SA-RESGS: SELF-AUGMENTED RESIDUAL 3D GAUSSIAN SPLATTING FOR NEXT BEST VIEW SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose Self-Augmented Residual 3D Gaussian Splatting (SA-ResGS), a novel framework for stabilizing uncertainty quantification and enhancing uncertainty-aware supervision in next-best-view (NBV) selection for active scene reconstruction. SA-ResGS improves both the reliability of uncertainty estimates and their effectiveness for supervision by generating Self-Augmented point clouds (SA-Points) via triangulation between a training view and a rasterized extrapolated view, enabling efficient scene coverage estimation. While improving scene coverage through physically guided view selection, SA-ResGS also addresses the challenge of under-supervised Gaussians, exacerbated by sparse and wide-baseline views, by introducing the first residual learning strategy tailored for 3D Gaussian Splatting. This targeted supervision enhances gradient flow in high-uncertainty Gaussians by combining uncertainty-driven filtering with dropout- and hard-negative-mining-inspired sampling. Our contributions are threefold: (1) a physically grounded view selection strategy that promotes efficient and uniform scene coverage; (2) an uncertainty-aware residual supervision scheme that amplifies learning signals for weakly contributing Gaussians, improving training stability and uncertainty estimation across scenes with diverse camera distributions; (3) an implicit unbiasing of uncertainty quantification as a consequence of constrained view selection and residual supervision, which together mitigate conflicting effects of wide-baseline exploration and sparse-view ambiguity in NBV planning. Experiments on active view selection demonstrate that SA-ResGS outperforms state-of-the-art baselines in both reconstruction quality and view selection robustness.

1 INTRODUCTION

Recent advances in neural rendering—particularly Neural Radiance Fields (NeRFs) (Mildenhall et al., 2020) and 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023)—have significantly advanced photorealistic scene reconstruction (Yu et al., 2024; Niedermayr et al., 2024; Kulhanek et al., 2024), enabling high-fidelity, real-time applications across diverse environments. Beyond static scene capture, these methods have spurred broader interest in tackling complex challenges, such as active view selection (Wang et al., 2023; Xiao et al., 2024; Chen et al., 2024), uncertainty quantification for next-best-view (NBV) selection (Wen et al., 2024). While pre-captured, dense-view training methods can achieve impressive reconstruction quality, in-situ (active) reconstruction—where views are selected and added progressively—remains challenging due to artifacts caused by shape-radiance ambiguity and further exacerbated by limited training views and dynamics of view addition strategy. Despite the inherent difficulty of estimating reliable uncertainty under such conditions, recent post-hoc approaches—such as Laplacian approximation-based, model-agnostic methods (Wen et al., 2024; Goli et al., 2024)—have shown promise by providing uncertainty signals without altering the rendering pipeline. However, several critical challenges remain unaddressed:

- **Disregarded physical constraints:** Computational uncertainty is often misaligned with physical plausibility of the reconstructed geometry.
- **Underutilized supervision:** Existing methods rarely convert uncertainty cues into actionable learning signals, leaving weakly contributing Gaussians under-supervised throughout training.
- **Performance dependency:** The reliability of uncertainty estimation remains tightly coupled with training dynamics, particularly in the early stages when scene coverage is incomplete.

In response to these challenges, we propose SA-ResGS, a Self-Augmented Residual 3D Gaussian Splatting framework that stabilizes uncertainty quantification and enhances uncertainty-aware supervision in next-best-view selection for progressive scene reconstruction, as shown in Fig. 1. SA-ResGS strategically decouples view selection from strong dependence on uncertainty estimates driven by internal learning dynamics, promoting more robust and geometry-aware surface coverage. Concretely, we physically prefilter a subset of candidate views based on their geometric dissimilarity and then apply uncertainty-based scoring within this subset—effectively implementing a physically grounded, uncertainty-informed selection strategy. To guide this process, we construct SA-Points by triangulating dense correspondences between a given training view and its rasterized extrapolated views at each view-selection step, after a fixed number of initial views have been trained. These SA-Points are then encoded using a hash-based scene representation, enabling efficient similarity measurement between candidate views and pre-selected views. We select the most dissimilar candidate views in the encoded space, encouraging coverage of previously unseen regions.

While the physically grounded and uncertainty-informed selection strategy enhances overall scene coverage, it inadvertently increases sparsity in multiview overlap—since more dissimilar views are less likely to observe shared regions, thereby weakening multi-view geometric constraints. To counterbalance this contradiction without sacrificing the benefits of diverse view selection, we introduce a residual supervision mechanism that reinforces training using ground truth RGB images. The proposed SA-ResGS—particularly its residual supervision module—is fully compatible with conventional 3DGS pipelines that rely on point-cloud-based rasterization and gradient-based optimization. Beyond the original supervision, SA-ResGS additionally rasterizes color images using a selected Gaussian subset: a small fraction of the most uncertain Gaussians is combined with a majority subset (e.g., 90%) of the original visible Gaussians. This strategy mirrors the principles of residual learning and draws conceptual inspiration from Dropout (Park et al., 2025; Srivastava et al., 2014) and Hard Negative Mining (Xuan et al., 2020; Jang et al., 2019), by amplifying supervision for Gaussians that typically receive weak gradients during backpropagation—similar to the role of skip connections in ResNet (He et al., 2016) in mitigating vanishing gradients. Our method thus provides additional learning signals specifically targeted at under-optimized Gaussians which are often overlooked due to their low contribution in 3DGS's rendering mechanism, receive repeated 2D photometric supervision, resulting in more stable optimization and improved reconstruction in sparse or ambiguous regions.

The main contributions of the SA-ResGS are threefold:

- **Physically grounded view selection:** We propose a geometry-aware strategy using Self-Augmented Points (SA-Points) to guide next-best-view selection, enforcing physical plausibility and promoting more balanced, coverage-oriented exploration.
- **Residual learning for 3DGS:** We introduce the first residual supervision framework specifically designed for 3D Gaussian Splatting, addressing the vanishing gradient problem by reinforcing weakly supervised Gaussians and improving both optimization stability and reconstruction quality.
- **Unbiased uncertainty quantification:** By jointly improving view distribution and supervising under-optimized Gaussians, SA-ResGS mitigates geometric sparsity and density bias, leading to fairer and more reliable uncertainty estimates throughout training.

2 Related Works

Next-best-view selection. NBV selection originated from the robotics community as a strategy to efficiently guide in-situ scene capture, where the goal was to incrementally select viewpoints that maximally reduce reconstruction ambiguity with minimal sensor movement (Connolly, 1985; Scott et al., 2003; Delmerico et al., 2018). Classical NBV methods typically followed rule-based paradigms, selecting views based on geometric coverage (Dunn and Frahm, 2009; Guédon et al., 2022; Guédon et al., 2023), viewpoint entropy (Vázquez et al., 2001), or visibility heuristics (Bircher et al., 2016; Sun et al., 2021). As NBV became increasingly relevant to 3D computer vision, especially under sparse-view constraints, learning-based approaches emerged to model scene-specific view policies via reinforcement learning or active learning frameworks (Wang et al., 2024). While these data-driven methods demonstrate improved adaptability over hand-crafted rules, they often rely on task-specific reward definitions and struggle to generalize across scene types. More recent efforts explore principled information-theoretic formulations, such as FisherRF (Wen et al., 2024), to formalize view selection in neural fields. In our work, we build on this line (Wen et al., 2024; Goli et al., 2024; Hanson et al.,

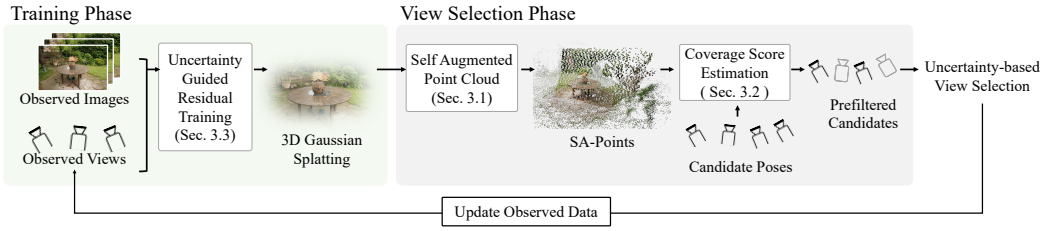

Figure 1: **Overview of SA-ResGS.** The framework alternates between view selection and training. At each NBV step, Self-Augmented Points are generated via triangulation from dense correspondences between a training view and its extrapolated render, enabling surface-aware coverage estimation (Sec.3.1). Candidate views are first physically filtered using hash-encoded feature dissimilarity, then ranked by uncertainty quantification scores for final selection (Sec.3.2). During training, residual supervision (Sec. 3.3) combines full and uncertainty-intensified renders to reinforce gradients toward weakly contributing Gaussians, improving training stability and reconstruction quality under sparse-view conditions.

2024; Wilson et al., 2025) by integrating physically grounded geometry priors to further stabilize early-stage view planning, particularly when minimal visual input is available.

Uncertainty quantification for 3DGS and neural rendering. Uncertainty quantification plays a pivotal role in active reconstruction, particularly for guiding view selection and supervision. In 3D Gaussian Splatting (3DGS), however, the high dynamic nature of primitive splitting and sensitivity to initialization often leads to unstable training, degrading the reliability of intermediate uncertainty signals. Earlier methods based on variational inference (Shen et al., 2021; Lee et al., 2025; Shen et al., 2022; Lyu et al., 2024) and ensemble-based estimates (Sünderhauf et al., 2023) enable stochastic or distributional reasoning, but they require costly model retraining or parallel inference and are typically incompatible with standard rendering pipelines. Recent post-hoc approaches (Wilson et al., 2025; Hanson et al., 2024) such as FisherRF (Wen et al., 2024) and BayesRays (Goli et al., 2024) estimate uncertainty using Laplacian approximations without altering model structure, and have demonstrated promising results on NeRF and 3DGS variants. Complementary to geometry-based methods, image-level approach (Wang et al., 2025) leverage perceptual quality on current rendering result as a proxy for uncertainty. However, these methods remain strongly coupled with the density of underlying Gaussians—leading to biased uncertainty estimates in early training stages when geometry is sparse or unevenly distributed, often misinterpreting under-observed regions as confident. This overlooked bias limits the reliability of NBV guidance when it's most needed. To mitigate limitation, we introduce residual learning with the help of using physically grounded view selection, enabling robust and a little loose-coupled uncertainty estimation in an early-stage of view selection while emphasizing the effect of targeting high-uncertainty Gaussian focused supervision.

