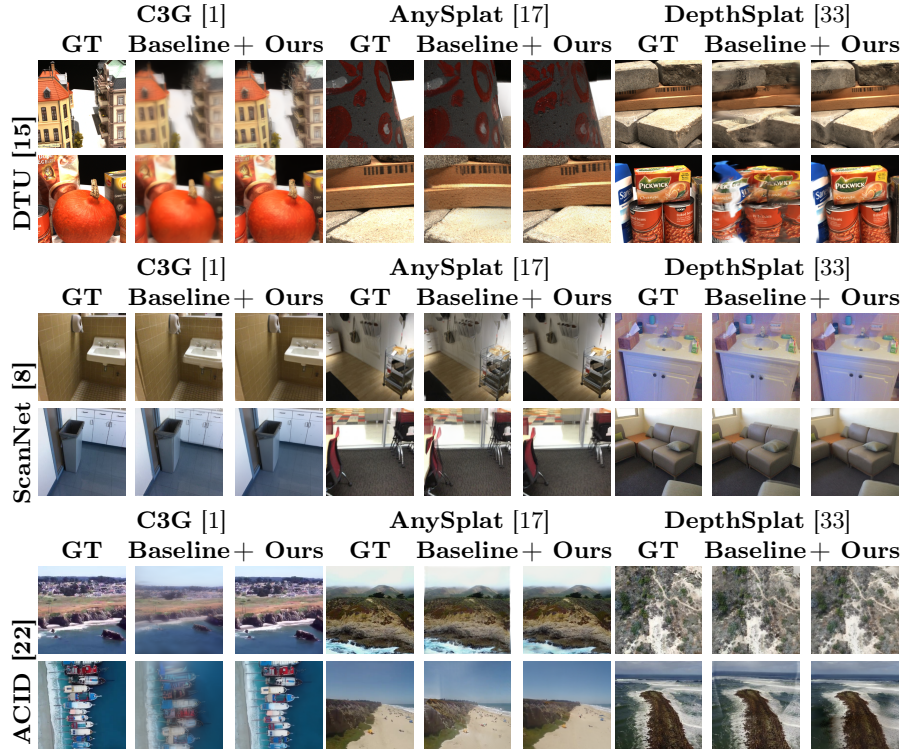


Fig. 4: Qualitative comparison on DTU [15], ScanNet [8], and ACID [22]. For each backbone, we show the ground truth, the original feed-forward baseline, and the corresponding plug-and-play integration with our method.

C3G and AnySplat) inherently suffer from severe blurring at object boundaries and loss of fine geometric details due to their static capacity bottlenecks. In contrast, our plug-and-play AGCA module effectively resolves these artifacts. By dynamically routing additional Gaussians to high-demand structural edges while safely suppressing generation in ambiguous, textureless regions, our approach recovers crisp boundaries and intricate textures without introducing noisy floaters. This significant visual enhancement aligns perfectly with our quantitative gains, demonstrating the optimal redistribution of representational capacity.

4.2 Ablation Study

We analyze the contributions of CAD and SRA on RE10K using the C3G backbone (Table 3). Starting from the baseline, CAD alone (Row 2) improves PSNR by **+0.55 dB**, while SRA alone (Row 3) adds **+0.42 dB**, validating their individual effectiveness. Enabling both with uniform gating (Row 4) yields a **+0.95 dB** gain, which is further enhanced to **20.44 dB** (a total gain of **+1.44 dB**) by our attention-derived shared gate $\mathbf{g}_i = \sigma(\text{MLP}([\hat{\mathcal{S}}_i, \hat{\mathcal{D}}_i]))$ (Row

Table 3: Ablation study of the core components. Shared Gate uses the same attention-derived signal for both branches.

CAD	SRA	Shared	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
			19.00	0.561	0.351
✓			19.55	0.578	0.338
	✓		19.42	0.574	0.341
✓	✓		19.95	0.593	0.325
✓	✓	✓	20.44	0.614	0.313

5). These results demonstrate that intrinsic sharpness $\mathcal{S}$ and disagreement $\mathcal{D}$ provide a superior routing criterion compared to naive adaptation. Notably, the full model outperforms the sum of individual gains, suggesting a synergistic interaction where CAD redistributes density while SRA enhances representation in demanding regions. Overall, this confirms that joint capacity allocation via a shared descriptor is more effective than independent or uniform strategies.

