PREVIEW: ACTIVE VIEW SELECTION FOR 3D RECONSTRUCTION WITHOUT WARM-UP

Anonymous authors
Paper under double-blind review

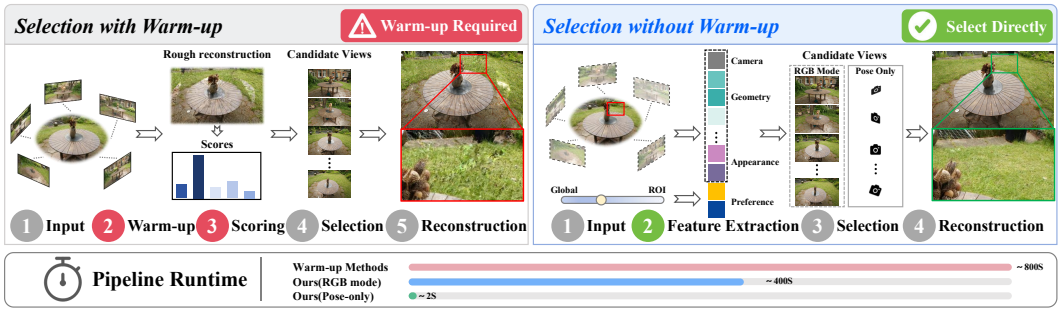


Figure 1: **View selection with and without reconstruction warm-up.** Reconstruction-driven methods first perform a warm-up reconstruction and then score candidate views before selection. In contrast, **PreView** selects views directly from available inputs without reconstruction warm-up, supporting both RGB-aware and pose-only modes as well as global–ROI preference control. This design substantially reduces the runtime required before the final view subset is determined.

ABSTRACT

Active view selection is a key component of 3D reconstruction, aiming to identify the most useful observations under a limited acquisition or reconstruction budget in practical scenarios. Effective view selection can reduce redundant data acquisition and reconstruction costs while preserving the quality of the final 3D model. However, many existing methods estimate the value of a view only after a warm-up reconstruction, where a preliminary scene representation must first be built from the initial observations. This paradigm not only introduces substantial computation before selection, but can also be unreliable when only a few initial observations are available, since the warm-up reconstruction itself may be incomplete or inaccurate.

We observe that the value of a candidate view can instead be inferred from the new evidence it contributes beyond the currently observed views, without first constructing a reliable 3D scene. Based on this insight, we introduce PreView, a framework for view selection before 3D reconstruction. PreView uses feed-forward visual and geometric priors to extract features from available observations and explicitly models interactions between candidate views and the selected set to estimate each candidate's marginal contribution. This formulation supports different information regimes: when candidate images are available, PreView exploits appearance and geometric cues; otherwise, its pose-only variant performs viewpoint selection before acquisition using camera poses and the context established by the initial observations. Because candidate value is estimated directly from available observations, task preferences can also be incorporated naturally, steering the reconstruction budget between global scene quality and user-specified regions of interest. Experiments across multiple scenes show that PreView selects effective views without reconstruction warm-up, remains competitive in the strict pose-only setting, supports controllable task-aware reconstruction, and substantially reduces computation before selection.

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

1 INTRODUCTION

Recent advances in neural scene representations have substantially improved the quality and efficiency of 3D reconstruction. Neural Radiance Fields (NeRF) represent scenes as continuous radiance fields optimized from posed images (Mildenhall et al., 2021), while 3D Gaussian Splatting (3DGS) provides an explicit representation with efficient optimization and real-time rendering (Kerbl et al., 2023). Despite this progress, reconstruction quality still depends strongly on the available observations. Insufficient views may lead to missing structures and inaccurate geometry, whereas redundant observations increase acquisition and optimization cost without comparable gains.

Active view selection addresses this problem by identifying the most useful observations under a limited sensing or reconstruction budget. It is a key component of active 3D reconstruction and is particularly relevant to UAV mapping, robotic inspection, and embodied systems, where sensing decisions often need to be made before additional acquisition or scene optimization. Many existing methods first perform a warm-up reconstruction from the initial observations, obtaining a rough scene representation from which candidate views are subsequently evaluated. The value of a candidate may then be estimated from uncertainty, information gain, coverage, or the expected improvement of the reconstructed scene. For example, ActiveNeRF uses reconstruction uncertainty for view acquisition (Pan et al., 2022), FisherRF evaluates candidates through Fisher information (Jiang et al., 2024), and POp-GS extends information-based selection to 3D Gaussian Splatting (Wilson et al., 2025). Although effective, these methods typically require a scene representation to be built or updated before candidate views can be scored.

This warm-up dependency brings two limitations. First, intermediate reconstruction introduces substantial computation before selection. Second, when only a few initial observations are available, the resulting rough reconstruction may itself be incomplete or inaccurate, making subsequent view selection less reliable. This raises a simple question: ***Can useful views be selected before 3D reconstruction?*** Our key observation is that the marginal value of a candidate view can be estimated from the new evidence it contributes relative to the currently available observations, without first constructing a reliable 3D scene. Rather than deriving candidate value from an intermediate reconstruction, we model it through direct interactions between candidate information and the observations already available. Feed-forward visual and geometric priors provide useful evidence for this reasoning, while the selected set determines whether a candidate contributes complementary or redundant information.

Based on this observation, we introduce **PreView**, a framework for view selection before 3D reconstruction. PreView estimates candidate value directly from available observation cues and explicitly models interactions between candidate views and the currently selected set. A sequential selector then re-evaluates each candidate's marginal contribution as the selected set evolves, favoring complementary observations without requiring a warm-up NeRF, Gaussian scene, or reconstruction-derived uncertainty.

PreView also supports different information and task conditions within the same framework. When candidate RGB images are available, it uses appearance and geometric cues for offline subset selection. When candidate images are unavailable, **PreView-Pose** represents unseen candidates only by camera pose and evaluates them using the context established by the initial observations, enabling pre-acquisition viewpoint recommendation without candidate RGB. The same formulation also allows task information to condition candidate value. Given a query image with user-defined regions of interest, PreView can allocate the limited view budget between global scene quality and task-relevant regions.

Our contributions are summarized as follows:

- We formulate **view selection before 3D reconstruction** as a setting in which informative observations are selected without relying on an optimized intermediate scene representation.
- We introduce **PreView**, which estimates each candidate's marginal contribution through direct interactions between available observations, candidate information, and the evolving selected set, and performs sequential set-aware selection without reconstruction warm-up.
- We show that the same framework supports RGB-aware offline selection, strict pose-only pre-acquisition recommendation, and task-aware reconstruction budget allocation, while substantially reducing computation before selection.

2 Related Work

Neural and feed-forward 3D reconstruction. Neural scene representations have greatly advanced novel-view synthesis and 3D reconstruction. NeRF represents scenes as continuous radiance fields optimized from posed images (Mildenhall et al., 2021), while subsequent extensions improve quality in challenging scenes (Barron et al., 2022). More recently, 3D Gaussian Splatting (3DGS) provides an explicit representation with efficient optimization and real-time rendering (Kerbl et al., 2023).

In parallel, feed-forward models have emerged as strong sources of visual and geometric priors. DUSt3R predicts dense 3D point maps from images (Wang et al., 2024), MASt3R extends this formulation with 3D-aware matching (Leroy et al., 2024), and recent multi-view models further scale feed-forward geometry estimation (Cabon et al., 2025; Yang et al., 2025; Wang et al., 2025). DINOv2 provides transferable visual representations from large-scale self-supervised learning (Oquab et al., 2023). Rather than using these models as final reconstruction systems, PreView uses their features as observation evidence for estimating view utility before reconstruction.

View selection and active reconstruction. View selection and next-best-view planning aim to identify informative observations under limited sensing or reconstruction budgets. ActiveNeRF uses uncertainty to guide acquisition for NeRF (Pan et al., 2022), while FisherRF formulates view utility through Fisher information of radiance fields (Jiang et al., 2024). For 3DGS, POp-GS estimates information gain from Gaussian parameters through optimal experimental design (Wilson et al., 2025), and Next Best Sense applies reconstruction-aware view selection to robotic sensing (Strong et al., 2025). These approaches demonstrate the effectiveness of uncertainty and information derived from a reconstructed scene, but require a scene representation to be available during selection.

