EVI-GS: EVIDENCE-VALIDATED EXTRACTION OF OBJECTS FROM 3D GAUSSIAN SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Extracting independently renderable objects from pretrained 3D Gaussian Splatting scenes is not equivalent to fitting object labels in the complete scene. Removing non-target Gaussians changes transmittance and may expose previously occluded erroneous primitives, while a high foreground contribution ratio can result from insufficient counter-evidence rather than reliable object support. We refer to these failure modes as extraction inconsistency and evidence insufficiency. We propose **EVI-GS**, a post-hoc object extraction framework that operates on a frozen pretrained 3DGS scene. EVI-GS first constructs scene-conditioned observational evidence and a fixed Gaussian association score, without directly committing to binary object labels. It then calibrates selection through a differentiable relaxation of independent-object rendering and validates the resulting candidates using absolute support, frozen-scene surface consistency, cross-view evidence, and local connectivity to reliable foreground seeds. We further introduce SNV-Eval, an auxiliary denser-view stress test based on interpolated cameras and shared pseudo-references. EVI-GS achieves 90.29% mIoU under the standard protocol and reduces per-object extraction time from 36.9 s to 30.7 s relative to COB-GS in our evaluation, without modifying the pretrained geometry, appearance, or opacity parameters. The code is available at `https://anonymous.4open.science/r/EVI-GS-17B5`.

1 INTRODUCTION

3D Gaussian Splatting (3DGS) Kerbl et al. (2023) provides a high-quality, real-time renderable explicit representation of complex three-dimensional scenes. As scene representations evolve from holistic reconstruction toward object-level understanding and editing, a central problem arises: given a fully trained 3DGS scene together with multi-view two-dimensional observations of a target object, how can one reliably extract, from the complete scene, a subset of Gaussians that can be rendered independently.

We study post-reconstruction object extraction from a pretrained 3DGS scene without modifying its geometry, appearance, or opacity. Multi-view 2D masks serve only as target observations. Prior Ye et al. (2024); Wu et al. (2024); Zhu et al. (2025b) work addresses related problems through feature learning, identity propagation, geometric adjustment, or rendering-contribution analysis. However, directly converting scene-conditioned Gaussian memberships into standalone objects introduces two key issues.

Scene-conditioned label fitting and independent object extraction are different inference problems. Removing non-target Gaussians changes α-compositing and may expose primitives previously suppressed by occlusion; we call this extraction inconsistency. Moreover, a high foreground contribution ratio may result from weak foreground support with little background counter-evidence rather than reliable object evidence, which we call evidence insufficiency. Thus, foreground compatibility alone is insufficient for final asset admission.

EVI-GS therefore separates observational attribution from export-time selection and evidence-based admission. **Rendering-Contribution Attribution (RCA)** constructs a fixed association score from observed foreground and background contributions, leaving the final selection unresolved. **Extraction-Aligned Calibration (EAC)** determines the selection boundary through independent-object rendering, while **Evidence-Validated Admission (EVA)** checks whether the selected candi-

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.

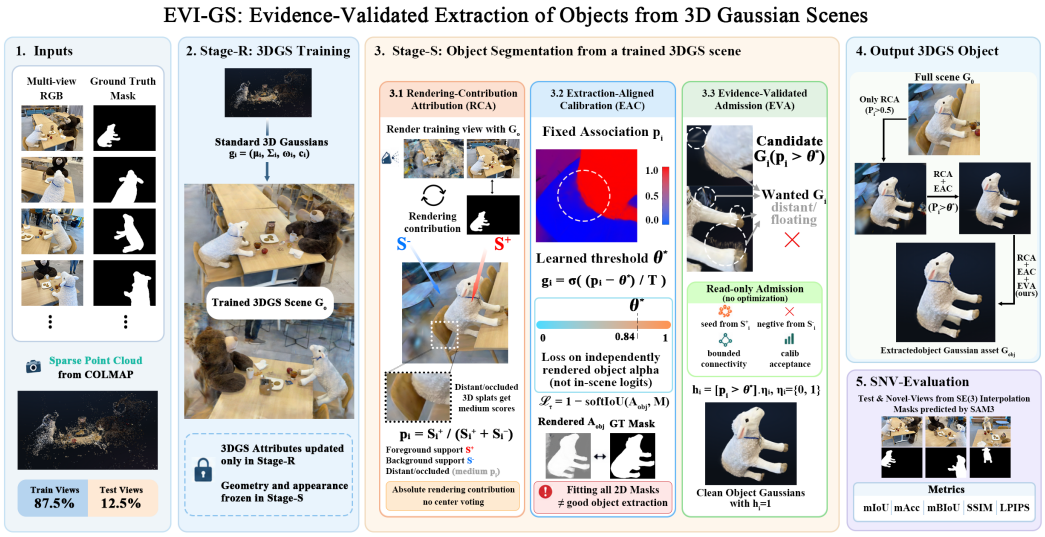


Figure 1: Overview of EVI-GS. Given a pretrained 3DGS scene and multi-view target masks, RCA constructs observational evidence and a fixed Gaussian association score. EAC calibrates the selection boundary through independent-object rendering, and EVA validates the candidates using contribution support, surface consistency, and local connectivity without updating the association score. The retained Gaussian subset forms an independently renderable object asset. SNV-Eval is used only for auxiliary evaluation and does not participate in extraction.

dates have sufficient observational and surface support. RCA thus provides the association on which extraction operates, rather than treating contribution dominance as sufficient grounds for inclusion in a standalone asset.