Residual supervision in 3DGS. While accurate uncertainty estimation helps localize regions requiring stronger supervision, it alone does not guarantee that gradients effectively reach under-optimized Gaussians in the 3DGS pipeline. Residual learning, as popularized by ResNet (He et al., 2016), has proven effective in mitigating vanishing gradients and improving training stability through skip connections and additive refinement, yet it remains underexplored in the context of 3D Gaussian Splatting. Existing 3DGS methods mainly rely on direct photometric losses (Kerbl et al., 2023) or external depth priors (Li et al., 2024; Xu et al., 2024), which often fail to sufficiently supervise Gaussians with low opacity or minimal rendering contributions. Recent studies such as pixelSplat (Charatan et al., 2024), PAPR (Zhang et al., 2023), and PAPR-in-Motion (Peng et al., 2024) explicitly discuss the vanishing gradient issue and propose solutions including differentiable parameterization of Gaussians, proximity attention-based differentiable renderer, adaptive updates, and activation tuning. Despite various strategies proposed to mitigate the vanishing gradient problem, prior approaches lack an explicit mechanism for correcting weakly supervised Gaussians, which remain largely unresolved due to insufficient gradient signals. Although dropout-based approaches (Park et al., 2025) help increase gradient diversity, they operate stochastically and do not target supervision to the most uncertain or least updated Gaussians. Our method addresses these limitations by introducing the first residual supervision strategy for 3DGS, applying uncertainty-guided rendering to intentionally amplify gradients for under-supervised Gaussians—without altering underlying rasterization process.

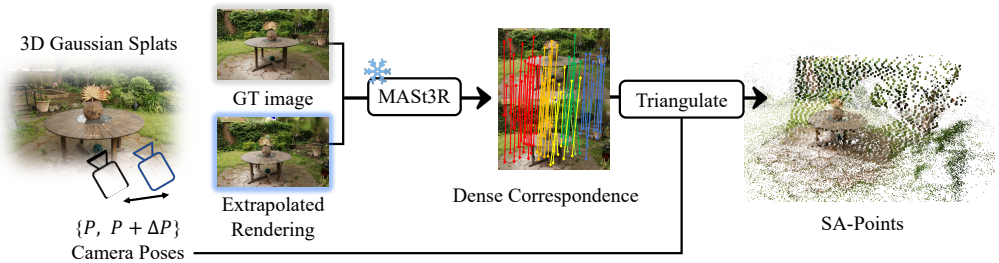

Figure 2: **SA-Points Generation.** An extrapolated image is rendered from a perturbed camera pose. Dense correspondences with the reference image are predicted using MASt3R, and triangulated to produce SA-Points, which are filtered by reprojection error for reliable surface geometry.

3 Self-Augmented Residual 3D Gaussian Splatting

The proposed SA-ResGS framework is illustrated in Fig. 1. SA-ResGS builds upon the state-of-the-art next-best-view selection method, FisherRF (Wen et al., 2024), extending it with SA-Points to support two core ideas: (1) guiding physically grounded view selection with reduced reliance on uncertainty estimation, and (2) applying residual supervision to uncertain Gaussians, mitigating the vanishing gradient problem, wherein weakly contributing Gaussians—those with minimal impact on rasterized pixel values—receive insufficient gradients during backpropagation. The physically grounded view selection is enabled by a geometrically encoded surface representation, constructed using SA-Points derived from a single training view. To construct SA-Points, we employ the 3D vision foundational model, MASt3R (Leroy et al., 2024), to predict dense correspondences between a given training image and a rasterized extrapolated view rendered from a nearby camera pose. The resulting 2D correspondences are then triangulated to produce 3D SA-Points.

3.1 Self-Augmented Points Generation

Given a reference image I_r with camera pose $\mathbf{T}_r = [\mathbf{R}_r \mid \mathbf{t}_r]$, we render an extrapolated image I_e from a perturbed pose $\mathbf{T}_e = [\mathbf{R}_r \mid \mathbf{t}_r + \Delta\mathbf{t}]$ using 3D Gaussian Splatting (Kerbl et al., 2023). Dense correspondences $\{(\mathbf{p}_r^i, \mathbf{p}_e^i)\}$ between I and I_e are predicted using the pretrained MASt3R model (Leroy et al., 2024), which is robust to moderate viewpoint changes and capable of producing contextually meaningful matches even in the presence of minor geometric distortions. Each SA-Point $\mathbf{X}^i$ is triangulated from a 2D correspondence pair using the projection matrices $\mathbf{P}_r$ and $\mathbf{P}_e$ derived from COLMAP, including intrinsic matrices over each extrinsic pose $\mathbf{T}$. However, because triangulation is performed repeatedly during training—while the model is still fitting to a sparse and incomplete geometry—rasterized extrapolated images may occasionally contain rendering noise due to inaccurately placed Gaussians. To ensure reliable geometry while fully leveraging the generalization capability of MASt3R, we apply reprojection error-based filtering:

$$\varepsilon^i = \frac{1}{2}\left(\left\|\mathbf{p}_r^i - \pi(\mathbf{P}_r\mathbf{X}^i)\right\| + \left\|\mathbf{p}_e^i - \pi(\mathbf{P}_e\mathbf{X}^i)\right\|\right), \quad \text{retain if } \varepsilon^i < \tau. \tag{1}$$

This filtering step discards geometrically inconsistent points while preserving accurate SA-Points from dense, context-aware matches—even when the extrapolated image is noisier than the original training view. Compared to prior methods such as CoMapGS (Jang and Pérez-Pellitero, 2025) or MP-SfM (Pataki et al., 2025), our triangulation pipeline produces scale-consistent, surface-aware geometry from a single image by leveraging extrapolated viewpoints rather than requiring multiview input or monocular depth estimates. The overall steps are visualized in Fig. 2.

3.2 Physically Grounded View Selection Algorithm

We present our physically grounded view selection algorithm for next-best-view (NBV) selection, illustrated in Fig. 3. As discussed in Sec. 2, NBV selection in 3D Gaussian Splatting (3DGS) is particularly challenging due to the tight coupling between uncertainty estimation and the quality of reconstructed geometry—both of which are highly sensitive to the sparsity and distribution of Gaussian splats. Under sparse-view settings, where reconstruction begins with as few as four images and new

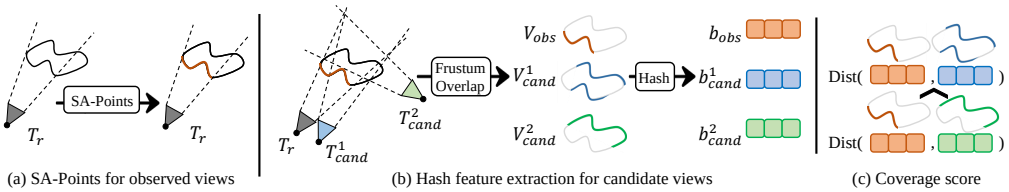

Figure 3: **Physically grounded candidate view selection via surface coverage.** (a) SA-Points from training views define observed voxels $\mathcal{V}_{\text{obs}}$. (b) Each candidate view generates a binary hash-encoded feature $\mathbf{b}$, via frustum-based visibility estimation. (c) Normalized Hamming distance between hash-encoded features quantifies coverage dissimilarity, enabling efficient selection of geometrically complementary views without rendered images or uncertainty scores.

views are incrementally added every 100 training iterations, uncertainty-based NBV strategies often become unreliable. This is because uncertainty signals are inherently biased or unstable when the geometry is incomplete or under-constrained. To address this, we introduce a surface-aware guidance mechanism based on SA-Points, enabling view selection to operate independently of the computed uncertainty quantification. By decoupling view selection from the internal training dynamics of 3DGS, our method provides more stable and physically meaningful candidate views during the early reconstruction phase—even before the model has accumulated sufficient confidence to produce reliable uncertainty maps.

We begin by discretizing the 3D scene into a voxel grid $\mathcal{V} = \{v_k\}_{k=1}^{K}$, where each voxel represents a unit volume in the scene. The bounding volume of $\mathcal{V}$ is defined by the sparse point cloud obtained from structure-from-motion (SfM). A voxel $v_k \in \mathcal{V}$ is marked as observed if it intersects with any SA-Point $\mathbf{X}^i$ (Sec. 3.1), forming the subset $\mathcal{V}_{\text{obs}} \subset \mathcal{V}$. To account for potential localization errors and promote coverage continuity, we dilate each occupied voxel using a 3D kernel $\mathcal{K}_r$ of radius r:

$$\tilde{\mathcal{V}}_{\text{obs}} = \bigcup_{v_k \in \mathcal{V}_{\text{obs}}} \mathcal{K}_r(v_k), \tag{2}$$

where $\tilde{\mathcal{V}}_{\text{obs}}$ denotes the dilated observed region for the current set of training views.

For each candidate view j, we compute a frustum $\mathcal{F}_j \subset \mathcal{V}$, derived from camera intrinsics (field of view) and near/far planes estimated from the SfM point distribution. A voxel is considered potentially visible from view j if its center lies within the frustum:

$$\mathcal{V}^{(j)}_{\text{cand}} = \{v_k \in \mathcal{V} \mid v_k \in \mathcal{F}_j\}. \tag{3}$$

To estimate the geometric dissimilarity between the current coverage and a candidate view (see Fig. 3), we compute the normalized Hamming distance:

$$d_j = \frac{1}{K} \left\| \mathbf{b}_{\text{obs}} \oplus \mathbf{b}^{(j)}_{\text{cand}} \right\|_1, \tag{4}$$

where $\oplus$ denotes the element-wise XOR operation between binary vectors , and $\|\cdot\|_1$ is the ℓ_1 norm (i.e., the number of differing entries). Here, $\mathbf{b}$ denotes a binary occupancy vector obtained by mapping voxel coordinates through a fixed random hashing function, following the spatial hashing strategy introduced in Instant-NGP (Müller et al., 2022). The resulting value $d_j \in [0, 1]$ measures the proportion of voxels with inconsistent occupancy status between the currently observed volume and the candidate view. Candidate views are then ranked in descending order of their normalized Hamming distances d_j, and the top $N\%$ (e.g., $N = 20$) are retained to form the physically filtered candidate set $\mathcal{C}'$.

We apply uncertainty quantification only within $\mathcal{C}'$, finalizing the view selection with finer-level scoring. This two-stage pipeline follows a coarse-to-fine strategy: it first expands the observed surface area using explicit geometric cues from SA-Points, then refines the choice using uncertainty-aware reasoning. By restricting uncertainty estimation to a smaller candidate pool, this approach improves computational efficiency while maintaining scene-aware diversity in the selected views. This strategy enables balanced, physically grounded exploration of unobserved regions.

