ANYGRANGS: AGENTIC 3DGS SEGMENTATION WITH VERSATILE GRANULARITY

Anonymous authors
Paper under double-blind review

ABSTRACT

3D Gaussian Splatting (3DGS) enables language-guided 3D scene understanding, while practical queries may target varying semantic granularities. Recent agentic methods leverage MLLMs for target reasoning and refinement, yet versatile-granularity segmentation remains challenged by (i) inflexible view acquisition, where predefined views may insufficiently observe fine-grained targets, and (ii) cross-view spatial ambiguity, where object-centric spatial relations can be inconsistently interpreted across viewpoints. We propose AnyGranGS, an agentic framework addressing both challenges. First, Target-centric Adaptive Observation (TAO) locates a coarse target proxy and constructs target-specific virtual views, while Visibility Gain-based Selection (VGS) selects complementary candidate views based on Gaussian visibility gain. Second, Semantic Axis Constraints (SAC) establish object-centric semantic axes for consistent part localization across views. Multi-view masks are then aggregated through visibility-weighted Gaussian evidence for target refinement. We further construct Gran-LERF and Gran-3D-OVS for versatile-granularity evaluation. Experiments on Gran-LERF, Gran-3D-OVS, and Ref-LERF demonstrate the effectiveness of AnyGranGS across diverse query granularities, improving the average mIoU over the strongest baselines by **15.37, 21.61, and 12.58 points**, respectively.

1 INTRODUCTION

3D Gaussian Splatting (3DGS) has recently been extended from photorealistic reconstruction to language-guided scene understanding (Kerbl et al., 2023; Qin et al., 2024; Shi et al., 2024; Wu et al., 2024), enabling users to identify and segment 3D content through natural-language queries. In practice, queried targets may vary substantially in semantic granularity, from entire objects to semantic parts, part sets, or spatially conditioned parts (Liu et al., 2025a; Wang et al., 2026). This motivates versatile-granularity 3DGS segmentation, which requires identifying not only the referred object, but also the semantic extent to be selected.

Earlier 3DGS segmentation methods encode masks or language priors into Gaussian representations (Qin et al., 2024; Wu et al., 2024; Cen et al., 2025a;b; Kim et al., 2025), supporting simple open-vocabulary queries while largely tying target granularity to the granularity of supervision masks. Referring methods (He et al., 2025; Tao et al., 2025; 2026) extend this capability to more complex queries, but often rely on per-scene referring annotations, limiting generalization. Recent agentic methods (Shi et al., 2026) leverage the rich semantic priors of MLLMs for complex query reasoning and global-to-local target refinement. However, as shown in Fig. 1, extending this paradigm to versatile-granularity segmentation remains challenging due to two key issues. **(i) Inflexible View Acquisition.** Existing methods primarily select observations from predefined camera views, lacking target-specific viewpoints for further refinement. This limitation becomes more pronounced for fine-grained targets, which are more likely to be insufficiently observed (Ye et al., 2024; Cen et al., 2025b; Shi et al., 2026). **(ii) Cross-view Spatial Ambiguity.** Spatial relations defined in an object-centric frame, such as *"The front-left leg of Jake the Dog."* remain fixed in 3D but become ambiguous when interpreted independently across different camera views, leading to inaccurate part segmentation (Chen et al., 2026).

To address these issues, our main ideas are two-fold. First, we locate a coarse target proxy from the semantic Gaussian field and use it to guide target-specific view acquisition for local refinement.

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.

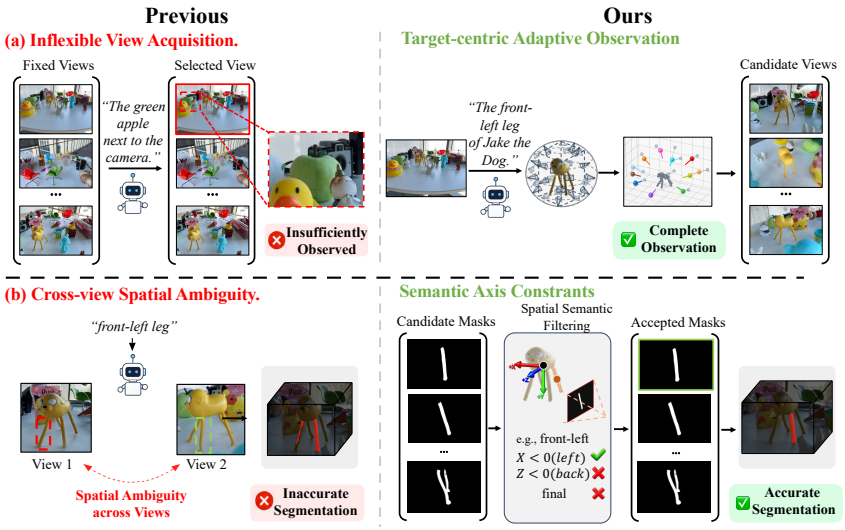


Figure 1: Illustration of the two key challenges in versatile-granularity 3DGS segmentation and our corresponding solutions.

Second, we maintain a shared object-centric spatial frame across views to disambiguate spatial relations. Together, these designs enable versatile-granularity target selection, including challenging spatially conditioned part queries.

Motivated by these ideas, we propose **AnyGranGS**, an agentic framework for versatile-granularity 3DGS segmentation. For adaptive target observation, we introduce **Target-centric Adaptive Observation (TAO)**, which first locates a coarse target proxy from the semantic Gaussian field and constructs target-centered virtual observations for local refinement. **Visibility Gain-based Selection (VGS)** then selects complementary candidate views according to the visibility gain of the coarse target proxy. To ensure consistent spatial reasoning across views, we introduce **Semantic Axis Constraints (SAC)**, which establish an object-centric spatial frame at each selected candidate view from the target proxy and apply consistent semantic axis constraints to disambiguate spatially conditioned part candidates across views. Finally, Multi-view Refinement aggregates the multi-view masks through visibility-weighted Gaussian evidence to refine the coarse target proxy to the granularity specified by the query.

We further construct **Gran-LERF** and **Gran-3D-OVS**, two benchmarks covering queries at different semantic granularities, with Gran-LERF additionally including challenging spatially conditioned part queries.