5 Conclusion

We presented AGCA, a plug-and-play framework designed to address the static capacity bottleneck in feed-forward 3D Gaussian Splatting. By leveraging intrinsic signals such as Attention Sharpness and Cross-head Disagreement, our method identifies structurally demanding regions without the need for auxiliary supervision or test-time optimization. This approach jointly improves spatial and representational capacity in a unified manner, effectively reducing common artifacts like boundary blur and floaters. While our work currently focuses on static reconstruction, extending these cues to dynamic 4D scenarios remains a challenge because temporal consistency and non-rigid deformations introduce significant complexity. We hope this attention-driven perspective provides a scalable path toward more adaptive and robust 3D reconstruction.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2022-0-00022, RS-2022-II220022, Development of immersive video spatial computing technology for ultra-realistic metaverse services) (40%); the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2026 (Project Name: Multimodal Storyverse: A Multi-Agent Collaborative Platform for Expansive Narrative Creation, Project Number: RS-2026-25527029, Contribution Rate: 40%); and the Artificial Intelligence Innovation Graduate School (Sogang University) grant funded by the Korea government (MSIT) (RS-2026-25547954) (20%).

References

1. An, H., Jung, J., Kim, M., Hong, S., Kim, C., Fukuda, K., Jeon, M., Han, J., Narihira, T., Ko, H., et al.: C3g: Learning compact 3d representations with 2k gaussians. arXiv preprint arXiv:2512.04021 (2025)
2. Baron-Lis, K., Rottmann, M., Mütze, A., Honari, S., Fua, P., Salzmann, M.: At-entropy: On the generalization ability of supervised semantic segmentation transformers to new objects in new domains. In: BMVC (2024)
3. Cen, J., Fang, J., Yang, C., Xie, L., Zhang, X., Shen, W., Tian, Q.: Segment any 3d gaussians. In: Proceedings of the AAAI conference on artificial intelligence. vol. 39, pp. 1971–1979 (2025)
4. Cerkezi, L., Davtyan, A., Sameni, S., Favaro, P.: Mirage: Unsupervised single image to novel view generation with cross attention guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3740–3750 (2025)
5. Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: 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)
6. Chen, Q., Wang, C., Wang, D., Zhang, T., Li, W., He, X.: Lifelong knowledge editing for vision language models with low-rank mixture-of-experts. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9455–9466 (2025)
7. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In: European conference on computer vision. pp. 370–386. Springer (2024)
8. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Fan, Z., Cong, W., Wen, K., Wang, K., Zhang, J., Ding, X., Xu, D., Ivanovic, B., Pavone, M., Pavlakos, G., et al.: Instantsplat: Sparse-view gaussian splatting in seconds. arXiv preprint arXiv:2403.20309 (2024)
11. Hong, S., Jung, J., Shin, H., Han, J., Yang, J., Luo, C., Kim, S.: Pf3plat: Pose-free feed-forward 3d gaussian splatting for novel view synthesis. In: Forty-second International Conference on Machine Learning (2025)
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
13. Huang, S.Y., Chou, Z.T., Wang, Y.C.F.: 3d gaussian inpainting with depth-guided cross-view consistency. In: Proceedings of the computer vision and pattern recognition conference. pp. 26704–26713 (2025)
14. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: International Conference on Learning Representations (2017)
15. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanaes, H.: Large scale multi-view stereopsis evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2014)