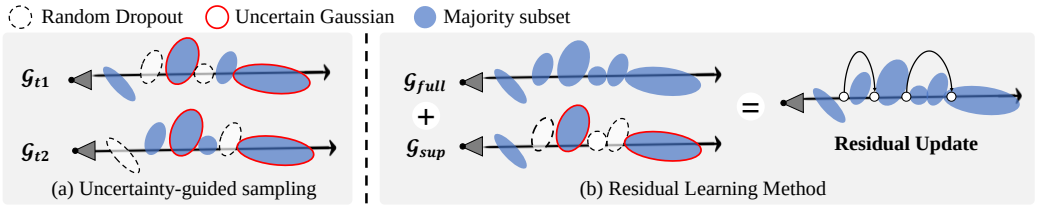

Figure 4: **Residual supervision in 3DGS.** (a) At each iteration (t_1, t_2), $\mathcal{G}_{\text{sup}}$ combines random and top-uncertain Gaussians; (b) residual supervision in 3DGS mimics ResNet-style skip connections.

3.3 Uncertainty-Guided Residual Learning in 3DGS

We propose the first residual learning for 3DGS to address the vanishing gradient issue affecting weakly contributing Gaussians, as shown in Fig. 4. These Gaussians often receive insufficient supervision during training due to their limited impact on rasterized pixels—particularly in sparse or ambiguous regions. While ResNet (He et al., 2016) mitigates similar issues via skip connections, such mechanisms are infeasible in 3DGS due to the dynamic and view-dependent nature of Gaussian properties. Instead, we propose a rasterizer-agnostic strategy that enhances gradient flow by generating auxiliary renders that emphasize high-uncertainty Gaussians. These images are supervised with ground-truth RGB images, forming the basis for a residual supervision scheme detailed below.

Residual supervision. To reinforce under-supervised Gaussians, we introduce a residual supervision scheme that leverages two rasterized images from the same training viewpoint: one using the full set of Gaussians $\mathcal{G}$ and another from a guided subset $\mathcal{G}_{\text{sup}}$, as shown in Fig. 4(a). We define this subset as:

$$\mathcal{G}_{\text{sup}} = \mathcal{G}_{\text{rand}} \cup \mathcal{G}_{\text{uncertain}}, \tag{5}$$

where $\mathcal{G}_{\text{rand}}$ is a random sample comprising $\alpha\%$ of $\mathcal{G}$ (e.g., α=90), and $\mathcal{G}_{\text{uncertain}}$ contains the top-β most uncertain Gaussians (e.g., β=10). To estimate uncertainty, we analyze two per-Gaussian attributes: opacity and scale. Gaussians with low opacity contribute minimally to alpha blending during rasterization, while those with large scale blur across pixels and tend to dominate ambiguous or low-texture regions. This rank identifies Gaussians that are both visually suppressed and spatially diffuse—making them key targets for correction. The combination $\mathcal{G}_{\text{sup}}$ ensures that $\mathcal{G}_{\text{rand}}$ maintains overall scene fidelity, while $\mathcal{G}_{\text{uncertain}}$ provides targeted supervision to gradient-deficient areas.

We compute two rendered images: I_{full} using the full set of Gaussians $\mathcal{G}$, and I_{sup} using the uncertainty-intensified subset $\mathcal{G}_{\text{sup}}$. Each image is supervised independently against the ground-truth image I_{gt} using ℓ_1 and SSIM losses:

$$\mathcal{L} = \sum_{i \in \{\text{full},\text{sup}\}} \lambda_i \left[\mathcal{L}_{\text{rgb}}(I_i, I_{\text{gt}}) + \mathcal{L}_{\text{ssim}}(I_i, I_{\text{gt}})\right], \tag{6}$$

where $\lambda_{\text{full}} + \lambda_{\text{sup}} = 1$, and we set both to 0.5 in practice. The uncertainty-intensified rasterization strategy is conceptually inspired by Dropout (Park et al., 2025; Srivastava et al., 2014) and Hard Negative Mining (Xuan et al., 2020; Jang et al., 2019). Randomly sampling $\mathcal{G}_{\text{rand}}$ provides stochastic diversity, allowing weakly contributing Gaussians to be supervised when dominant ones are excluded. Meanwhile, the deterministic inclusion of $\mathcal{G}_{\text{uncertain}}$ ensures consistent gradient flow to Gaussians that are persistently under-optimized. This dual mechanism reinforces learning in uncertain or ambiguous regions without modifying the rasterization process, and complements full-image supervision to maintain global photometric fidelity.

By supervising both the full and uncertainty-intensified images, we promote stronger gradient flow toward uncertain or low-opacity Gaussians without compromising photometric quality. This strategy mirrors the role of residual skip connections in ResNet (He et al., 2016) (Fig. 4(b)), supporting more stable convergence and mitigating overfitting in sparse or wide-baseline training settings. It is particularly effective in the early stages of next-best-view selection, where reconstruction is sensitive to both sparsely initialized regions and supervision bias caused by overfitting to limited views.

Category	Methods	PSNR↑	SSIM↑	LPIPS↓
Rule-based	Random	19.580	0.574	0.460
	ACP	20.458	0.598	0.446
2D-based	MUSIQ	19.858	0.577	0.464
	CrossScore	20.981	0.613	0.441
3D-based	ActiveNeRF	17.889	0.533	**0.414**
	FisherRF	20.814	0.601	0.447
	Ours	**21.325**	**0.610**	0.450

(a) Mip-NeRF 360 dataset

Category	Methods	PSNR↑	SSIM↑	LPIPS↓
Rule-based	Random	18.903	0.695	0.386
	ACP	19.766	0.714	0.374
2D-based	MUSIQ	18.424	0.687	0.404
	CrossScore	19.851	**0.728**	**0.364**
3D-based	FisherRF	19.777	0.722	0.368
	Ours	**20.189**	0.726	0.370

(b) Deep Blending & Tank and Temples dataset

Table 1: **Quantitative results for the Active View Selection.** We compare our model with (1) Rule-based models (Random, ACP (Kopanas and Drettakis, 2023)), (2) 2D-based models (MUSIQ, CrossScore (Chen et al., 2024)), and 3D-based models (ActiveNeRF (Pan et al., 2022), FisherRF (Wen et al., 2024)). Results are averaged over 9 scenes from the Mip-NeRF 360 and 7 additional scenes from the Deep Blending and Tanks and Temples.

4 Experimental Results

Dataset. We evaluate our approach on benchmark dataset: Mip-NeRF 360 (Barron et al., 2022). While Mip-NeRF 360 comprises nine real-world scenes with full 360-degree coverage, its uniform and curated camera trajectories provide limited challenge for active view selection, since even simple heuristics (e.g., furthest-distance selection) already perform reliably under balanced coverage (Xiao et al., 2024). To address this limitation, we carefully curate an extended benchmark dataset inlucindg seven diverse scenes from Deep Blending (Hedman et al., 2018) and Tanks and Temples (Knapitsch et al., 2017), which introduce unbalanced view distributions and varied scene scales that better reflect practical conditions. All experiments are conducted using images at their original resolutions, and further details on dataset curation are provided in Sec. A of the Appendix.

Counterparts. We compare our method quantitatively and qualitatively against several active 3DGS baselines which operate solely on RGB images: FisherRF (Wen et al., 2024), ActiveNeRF (Pan et al., 2022), ACP (Kopanas and Drettakis, 2023), and random view selection. We also included 2D based view selection methods adopted from Active View Selector framework (Wang et al., 2025). Following this framework, we incorporated two image quality assessment (IQA) models (MUSIQ and CrossScore) to evaluate perceptual quality. Both models were re-implemented according to the authors' official instructions and publicly available code.

4.1 Active View Selection

Experimental settings. Following the experimental protocols outlined by Wen et al. (2024), we adopt their prescribed initial view configurations and view selection schedules. Specifically, our experiments initiate with four uniformly distributed views, subsequently selecting an additional view every 100 epochs until reaching a total of 20 training views (For fewer training views, see Appendix Sec. C). We apply the same active selection strategy consistently across all datasets. For consistency, each model is initialized with the same random seed and trained for 20,000 iterations. All other settings remain unchanged across experiments, with exception of the view selection algorithms.

Results. Quantitative and qualitative results for the Mip-NeRF 360 dataset and additional scenes from the Deep Blending and Tank and Temples datasets are summarized in Table 1 and Fig. 5. The counterparts exhibit limited performance in 3D reconstruction, particularly in regions with sparse observations, due primarily to biased view selection and overfitting caused by vanishing gradients. This results in incomplete reconstructions, characterized by floating artifacts and missing geometry, *i.e.* holes and missing objects. In contrast, our method consistently delivers improved reconstruction quality, ranking first in PSNR and SSIM metrics and second in LPIPS for the Mip-NeRF 360 dataset. Our uncertainty-guided residual learning approach, based on dropout effects, produces smoother reconstructions in uncertain regions without compromising comparative performance. For camera view distribution, see Appendix Sec. B.

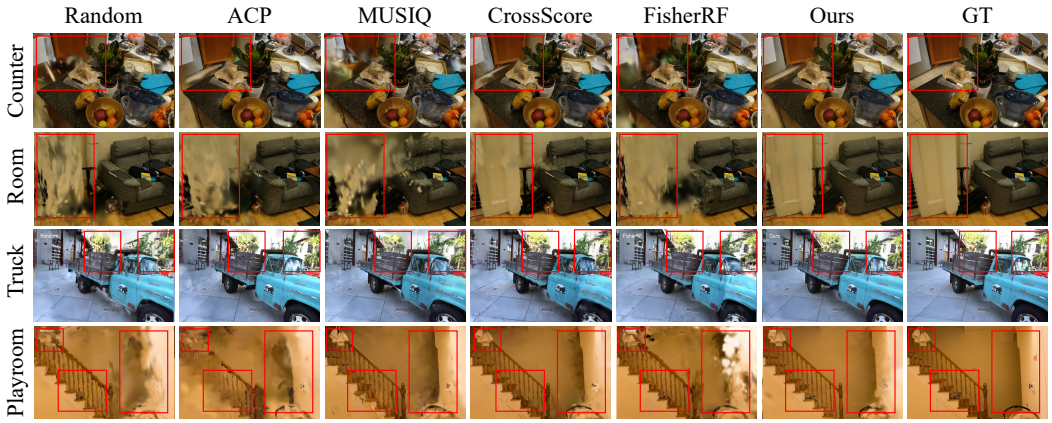

Figure 5: **Qualitative Comparison of Active View Selection.** Reconstruction from 20 selected views per scene. Our method shows improved completeness and fewer artifacts compared to baselines. For multi-view visualization, please refer the Appendix. Sec. C and supplementary video.