Our contributions are summarized as follows:

- We propose **AnyGranGS**, an agentic framework for versatile-granularity 3DGS segmentation, integrating target-centric adaptive observation, object-centric spatial reasoning, and visibility-weighted Gaussian refinement.
- We construct **Gran-LERF** and **Gran-3D-OVS**, two benchmarks for evaluating versatile-granularity 3DGS segmentation, including fine-grained and spatially conditioned part queries.
- Extensive experiments on **Gran-LERF, Gran-3D-OVS, and Ref-LERF** demonstrate the effectiveness of AnyGranGS across diverse query granularities, with average mIoU improvements of **15.37, 21.61, and 12.58 points** over the strongest baselines, respectively.

2 Related Work

Open-vocabulary Segmentation. Earlier methods encode 2D masks or language priors into Gaussian representations for segmentation (Ye et al., 2024; Qin et al., 2024; Wu et al., 2024; Cen et al., 2025a; Kim et al., 2025; Ye et al., 2026). To support different semantic granularities, SAGA (Cen et al., 2025a) and LangSplat (Qin et al., 2024) learn from multi-granularity or hierarchical masks,

while subsequent works improve language representation and cross-view semantic consistency (Li et al., 2025; Cen et al., 2025b). ReLaGS (Xie et al., 2026) further constructs a mask-guided hierarchical Gaussian representation for adaptive object-to-part retrieval. However, these methods mainly handle short, direct queries, while their available granularities remain constrained by mask supervision or preconstructed semantic hierarchies.

Referring Segmentation. Referring methods extend open-vocabulary segmentation to complex queries involving attributes, context, and spatial relations. ReferSplat (He et al., 2025) learns position-aware referring features, while subsequent works improve camera-aware cross-view alignment and explicit geometric reasoning (Tao et al., 2025; 2026). These methods follow the scene-specific referring setting of Ref-LERF, requiring referring expressions for each reconstructed scene and scene-specific optimization, which limits their generalization ability. To improve this, recent methods leverage MLLMs to reason over implicit or complex instructions (Liu et al., 2025b; Ding et al., 2026a;b), while REALM (Shi et al., 2026) further develops this direction into an *agentic* framework for reasoning-based 3D segmentation. However, these approaches still rely on predefined observations and lack consistent object-centric spatial reasoning across viewpoints. SAGO (Liao et al., 2026) partially addresses the former by acquiring virtual views around the target, but its viewpoints remain tied to the initial interaction distance and the latter ambiguity remains unresolved.

Differently, in this work, we propose AnyGranGS, an agentic framework for versatile-granularity 3DGS segmentation with target-centric adaptive observation and consistent object-centric spatial reasoning across views.

3 PRELIMINARY

3D Gaussian Splatting. A 3DGS scene is represented by a set of anisotropic 3D Gaussians (Kerbl et al., 2023),

$$g_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i, \mathbf{c}_i), \qquad \mathcal{G} = \{g_i\}_{i=1}^{N}, \tag{1}$$

where $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$, α_i, and $\mathbf{c}_i$ denote the center, covariance, opacity, and appearance of g_i, respectively. Following SAGA (Cen et al., 2025a), each Gaussian is additionally associated with a learned affinity feature $\mathbf{f}_i \in \mathbb{R}^D$, which supports clustering and association of Gaussian primitives.

Given a camera view v, projected Gaussians are depth-sorted and alpha-composited following 3DGS rasterization (Kerbl et al., 2023). The rendering contribution of Gaussian g_i to pixel u is:

$$w_{i,v}(u) = T_{i,v}(u)\alpha_{i,v}(u), \qquad T_{i,v}(u) = \prod_{j<i} \left(1 - \alpha_{j,v}(u)\right), \tag{2}$$

where $\alpha_{i,v}(u)$ is the projected opacity, $T_{i,v}(u)$ is the accumulated transmittance, and $j < i$ denotes front-to-back order for the current pixel and view. The weight $w_{i,v}(u)$ measures the visibility-aware contribution of g_i to pixel u.

Problem Formulation. We consider an interactive setting where a user observes a pretrained 3DGS scene $\mathcal{G}$ from a reference camera view v_0, obtaining the corresponding reference image I_0, and issues a natural-language query q about a target in the scene. Our goal is to identify the Gaussians corresponding to the queried target at its specified semantic granularity, including objects, parts, part sets, and spatially conditioned parts. We formulate the task as:

$$\mathcal{G}_{\text{fine}} = \text{AnyGranGS}(\mathcal{G}, q, I_0, v_0), \qquad \mathcal{G}_{\text{fine}} \subseteq \mathcal{G}, \tag{3}$$

where $\mathcal{G}_{\text{fine}}$ denotes the final set of Gaussians corresponding to the queried target.

4 METHOD

Overview. As illustrated in Fig. 2, AnyGranGS first parses the user query and obtains a coarse target mask. To address inflexible view acquisition, Target-centric Adaptive Observation (TAO; § 4.1) establishes a coarse target proxy and constructs target-centric virtual observations, from which Visibility Gain-based Selection (VGS) selects a compact set of candidate views according to their marginal visibility gain. For spatially conditioned part queries, Semantic Axis Constraints (SAC; § 4.2) address cross-view spatial ambiguity by establishing object-centric semantic axes and filtering