16. Jiang, H., Tan, H., Wang, P., Jin, H., Zhao, Y., Bi, S., Zhang, K., Luan, F., Sunkavalli, K., Huang, Q., et al.: Rayzer: A self-supervised large view synthesis model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4918–4929 (2025)
17. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., et al.: Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *ACM Transactions on Graphics (TOG)* **44**(6), 1–16 (2025)
18. Jin, H., Jiang, H., Tan, H., Zhang, K., Bi, S., Zhang, T., Luan, F., Snavely, N., Xu, Z.: LVSM: A large view synthesis model with minimal 3d inductive bias. In: The Thirteenth International Conference on Learning Representations (2025)
19. Kang, G., Yoo, J., Park, J., Nam, S., Im, H., Shin, S., Kim, S., Park, E.: Selfsplat: Pose-free and 3d prior-free generalizable 3d gaussian splatting. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22012–22022. IEEE (2025)
20. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **42**(4), 139–1 (2023)
21. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
22. Liu, A., Tucker, R., Jampani, V., Makadia, A., Snavely, N., Kanazawa, A.: Infinite nature: Perpetual view generation of natural scenes from a single image. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14458–14467 (2021)
23. Liu, M., Shi, R., Chen, L., Zhang, Z., Xu, C., Wei, X., Chen, H., Zeng, C., Gu, J., Su, H.: One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10072–10083 (2024)
24. Liu, S.Y., Wang, C.Y., Yin, H., Molchanov, P., Wang, Y.C.F., Cheng, K.T., Chen, M.H.: Dora: Weight-decomposed low-rank adaptation. In: Forty-first International Conference on Machine Learning (2024)
25. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017)
26. Miao, S., Huang, J., Bai, D., Yan, X., Zhou, H., Wang, Y., Liu, B., Geiger, A., Liao, Y.: Evolsplat: Efficient volume-based gaussian splatting for urban view synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11286–11296 (2025)
27. Moreau, A., Shaw, R., Nazarczuk, M., Shin, J., Tanay, T., Zhang, Z., Xu, S., Pérez-Pellitero, E.: Off the grid: Detection of primitives for feed-forward 3d gaussian splatting. *arXiv preprint arXiv:2512.15508* (2025)
28. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023)
29. Tian, Q., Tan, X., Xie, Y., Ma, L.: Drivingforward: Feed-forward 3d gaussian splatting for driving scene reconstruction from flexible surround-view input. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7374–7382 (2025)
30. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)

31. Wang, Y., Chai, L., Luo, X., Niemeyer, M., Lagunas, M., Lombardi, S., Tang, S., Sun, T.: Learning efficient fuse-and-refine for feed-forward 3d gaussian splatting. arXiv preprint arXiv:2503.14698 (2025)
32. Weinzaepfel, P., Leroy, V., Lucas, T., Brégier, R., Cabon, Y., Arora, V., Antsfeld, L., Chidlovskii, B., Csurka, G., Revaud, J.: Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems* **35**, 3502–3516 (2022)
33. Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., Geiger, A., Pollefeys, M.: Depth-splat: Connecting gaussian splatting and depth. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 16453–16463 (2025)
34. Xu, Y., Shi, Z., Yifan, W., Chen, H., Yang, C., Peng, S., Shen, Y., Wetzstein, G.: Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In: *European Conference on Computer Vision*. pp. 1–20. Springer (2024)
35. Ye, B., Liu, S., Xu, H., Li, X., Pollefeys, M., Yang, M.H., Peng, S.: No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In: *International Conference on Learning Representations*. vol. 2025, pp. 54009–54033 (2025)
36. Yoon, Y., Yoon, K.J.: Resplat: Degradation-agnostic feed-forward gaussian splatting via self-guided residual diffusion. In: *The Fourteenth International Conference on Learning Representations* (2026)
37. Zhai, S., Likhomanenko, T., Littwin, E., Busbridge, D., Ramapuram, J., Zhang, Y., Gu, J., Susskind, J.M.: Stabilizing transformer training by preventing attention entropy collapse. In: *International Conference on Machine Learning*. pp. 40770–40803. PMLR (2023)
38. Zhang, K., Bi, S., Tan, H., Xiangli, Y., Zhao, N., Sunkavalli, K., Xu, Z.: Gs-lrm: Large reconstruction model for 3d gaussian splatting. In: *European Conference on Computer Vision*. pp. 1–19. Springer (2024)
39. Zhang, Q., Chen, M., Bukharin, A., Karampatziakis, N., He, P., Cheng, Y., Chen, W., Zhao, T.: Adaptive budget allocation for parameter-efficient fine-tuning. In: *The Eleventh International Conference on Learning Representations* (2023)
40. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph.* **37**(4), 65:1–65:12 (2018)
41. Zou, Z.X., Yu, Z., Guo, Y.C., Li, Y., Liang, D., Cao, Y.P., Zhang, S.H.: Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 10324–10335. IEEE Computer Society (2024)