Additionally, experimental outcomes on the additional datasets validate the generalizability of our method. The Deep Blending dataset demonstrates our method's effectiveness in handling real-world-like diverse camera distributions. Likewise, the extensive outdoor scenes of the Tank and Temples dataset further illustrate our enhanced coverage, exemplified by the Truck scene in Figure 5.

4.2 Comparison on Uncertainty Estimation

Experimental settings. We evaluate the effectiveness of ResGS (Sec. 3.3) for improving uncertainty estimation accuracy. Specifically, we examine whether incorporating residual loss enhances the alignment between depth errors and predicted uncertainties under otherwise identical conditions. To measure this, we utilize the Area Under the Sparsification Error (AUSE) metric—a standard for evaluating uncertainty calibration adopted in (Wen et al., 2024; Goli et al., 2024; Shen et al., 2022)—where lower scores indicate better alignment between uncertainty and actual error. Thus, a lower AUSE score indicates better alignment between uncertainty predictions and actual errors, reflecting superior uncertainty calibration.

Following the approach used in CF-NeRF Shen et al. (2022), we employ depth maps from the NerfingMVS Wei et al. (2021) network, optimized using stereo depth from COLMAP at test time. Experiments are conducted on the Bonsai scene of Mip-NeRF 360 under identical view selection and evaluated using 19 uniformly sampled test images to ensure comparability.

Results. By incorporating residual learning applied in ResGS, we observed a reduction in AUSE from 0.328 to 0.317, indicating improved uncertainty calibration. The results signify enhanced alignment between predicted uncertainty and actual depth errors (see Fig. 6). This improvement is attributed to the ResGS approach, which increases certainty in regions with low prediction error while enabling continued learning through skip connections in uncertain regions. Consequently, our model achieves both structurally and quantitatively superior uncertainty calibration, leading to improved accuracy in uncertainty-driven tasks such as active mapping.

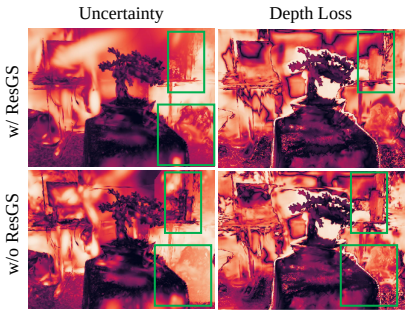

Figure 6: **Uncertainty Comparison.**

4.3 Ablation Studies

The ablation study in Table 2 highlights the contributions of each component in our proposed SA-ResGS framework. Incorporating residual learning (ResGS) alone improves reconstruction

Methods	Our proposed methods		Metrics		
	Sec. 3.2	Sec. 3.3	PSNR	SSIM	LPIPS
FisherRF†	-	-	20.991	0.618	0.454
‡+ResGS	-	✓	21.194	0.622	0.458
†+ResGS	-	✓	20.751	0.610	0.469
†+SA-HashGS	✓	-	21.254	0.623	**0.449**
†+SA-ResGS	✓	✓	**21.559**	**0.628**	0.456

Table 2: **Ablation using Mip-NeRF 360 dataset.** ‡ denotes fixed-order view selection, replicating the original selection sequence from FisherRF†, whereas † indicates dynamically updated view selection based on the model's progressive training status.

Process	MASt3R	Triangulation	prefilter	Fisher	Total	Raster.	GPU
FisherRF	—	—	—	28.00 s	28.00 s	0.005 s/iter	∼8K
Ours	0.40s	1.49 s	5.00s	5.60 s	12.50 s	0.027 s/iter	∼10.5K

Table 3: **Runtime comparison of FisherRF and SA-ResGS on the Bonsai scene.** Breakdown of view selection and training costs. SA-ResGS introduces additional prefiltering steps but reduces total runtime by 55%, with only modest increases in per-iteration cost and GPU memory usage.

stability, particularly in ambiguous or sparse regions (‡+ResGS). However, it shows limitations when view selection remains uncontrolled (†+ResGS), indicating that training improvements alone are insufficient, especially under high computational uncertainty quantification errors. As shown in Sec. 4.2, our training module effectively aligns predicted uncertainty with actual losses, yet purely uncertainty-driven view selection strategies remain vulnerable to bias from internal learning dynamics

In contrast, self-augmented prefiltering (SA-HashGS) independently improves geometric coverage through physically grounded view selection. Notably, combining SA-HashGS with ResGS yields substantial synergistic improvements, confirming the complementarity between robust view selection and residual supervision in optimizing both uncertainty quantification and reconstruction quality. For visual comparison please see Appendix Sec. D.

4.4 Computation Efficiency Analysis

A key challenge in active view selection is computational cost, as FisherRF computes per-Gaussian Fisher information via backpropagation across all candidate views, creating bottlenecks in large-scale datasets. To evaluate our prefiltering strategy, we conducted a runtime analysis on the Bonsai scene using a mid-range GPU (38 TFLOPS fp32), summarized in Table 3. SA-ResGS replaces exhaustive Fisher evaluation with a four-step process: dense correspondence prediction (MASt3R), triangulation for SA-Points, voxel-based prefiltering, and Fisher computation on only 20% of views. Despite these extra steps, view selection is 55% faster (28.0s $\rightarrow$ 12.5s), with only a modest increase in per-iteration cost (0.005s $\rightarrow$ 0.027s). Unlike FisherRF, whose cost grows with candidate views and becomes impractical for large datasets such as ScanNet (about 1600 images), SA-ResGS maintains fixed training cost once the schedule is set. GPU memory usage rises slightly but remains within standard limits, making our approach both scalable and practical.

5 Conclusion

This paper presents SA-ResGS, a Self-Augmented Residual 3D Gaussian Splatting framework for stabilize uncertainty quantification and enhance uncertainty-aware supervision in next-best-view (NBV) selection for active scene reconstruction. We introduce Self-Augmented Points, triangulated from a training view and a rasterized extrapolated view. These points enable physically grounded view selection and help mitigate erroneous uncertainty bias, while supporting targeted 3D supervision in high-uncertainty regions. Furthermore, we propose the first residual learning strategy tailored to 3D Gaussian Splatting, enabling effective supervision for both uncertain image regions and weakly contributing Gaussian splats. This leads to improved photometric reconstruction in novel view synthesis. Experimental results demonstrate the effectiveness of SA-ResGS across a range of realistic scenes, encompassing both indoor and outdoor environments and varying scene scales.

LIMITATIONS AND FUTURE WORK

SA-ResGS has several limitations. The method also relies on hyperparameter choices, potentially affecting performance to specific scenarios. Method to probabilistically choosing parameter would have potential to handle such limitations. Current method relies on 2D feedforward matching models, which show robust correspondence matching in most cases, application with more stronger model would further enhance the performance of the proposed methods. Extending SA-ResGS to dynamic environments involving moving objects or illumination changes remains an open challenge.

ETHIC STATEMENT

Our work focuses on advancing active 3D scene reconstruction through residual supervision and physically guided view selection in 3D Gaussian Splatting. The proposed method is entirely built upon publicly available datasets such as MipNeRF360 (Barron et al., 2022), Deep Blending (Hedman et al., 2018), and Tank and Temples dataset (Knapitsch et al., 2017), which are widely used in the vision community. No personally identifiable or sensitive information is included in our experiments. However, as with all vision datasets, there may exist inherent biases in scene distributions (e.g., indoor environments being more common than outdoor scenes). These biases could affect generalization in certain deployment scenarios. We encourage users to be aware of such limitations when applying our method to broader real-world applications.

RREPRODUCIBILITY STATEMENT

We provide detailed descriptions of our algorithmic components in Sec. 3, including uncertainty-guided supervision, view selection strategies, and residual learning for 3DGS. The datasets used (MipNeRF360 (Barron et al., 2022), DeepBlending (Hedman et al., 2018), Tanks and Temples datasets (Knapitsch et al., 2017)) and the evaluation protocols are specified in Sec. 4. To facilitate reproducibility, the hyperparameters, training recipes, and implementation details are documented in Appendix B. All experiments were conducted with standard PyTorch libraries and commonly available GPUs, ensuring that the results can be reproduced with reasonable computational resources.

REFERENCES

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

Andreas Bircher, Mina Kamel, Kostas Alexis, Helen Oleynikova, and Roland Siegwart. Receding horizon "next-best-view" planner for 3d exploration. In *ICRA*, 2016.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

Xiao Chen, Quanyi Li, Tai Wang, Tianfan Xue, and Jiangmiao Pang. Gennbv: Generalizable next-best-view policy for active 3d reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (IEEE Conf. Comput. Vis. Pattern Recog.)*, 2024.

C. I. Connolly. The determination of next best views. In *Proceedings. 1985 IEEE International Conference on Robotics and Automation*, volume 2, 1985.

Jeffrey Delmerico, Stefan Isler, Reza Sabzevari, and Davide Scaramuzza. A comparison of volumetric information gain metrics for active 3d object reconstruction. *Autonomous Robots*, 42(2), 2018.

Enrique Dunn and Jan-Michael Frahm. Next best view planning for active model improvement. In *British Machine Vision Conference*, 2009. URL https://api.semanticscholar.org/CorpusID:1345419.

Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. Bayes' Rays: Uncertainty quantification in neural radiance fields. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

Antoine Guédon, Pascal Monasse, and Vincent Lepetit. Scone: surface coverage optimization in unknown environments by volumetric integration. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022.

Antoine Guédon, Tom Monnier, Pascal Monasse, and Vincent Lepetit. Macarons: Mapping and coverage anticipation with rgb online self-supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

Alex Hanson, Allen Tu, Vasu Singla, Mayuka Jayawardhana, Matthias Zwicker, and Tom Goldstein. Pup 3d-gs: Principled uncertainty pruning for 3d gaussian splatting. *arXiv*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. 37(6), 2018.

Youngkyoon Jang and Eduardo Pérez-Pellitero. Comapgs: Covisibility map-based gaussian splatting for sparse novel view synthesis, 2025.

Youngkyoon Jang, Hatice Gunes, and Ioannis Patras. Registration-free face-ssd: Single shot analysis of smiles, facial attributes, and affect in the wild. *Computer Vision and Image Understanding*, 182, 2019.