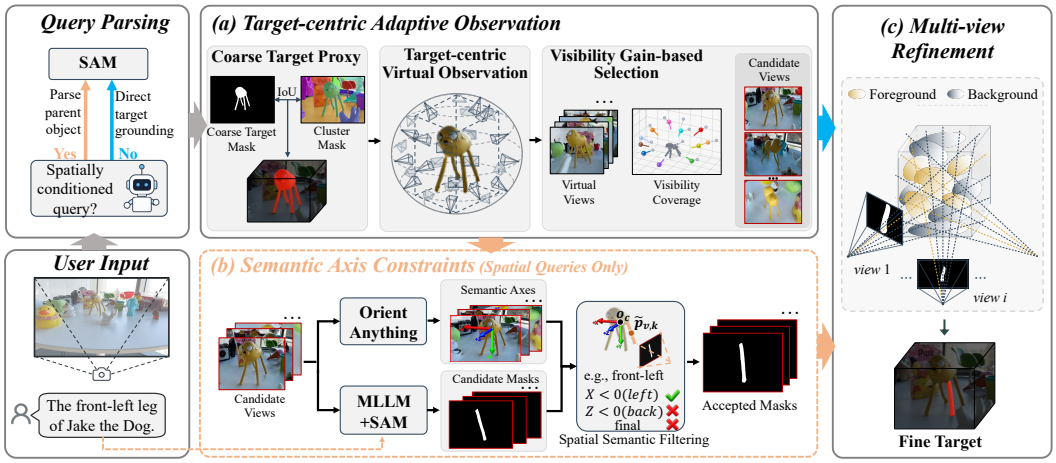


Figure 2: **Overview of AnyGranGS.** Given the user input, AnyGranGS first parses the query and obtains a coarse target mask, grounding the parent object for spatially conditioned queries and the queried target otherwise. (a) Target-centric Adaptive Observation (TAO) establishes a coarse target proxy, constructs target-centric virtual observations, and applies Visibility Gain-based Selection (VGS) to obtain candidate views. (b) For spatial queries, Semantic Axis Constraints (SAC) construct semantic axes and filter candidate sasks according to their object-relative 3D positions, yielding accepted masks. (c) Multi-view Refinement aggregates foreground and background Gaussian evidence from the multi-view masks through visibility-weighted evidence aggregation to obtain the fine target.

candidate masks according to their object-relative 3D positions, yielding accepted masks. Finally, multi-view refinement (§ 4.3) aggregates foreground and background Gaussian evidence from the multi-view masks through visibility-weighted evidence aggregation to obtain the fine target Gaussian set.

Query Parsing. We first determine whether the query contains an object-centric spatial relation. For a spatially conditioned query, the spatial relation is defined relative to its parent object. For example, given *"the front-left leg of Jake the Dog,"* we parse and ground *Jake the Dog* as the parent object, which provides the reference for interpreting *front-left*. Otherwise, the queried target is directly grounded. In both cases, MLLM localizes the grounded region with a bounding box, which is passed to SAM 2 to obtain the coarse target mask for subsequent processing.

4.1 Target-centric Adaptive Observation

Coarse Target Proxy. To construct the coarse target proxy, we associate the coarse target mask with the pretrained semantic Gaussian field. Following SAGA Cen et al. (2025a), we learn an instance-level Gaussian affinity feature field through contrastive learning and apply HDBSCAN McInnes et al. (2017) to group the learned features into instance-level Gaussian clusters. We then project each cluster onto the reference view and select the one whose projected mask has the highest IoU with the coarse target mask, yielding an initial Gaussian set $\mathcal{G}_c'$. We further refine $\mathcal{G}_c'$ by removing Gaussians whose rendered contributions are dominated by regions outside the coarse target mask, followed by spatial outlier filtering, yielding the final coarse target proxy $\mathcal{G}_c$. Details of the two filtering steps are provided in the Appendix.

Target-centric Virtual Observation. Then, we construct a set of target-centric virtual views around the coarse target proxy $\mathcal{G}_c$, providing closer observations. Specifically, We first compute a world-axis-aligned bounding box (AABB) of $\mathcal{G}_c$ from the coordinate-wise minima and maxima of its Gaussian centers, and slightly expand the box. We then enclose the expanded box with a bounding sphere to characterize the target extent. Let $\mathbf{b}_{\min}$ and $\mathbf{b}_{\max}$ denote the minimum and maximum corners of the expanded box. The bounding sphere is centered at $\mathbf{o}_c = (\mathbf{b}_{\min} + \mathbf{b}_{\max})/2$ with radius $r_c = \|\mathbf{b}_{\max} - \mathbf{b}_{\min}\|_2/2$.

To determine the virtual-camera distance, we use the horizontal and vertical fields of view of the reference camera, denoted by θ_x and θ_y. Let ρ denote the distance from each virtual camera to the target center $\mathbf{o}_c$. We choose ρ such that the target bounding sphere just fits within the narrower field of view:

$$\rho = \frac{r_c}{\sin\left(\frac{1}{2}\min(\theta_x, \theta_y)\right)}. \tag{4}$$

Then, to obtain uniformly distributed viewing directions, we sample J points on the unit sphere using Fibonacci sphere sampling Keinert et al. (2015). The samples progress from the lower to the upper part of the sphere while rotating around the vertical axis, forming a spiral-like distribution over the sphere. Each sampled point defines a unit direction $\mathbf{d}_n$, which is scaled by ρ and translated to the target center, giving the virtual-camera position $\mathbf{t}_n = \mathbf{o}_c + \rho\mathbf{d}_n$. Each camera is oriented toward $\mathbf{o}_c$, yielding the virtual view set $\mathcal{V} = \{v_n\}_{n=1}^{J}$. Details of the sphere sampling are provided in the Appendix.

Visibility Gain-based Selection. Although the sampled virtual views provide diverse viewing directions, processing all of them with an MLLM would introduce unnecessary visual input and reasoning overhead. Moreover, some views may observe the target only weakly because of occlusion, while others may repeatedly observe similar regions of the target. We therefore introduce Visibility Gain-based Selection (VGS) to select a compact set of candidate views that jointly provides broad and complementary visibility of the coarse target proxy $\mathcal{G}_c$.

For each Gaussian $g_i \in \mathcal{G}_c$ and virtual view $v \in \mathcal{V}$, we first measure its visibility by accumulating its rendering contribution over the pixels covered by its projected footprint:

$$\tau_{i,v} = \sum_{u \in \Omega_{i,v}} w_{i,v}(u), \tag{5}$$

where $\Omega_{i,v}$ denotes the pixels to which g_i contributes in view v, and $w_{i,v}(u)$ reflects its actual rendering contribution after accounting for occlusion by other Gaussians. Since different Gaussians may have different contribution scales, we normalize the visibility as $\hat{\tau}_{i,v} = \tau_{i,v} / \max_{v' \in \mathcal{V}} \tau_{i,v'}$.