GenNBV instead learns a generalizable next-best-view policy and reduces reliance on per-scene optimized representations (Chen et al., 2024). Its main goal is active control and cross-scene generalization, using reinforcement learning to select viewpoints in a continuous action space. In contrast, PreView focuses on *view selection before 3D reconstruction* itself: it directly estimates the marginal utility of candidate observations relative to the currently available set, and supports different observation conditions without requiring an optimized intermediate scene.

Task-aware and ROI-guided reconstruction. Reconstruction quality is not always equally important across a scene. Recent methods therefore use object or region information to focus reconstruction resources on task-relevant content. ROI-GS combines object-guided camera selection with targeted Gaussian optimization (Bui et al., 2025), while OUGS uses semantic information together with Gaussian uncertainty for object-aware view selection (Li et al., 2026).

These works show that view utility depends on what the final reconstruction is expected to preserve. However, task objectives are typically introduced within an existing reconstruction or uncertainty model. PreView instead incorporates task information directly into view utility before reconstruction. A single annotated query image can specify regions of interest, allowing the same selector to shift a limited view budget between global reconstruction quality and task-relevant regions.

3 Method

We study view selection under a limited reconstruction budget. Let $\mathcal{S}_0$ denote the initially observed set, which contains RGB images with known camera parameters and includes an annotated query view (I_q, π_q) with user-defined regions $\mathcal{R} = \{R_m\}_{m=1}^{M}$. Given a candidate pool $\mathcal{C}$, our goal is to select K additional views that best support the reconstruction objective specified by the user. View utility depends on both the requested preference between global and regional fidelity and the information available for each candidate. Our framework handles these factors with a shared selector and does not optimize a NeRF or Gaussian scene representation before selection.

Design rationale. Sparse initial views can yield unreliable warm-up reconstructions, while the value of a candidate is inherently dependent on what has already been observed. PreView therefore estimates marginal value directly from candidate-to-set interactions and updates it sequentially as the selected set evolves.

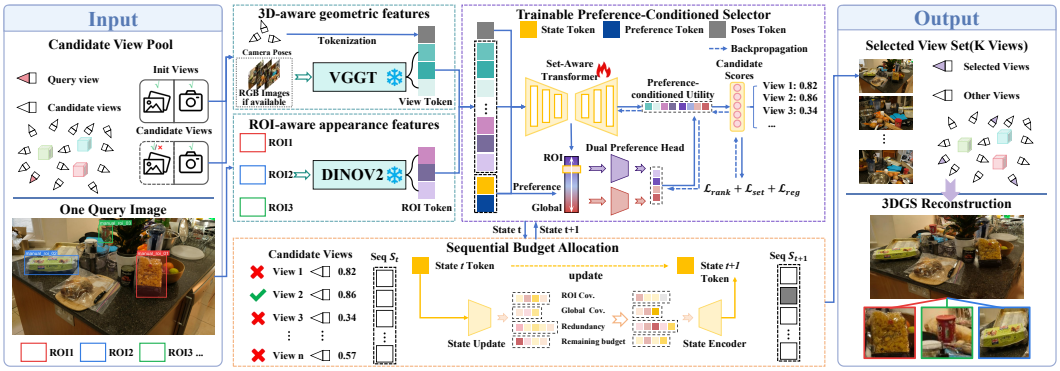


Figure 2: **Overview of PreView for view selection without reconstruction warm-up.** The annotated query view belongs to the initially observed set, whose RGB images provide scene and ROI context before selection. In the RGB-aware regime, candidate images provide additional candidate-specific evidence. In the pose-only regime, each unseen candidate is represented only by its camera pose, with no access to candidate RGB or any feature derived from it. A preference-conditioned, set-aware selector then sequentially allocates the limited reconstruction budget before the selected observations are passed to the final 3DGS reconstruction.

3.1 Problem Formulation

Camera parameters are assumed known for all candidates. In the primary RGB-aware regime, each candidate additionally has an observed RGB image, $\mathcal{C} = \{(I_i, \pi_i)\}_{i=1}^{N}$. We also consider a stricter pose-only regime, in which RGB observations are available only for the initially observed set $\mathcal{S}_0$. During the entire selection process, each unseen candidate is represented solely by its camera pose π_i. No candidate RGB image or any candidate-image-derived appearance, depth, visibility, confidence, or geometric feature is accessed. The two regimes therefore share the same selection objective and differ only in the candidate-side evidence available to the selector.

We seek

$$\mathcal{A}^* = \arg \max_{\substack{\mathcal{A} \subseteq \mathcal{C} \setminus \mathcal{S}_0 \\ |\mathcal{A}| = K}} J_\lambda(\mathcal{S}_0 \cup \mathcal{A}; \mathcal{R}), \tag{1}$$

where $\lambda \in [0, 1]$ denotes the requested preference between global and ROI reconstruction. Since J_λ cannot be evaluated before reconstruction, we estimate its marginal utility from feed-forward evidence available under the corresponding observation regime. This shifts view selection from reasoning over an optimized scene representation to reasoning directly from currently available observations.

3.2 Feature Extraction Before Reconstruction

Instead of deriving uncertainty or information gain from a warm-up reconstruction, we construct selection evidence from frozen feed-forward priors. DINOv2 provides appearance descriptors for the annotated ROIs, while VGGT provides dense geometry and confidence from the currently observed RGB images. In the pose-only regime, both encoders operate only on $\mathcal{S}_0$; unseen candidate images are never passed to either encoder. These features establish scene and ROI context before candidate scoring.

Candidate pose information is available in both regimes. At selection step t, candidate i is encoded using its camera geometry together with pose-derived relations to the scene context and the current selected set:

$$\mathbf{u}_i^{(t)} = E_{\text{pose}}\big([\mathbf{x}_i^{\text{pose}}; \mathbf{d}_i^{\text{pose}}(\mathcal{S}_t)]\big) + \mathbf{e}_i, \tag{2}$$

where $\mathbf{d}_i^{\text{pose}}(\mathcal{S}_t)$ contains pose-based relations between candidate i and the current selected set, including coverage and novelty cues computed without accessing candidate RGB. $\mathbf{e}_i$ is the candidate identity embedding.

When candidate RGB is available, the RGB-aware regime additionally extracts candidate-specific appearance and geometric evidence, including ROI relevance, visibility, projected support, and geometric consistency. This branch is disabled when candidate RGB is unavailable:

$$\widetilde{\mathbf{v}}_i^{(t)} = \begin{cases} E_{\text{view}}([\mathbf{x}_i^{\text{rgb}}; \mathbf{d}_i^{\text{rgb}}(\mathcal{S}_t)]) + \mathbf{e}_i, & a_i = 1, \\ \mathbf{e}_{\text{miss}} + \mathbf{e}_i, & a_i = 0, \end{cases} \tag{3}$$

where a_i indicates candidate-image availability. For pose-only selection, $a_i = 0$ for every unseen candidate throughout the complete selection process. Candidate-specific information is therefore restricted to the known camera pose, while all candidate-image-derived cues are unavailable. Observation availability is treated as an input condition of a shared selector rather than defining separate algorithms.

3.3 Preference-Conditioned Utility Estimation

Candidate utility depends jointly on the candidate, the reconstruction preference, and the views already selected. We therefore combine candidate tokens with ROI, selected-set state, and preference tokens:

$$\begin{aligned} \mathbf{r}_m^{(t)} &= E_{\text{roi}}([\mathbf{z}_m; c_m(\mathcal{S}_t)]), \\ \mathbf{s}^{(t)} &= E_{\text{state}}(\mathbf{q}_t), \qquad \mathbf{p}_\lambda = E_{\text{pref}}(\lambda), \\ H_t &= \mathcal{T}_\theta([U_t; \widetilde{V}_t; R_t; \mathbf{s}^{(t)}; \mathbf{p}_\lambda]). \end{aligned} \tag{4}$$

Here, $c_m(\mathcal{S}_t)$ measures the current support of ROI m, while $\mathbf{q}_t$ summarizes global and regional coverage, redundancy, geometric reliability, and remaining budget. Candidate–candidate interaction captures redundancy and complementarity, while candidate–ROI interaction relates the available candidate evidence to the regions specified by the user.