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 Transactions on Graphics*, 36(4), 2017.

Georgios Kopanas and George Drettakis. Improving NeRF Quality by Progressive Camera Placement for Free-Viewpoint Navigation. In Michael Guthe and Thorsten Grosch, editors, *Vision, Modeling, and Visualization*. The Eurographics Association, 2023. ISBN 978-3-03868-232-5. doi: 10.2312/vmv.20231222.

Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. WildGaussians: 3D gaussian splatting in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Sibaek Lee, Kyeongsu Kang, Seongbo Ha, and Hyeonwoo Yu. Bayesian nerf: Quantifying uncertainty with volume density for neural implicit fields. *IEEE Robotics Autom. Lett.*, 10(3), 2025.

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

Jiahe Li, Xiao Bai Jiawei Zhang, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Dngaussian: Depth-regularized neural gaussian splatting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

Linjie Lyu, Ayush Tewari, Marc Habermann, Shunsuke Saito, Michael Zollhöfer, Thomas Leimküehler, and Christian Theobalt. Manifold sampling for differentiable uncertainty in radiance fields. In *SIGGRAPH Asia Conference Proceedings*, 2024. doi: 10.1145/3680528.3687655.

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. In *Eur. Conf. Comput. Vis.*, 2020.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4), July 2022. doi: 10.1145/3528223.3530127. URL https://doi.org/10.1145/3528223.3530127.

Simon Niedermayr, Josef Stumpfegger, and Rüdiger Westermann. Compressed 3d gaussian splatting for accelerated novel view synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2024.

Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. Activenerf: Learning where to see with uncertainty estimation. In *Eur. Conf. Comput. Vis.* Springer, 2022.

Hyunwoo Park, Gun Ryu, and Wonjun Kim. Dropgaussian: Structural regularization for sparse-view gaussian splatting, 2025. URL https://arxiv.org/abs/2504.00773.

Zador Pataki, Paul-Edouard Sarlin, Johannes L. Schönberger, and Marc Pollefeys. MP-SfM: Monocular Surface Priors for Robust Structure-from-Motion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025.

Shichong Peng, Yanshu Zhang, and Ke Li. Papr in motion: Seamless point-level 3d scene interpolation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

Johannes L Schonberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Eur. Conf. Comput. Vis.*, 2016.

William R. Scott, Gerhard Roth, and Jean-François Rivest. View planning for automated three-dimensional object reconstruction and inspection. *ACM Computing Surveys (CSUR)*, 35(1), 2003.

Jianxiong Shen, Adria Ruiz, Antonio Agudo, and Francesc Moreno-Noguer. Stochastic neural radiance fields: Quantifying uncertainty in implicit 3d representations. In *International Conference on 3D Vision*, 2021.

Jianxiong Shen, Antonio Agudo, Francesc Moreno-Noguer, and Adria Ruiz. Conditional-flow nerf: Accurate 3d modelling with reliable uncertainty quantification. In *Eur. Conf. Comput. Vis.*, 2022.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), 2014. ISSN 1532-4435.

Yifan Sun, Qixing Huang, Dun-Yu Hsiao, Li Guan, and Gang Hua. Learning view selection for 3d scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2021.

Niko Sünderhauf, Jad Abou-Chakra, and Dimity Miller. Density-aware nerf ensembles: Quantifying predictive uncertainty in neural radiance fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

Pere-Pau Vázquez, Miquel Feixas, Mateu Sbert, and Wolfgang Heidrich. Viewpoint selection using viewpoint entropy. In *Proceedings of the Vision Modeling and Visualization Conference 2001*, 2001.

Tao Wang, Weibin Xi, Yong Cheng, Hao Han, and Yang Yang. Rl-nbv: A deep reinforcement learning based next-best-view method for unknown object reconstruction. *Pattern Recognition Letters*, 184, 2024.

Zihang Wang et al. Uncertainty-aware ensemble nerf for active view planning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

Zirui Wang, Yash Bhalgat, Ruining Li, and Victor Adrian Prisacariu. Active view selector: Fast and accurate active view selection with cross reference image quality assessment. 2025.

Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Int. Conf. Comput. Vis.*, 2021.

Jiang Wen, Lei Boshu, and Daniilidis Kostas. Fisherrf: Active view selection and uncertainty quantification for radiance fields using fisher informations. In *Eur. Conf. Comput. Vis.*, 2024.

Joey Wilson, Marcelino Almeida, Sachit Mahajan, Martin Labrie, Maani Ghaffari, Omid Ghasemalizadeh, Min Sun, Cheng-Hao Kuo, and Arnab Sen. Pop-gs: Next best view in 3d-gaussian splatting with p-optimality. 2025.

Wenhui Xiao, Rodrigo Santa Cruz, David Ahmedt-Aristizabal, Olivier Salvado, Clinton Fookes, and Leo Lebrat. Nerf director: Revisiting view selection in neural volume rendering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2024.

Wangze Xu, Huachen Gao, Shihe Shen, Jianbo Jiao Rui Peng, and Ronggang Wang. Mvpgs: Excavating multi-view priors for gaussian splatting from sparse input views. In *Eur. Conf. Comput. Vis.*, 2024.

Hong Xuan, Abby Stylianou, Xiaotong Liu, and Robert Pless. Hard negative examples are hard, but useful. In *Eur. Conf. Comput. Vis.*, 2020.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2024.

Yanshu Zhang, Shichong Peng, Alireza Moazeni, and Ke Li. Papr: Proximity attention point rendering. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

APPENDIX OVERVIEW

This supplementary document provides additional implementation details and results that support and extend the main paper. It is organized as follows:

- Sec. A. Implementation Details.
- Sec. B. Camera Distribution Analysis.
- Sec. C. Extended Qualitative Comparisons.
- Sec. D. Qualitative Results for Ablation Studies.
- Sec. E. Additional Ablation Analysis.
- Sec. F. Supplementary Video Overview.

A IMPLEMENTATION DETAILS

A.1 COVERAGE ESTIMATION

We follow the baseline setup of FisherRF (Wen et al., 2024), initializing our experiments with a sparse point cloud reconstructed via COLMAP (Schonberger and Frahm, 2016; Schonberger et al., 2016). This SfM output defines the axis-aligned bounding volume for discretizing the scene into voxel grids. To enable physically grounded view selection, we construct Self-Augmented Points (SA-Points) via triangulation between a training view and its extrapolated view. SA-Points are used to estimate surface occupancy and to encode observed geometry into binary voxel features. Candidate views are scored based on their voxel-level dissimilarity to the encoded training views, supporting robust coverage estimation without relying on early uncertainty signals.

In the following, we detail (1) Self-Augmented Points (SA-Points) Generation, and (2) Observed Surface Coverage Estimation and View Frustum Construction.

A.1.1 SELF-AUGMENTED POINTS (SA-POINTS) GENERATION

Extrapolated View Generation. To synthesize a novel rasterized view while maintaining sufficient scene overlap, we perturb the original camera center by ± 0.25 units along the x and y axes and translate it backward by 0.5 units along the z-axis. This backward-only perturbation ensures that the extrapolated view retains a high degree of visibility overlap with the original training view, keeping most scene content within the shared frustums. As a result, the extrapolated view covers a large portion of the original training image while still providing a novel perspective of the same surfaces, enabling reliable correspondence estimation for SA-Point generation.

Dense Correspondence Matching. We compute dense correspondences between the ground truth image of original training view and the rendering from extrapolated view using the pretrained MASt3R model (Leroy et al., 2024) (MASt3R_ViTLarge_BaseDecoder_512_catmlpdpt_metric), which is robust to moderate viewpoint perturbations. As discussed in the main paper, this 3D vision foundation model produces context-aware dense correspondences by capturing structured scene semantics over 16×16 local patches, even when the extrapolated view includes geometric artifacts. This robustness allows us to extract reliable matches despite distortions in the rasterized extrapolated images, enabling consistent 3D reconstruction from single-view observations. Consequently, we can leverage extrapolated viewpoints to augment sparse training views without requiring multiview supervision.

Triangulation and Reprojection Filtering. To ensure geometric consistency, we triangulate 3D points from matched correspondences and filter them based on reprojection error, discarding points with a bidirectional reprojection error exceeding 0.5 pixels. For computational efficiency, we parallelize the triangulation process across multiple CPU threads and apply a spatial stride of 5 pixels in both x and y directions to subsample correspondence pairs. These filtering strategies significantly accelerate processing while preserving high-fidelity geometric structure.

A.1.2 Observed Surface Coverage Estimation and View Frustum Construction

Voxel Grid Construction. Given the sparse SfM points, we define an axis-aligned bounding box (AABB) that encapsulates the entire 3D point cloud. The AABB is computed from the minimum and maximum bounds of the reconstructed points.

Initial Occupancy Estimation. We discretize the scene into a voxel grid and mark a voxel as occupied if it contains a minimum number of SfM points. To better represent scene geometry, we apply N-fold upsampling to the occupied voxels. For the minimum number for determining occupancy, we use 2 for outdoor scenes, and 5 for indoor cases.

Observed Region Calculation. SA-Points are mapped to their nearest voxels to define the observed surface region. To account for possible triangulation errors and improve spatial robustness, we apply a 3D dilation operation to the occupied voxels. For all cases, a dilation radius of 2 is applied.

View Frustum Determination. To evaluate candidate views, we define view frustums using camera intrinsics and the global maximum bounds computed from the SfM point cloud. These frustums ignore visibility constraints but serve as a conservative estimate of potential scene coverage. Coverage scores for physically grounded view selection are then computed by measuring voxel-level intersections between the candidate frustums and the observed surface, represented via hash-encoded voxel occupancy representation.

A.2 Experimental setup

All models presented in the main manuscript, including the ablation variants of our proposed method, were trained and evaluated on a single NVIDIA V100 GPU with 32 GB of memory. CPU-based components—such as SA-Points generation, voxel grid processing, and triangulation—were parallelized across 8 threads for efficiency. Each experiment was run once with a fixed random seed of 0 to ensure reproducibility.