We then progressively select a subset of virtual views that jointly covers the coarse target proxy. Let $\mathcal{S} \subseteq \mathcal{V}$ denote the views selected so far. We define their visibility coverage as

$$C(\mathcal{S}) = \sum_{g_i \in \mathcal{G}_c} \max_{v \in \mathcal{S}} \hat{\tau}_{i,v}, \tag{6}$$

with $C(\varnothing) = 0$. For each Gaussian, the coverage depends on its highest visibility among the currently selected views. Then, for each remaining virtual view $v \in \mathcal{V} \setminus \mathcal{S}$, we define its marginal visibility gain as the additional visibility coverage it provides over the current selection:

$$\Delta(v \mid \mathcal{S}) = C(\mathcal{S} \cup \{v\}) - C(\mathcal{S}) \quad = \sum_{g_i \in \mathcal{G}_c} \max\left(\hat{\tau}_{i,v} - \max_{v' \in \mathcal{S}} \hat{\tau}_{i,v'}, 0\right). \tag{7}$$

When $\mathcal{S} = \varnothing$, the inner maximum is defined as zero. The derivation is provided in the Appendix.

If a Gaussian is already highly visible in one of the selected views, another view contributes additional gain for that Gaussian only when it provides even higher visibility. In contrast, Gaussians that remain weakly observed by the current selection can contribute larger gains when a new view reveals them more clearly. Therefore, maximizing the marginal visibility gain naturally favors views that complement the regions already covered by $\mathcal{S}$. Starting from $\mathcal{S} = \varnothing$, we greedily add the remaining view with the largest marginal visibility gain until K_c views are selected. The final selected set forms the candidate views for subsequent processing.

4.2 Semantic Axis Constraints

For spatially conditioned part queries, candidate masks may remain ambiguous across viewpoints. We therefore introduce Semantic Axis Constraints (SAC). For each candidate view, the MLLM and SAM generate part candidate masks, while Orient Anything (Wang et al., 2025) estimates the semantic axes of the parsed parent object. Then, sptial semantic filtering selects the candidate masks according to their 3D positions relative to these axes.

Candidate Mask Generation. For each candidate view v, we feed the rendered image together with the spatial query q to the MLLM. The MLLM proposes segmentation prompts for the visible part candidates, which are passed to SAM to obtain $\{M_{v,k}\}_{k=1}^{N_v}$, where N_v denotes the number of candidate masks in view v.

Semantic Axis Construction. In parallel, we apply Orient Anything (Wang et al., 2025) to each candidate view to estimate the semantic orientation of the parsed parent object. The predicted semantic axes are transformed from the camera coordinate system to the world coordinate system using the camera pose, yielding a rotation matrix $\mathbf{R}_v$. We anchor these axes at the center $\mathbf{o}_c$ of the coarse parent-object proxy, forming an object-centric frame whose axes correspond to the semantic right, down, and front directions.

Spatial Semantic Filtering. For each candidate mask $M_{v,k}$, we estimate its 3D center from the contributing Gaussians:

$$\mathbf{p}_{v,k} = \frac{\sum_i \omega_{i,v,k} \boldsymbol{\mu}_i}{\sum_i \omega_{i,v,k}}, \tag{8}$$

where $\omega_{i,v,k}$ denotes the contribution of Gaussian g_i to $M_{v,k}$. Its coordinates in the semantic frame are

$$\widetilde{\mathbf{p}}_{v,k} = \mathbf{R}_v^\top (\mathbf{p}_{v,k} - \mathbf{o}_c). \tag{9}$$

We translate each spatial relation $r \in \mathcal{R}(q)$ into a semantic axis $a_r \in \{X, Y, Z\}$ and a required side $s_r \in \{-1, +1\}$. A candidate is accepted only if its 3D center satisfies all corresponding sign constraints:

$$\phi_{v,k}(q) = \bigwedge_{r \in \mathcal{R}(q)} \left[s_r \, \widetilde{p}_{v,k}^{(a_r)} > 0 \right]. \tag{10}$$

For example, a *front-left* part requires a negative coordinate along the semantic right axis and a positive coordinate along the semantic front axis. Candidates satisfying all constraints are retained as the accepted masks M_v.

4.3 Multi-view Refinement

Given the multi-view target masks, we aggregate Gaussian rendering contributions inside and outside these masks to refine the coarse target proxy. For spatially conditioned queries, these masks are the accepted masks produced by SAC; otherwise, they are obtained directly from the candidate views. Inspired by FlashSplat (Shen et al., 2025), we perform the refinement only within the expanded AABB of the coarse target proxy, while computing rendering contributions with full-scene alpha compositing to preserve occlusion from surrounding geometry.

Let $\mathcal{V}_{\text{mask}}$ denote the set of views associated with the multi-view target masks. For each Gaussian g_i inside the expanded AABB, we accumulate foreground and background evidence across these views:

$$E_i^{\text{fg}} = \sum_{v \in \mathcal{V}_{\text{mask}}} \sum_u M_v(u)\, w_{i,v}(u), \qquad E_i^{\text{bg}} = \sum_{v \in \mathcal{V}_{\text{mask}}} \sum_u \big(1 - M_v(u)\big)\, w_{i,v}(u), \tag{11}$$

where M_v denotes the target mask in view v, and $w_{i,v}(u)$ is the visibility-aware rendering contribution defined in Eq. (2). Thus, each view contributes evidence in proportion to the visible rendering mass of the Gaussian.

We then update Gaussian membership by comparing foreground and background evidence. A Gaussian is included when foreground evidence dominates and excluded when background evidence dominates; in ambiguous cases, we retain its original membership in $\mathcal{G}_c$. Gaussians outside the expanded AABB remain unchanged. This refinement can recover missing target Gaussians while suppressing false positives in the coarse proxy, yielding the final fine-grained Gaussian set $\mathcal{G}_{\text{fine}}$.

5 Experiments

5.1 Experimental Setup

Benchmarks. We evaluate AnyGranGS on Gran-LERF, Gran-3D-OVS, and Ref-LERF (He et al., 2025). Gran-LERF extends LERF-Mask (Ye et al., 2024) with versatile-granularity annotations,

Table 1: Multi-granularity segmentation on Gran-LERF (%). Best and second-best results are shown in **bold** and underlined, respectively.