To make the requested reconstruction objective explicit, two utility heads predict global- and ROI-oriented corrections. Their contributions are routed by the preference token:

$$\begin{aligned} (a_G, a_R) &= \text{softmax}(g_\theta(\mathbf{p}_\lambda)), \\ \delta_i^{(t)} &= a_G h_\theta^G(\mathbf{h}_i^{(t)}) + a_R h_\theta^R(\mathbf{h}_i^{(t)}). \end{aligned} \tag{5}$$

The same contextualized candidate can therefore receive different utility under different reconstruction objectives. The two endpoint preferences correspond to Ours-Global and Ours-ROI, while intermediate preference values continuously change the allocation behavior without modifying the selector architecture.

3.4 Sequential View Selection

View utility is set-dependent: once a region is sufficiently supported, another similar observation provides limited additional value. We therefore select views sequentially according to their marginal contribution to the current selected set rather than ranking all candidates independently.

Let $G(\mathcal{S})$ denote global support, $C_m(\mathcal{S})$ the support of ROI m, and $D(\mathcal{S})$ selected-view redundancy. We define average regional support as $A(\mathcal{S}) = \frac{1}{M}\sum_m C_m(\mathcal{S})$ and use a complementary balanced term

$$B(\mathcal{S}) = \sum_{m=1}^{M} w_m \log(\epsilon + C_m(\mathcal{S})), \tag{6}$$

whose concavity assigns larger marginal value to poorly supported ROIs. This discourages repeatedly allocating observations to an already well-covered region while leaving another task region under-observed.

For $\Delta F_i = F(\mathcal{S}_t \cup \{v_i\}) - F(\mathcal{S}_t)$, the stable reconstruction-free proxy combines scene support, ROI support, multi-ROI balance, and redundancy:

$$u_i^{(t)} = \lambda_g \Delta G_i + \lambda_a \Delta A_i + \lambda_b \Delta B_i - \lambda_v \Delta D_i + u_{\text{stab}}(v_i \mid \mathcal{S}_t). \tag{7}$$

The evidence used to evaluate these terms follows the observation regime. RGB-aware selection can use candidate-specific appearance, visibility, and geometric consistency. Pose-only selection removes

all candidate-image-derived terms; candidate utility is estimated from the camera pose relative to scene and ROI context inferred from the initially observed views.

The learned selector provides a bounded correction to this stable proxy:

$$s_i^{(t)} = u_i^{(t)} + \alpha \, \mathrm{clip}(\delta_i^{(t)}, -\rho, \rho), \tag{8}$$

so that learned task interactions refine rather than replace the interpretable reconstruction-free evidence.

Starting from $\mathcal{S}_0$, we iteratively choose

$$v_t^* = \arg\max_{v_i \notin \mathcal{S}_t} s_i^{(t)}, \qquad \mathcal{S}_{t+1} = \mathcal{S}_t \cup \{v_t^*\}, \tag{9}$$

and update global support, ROI coverage, redundancy, and remaining budget before rescoring the remaining candidates. Because candidate utilities are recomputed after every selected view, redundant observations are naturally suppressed as their marginal contribution decreases. The procedure therefore performs sequential reconstruction-budget allocation rather than static top-K selection.

Supervision without reconstruction warm-up. The selector is calibrated using proxy utilities that are available before scene reconstruction. For a selected set, the proxy teacher follows the same global, regional, balanced-coverage, and redundancy terms:

$$\widetilde{J}_\lambda(\mathcal{S}) = \lambda_g G(\mathcal{S}) + \lambda_a A(\mathcal{S}) + \lambda_b B(\mathcal{S}) - \lambda_v D(\mathcal{S}). \tag{10}$$

Its marginal gains supervise candidate ordering and selected-set quality. The teacher never uses final PSNR or LPIPS, held-out evaluation views, or an optimized NeRF or Gaussian representation.

To train the shared selector for pose-only operation, we additionally mask the candidate-image branch during calibration. For these samples, $\widetilde{\mathbf{v}}_i^{(t)}$ is replaced by $\mathbf{e}_{\mathrm{miss}} + \mathbf{e}_i$, and all candidate-image-derived proxy terms are disabled. The remaining information exactly matches the pose-only inference condition: RGB context from $\mathcal{S}_0$, candidate camera poses, selected-set state, and task preference. Thus, pose-only selection is learned without using candidate RGB as either selector input or proxy evidence. Detailed proxy construction and training losses are provided in the supplementary material.

After selection, the chosen observations are passed to the common 3DGS backend for final reconstruction.

4 Experiments

4.1 Experimental Setup

We evaluate PreView on seven Mip-NeRF 360 scenes (Barron et al., 2022), with additional cross-dataset evaluation on LLFF and two self-collected outdoor UAV scenes in the supplementary material. Unless otherwise specified, all methods start from 10 initial views, select $K = 5$ additional views, and use the same 3DGS backend for 10,000 iterations. We compare with Random+3DGS, FisherRF (Jiang et al., 2024), ROI-GS (Bui et al., 2025), OUGS (Li et al., 2026), and POpGS-DOpt (Wilson et al., 2025).

We evaluate both RGB-aware selection and a strict pose-only regime in which candidate RGB is withheld throughout selection. Metrics include global PSNR/LPIPS and average/minimum ROI PSNR and LPIPS. Additional sensitivity, preference, cross-dataset, and implementation results are provided in the supplementary material.

4.2 View Selection Before Reconstruction

We first test whether useful views can be selected without an optimized intermediate reconstruction. To isolate this factor, we fix PreView to the global objective and compare the RGB-aware and pose-only settings.

Table 1 shows that PreView achieves strong global reconstruction quality in both regimes without reconstruction warm-up. FisherRF and POpGS-DOpt can score unseen poses, but require an optimized scene representation from the initial observations. ROI-GS and OUGS additionally depend on candidate images and therefore do not apply to the pose-only setting.

Table 1: **Selection requirements and downstream reconstruction quality.** Filled circles indicate that the corresponding input or computation is not required during selection. PreView achieves the strongest global reconstruction quality while removing reconstruction warm-up, with PreView-Pose further requiring no candidate RGB.

Method	No Warm-up Recon.	No Candidate RGB	G-PSNR ↑	G-LPIPS ↓
Random + 3DGS	●	●	18.324	0.326
FisherRF	○	●	19.821	<u>0.276</u>
ROI-GS	○	○	18.810	0.309
OUGS	○	○	19.339	0.295
POpGS-DOpt	○	●	<u>20.019</u>	0.278
Ours-Pose	●	●	20.606	0.254
Ours-RGB	●	○	**20.676**	**0.253**

Table 2: Task-conditioned reconstruction quality in the RGB-aware setting. **The global preference provides the strongest overall reconstruction, while the ROI preference achieves the best average and worst-case regional quality with competitive global fidelity.**

Method	G-PSNR ↑	Avg. ROI PSNR ↑	Min ROI PSNR ↑	Avg. ROI LPIPS ↓
Random + 3DGS	18.324	24.744	21.492	0.00503
FisherRF	19.821	25.349	22.229	0.00442
ROI-GS	18.810	25.662	22.480	0.00386
OUGS	19.339	24.782	21.875	0.00504
POpGS-DOpt	<u>20.019</u>	25.680	22.686	0.00460
Ours-RGB-Global	**20.676**	<u>26.519</u>	<u>23.613</u>	<u>0.00320</u>
Ours-RGB-ROI	20.017	**27.005**	**24.356**	**0.00301**

The small gap between our RGB-aware and pose-only variants further shows that candidate appearance is helpful but not essential. Even without candidate RGB, the relation between candidate poses and the context established by the initial observations provides an effective signal for view selection. Results under sparser initial observations are reported in the supplementary material.

4.3 Task-Conditioned Reconstruction

We next examine whether the same selector can adapt its decisions to different reconstruction objectives. In the RGB-aware setting, we compare the global and ROI-oriented preference endpoints while keeping the selector and candidate pool unchanged.