A.2.1 Dataset Detail

We evaluate our method and baselines on two types of datasets: (1) Mip-NeRF 360, and (2) Extended NBV benchmark datasets. Mip-NeRF 360 consists of nine real-world scenes with dense 360-degree camera coverage, captured across both indoor and outdoor scenarios. While this dataset provides a controlled and well-curated benchmark, its uniform view distribution limits the difficulty of next-best-view (NBV) evaluation. Such settings often make next-best-view (NBV) strategies appear less critical, since even simple heuristics (e.g., furthest-distance selection) already perform reliably under balanced coverage (Xiao et al., 2024)

To address this, we additionally construct an Extended NBV benchmark by selecting seven challenging scenes from Deep Blending and Tanks and Temples, characterized by irregular camera trajectories and diverse scene scales. This curated set introduces more realistic and unbalanced conditions, offering a complementary testbed for assessing robustness in active view selection.

These datasets contain large-scale and geometrically complex scenes with irregular camera distributions, but their difficulty also means that many methods fail outright. To filter such degenerate cases, we trained the FisherRF baseline under the standard scheme and retained only scenes where it achieved at least 17 dB PSNR, ensuring that the comparisons remained fair and informative.

This procedure produced seven representative scenes: Horse, Truck, Francis, Ballroom, Barn, Ponche, and Playroom. The selected set spans both indoor and outdoor environments, and includes highly complex camera distributions (e.g., Ballroom, Ponche, and Playroom) that deviate substantially from the curated coverage of Mip-NeRF 360. These characteristics create more realistic stress tests for NBV strategies by introducing occlusions, scale variations, and unbalanced observations. A brief visualization of scene configurations and camera trajectories is provided in Sec.2 of this supplementary material.

B Camera Distribution Analysis

To evaluate the effectiveness of our view selection strategy, we compare the camera distributions produced by different methods. Fig. S1, S2, and S3 visualize these distributions from both bird's-eye and side perspectives. As discussed in the main paper, FisherRF tends to produce clustered view selections due to its tight coupling with the internal 3DGS learning dynamics—visibly highlighted in the semi-transparent yellow regions. In contrast, our method yields a more spatially uniform and well-dispersed distribution of viewpoints.

This distinction is further illustrated in Fig. S4, where in the Room scene from the Mip-NeRF 360 dataset, FisherRF often selects redundant or near-parallel views. Our method instead promotes angular diversity, leading to broader scene exploration. A similar trend is observed in the Deep Blending dataset, where our approach selects viewpoints across a wider vertical range. These comparisons reinforce our claim that SA-ResGS facilitates physically grounded and geometrically diverse view selection, contributing to improved scene coverage.

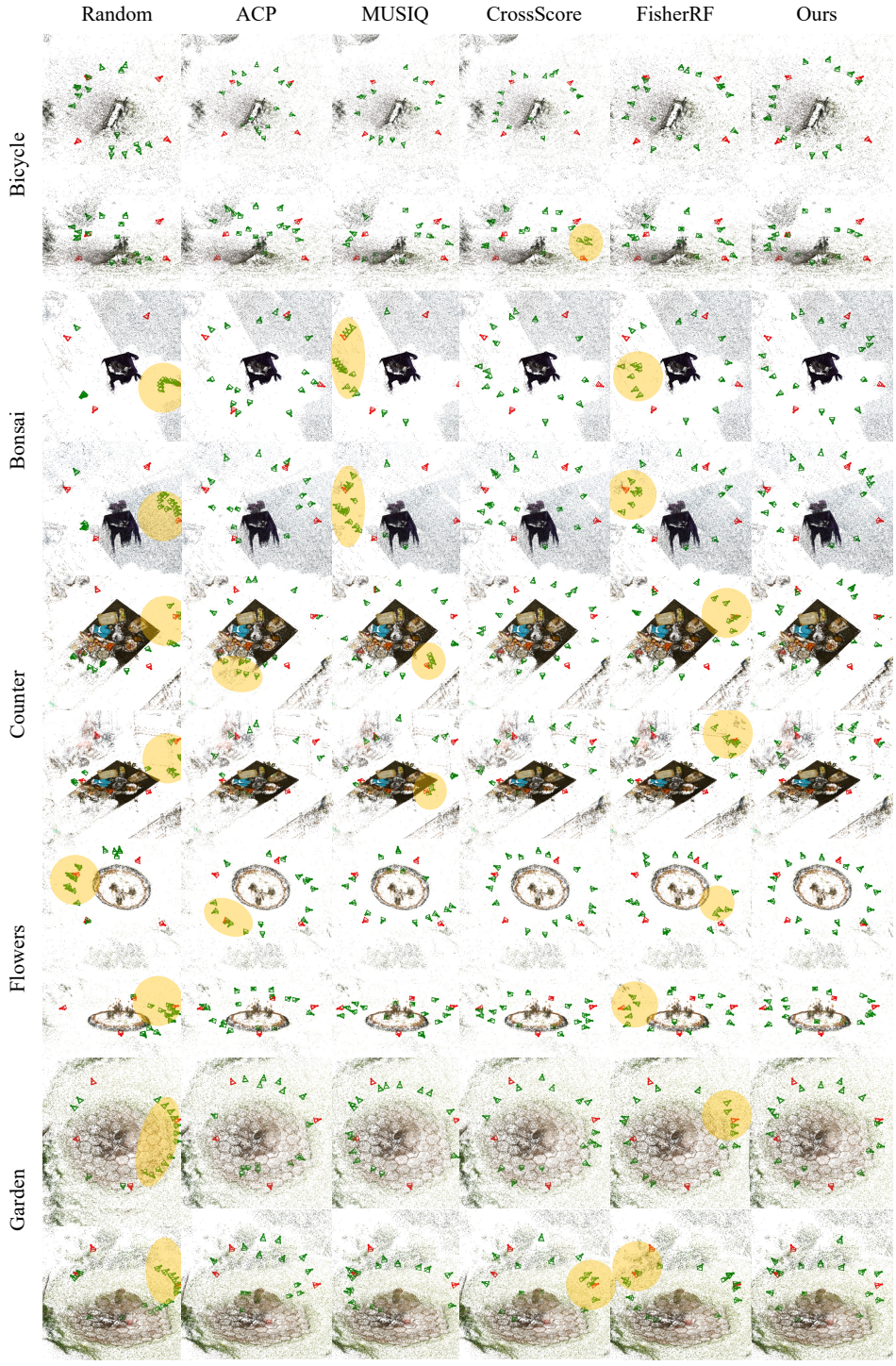

Figure S1: **Camera View Distribution on Mip-NeRF 360.** Visualization of camera poses selected by each method on the Mip-NeRF 360 dataset. Red frustums indicate the initial views, while green frustums denote views added during active selection. Our method produces a more uniformly distributed set of viewpoints, while both baselines exhibit clustered view selections (highlighted in yellow circle), particularly in FisherRF due to its reliance on uncertainty signals entangled with 3DGS training dynamics.

Figure S2: **Camera View Distribution on Mip-NeRF 360.** Visualization of camera poses selected by each method on the Mip-NeRF 360 dataset. Red frustums indicate the initial views, while green frustums denote views added during active selection. Our method produces a more uniformly distributed set of viewpoints, while both baselines exhibit clustered view selections (highlighted in yellow circle), particularly in FisherRF due to its reliance on uncertainty signals entangled with 3DGS training dynamics.

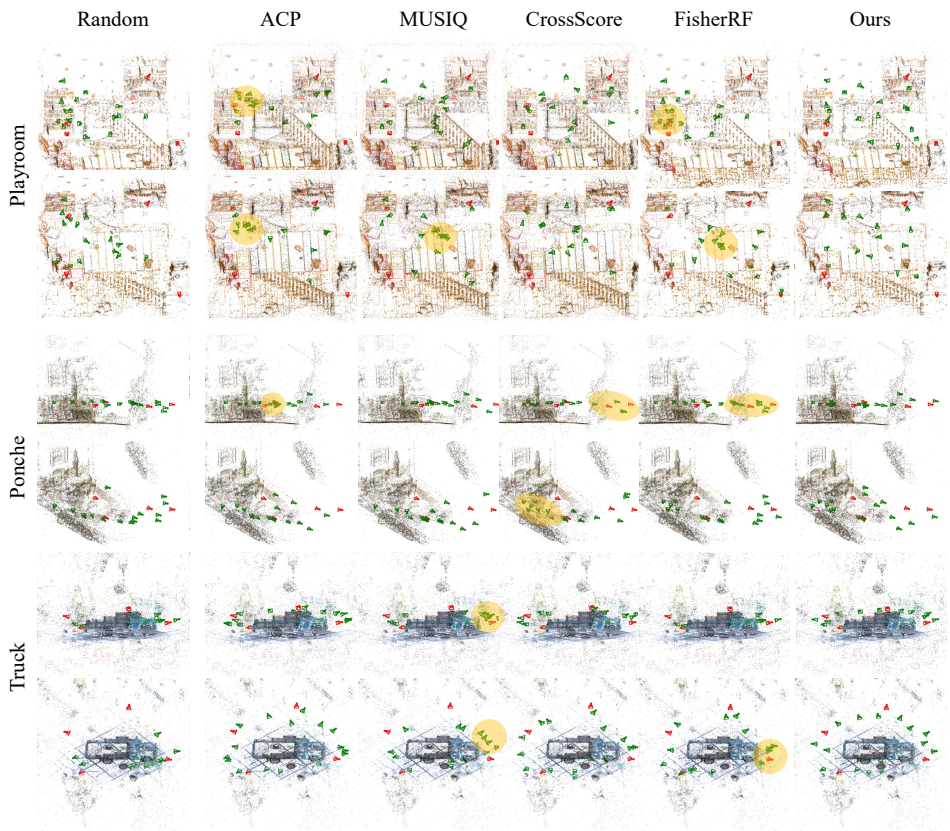

Figure S3: **Camera View Distribution on Deep Blending and Tanks & Temples** View distributions on additional datasets, following the same color and annotation scheme as Fig. S1. Our method maintains broader spatial coverage, while baseline methods often exhibit clustering (yellow circles), consistent with the biases observed in Mip-NeRF 360.