Method	Figurines		Ramen		Teatime		Avg.	
	mIoU↑	mBIoU↑	mIoU↑	mBIoU↑	mIoU↑	mBIoU↑	mIoU↑	mBIoU↑
SAGA	38.15	37.43	18.85	16.19	30.32	24.86	29.11	26.16
LaGa	38.89	38.15	52.16	45.62	29.40	27.01	40.15	36.93
Dr. Splat	32.12	31.52	29.47	21.74	28.67	25.37	30.09	26.21
ReLaGS	37.95	37.10	53.05	42.55	39.69	38.68	43.56	39.45
REALM	61.40	60.51	61.81	58.68	60.57	56.05	61.26	58.41
AnyGranGS	**80.45**	**79.83**	**77.18**	**70.06**	**72.25**	**68.59**	**76.63**	**72.83**

Table 2: Results on Gran-3D-OVS (mIoU, %).

Method	Bed	Bench	Lawn	Room	Sofa	Avg.
SAGA	34.90	27.22	11.06	18.25	26.21	23.53
SegAnyGaussians	46.67	57.76	57.32	29.66	37.80	45.84
Dr. Splat	22.76	55.90	51.93	25.79	40.99	39.47
LaGa	64.07	58.99	66.44	56.65	38.83	57.00
ReLaGS	71.52	57.90	67.66	52.96	49.36	59.88
REALM	67.46	48.39	60.06	37.66	62.87	55.29
AnyGranGS	**82.15**	**84.96**	**81.76**	**85.20**	**73.40**	**81.49**

Table 3: Results on Ref-LERF (mIoU, %).

Method	Fig.	Ramen	Tea.	Avg.
LangSplat†	17.9	12.0	7.6	12.5
GS-Grouping†	8.6	27.9	14.8	17.1
GOI†	16.5	27.1	22.9	22.2
ReferSplat†	25.7	35.2	31.3	30.7
ReLaGS‡	27.61	5.72	1.79	11.71
REALM‡	**63.82**	53.31	67.18	61.44
AnyGranGS	58.10	**78.91**	**85.05**	**74.02**

†From ReferSplat (He et al., 2025); ‡Our reproduced.

while Gran-3D-OVS extends 3D-OVS (Liu et al., 2023). They contain 87 queries with 252 masks and 51 categories with 254 masks, respectively. Ref-LERF He et al. (2025) contains 42 referring expressions and 198 masks involving object attributes and spatial relations. For each query, we infer the 3D Gaussian mask once from a fixed reference view and render the frozen mask to all evaluation views.

Baselines. We compare AnyGranGS with a broad range of 3DGS-based segmentation and scene understanding methods, including LangSplat Qin et al. (2024), GS-Grouping Ye et al. (2024), SAGA (Cen et al., 2025a), GOI Qu et al. (2024), LaGa (Cen et al., 2025b), Dr. Splat (Kim et al., 2025), ReferSplat (He et al., 2025), ReLaGS (Xie et al., 2026), and REALM (Shi et al., 2026). For Gran-LERF and Gran-3D-OVS, we report results obtained by running the publicly released implementations of open-source baselines, while for Ref-LERF, we use the results reported in the original paper He et al. (2025), unless otherwise specified.

Evaluation Metrics. Following previous work (Piekenbrinck et al., 2025), we render each predicted 3D Gaussian mask to the annotated views and report mean Intersection over Union (mIoU) and mean Boundary IoU (mBIoU).

Implementation Details. We implement our method in PyTorch and keep the reconstructed Gaussian geometry and affinity feature field fixed throughout inference. In the experiment, we utilize Gemini-3.6-Flash as default MLLM and SAM 2.1 Hiera-L (Ravi et al., 2025) produces target masks. TAO constructs 48 target-centric Virtual Observations, and VGSselects 12 candidate views. The reference view initializes the coarse target proxy and is excluded from S. All experiments are run with PyTorch 2.7.1 and CUDA 12.8 on a single NVIDIA RTX PRO 6000 Blackwell GPU.

5.2 Quantitative Comparisons

Versatile-granularity Segmentation. As shown in Tables 1 and 2, AnyGranGS consistently achieves the best performance on both multi-granularity benchmarks. On Gran-LERF, our method obtains 76.63% mIoU and 72.83% mBIoU, outperforming the strongest baseline REALM by 15.37 and 14.42 points, respectively. The advantage further extends to Gran-3D-OVS, where AnyGranGS ranks first across all five scenes and achieves 81.49% average mIoU, surpassing the second-best ReLaGS by 21.61 points. These results demonstrate the effectiveness and robustness of AnyGranGS for segmenting targets across varying semantic granularities and diverse scenes.

Referring Segmentation. We further evaluate AnyGranGS on Ref-LERF, which contains complex referring expressions involving object attributes and spatial relations. As shown in Table 3, AnyGranGS achieves the best average mIoU of 74.02%, outperforming the strong agentic baseline



Figure 3: Multi-granularity segmentation results on our Gran-LERF and Gran-3D-OVS datasets.

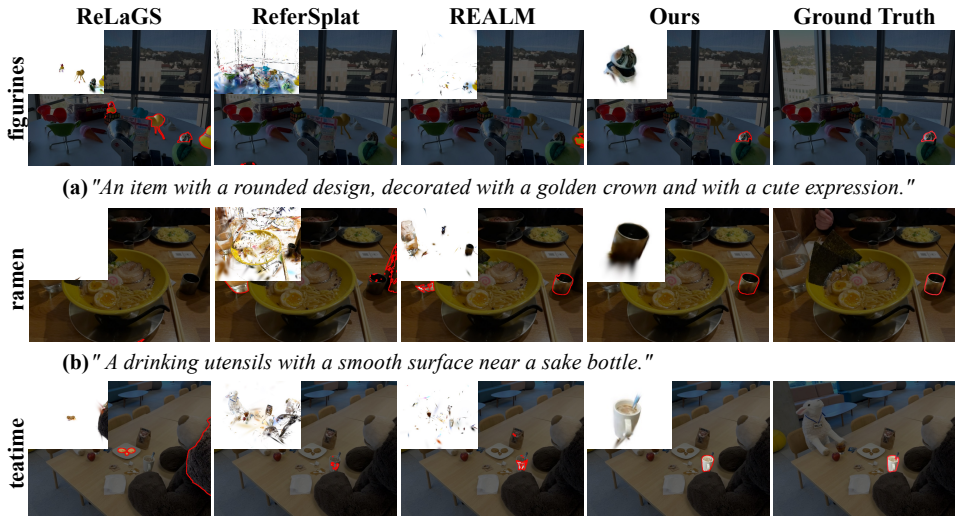


Figure 4: Qualitative comparison of R3DGS on Ref-LERF.