Table 2 shows a clear shift in reconstruction behavior. The global preference provides the strongest scene-wide quality, whereas the ROI preference improves both average ROI fidelity and the quality of the weakest reconstructed region while retaining competitive global performance. Since the two operating points use the same selector, candidate utility changes with the requested reconstruction objective rather than being fixed for each view. This behavior supports our view of candidate value as objective-dependent rather than intrinsic to the viewpoint itself.

Figure 3 visualizes the global–regional trade-off between the evaluated preference settings. Additional intermediate preference samples are provided in the supplementary material, where we further study continuous control of reconstruction priorities.

Figure 4 further confirms that ROI-oriented selection improves local reconstruction in task-relevant regions while preserving coherent surrounding structure.

4.4 Selection Efficiency

Table 3 reports all computation required before the final subset is determined, excluding the common final 3DGS reconstruction. Reconstruction-guided methods incur substantial warm-up optimization before scoring, whereas PreView removes this stage. The pose-only variant further avoids candidate-image processing, reducing both computation and required observations.

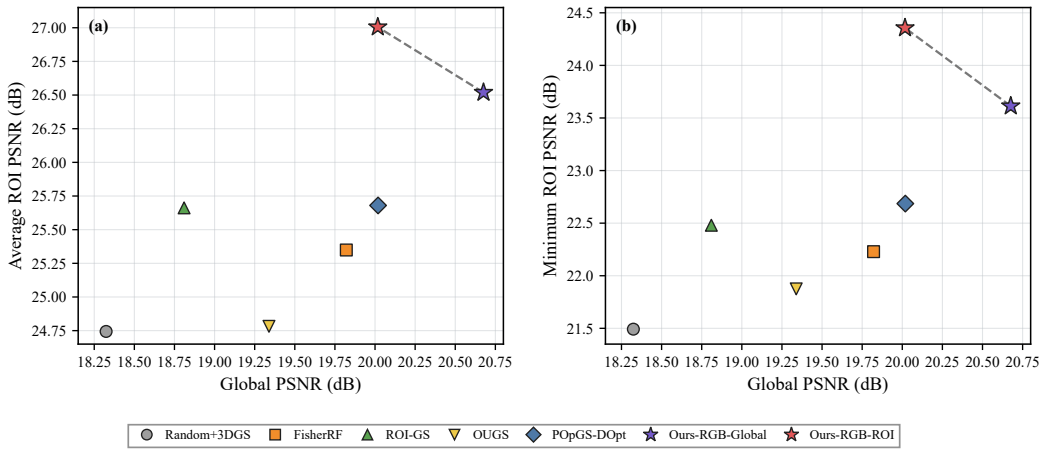


Figure 3: **Global–regional reconstruction trade-off.** Changing the reconstruction preference shifts the selected view budget between scene-wide quality and regional fidelity, demonstrating controllable task-conditioned selection.

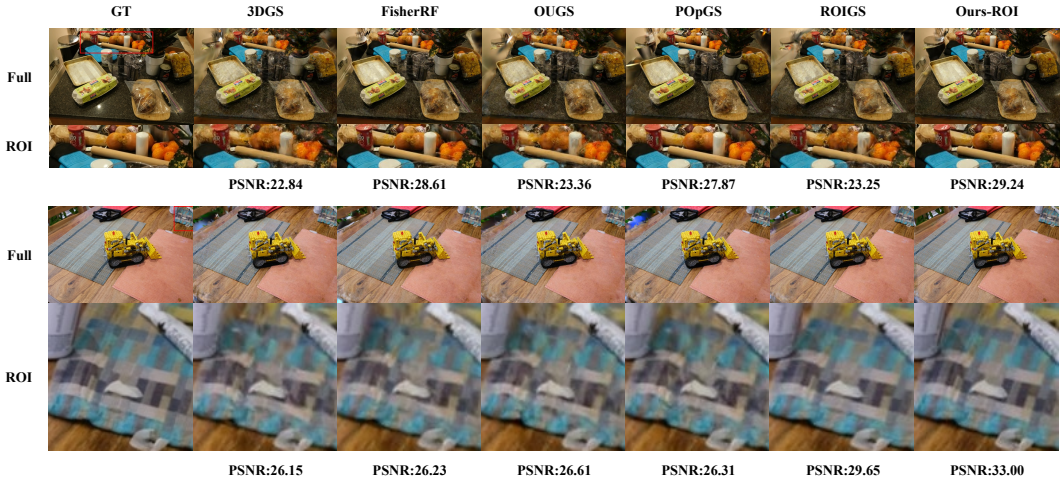


Figure 4: **Qualitative reconstruction comparison on representative scenes.** ROI-oriented PreView recovers finer details in user-specified regions while maintaining coherent surrounding structure, consistent with the regional improvements in Table 2. Boxes indicate the enlarged regions.

4.5 Ablation Study

We ablate the main components for observation evidence, task conditioning, and set-aware selection. All variants follow the same experimental protocol.

Table 4 shows that feed-forward geometry improves both global and regional reconstruction, while ROI-guided tokenization and preference routing are essential for producing preference-dependent selections. The ROI geometry gate further improves regional reliability. Removing sequential set optimization mainly reduces global quality, indicating that static ranking can retain locally useful views but fails to suppress candidates that become redundant after earlier selections.

Additional sensitivity and calibration analyses are provided in the supplementary material.

Table 3: Pre-selection computation cost. *Front-end Prep.* includes COLMAP/input preparation for competing methods and feed-forward evidence/token construction for ours. *Warm-up Recon.* denotes intermediate scene optimization before scoring. **Our main efficiency gain comes from removing reconstruction warm-up from the decision loop.**

Method	Front-end Prep. (s) ↓	Warm-up Recon. (s) ↓	Selector (s) ↓	Total (s) ↓
FisherRF	813.47	243.25	51.88	1108.60
ROI-GS	524.73	243.25	0.80	768.77
OUGS	524.73	243.25	1.02	768.99
POpGS-DOpt	524.73	243.25	68.84	836.82
Ours-RGB-Global	482.80	**0**	3.26	**486.06**
Ours-RGB-ROI	482.80	**0**	4.50	**487.30**
Ours-PoseOnly-Global	1.950	**0**	0.221	**2.171**
Ours-PoseOnly-ROI	1.950	**0**	0.221	**2.171**

Table 4: **Ablation of the main selector components.** Global PSNR is measured under the global preference and ROI metrics under the ROI preference; "Distinct G/R" counts scenes with different selected subsets. Geometry, preference conditioning, and sequential reasoning contribute complementary improvements to the final selection.

Variant	Global PSNR ↑	ROI Avg. PSNR ↑	ROI Min. PSNR ↑	Distinct G/R
Full model	20.676	27.005	24.356	7/7
w/o Feed-Forward Geometry	20.441	26.241	23.929	7/7
w/o ROI-Guided Tokenization	20.419	25.610	23.157	0/7
w/o Preference Routing	20.336	25.444	22.518	0/7
w/o Sequential Set Optimization	19.693	26.344	24.690	6/7
w/o ROI Geometry Consistency Gate	20.463	25.917	23.417	7/7

5 Conclusion and Limitations

We presented PreView, a reconstruction-free framework for view selection before 3D reconstruction. Instead of estimating view utility from a warm-up scene reconstruction, our method reasons directly from feed-forward observation evidence and user-specified task information before reconstruction. The same framework supports both RGB-aware and pose-only selection, and can adapt the limited view budget between global scene quality and ROI-oriented fidelity through preference conditioning and sequential set reasoning. Experiments show that this formulation maintains strong reconstruction quality, supports controllable global–regional trade-offs, and reduces the computation required before the final view subset is determined.

Several limitations remain. Task-conditioned selection becomes more difficult when ROIs are extremely small, heavily occluded, or poorly supported by the initial observations. In addition, the pose-only setting should be interpreted as pre-acquisition viewpoint recommendation rather than a complete active reconstruction system, since trajectory planning, online camera control, and closed-loop acquisition are beyond the current scope. Future work will explore more robust reasoning under weak ROI evidence and extensions toward closed-loop active reconstruction.

AI Use Statement

Generative AI tools were used to assist with drafting, revising, and polishing parts of the manuscript. All AI-assisted text was reviewed and edited by the authors, who take full responsibility for the final content.