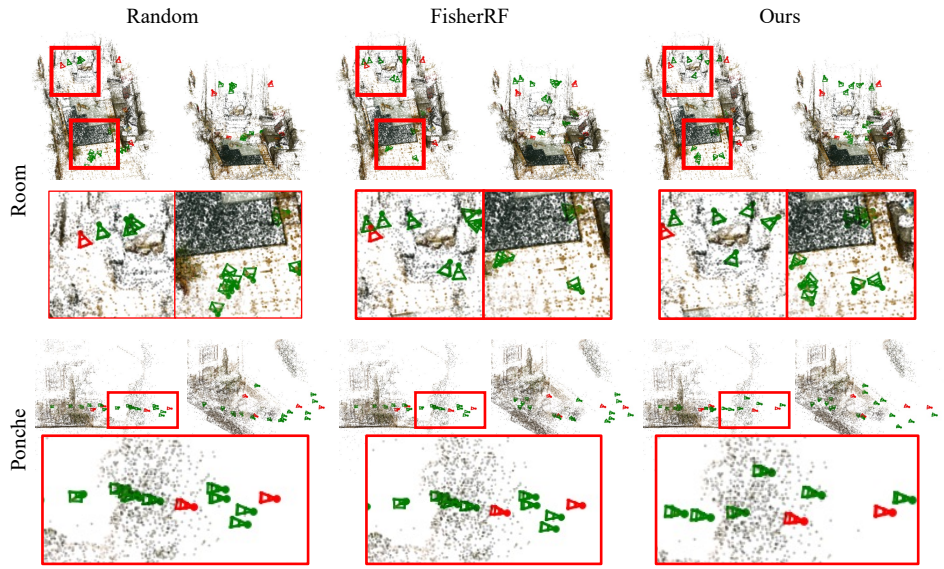

Figure S4: **Diversity of Selected Views (Zoomed-in Analysis)** We highlight the spatial diversity of selected camera poses by presenting zoomed-in regions corresponding to the boxed areas above. Compared to the baselines, our method selects views from a broader range of angles (Room, top) and elevations (Ponche, bottom), resulting in more comprehensive scene coverage.

C Extended Qualitative Comparisons

The qualitative results presented in the main paper are limited by space constraints, which may obscure the full advantages of our method. To address this, we provide extended visualizations from multiple test viewpoints. Fig. S5 to S9 display results from five scenes across the Mip-NeRF 360, Deep Blending, and Tanks and Temples datasets, with six to eight novel test views per scene. Our method consistently achieves broader and more complete scene coverage compared to the FisherRF baseline.

As discussed in the main paper, our residual supervision strategy further improves geometric consistency and reconstruction robustness, particularly in sparse or limited-view scenarios. This is especially beneficial under the standard protocol of active or next-best-view selection, where training begins with a small number of views (e.g., 4) and progressively adds new views (typically one at a time). The synergy between physically grounded view selection and residual learning enables high-fidelity reconstruction even from limited initial observations. As mentioned in Sec. F, we also include a supplementary video with 360-degree novel view renderings.

We encourage reviewers to view this video to better appreciate the improvements in coverage and structural accuracy provided by our proposed SA-ResGS.

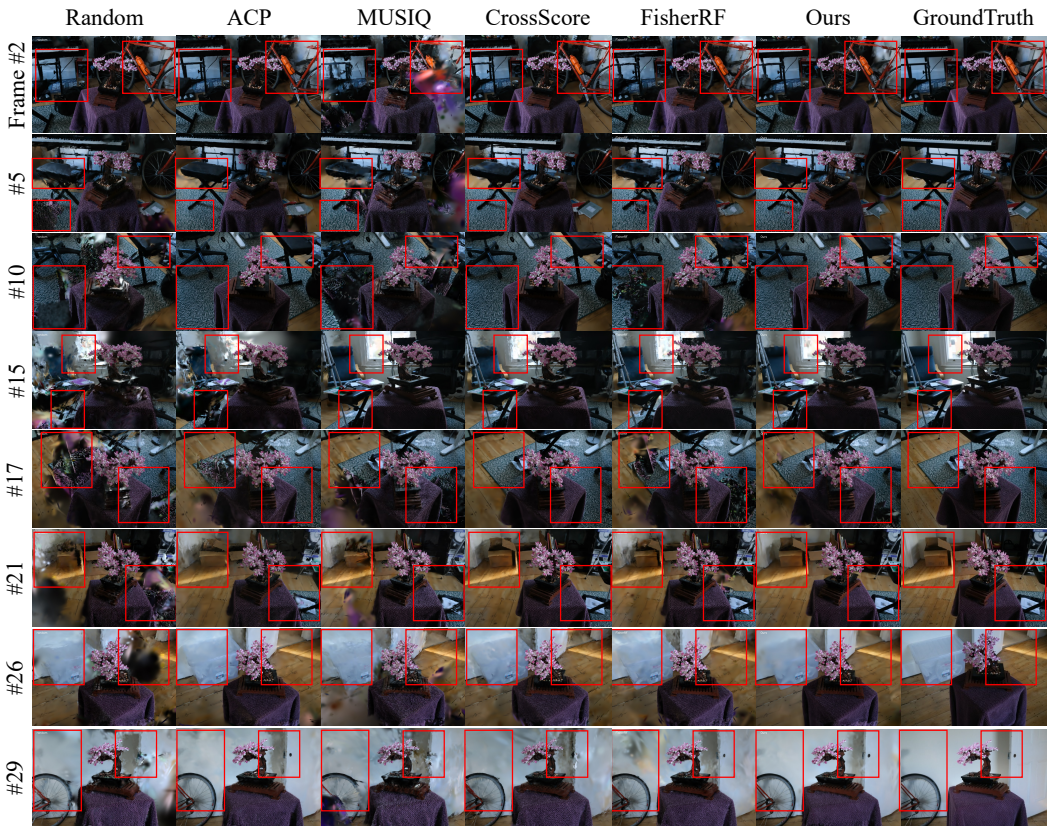

Figure S5: **Qualitative Comparison Across Multiple Test-Time Views (Bonsai Scene).** We compare novel view synthesis results of our method and baseline models (Random, FisherRF) on eight test views from the Bonsai scene. Our method consistently yields more complete reconstructions, particularly in under-observed or occluded regions such as table edges, reflective surfaces, and background structures (see red boxes). In contrast, baseline methods often exhibit structural artifacts, missing geometry, or texture blurring due to biased or clustered view selection.

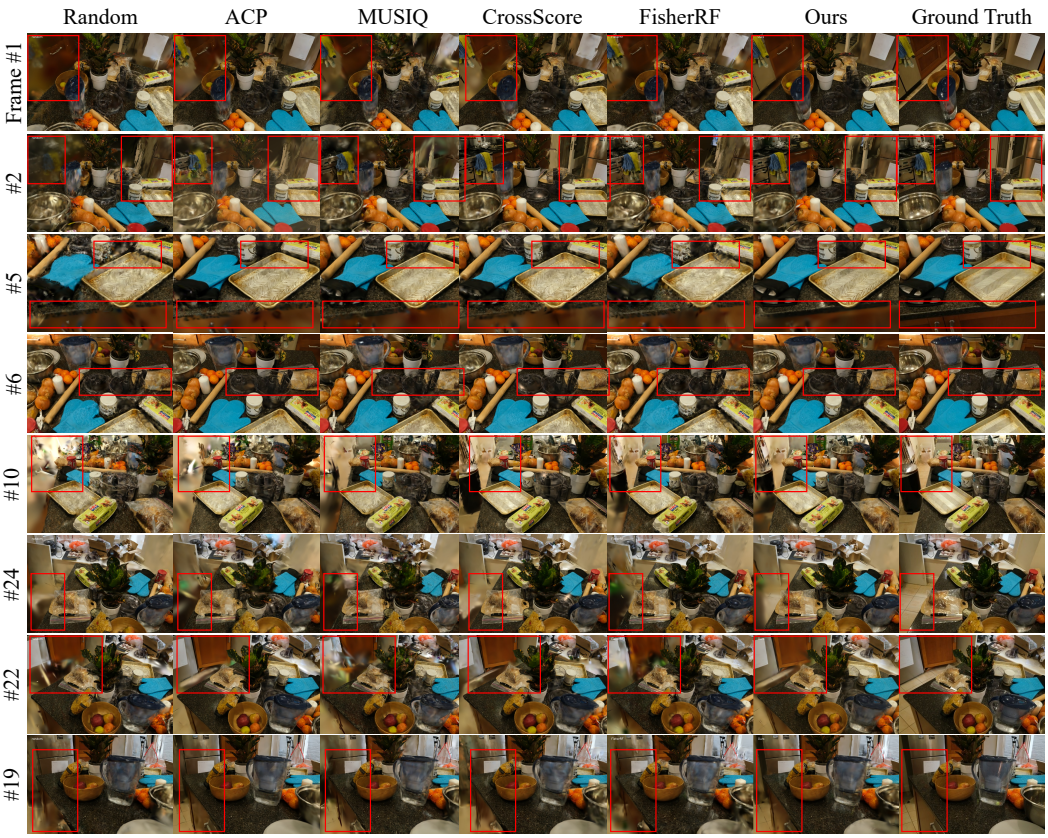

Figure S6: **Qualitative Comparison Across Multiple Test-Time Views (Counter Scene).** Following the format in Fig. S5, we compare novel view synthesis results across eight test views for the Counter scene. Our method achieves more complete and consistent reconstructions in occluded and sparsely observed regions (red boxes), while baseline methods (Random, FisherRF) often suffer from structural artifacts and blurring due to biased or clustered view selections.

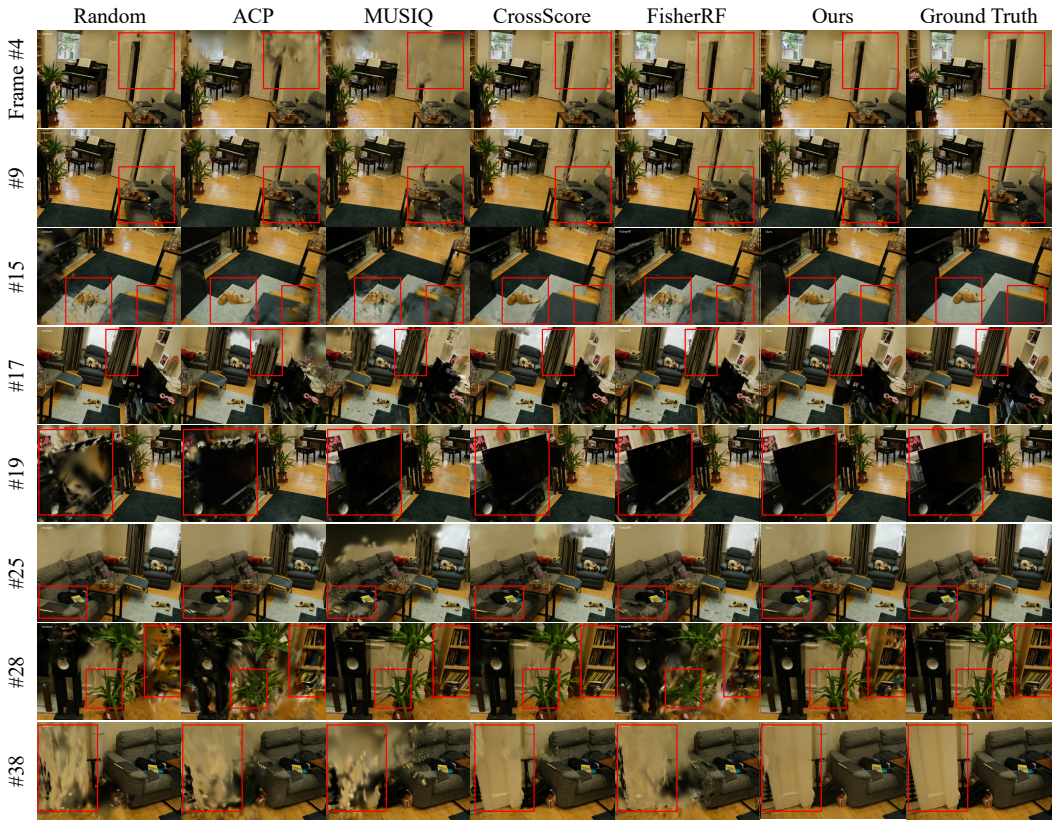

Figure S7: **Qualitative Comparison Across Multiple Test-Time Views (Room Scene).** As in Fig. S5, we compare novel view synthesis results across eight test views. Our method produces more complete and accurate reconstructions in occluded or sparsely covered regions (red boxes), whereas baseline methods often exhibit geometric artifacts and texture degradation due to biased view selection.