REALM by 12.58 points. These results demonstrate that AnyGranGS generalizes beyond multi-granularity open-vocabulary queries and remains effective for complex referring segmentation.

5.3 Qualitative Comparisons

Qualitative comparisons are shown in Figs. 3 and 4. On Gran-LERF and Gran-3D-OVS, AnyGranGS produces more complete and spatially accurate masks for fine-grained and spatially conditioned targets, while competing methods often include nearby distractors or miss parts of the queried region. The advantage is especially clear for object-part queries involving directional relations, where our predictions better preserve the intended object-relative semantics. On Ref-LERF, AnyGranGS also localizes targets more precisely under complex referring expressions and yields masks that align more closely with the ground truth.

5.4 Ablation Studies

We conduct ablation studies on *Gran-LERF Figurines* to evaluate the key designs of Target-centric Adaptive Observation (TAO) and Semantic Axis Constraints (SAC).

Ablation of TAO. We separately examine target-centric virtual observation and Visibility Gain-based Selection (VGS). *Fixed views + Random* removes both target-centric virtual observation and VGS, randomly selecting observations from the original fixed-view pool. *Fixed views + VGS* retains VGS but replaces the target-centric virtual-view pool with fixed views. To isolate the effect of VGS, we retain the target-centric virtual-view pool and replace VGS with either random selection (*w/o VGS (Random)*) or top-K selection based on individual view visibility (*w/o VGS (MAX)*).

As shown in Table 4, target-centric observation consistently improves over fixed-view acquisition, while VGS further outperforms Random and MAX by accounting for complementary visibility across selected views.

Table 4: **Ablation of TAO.**

Setting	mIoU↑	mBIoU↑
Full	**80.45**	**79.83**
Fixed views + Random	71.01	70.37
Fixed views + VGS	75.70	74.80
w/o VGS (Random)	74.88	73.81
w/o VGS (MAX)	76.67	76.03

We further analyze the effect of the number of candidate views K_c on different view-selection strategies. As shown in Fig. 5, Random uniformly samples views, whereas MAX selects the top-K views according to their individual visibility scores without considering redundancy across views. In contrast, VGS greedily selects views according to their marginal visibility gain, thereby encouraging complementary target coverage. VGS consistently achieves higher segmentation performance across different values of K_c, while the improvement gradually saturates as more views are selected.

Ablation of SAC. We evaluate SAC across different MLLMs to examine whether its effect is consistent across different spatial reasoning capabilities.

For SAC, we evaluate the six spatially conditioned part queries in *Figurines*, comprising 19 annotated ground-truth masks. The *w/o SAC* setting uses the same MLLM- and SAM-generated candidate masks but removes semantic-axis-based spatial filtering. The *w/ SAC* setting additionally filters candidate masks according to their object-relative 3D positions under the constructed semantic axes. As shown in Table 5, SAC consistently improves segmentation performance across different MLLMs. The gains are particularly pronounced for weaker MLLMs, while SAC remains beneficial even with stronger models.

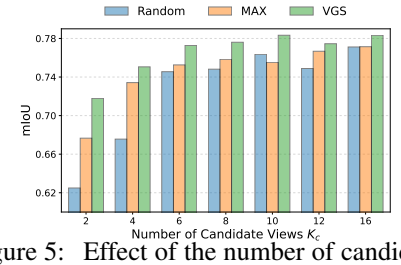


Figure 5: Effect of the number of candidate views on different view-selection strategies.

6 Conclusion

We present AnyGranGS, an agentic framework for versatile-granularity 3DGS segmentation. Target-centric Adaptive Observation acquires informative candidate views around a coarse target proxy, while Semantic Axis Constraints resolve cross-view spatial ambiguity for spatially conditioned queries. Multi-view Refinement then aggregates visibility-weighted Gaussian evidence to obtain the fine target region. Together, these components enable adaptive observation, consistent spatial reasoning, and fine-grained target refinement across varying query granularities, including challenging spatially conditioned queries.

Table 5: **SAC ablation across different MLLMs** on *Gran-LERF Figurines*.

MLLM	w/o SAC	w/ SAC
Qwen3-VL-32B	15.2	**41.0**
Qwen3.7-Plus	43.1	**66.7**
Gemini-2.5-Pro	54.3	**73.0**
Gemini-3.6-Flash	71.0	**75.9**

Limitations. Selection depends on the coarse parent, estimated orientation, and depth support of the candidate masks. Symmetry can make semantic orientation ambiguous, while small parts and missing geometry can impair localization. The spherical sampler may under-observe views below the scene world-up plane, and non-spatial refinement masks can merge repeated instances. The prompt set assumes a visible target, and the evaluation focuses on explicit queries.

AI USE STATEMENT

Generative AI tools were employed in this study primarily to refine the manuscript's language and presentation, and, to a limited extent, to support code troubleshooting and implementation. Their use did not extend to the creation or alteration of experimental data or results, nor were they relied upon to develop scientific interpretations, claims, or conclusions. Decisions concerning the design of the method, assessment of its performance, and interpretation of the findings remained entirely with the authors. Any AI-supported text or code was subsequently examined and, where necessary, revised by the authors, with code outputs checked through appropriate testing and all technical statements and experimental findings independently verified. The authors assume full responsibility for the accuracy, validity, and final presentation of all content included in this work.

REFERENCES

Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3D gaussians. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1971–1979, 2025a. doi: 10.1609/aaai.v39i2.32193.

Jiazhong Cen, Xudong Zhou, Jiemin Fang, Changsong Wen, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Tackling view-dependent semantics in 3D language gaussian splatting. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 7013–7034, 2025b.

Zeren Chen, Xiaoya Lu, Zhijie Zheng, Pengrui Li, Lehan He, Yijin Zhou, Jing Shao, Bohan Zhuang, and Lu Sheng. Geometrically-constrained agent for spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38689–38699, 2026.