Reproducibility Statement

Implementation details, dataset splits, ROI annotation and evaluation protocols, pose-only restrictions, runtime measurement, and selector calibration are provided in the appendix. We additionally provide anonymized code for the core view-selection and evaluation pipeline, together with configuration files and experimental metadata, to facilitate reproduction of the reported results.

References

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5460–5469. IEEE, 2022.

Quoc-Anh Bui, Gilles Rougeron, Géraldine Morin, and Simone Gasparini. Roi-gs: interest-based local quality 3d gaussian splatting, 2025.

Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1050–1060. IEEE, 2025.

Xiao Chen, Quanyi Li, Tai Wang, Tianfan Xue, and Jiangmiao Pang. Gennbv: Generalizable next-best-view policy for active 3d reconstruction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16436–16445. IEEE, 2024.

Wen Jiang, Boshu Lei, and Kostas Daniilidis. Fisherrf: Active view selection and mapping with radiance fields using fisher information. In *European conference on computer vision*, pp. 422–440. Springer, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. volume 42, pp. 139–1, 2023.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Haiyi Li, Qi Chen, Denis Kalkofen, and Hsiang-Ting Chen. Ougs: Active view selection via object-aware uncertainty estimation in 3dgs, 2026.

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. volume 65, pp. 99–106. ACM New York, NY, USA, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision, 2023.

Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. Activenerf: Learning where to see with uncertainty estimation. In *European Conference on Computer Vision*, pp. 230–246. Springer, 2022.

Matthew Strong, Boshu Lei, Aiden Swann, Wen Jiang, Kostas Daniilidis, and Monroe Kennedy. Next best sense: Guiding vision and touch with fisherrf for 3d gaussian splatting. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3204–3210. IEEE, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 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. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3646–3655. IEEE, 2025.

Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935. IEEE, 2025.

Appendix

A Sequential Reconstruction-Budget Allocation

The main paper treats view selection as a sequential budget-allocation problem rather than independent candidate ranking. This appendix provides additional intuition and details on this formulation.

A.1 Set-Dependent View Utility

A candidate view does not have a fixed utility independent of the selected set. For example, a view that strongly observes an annotated ROI can be valuable when that region is poorly supported, but becomes less useful after several similar views have already been selected. Likewise, a viewpoint that initially improves global camera diversity may become redundant once nearby observations are included.

This effect is particularly important under a small view budget. A static top-K strategy may repeatedly select individually strong but mutually redundant views, or over-allocate observations to an easy ROI while leaving other task regions weakly supported. We therefore evaluate each candidate by its marginal contribution conditioned on the current selected set $\mathcal{S}_t$.

Let $G(\mathcal{S})$ denote global scene support, $A(\mathcal{S})$ average ROI support, $B(\mathcal{S})$ balanced multi-ROI support, and $D(\mathcal{S})$ redundancy. For candidate v_i, we define

$$\Delta F_i = F(\mathcal{S}_t \cup \{v_i\}) - F(\mathcal{S}_t), \tag{11}$$

and compute the reconstruction-free proxy utility

$$u_i^{(t)} = \lambda_g \Delta G_i + \lambda_a \Delta A_i + \lambda_b \Delta B_i - \lambda_v \Delta D_i + u_{\text{stab}}(v_i \mid \mathcal{S}_t). \tag{12}$$

The balanced regional term is

$$B(\mathcal{S}) = \sum_{m=1}^{M} w_m \log(\epsilon + C_m(\mathcal{S})), \tag{13}$$

where $C_m(\mathcal{S})$ measures the support of ROI m. Because the logarithm is concave, additional support receives a larger marginal value for poorly observed regions. This complements average ROI coverage and helps prevent the limited budget from concentrating on only one easy region.

A.2 Sequential Selection

The learned selector refines the stable reconstruction-free proxy through a bounded correction:

$$s_i^{(t)} = u_i^{(t)} + \alpha \ \text{clip}\left(\delta_i^{(t)}, -\rho, \rho\right). \tag{14}$$

The proxy preserves stable geometric, coverage, and redundancy cues, while $\delta_i^{(t)}$ models interactions among the candidate, annotated regions, task preference, and the views already selected.

Starting from the initial set $\mathcal{S}_0$, we repeatedly select

$$v_t^* = \arg\max_{v_i \notin \mathcal{S}_t} s_i^{(t)}, \qquad \mathcal{S}_{t+1} = \mathcal{S}_t \cup \{v_t^*\}. \tag{15}$$

After each selection, global support, ROI coverage, redundancy, and the remaining budget are updated before the remaining candidates are rescored. Consequently, the marginal value of a candidate decreases when its useful content has already been covered by earlier selections.

The same sequential procedure is used for both Ours-Global and Ours-ROI. The two operating points share the candidate evidence, selector parameters, and state-update mechanism; they differ only in the requested preference and the corresponding allocation weights. Hence, preference conditioning determines *what* the reconstruction budget should prioritize, while sequential reasoning determines *how* that budget should be distributed across complementary observations.

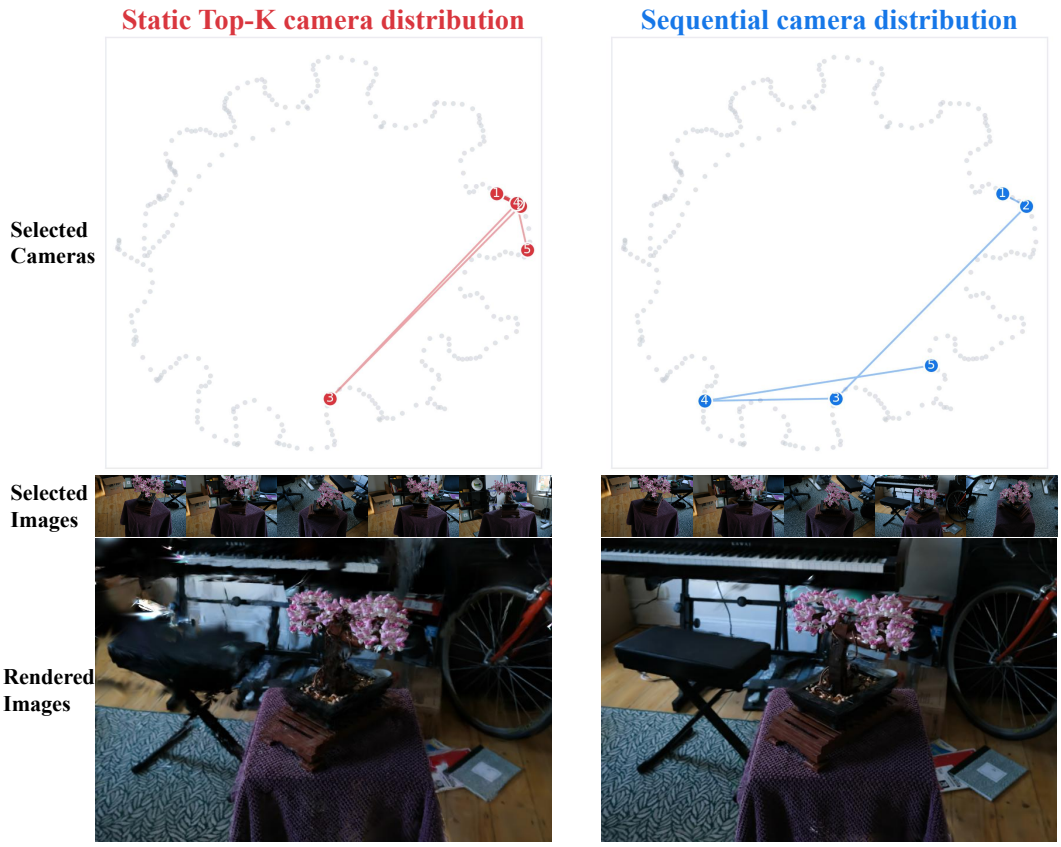


Figure 5: Static Top-K versus sequential set-aware view selection. Static ranking produces several clustered viewpoints with overlapping observations, while sequential rescoring yields a more complementary camera distribution. The selected images and corresponding reconstruction results are shown below.

A.3 Relation to Static Top-K Ranking

A static selector assigns each candidate an independent score and directly chooses the top-K views. Such a formulation cannot account for the fact that a view may become redundant after another observation has been selected. Our sequential formulation instead recomputes marginal utility after every selection.