Figure S8: **Qualitative Comparison Across Multiple Test-Time Views (Truck Scene).** Similar to Fig. S5, we compare novel view synthesis results across eight test views. Our method yields more complete and consistent reconstructions, particularly in occluded or sparsely observed regions (red boxes), while baseline methods exhibit artifacts and texture degradation due to limited view diversity.

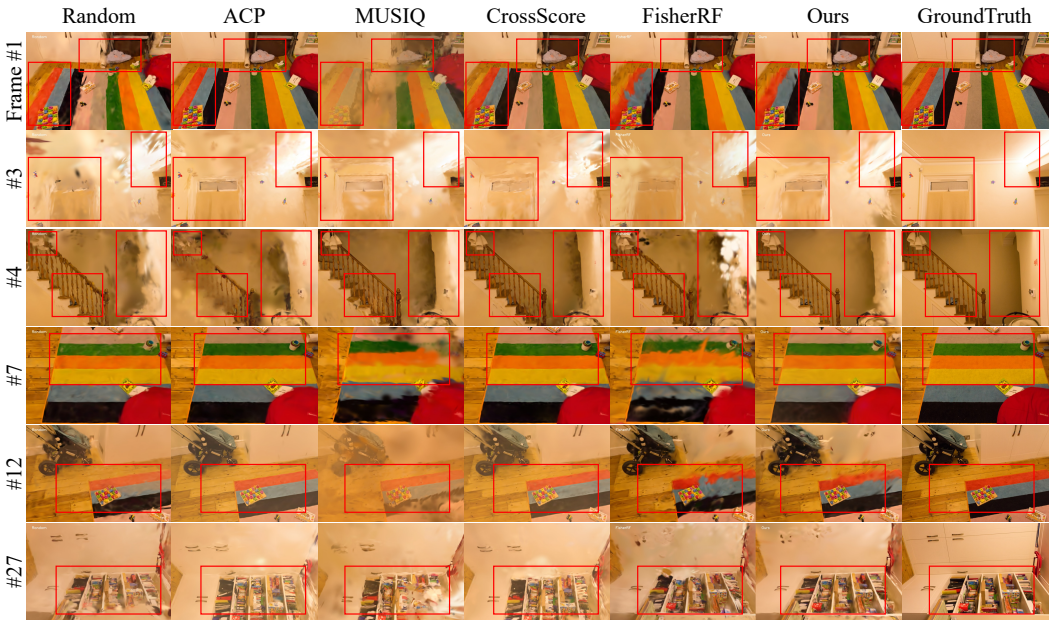

Figure S9: **Qualitative Comparison Across Multiple Test-Time Views (Playroom Scene).** Following the analysis in Fig. S5, we show results for six novel test views. Our method consistently improves reconstruction completeness in occluded and sparsely observed regions (red boxes), where baselines exhibit structural artifacts or blurred textures due to limited view diversity.

D Qualitative Results for Ablation Studies

To supplement the quantitative results presented in the ablation study, we provide extended qualitative comparisons in Fig. S10, evaluating different model variants on the **Room** and **Counter** scenes. The variants include the baseline FisherRF, residual supervision with fixed-order view selection (‡+ResGS), our physically grounded view selection method (SA-HashGS), and the full SA-ResGS model combining both components. In the **Room** scene, the baseline FisherRF exhibits missing geometry and artifacts near occluded regions and object boundaries (orange boxes). SA-HashGS mitigates these issues by selecting geometrically diverse viewpoints, leading to improved surface coverage. Residual supervision (‡+ResGS) further refines local details—even in regions already observed by the baseline—demonstrating enhanced reconstruction without additional view coverage, as evidenced by improved toy geometry and floor textures (red boxes). The full SA-ResGS model combines both benefits, producing reconstructions that closely align with ground-truth in terms of both structural completeness and fine detail.

In the **Counter** scene, residual supervision (‡+ResGS) reduces floating artifacts and improves under-optimized regions caused by occlusion, such as the shadow-like area on the left side of the countertop (first row). SA-HashGS enhances global coverage, alleviating severe blurring and recovering geometry in previously unobserved areas (red boxes). The full SA-ResGS model integrates both benefits, yielding sharper and more complete reconstructions by jointly leveraging physically grounded view selection and residual learning. These results validate the complementary roles of physically grounded view selection and uncertainty-guided residual learning. Their integration in SA-ResGS consistently improves reconstruction fidelity under sparse-view settings.

Additional 360-degree renderings for these ablation results are included in the supplementary video and are recommended for further comparison.

Figure S10: **Qualitative Self-Comparison for Ablation Study.** Results on the **Room** and **Counter** scenes comparing FisherRF, ‡+ResGS, SA-HashGS, and SA-ResGS. Orange boxes highlight improvements from coverage-guided view selection, while red boxes emphasize the effects of residual supervision. Residual learning enhances geometric stability and reduces artifacts (e.g., jittering surfaces), whereas SA-HashGS recovers unobserved (e.g., occluded or shadowed areas). The full SA-ResGS combines both benefits, yielding the most complete and and structurally faithful reconstructions.

E Additional Ablation Analysis

Robustness to Correspondence Noise. SA-Points are obtained through triangulation of dense correspondences, so their accuracy can influence reconstruction and view selection. To examine robustness, we conducted experiments with synthetic correspondence noise and different reprojection thresholds.

Adding uniform noise (0.0–10.0 pixels) showed that reconstruction quality remains stable up to moderate levels (≤ 5.0 pixels) when combined with a 1-pixel reprojection filter, as shown in Table. S1,

with only larger perturbations causing noticeable decline. This indicates that the system tolerates moderate correspondence errors without significant degradation.

Noise Amplitude	0.0	0.5	1.0	5.0	10.0
PSNR	24.441	24.199	24.117	24.311	23.121

Table S1: **Robustness to synthetic correspondence noise.** PSNR remains stable under moderate perturbations (≤ 5 pixels), showing SA-Points are resilient to realistic levels of noise.

Varying the reprojection error threshold further confirmed this robustness. As shown in Table. S2, very strict filtering (0.5 pixels) reduced coverage, while very loose filtering (2.0 pixels) introduced slight inaccuracies. A threshold of 1.0 pixel provided the best trade-off, ensuring sufficient coverage and reliable geometry.

These results demonstrate that SA-Points are resilient to moderate correspondence errors. Our combination of reprojection filtering and voxel dilation effectively balances coverage and accuracy, enabling stable and reliable NBV selection in practice.

Reprojection error	0.5	1.0	2.0
PSNR	22.634	24.441	24.244
SSIM	0.817	0.838	0.834
LPIPS	0.332	0.317	0.318
Coverage	57.74%	94.11%	94.89%

Table S2: **Effect of reprojection filtering thresholds.** A 1-pixel threshold offers the best trade-off between accuracy and coverage, confirming that accurate filtering is essential for stable SA-Points.

Effects of number of selected views. The number of available input views is a critical factor in reconstruction quality, particularly under sparse-view settings. To analyze this effect, we conducted an ablation on the Bonsai scene, starting from 4 fixed initial views and incrementally adding views using our selection strategy. For clarity, we uniformly subsampled training views and report PSNR across different totals.

Results in Table S3 that under sparse conditions (7–10 views), our method reaches higher PSNR with fewer inputs compared to FisherRF, demonstrating the benefit of physically grounded filtering in stabilizing early-stage. As more views are added, the initial advantage narrows but new gains appear from 13 views onward, where residual learning further refines Gaussian parameters and improves geometry.

Overall, the method excels in low-view regimes while continuing to scale effectively with more observations, validating its robustness across varying view counts.

Selected Views	4	7	10	13	16	19
random	15.929	16.485	17.573	17.694	18.233	18.400
FisherRF	15.985	17.269	18.773	20.188	21.722	22.650
Ours	15.970	17.606	18.895	20.159	23.229	24.064

Table S3: **Performance scaling with the number of selected views.** Our method yields higher PSNR in sparse-view regimes (7–10 views) and continues to improve as more views are added, highlighting both early stability and long-term scalability.

F Supplementary Video Overview

To complement the static visualizations provided in this document, we include a supplementary video that offers dynamic and comprehensive renderings of our results. This video is intended to provide a

deeper visual understanding of the improvements achieved by our proposed SA-ResGS framework across various evaluation scenarios.

The video includes:

- **Novel View Rendering.** We present extended 360-degree novel view trajectories, captured along spiral and circular camera paths, which go beyond the discrete test views shown in the main paper and supplementary figures. These renderings highlight the effectiveness of our physically grounded view selection and residual supervision in preserving structural consistency and photometric quality across challenging viewpoints.
- **Ablation Study Comparisons.** To illustrate the impact of each component, we show side-by-side comparisons of different model variants under continuous camera movement. These scenes demonstrate the robustness and fidelity improvements from residual supervision and surface-aware physically grounded view selection, especially under sparse-view or occluded regions.

USE OF LARGE LANGUAGE MODELS

All substantive content was authored by the researchers, with the Large Language Models employed solely for auxiliary polishing of sentence-level expressions for better readability.