Jiayu Ding, Meilu Song, Yun Chen, Wei Gao, and Ge Li. CausalSplat: Towards comprehensive hierarchical reasoning in 3D gaussian splatting, 2026a.

Jiayu Ding, Meilu Song, Xiaoyi Zhang, Hongbo Jin, Yichen Jin, and Xiangtian Si. ZeroSplat: Generalized referring segmentation in 3D gaussian splatting. In *European Conference on Computer Vision (ECCV)*, 2026b.

Shuting He, Guangquan Jie, Changshuo Wang, Yun Zhou, Shuming Hu, Guanbin Li, and Henghui Ding. ReferSplat: Referring segmentation in 3D gaussian splatting. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2025.

Benjamin Keinert, Matthias Innmann, Michael Sänger, and Marc Stamminger. Spherical fibonacci mapping. *ACM Trans. Graph.*, 34(6):193:1–193:7, 2015. doi: 10.1145/2816795.2818131. URL https://doi.org/10.1145/2816795.2818131.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433.

Jun-Seong Kim, GeonU Kim, Yu-Ji Kim, Yu-Chiang Frank Wang, Jaesung Choe, and Tae-Hyun Oh. Dr. splat: Directly referring 3D gaussian splatting via direct language embedding registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14137–14146, 2025.

Wanhua Li, Yujie Zhao, Minghan Qin, Yang Liu, Yuanhao Cai, Chuang Gan, and Hanspeter Pfister. Langsplatv2: High-dimensional 3d language gaussian splatting with 450+ fps. *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2507.07136.

Liwei Liao, Rongjie Wang, and Ronggang Wang. Online segment 3d gaussians via launching virtual drones, 2026. URL https://arxiv.org/abs/2607.01628.

Jing Liu, Wenxuan Wang, Yisi Zhang, Yepeng Tang, Xingjian He, Longteng Guo, Tongtian Yue, and Xinlong Wang. Towards unified referring expression segmentation across omni-level visual target granularities. *arXiv preprint arXiv:2504.01954*, 2025a.

Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric P. Xing, and Shijian Lu. Weakly supervised 3D open-vocabulary segmentation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a76b693f36916a5ed84d6e5b39a0dc03-Abstract.html.

Zhenyang Liu, Yikai Wang, Sixiao Zheng, Tongying Pan, Longfei Liang, Yanwei Fu, and Xiangyang Xue. ReasonGrounder: LVLM-guided hierarchical feature splatting for open-vocabulary 3D visual grounding and reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3718–3727, 2025b.

Leland McInnes, John Healy, Steve Astels, et al. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

Jens Piekenbrinck, Christian Schmidt, Alexander Hermans, Narunas Vaskevicius, Timm Linder, and Bastian Leibe. OpenSplat3D: Open-vocabulary 3D instance segmentation using gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5285–5294, 2025.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. LangSplat: 3D language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20051–20060, 2024.

Yansong Qu, Shaohui Dai, Xinyang Li, Jianghang Lin, Liujuan Cao, Shengchuan Zhang, and Rongrong Ji. Goi: Find 3d gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 5328–5337, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025.

Qiuhong Shen, Xingyi Yang, and Xinchao Wang. FlashSplat: 2D to 3D gaussian splatting segmentation solved optimally. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 456–472, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72670-5. doi: 10.1007/978-3-031-72670-5_26.

Changyue Shi, Minghao Chen, Yiping Mao, Chuxiao Yang, Xinyuan Hu, Jiajun Ding, and Zhou Yu. REALM: An MLLM-agent framework for open world 3D reasoning segmentation and editing on gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16779–16788, 2026.

Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3D gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5333–5343, 2024.

Yuwen Tao, Kanglei Zhou, Xin Tan, and Yuan Xie. Camera-aware cross-view alignment for referring 3D gaussian splatting segmentation, 2025.

Yuwen Tao, Kanglei Zhou, Chang Li, and Liyuan Wang. Geometry-aware cross-modal graph alignment for referring segmentation in 3D gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20433–20442, 2026.

Yizhao Wang, Xinfa Wang, Jingbo Wang, Guantao Zhang, Yafeng Han, Guohong Gao, and Yuhe Xia. Beyond similarity matching: Structured reasoning for open-vocabulary referring segmentation in 3dgs. *arXiv preprint arXiv:2608.16103*, 2026.

Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything: Learning robust object orientation estimation from rendering 3D models. In *Forty-Second International Conference on Machine Learning*, 2025.

Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, and Jian Zhang. OpenGaussian: Towards point-level 3D gaussian-based open vocabulary understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 19114–19138, 2024.

Yaxu Xie, Abdalla Arafa, Alireza Javanmardi, Christen Millerdurai, Jia Cheng Hu, Shaoxiang Wang, Alain Pagani, and Didier Stricker. ReLaGS: Relational language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23826–23836, 2026.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3D scenes. In *European Conference on Computer Vision (ECCV)*, pp. 162–179, 2024.

Xulun Ye, Qin Zhang, and Kun Zhou. LangRef3DGS: Natural language-guided 3D referential segmentation from partial observations via 3D gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38595–38605, 2026.

APPENDIX

A COARSE TARGET PROXY REFINEMENT

The Gaussian set $\mathcal{G}'_c$ selected by projected-mask IoU may still contain Gaussians whose rendered contributions mainly fall outside the coarse target mask, as well as a small number of spatially isolated Gaussians. We therefore refine $\mathcal{G}'_c$ in two stages.

Visibility-aware Contribution Filtering. For each Gaussian $g_i \in \mathcal{G}'_c$, we accumulate its rendering contribution separately over pixels inside and outside the coarse target mask in the reference view. Let Ω_{fg} and Ω_{bg} denote these two pixel regions. Using the rendering weight $w_{i,v_0}(u)$ defined in Eq. 2, we compute

$$\begin{aligned} c_i^{\mathrm{fg}} &= \sum_{u \in \Omega_{\mathrm{fg}}} w_{i,v_0}(u), \\ c_i^{\mathrm{bg}} &= \sum_{u \in \Omega_{\mathrm{bg}}} w_{i,v_0}(u). \end{aligned} \tag{12}$$

We retain Gaussians whose foreground contribution is larger than their background contribution:

$$\mathcal{G}''_c = \left\{ g_i \in \mathcal{G}'_c \;\middle|\; c_i^{\mathrm{fg}} > c_i^{\mathrm{bg}} \right\}. \tag{13}$$

Since $w_{i,v_0}(u)$ is computed under full-scene alpha compositing, the comparison reflects the actual visible contribution of each Gaussian in the reference view.