As illustrated in Fig. 5, static ranking tends to retain several nearby viewpoints, whereas sequential rescoring spreads the limited budget across more complementary observations. This qualitative behavior is consistent with the ablation in Table 4, where removing sequential set optimization mainly degrades global reconstruction quality.

B Reconstruction-Free Proxy Supervision

The main paper calibrates the selector without supervision from an optimized scene reconstruction. This appendix clarifies how the proxy teacher is constructed and how the same training procedure supports both RGB-aware and pose-only selection.

B.1 PROXY TEACHER

Because view selection is performed before final 3DGS reconstruction, its supervision should also be available at selection time. We therefore construct the teacher from reconstruction-free signals describing global scene support, ROI support, multi-ROI balance, and view redundancy.

For a selected set $\mathcal{S}$, the proxy objective follows

$$\widetilde{J}_\lambda(\mathcal{S}) = \lambda_g G(\mathcal{S}) + \lambda_a A(\mathcal{S}) + \lambda_b B(\mathcal{S}) - \lambda_v D(\mathcal{S}), \tag{16}$$

using the same terms as in the main paper. At selection step t, candidate v_i receives the teacher score

$$y_i^{(t)} = \widetilde{J}_\lambda(\mathcal{S}_t \cup \{v_i\}) - \widetilde{J}_\lambda(\mathcal{S}_t). \tag{17}$$

The teacher therefore provides set-dependent marginal preferences rather than a fixed target for each candidate.

The evidence used to compute these quantities depends on the observation regime. RGB-aware supervision can use candidate-specific appearance and geometry available from the observed candidate image. For pose-only samples, all candidate-image-derived quantities are removed. The proxy is computed only from the scene and ROI context inferred from the initially observed RGB views, candidate camera poses, and the evolving selected-set state.

Importantly, neither regime uses final PSNR or LPIPS, held-out evaluation views, or an optimized NeRF/3DGS representation to construct the teacher.

B.2 SELECTOR CALIBRATION

The selector is trained to preserve the relative candidate preferences induced by the proxy teacher rather than directly regress final reconstruction quality. A pairwise ranking loss encourages a candidate with larger proxy gain to receive a higher score:

$$\mathcal{L}_{\text{rank}} = \sum_{(v^+, v^-)} \max\left(0,\, \mu - s_\theta(v^+) + s_\theta(v^-)\right), \tag{18}$$

where v^+ and v^- denote candidates with higher and lower teacher utility.

We additionally use a listwise alignment term $\mathcal{L}_{\text{list}}$ to match the predicted candidate distribution to the proxy distribution, and a set-level term $\mathcal{L}_{\text{set}}$ to encourage sets with larger proxy utility to receive larger accumulated scores. The learned correction is regularized by

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_v\left[\delta_\theta(v)^2\right], \tag{19}$$

so that it calibrates rather than replaces the stable reconstruction-free utility. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{rank}} + \eta_1 \mathcal{L}_{\text{list}} + \eta_2 \mathcal{L}_{\text{set}} + \eta_3 \mathcal{L}_{\text{reg}}. \tag{20}$$

B.3 CALIBRATION FOR POSE-ONLY SELECTION

Pose-only calibration is implemented by masking the candidate-image branch. For these samples, the candidate visual token is replaced by the learned missing-observation token, and all quantities derived from candidate RGB are disabled, including candidate appearance, depth, visibility, confidence, and geometric consistency.

The remaining information matches the pose-only inference condition exactly: RGB observations from the initial set $\mathcal{S}_0$, including the annotated query view, are used to establish scene and ROI context, while each unseen candidate is represented only by its camera pose. Candidate RGB is therefore used neither as selector input nor as proxy-teacher evidence for pose-only samples.

At inference time, this restriction is enforced throughout the complete selection process. Candidate images, although available in an offline benchmark, are withheld from the selector until the final subset has been determined. Only then are the selected RGB observations revealed and passed together with $\mathcal{S}_0$ to the common 3DGS reconstruction backend.

The RGB-aware and pose-only regimes therefore share the same selector, preference mechanism, and sequential allocation procedure. They differ only in whether candidate-side RGB observations are available at selection time.

C ROI Geometry and Visibility Details

The main paper uses ROI-level evidence to guide task-conditioned view selection. This appendix provides additional details on how ROI support is estimated and how unreliable geometric observations are filtered.

C.1 Refined ROI Visibility Estimation

A naive ROI visibility estimation can be obtained by projecting the annotated region from the query view to a candidate camera using the estimated scene geometry. However, such projection alone may overestimate the usefulness of a candidate view, since a projected region does not necessarily correspond to a reliably observed surface. Occlusion, inaccurate depth estimation, and low-confidence geometry can all introduce unreliable ROI evidence.

To address this issue, we estimate ROI support by jointly considering projected coverage, geometric consistency, and confidence of the reconstructed geometry. For ROI R_m and candidate view v_i, we first obtain projected ROI evidence from the feed-forward geometry estimated from the currently available observations. The resulting support is then refined by reliability weighting:

$$C_m(v_i) = \sum_{x \in R_m} w_{\mathrm{geo}}(x) w_{\mathrm{conf}}(x) w_{\mathrm{vis}}(x), \tag{21}$$

where w_{geo}, w_{conf}, and w_{vis} measure geometric consistency, confidence, and visibility reliability, respectively.

This refinement prevents the selector from assigning high utility to views that only provide apparent ROI coverage but unreliable observations. The same principle is used when computing marginal ROI gains during sequential selection.

C.2 Geometry Consistency Gate

Although feed-forward geometry provides useful scene priors before reconstruction, its uncertainty varies across different regions and views. Therefore, we introduce a geometry consistency gate to avoid over-trusting unstable ROI evidence.

For each candidate ROI observation, the gate evaluates whether the projected geometry agrees with the available geometric context. When the consistency is high, ROI evidence contributes normally to the candidate utility. When the observation is unreliable, the contribution is softly reduced rather than directly discarded:

$$g_i^m = \sigma\left(\phi(\mathbf{z}_i^m)\right), \tag{22}$$

where $\mathbf{z}_i^m$ contains geometric consistency cues and $\phi(\cdot)$ is a learned mapping.

The refined ROI contribution is then weighted by

$$\hat{C}_m(v_i) = g_i^m C_m(v_i). \tag{23}$$

The gate serves a different purpose from preference conditioning. Preference conditioning determines whether the reconstruction budget should emphasize global scene support or task regions, whereas the geometry gate determines whether the available ROI evidence is reliable enough to influence the decision.

It is important to note that this refinement is only applied when the required candidate-side geometric evidence is available. In the pose-only regime, candidate images and candidate-image-derived geometry are unavailable during selection; therefore, the selector relies on pose-based relations to the scene context inferred from the initial observations.

D Experimental Protocol and Implementation Details

This appendix provides additional details about the evaluation protocol, including dataset construction, ROI annotation, baseline adaptation, runtime measurement, and implementation settings.

D.1 Dataset and View Splits

We evaluate our method primarily on seven scenes from Mip-NeRF 360 (Barron et al., 2022), including *bicycle*, *bonsai*, *counter*, *garden*, *kitchen*, *room*, and *stump*. Additional evaluations are conducted on LLFF scenes and two self-collected outdoor scenes to examine cross-dataset generalization.

For each scene, all images are associated with known camera poses. We divide the views into three subsets: initial observations $\mathcal{S}_0$, candidate views $\mathcal{C}$, and held-out test views. The initial observations are available before selection and are used to construct the scene context. The selector chooses K additional views from the candidate pool, after which the selected views and the initial observations are used for final 3DGS reconstruction.

Unless otherwise specified, all experiments use 10 initial views, select $K = 5$ additional views, and optimize the final 3DGS model for 10,000 iterations.

D.2 Query and ROI Annotation

For each scene, one image from the initial observation set is selected as the query view and manually annotated with multiple regions of interest (ROIs). The query image is not an additional observation; it is part of the initially available RGB set.

The annotated ROIs specify user-interested regions that define the task objective. ROI annotations are fixed before view selection and are used by all methods under the same evaluation protocol. The candidate selection process does not access held-out test views or any information derived from final reconstruction quality.