Spatial Outlier Filtering. The contribution-filtered set $\mathcal{G}''_c$ may still contain spatially isolated Gaussians that can distort the estimated target extent. For each Gaussian $g_i \in \mathcal{G}''_c$, we compute its average Euclidean distance to its K_{nn} nearest neighbors:

$$d_i = \frac{1}{K_{\mathrm{nn}}} \sum_{j \in \mathcal{N}_{K_{\mathrm{nn}}}(i)} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2, \tag{14}$$

where $\mathcal{N}_{K_{\mathrm{nn}}}(i)$ denotes the set of K_{nn} nearest neighbors of g_i within $\mathcal{G}''_c$.

We then compute the median and median absolute deviation (MAD) of these local distances:

$$\begin{aligned} \bar{d} &= \mathrm{median}_{g_i \in \mathcal{G}''_c}\, d_i, \\ s_{\mathrm{MAD}} &= \mathrm{median}_{g_i \in \mathcal{G}''_c}\, |d_i - \bar{d}|. \end{aligned} \tag{15}$$

We retain Gaussians whose local distance does not exceed the MAD-based threshold:

$$\mathcal{G}_c = \left\{ g_i \in \mathcal{G}''_c \;\middle|\; d_i \le \bar{d} + \lambda s_{\mathrm{MAD}} \right\}, \tag{16}$$

where λ controls the filtering strength. The resulting set $\mathcal{G}_c$ is used as the final coarse target proxy.

B Fibonacci Sphere Sampling

Given the number of virtual views J, we generate J approximately uniform points on the unit sphere. For the n-th sample, $n = 1, \ldots, J$, its vertical coordinate is set to

$$z_n = \frac{2n-1}{J} - 1, \tag{17}$$

which progressively moves the samples from the lower to the upper part of the unit sphere.

Meanwhile, we rotate the sample around the vertical axis according to the inverse golden ratio $\varphi = (\sqrt{5}-1)/2$. The remaining coordinates are computed as

$$\begin{aligned} x_n &= \sqrt{1-z_n^2}\cos(2\pi n\varphi), \\ y_n &= \sqrt{1-z_n^2}\sin(2\pi n\varphi). \end{aligned} \tag{18}$$

Therefore, the sampled direction

$$\mathbf{d}_n = (x_n, y_n, z_n) \tag{19}$$

lies on the unit sphere. As n increases, the vertical coordinate changes progressively while the point continuously rotates around the vertical axis, producing a spiral-like distribution over the sphere and avoiding strong concentration near particular directions.

We then map each unit-sphere sample to the target-centered virtual-camera sphere by scaling it with the camera distance ρ and translating it by the target center $\mathbf{o}_c$:

$$\mathbf{t}_n = \mathbf{o}_c + \rho\mathbf{d}_n. \tag{20}$$

Each virtual camera is oriented toward $\mathbf{o}_c$, yielding the virtual view v_n.

C Derivation of Marginal Visibility Gain

Recall that the visibility coverage of a selected view set $\mathcal{S}$ is

$$C(\mathcal{S}) = \sum_{g_i \in \mathcal{G}_c} \max_{v' \in \mathcal{S}} \hat{\tau}_{i,v'}. \tag{21}$$

After adding a remaining view v, the coverage becomes

$$C(\mathcal{S} \cup \{v\}) = \sum_{g_i \in \mathcal{G}_c} \max\left(\max_{v' \in \mathcal{S}} \hat{\tau}_{i,v'}, \hat{\tau}_{i,v}\right). \tag{22}$$

Therefore, the marginal visibility gain is

$$\begin{aligned} \Delta(v \mid \mathcal{S}) &= C(\mathcal{S} \cup \{v\}) - C(\mathcal{S}) \\ &= \sum_{g_i \in \mathcal{G}_c} \left[\max\left(\max_{v' \in \mathcal{S}} \hat{\tau}_{i,v'}, \hat{\tau}_{i,v}\right) - \max_{v' \in \mathcal{S}} \hat{\tau}_{i,v'}\right]. \end{aligned} \tag{23}$$

Using the identity $\max(a,b) - a = \max(b-a, 0)$, we obtain

$$\Delta(v \mid \mathcal{S}) = \sum_{g_i \in \mathcal{G}_c} \max\left(\hat{\tau}_{i,v} - \max_{v' \in \mathcal{S}} \hat{\tau}_{i,v'}, 0\right). \tag{24}$$

For $\mathcal{S} = \varnothing$, we define the current visibility of each Gaussian as zero.

D Detailed Results for Candidate-View Selection

For completeness, in Tab. 6, we provide the exact numerical results corresponding to the candidate-view analysis in Fig. 5. The main paper reports the performance trend as the candidate-view pool size K_c varies; here we list the underlying mIoU values for all tested settings.

Table 6: Detailed results of candidate-view selection on *Gran-LERF Figurines* (mIoU, %).

K_c	**Random**	**MAX**	**VGS**
2	62.49	67.66	**71.79**
4	67.56	73.42	**75.06**
6	74.55	75.26	**77.28**
8	74.81	75.84	**77.61**
10	76.36	75.53	**78.33**
12	74.88	76.67	**77.45**
16	77.12	77.15	**78.33**

Table 7: Ablation of Multi-view Gaussian Evidence Refinement.

Setting	**mIoU** ↑	**mBIoU** ↑
Full	**80.45**	**79.83**
w/o Refinement	47.96	46.75

E Ablation of Multi-view Refinement

As shown in Table 7, removing multi-view Gaussian evidence refinement causes a substantial performance drop, with mIoU decreasing from 80.45 to 47.96 and mBIoU from 79.83 to 46.75. This demonstrates that aggregating complementary foreground and background evidence across views is critical for refining the coarse target into an accurate fine-grained segmentation.