ROI metrics are computed by projecting the annotated regions into evaluation views using the reconstructed geometry and measuring reconstruction quality only on valid projected regions. This avoids assigning scores to regions that are not visible in the evaluation view.

D.3 Pose-Only Evaluation Protocol

The pose-only setting is designed to evaluate whether informative viewpoints can be selected before candidate observations are acquired.

In this setting, RGB images are available only for the initial observation set $\mathcal{S}_0$, including the annotated query view. During the entire selection process, every candidate view is represented only by its camera pose. Candidate RGB images are strictly withheld from the selector, and no feature derived from candidate RGB is used, including appearance descriptors, candidate depth, visibility, confidence, or geometric consistency.

Although candidate images exist in the offline benchmark, they are revealed only after the view subset has been selected. The selected RGB observations are then combined with the initial observations for the common final 3DGS reconstruction and evaluation.

Therefore, the pose-only experiment evaluates viewpoint recommendation under limited observation availability rather than complete online trajectory planning or closed-loop active reconstruction.

D.4 Baseline Adaptation and Fair Comparison

All compared methods use the same initial observations, candidate pool, view budget, final 3DGS reconstruction backend, and held-out evaluation views.

For reconstruction-dependent baselines, including FisherRF and POpGS-DOpt, the methods are evaluated according to their original settings, where a warm-up scene representation is optimized before candidate scoring. For ROI-aware baselines, including ROI-GS and OUGS, the provided task information is adapted to our ROI-based evaluation protocol.

After view selection, all selected view sets are reconstructed using the same 3DGS implementation to ensure that differences mainly reflect the quality of the selected observations rather than reconstruction backend differences.

Table 5: Cross-dataset evaluation on LLFF and two self-collected scenes. LLFF results are averaged over eight scenes, while the two self-collected scenes are reported separately. Best and second-best results within each setting are shown in bold and underline, respectively.

Dataset	Method	G-PSNR ↑	Avg. ROI PSNR ↑	Min ROI PSNR ↑
LLFF Avg.	FisherRF	23.422	28.092	26.078
	ROI-GS	22.726	27.303	25.609
	OUGS	<u>24.138</u>	27.936	26.520
	POpGS-DOpt	24.006	27.931	<u>26.552</u>
	Ours-RGB-Global	**24.139**	27.918	26.017
	Ours-RGB-ROI	24.101	**28.437**	**26.955**
Self Scene 1	FisherRF	17.333	22.608	16.919
	ROI-GS	17.248	25.990	<u>24.837</u>
	OUGS	17.248	25.990	<u>24.837</u>
	POpGS-DOpt	<u>18.126</u>	21.834	16.697
	Ours-RGB-Global	**18.201**	<u>26.976</u>	24.669
	Ours-RGB-ROI	17.418	**27.776**	**26.488**
Self Scene 2	FisherRF	20.359	23.598	19.538
	ROI-GS	20.522	25.318	19.479
	OUGS	20.469	<u>25.331</u>	<u>20.312</u>
	POpGS-DOpt	20.355	24.325	18.888
	Ours-RGB-Global	**21.267**	23.964	18.519
	Ours-RGB-ROI	<u>20.674</u>	**25.541**	**20.381**

D.5 Runtime Measurement

We report selection-related runtime rather than final reconstruction time, because all methods share the same downstream 3DGS reconstruction backend.

Runtime is divided into three components: (1) front-end preparation, including feature extraction and scene context construction; (2) warm-up reconstruction, when required by a method before selection; and (3) view selection.

For reconstruction-free methods, the warm-up reconstruction stage is absent. For reconstruction-dependent methods, the optimization time required before candidate scoring is included in the reported selection pipeline cost.

All runtime measurements are performed using the same hardware configuration.

E Additional Experiments

E.1 Cross-Dataset Generalization

We further evaluate whether the behavior of PreView transfers beyond the Mip-NeRF 360 benchmark. We use eight LLFF scenes and two independently collected scenes with different scene content and camera distributions. The same selector, preference settings, and reconstruction protocol are used without dataset-specific modification. All methods start from 10 initial views, select $K = 5$ additional views, and use the same 3DGS backend for 10,000 iterations.

Table 5 shows that the preference-dependent behavior of PreView transfers across different scene distributions. On LLFF, PreView-RGB-Global achieves the strongest global reconstruction quality, while PreView-RGB-ROI obtains the best average and minimum ROI fidelity.

The two self-collected scenes show a similar separation between the two operating modes. PreView-RGB-Global achieves the highest G-PSNR on both scenes. PreView-RGB-ROI consistently shifts reconstruction quality toward the task regions, achieving the best average ROI PSNR on both scenes and the best minimum ROI PSNR on Scene 1 and Scene 2.

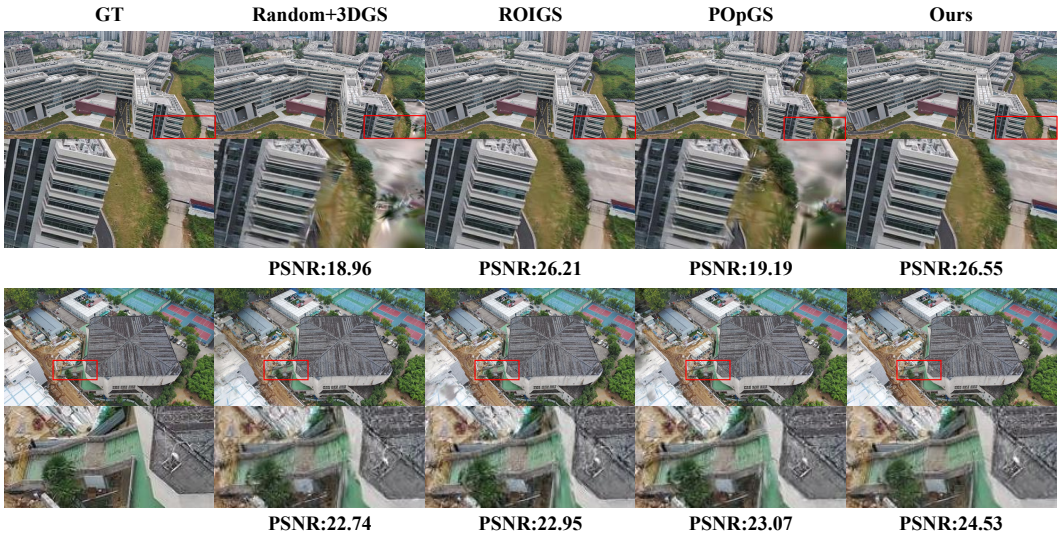


Figure 6: **Qualitative results on self-collected scenes.** Representative held-out views from two independently captured scenes. Compared with other view selection methods, PreView produces higher-fidelity reconstructions under the same 3DGS backend. The reported PSNR values correspond to the displayed test views.

These results suggest that the preference behavior is not specific to Mip-NeRF 360 or LLFF. Without dataset-specific modification, the global mode consistently favors scene-wide reconstruction, whereas the ROI mode reallocates the limited view budget toward user-specified regions.

Qualitative results on self-collected scenes. Figure 6 further shows representative held-out views from the two self-collected scenes. Under the same reconstruction backend, PreView preserves finer scene structures and produces reconstructions closer to the ground truth across both examples. These qualitative results complement the quantitative cross-dataset evaluation and further demonstrate that the selected views remain effective on independently captured scenes with camera distributions and scene content different from the primary Mip-NeRF 360 benchmark.

E.2 Sensitivity to View Budget and Initialization

We further study the sensitivity of PreView to the number of selected views and the amount of initial scene evidence. All experiments follow the same Mip-NeRF 360 protocol and use the same 3DGS reconstruction backend.

Effect of the view budget. We first fix the number of initial observations to 10 and vary the additional view budget $K \in \{3, 5, 10\}$. Each entry in Table 6 reports Global PSNR / Average ROI PSNR.

The two preference modes remain complementary across different budgets. Under the tightest budget, PreView-ROI already provides the strongest regional quality, while PreView-Global remains close to the strongest global baseline. As more views become available, the two modes become more clearly separated: PreView-Global increasingly favors scene-wide support, whereas PreView-ROI continues allocating views toward task-relevant regions. This trend indicates that preference control becomes more expressive when the budget provides greater freedom for view allocation.

Effect of initialization. We next fix the additional-view budget to $K = 5$ and vary the number of initial observations. The query-only setting contains no auxiliary multi-view context before selection, while the 5-view and 10-view settings progressively provide stronger scene evidence.

Table 6: View-budget analysis on Mip-NeRF 360 with 10 initial views and 10k 3DGS iterations. Each entry reports Global PSNR / Average ROI PSNR. Best and second-best results under each budget are shown in bold and underline, respectively.

Method	$K = 3$	$K = 5$	$K = 10$
3DGS	18.683 / 24.335	18.324 / 24.744	18.720 / 24.165
FisherRF	19.625 / 24.633	19.821 / 25.349	<u>21.588</u> / 25.763
ROI-GS	19.095 / 25.432	18.810 / 25.662	19.708 / 26.450
OUGS	19.449 / 24.380	19.339 / 24.782	20.203 / 24.705
POpGS-DOpt	<u>19.927</u> / 24.820	20.019 / <u>25.680</u>	21.286 / 25.602
Ours-Global	**19.984** / <u>25.471</u>	**20.676** / 26.519	**21.632** / <u>26.898</u>
Ours-ROI	19.276 / **25.827**	<u>20.017</u> / **27.005**	20.999 / **27.055**

Table 7: Initialization analysis on Mip-NeRF 360 with $K = 5$ and 10k 3DGS iterations. Each entry reports Global PSNR / Average ROI PSNR. "–" indicates that the required warm-up reconstruction could not be established.

Method	**Query only**	**5 initial views**	**10 initial views**
3DGS	–	15.373 / 23.429	18.324 / 24.744
FisherRF	–	17.578 / 24.544	19.821 / 25.349
ROI-GS	–	16.112 / 24.923	18.810 / 25.662
OUGS	–	17.129 / 24.397	19.339 / 24.782
POpGS-DOpt	–	<u>18.030</u> / **25.682**	<u>20.019</u> / 25.680
Ours-Global	**16.560** / <u>23.051</u>	**18.220** / 24.747	**20.676** / <u>26.519</u>
Ours-ROI	<u>16.127</u> / **23.210**	17.614 / <u>25.461</u>	20.017 / **27.005**

The initialization sweep highlights the different dependence of the compared methods on early scene evidence. In the query-only setting, methods that require an initialized reconstruction or multi-view scene context cannot establish the state needed for utility estimation, whereas PreView remains operational. With 5 and 10 initial views, all methods benefit from stronger scene context, while the intended preference behavior of PreView becomes more stable as the available evidence increases.

These results are consistent with the motivation for selecting views before reconstruction: PreView does not require a reliable optimized scene to begin selection, while stronger initial observations can further improve the quality of the available context.

E.3 Robustness to Initialization and ROI Annotation

We further evaluate the robustness of PreView under variations in both scene initialization and task specification. We repeat the experiment five times, where each trial re-samples the initial views, performs view selection again, and independently specifies the ROI annotations. All other settings follow the main experimental protocol. We report the mean and standard deviation across the five trials.

Table 8 shows that PreView remains stable despite changes in the initial observations and ROI annotations. Both preference modes exhibit small variation in global reconstruction quality across repeated trials. More importantly, their intended behavior is consistently preserved: the global preference maintains stronger scene-wide reconstruction, whereas the ROI preference achieves higher average and minimum regional fidelity. This indicates that preference-conditioned view allocation is not tied to a particular initialization or a single ROI annotation.

E.4 Preference Conditioning under Pose-Only Selection

We further examine whether task preference remains effective when candidate RGB is unavailable. This experiment forms a 2×2 comparison over candidate information availability (RGB-aware vs. pose-only) and reconstruction preference (Global vs. ROI). Each entry reports Global PSNR / Average ROI PSNR.

Table 8: **Robustness across five repeated trials.** Each trial uses re-sampled initial views, newly selected views, and independently specified ROI annotations. PreView exhibits low variance across trials while preserving the intended global–ROI preference behavior.

Method	G-PSNR ↑	G-LPIPS ↓	Avg. ROI PSNR ↑	Min ROI PSNR ↑	Avg. ROI LPIPS ↓
Ours-RGB-Global	20.691 ± 0.036	0.25219 ± 0.00044	26.162 ± 0.274	23.537 ± 0.353	0.003273 ± 0.000050
Ours-RGB-ROI	20.061 ± 0.038	0.27215 ± 0.00033	26.595 ± 0.243	24.346 ± 0.150	0.002998 ± 0.000060

Table 9: Preference conditioning under different observation regimes. Each entry reports Global PSNR / Average ROI PSNR.

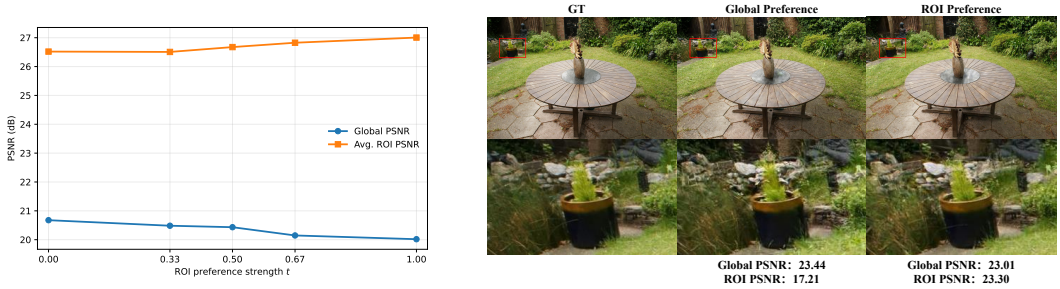
	Reconstruction Preference	
Candidate Evidence	**Global**	**ROI**
RGB-aware	**20.676** / 26.519	20.017 / **27.005**
Pose-only	**20.606** / 26.564	19.756 / **26.814**

The same preference behavior is preserved when candidate RGB is removed. Under both observation regimes, the Global preference yields higher global reconstruction quality, whereas the ROI preference shifts quality toward the task regions. Importantly, this behavior remains visible in the pose-only setting, where unseen candidates are represented only by camera pose during selection.

The result indicates that preference conditioning does not depend on accessing candidate appearance. Candidate RGB provides additional evidence for utility estimation, but the task objective can still guide view allocation through the query and the scene context established by the initial observations.

E.5 Continuous Preference Control

Beyond the two endpoint preferences, we evaluate whether the same selector can adjust its behavior as the reconstruction objective changes. We sample five preference strengths from the global-oriented endpoint ($t = 0$) to the ROI-oriented endpoint ($t = 1$), while keeping the initial views, candidate pool, view budget, and reconstruction backend unchanged.



(a) Preference sweep (b) Qualitative effect

Figure 7: Effect of reconstruction preference. (a) Sampled preference strengths reveal a controllable global–regional trade-off: increasing ROI preference generally improves regional fidelity while reducing global reconstruction quality. (b) A representative example shows the same behavior, where the ROI-oriented preference improves the annotated region while the global-oriented preference preserves stronger scene-wide reconstruction.

Table 10 and Fig. 7 show a consistent shift in reconstruction behavior as the preference moves from the global-oriented endpoint toward the ROI-oriented endpoint. Average ROI PSNR generally increases with stronger ROI preference, while global PSNR gradually decreases. Although only a small number of intermediate values are sampled, the trend shows that preference conditioning provides controllable budget allocation rather than switching between two independently designed selection strategies.

Table 10: Preference sampling on Mip-NeRF 360 with 10 initial views, $K = 5$, and 10k 3DGS iterations. Larger t places greater emphasis on ROI reconstruction.

Preference	**G-PSNR ↑**	**Avg. ROI PSNR ↑**
$t = 0$	20.676	26.519
$t = 0.33$	20.484	26.507
$t = 0.50$	20.433	26.675
$t = 0.67$	20.147	26.826
$t = 1$	20.017	27.005

The qualitative example in Fig. 7(b) provides a visual counterpart to this trend. The global-oriented mode preserves stronger scene-wide quality, whereas the ROI-oriented mode reallocates the limited view budget toward the annotated region and improves its local reconstruction.