Sparse-view projection metrics may miss floaters, elongated spikes, or detached residuals that remain occluded under the evaluated cameras, while SSIM Wang et al. (2004) and LPIPS Zhang et al. (2018) mainly assess visible-region appearance. We therefore introduce Synthetic Novel-View Auxiliary Evaluation (SNV-Eval), which interpolates test-camera poses, renders the frozen scene, and generates shared SAM 3 Carion et al. (2025) pseudo-references. All methods are evaluated under identical auxiliary views, increasing viewpoint density and sensitivity to view-dependent residuals. Our main contributions are as follows:

- We identify an extraction inconsistency in post-hoc 3DGS object extraction: Gaussian memberships estimated under full-scene transmittance do not necessarily remain valid after non-target primitives are removed and visibility is recomputed. This motivates treating independently renderable object extraction as a distinct problem from scene-conditioned label fitting.

- We introduce extraction-aligned calibration, which optimizes the Gaussian selection threshold directly through the independent-object α-compositing operator. The decision boundary is optimized through a continuous relaxation of export-time selection, rather than through auxiliary scene-conditioned label rendering.

- We decouple contribution-based association from evidence-based asset admission. RCA supplies a fixed observational score for extraction-aligned calibration, while EVA evaluates candidate validity using absolute support, cross-view surface consistency, and connectivity to reliable seeds. This separation addresses candidates whose high foreground association is not backed by sufficient object evidence.

- We analyze a coverage limitation of sparse-view object-extraction evaluation and introduce SNV-Eval as an auxiliary stress test that increases viewpoint density using interpolated cameras and shared pseudo-references.

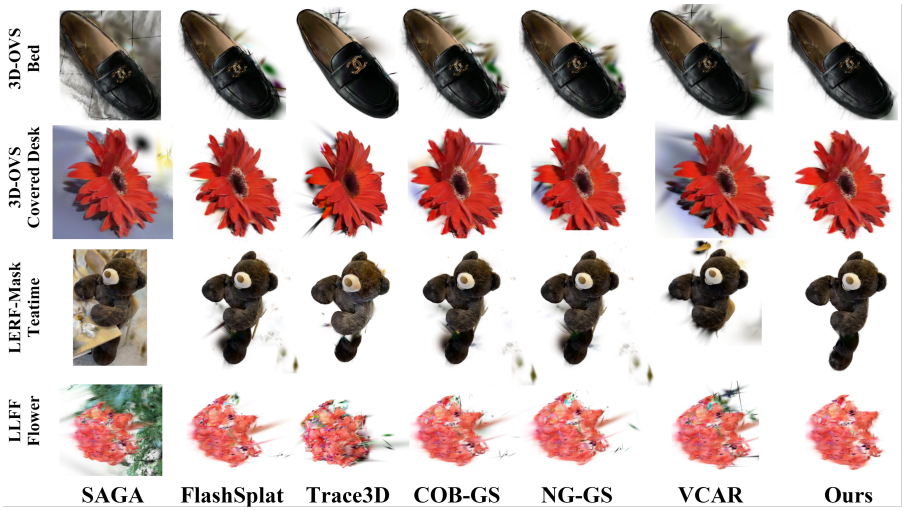


Figure 2: Qualitative comparison on 3D-OVS Liu et al. (2023a), LERF-Mask Ye et al. (2024), and LLFF Mildenhall et al. (2019). EVI-GS produces cleaner independently renderable object representations, particularly under viewpoints that expose boundary artifacts, elongated residuals, and detached Gaussians.

2 Related Work

2.1 3D Gaussian Scene Segmentation

Existing 3DGS-based scene segmentation methods He et al. (2026a) can be grouped by **the stage at which segmentation is introduced**. The first category establishes object identities **before the scene representation is constructed** and maintains object-label throughout reconstruction. Segment-then-Splat Lu et al. (2025) and ObjectGSZhu et al. (2025a) attach object labels to the input initial point cloud; newly spawned Gaussians inherit labels from their parents, and the voting mechanism of ObjectGS likewise occurs before 3DGS scene training. Such methods typically struggle to produce object representations with sharp boundaries Marrie et al. (2025) and require modifying the original scene reconstruction pipeline.

The second category **jointly optimizes geometry, appearance, and instance features during reconstruction**. GaussianGrouping Ye et al. (2024) and Unified-Lift Zhu et al. (2025b) attach high-dimensional instance features to Gaussians and lift two-dimensional semantic information into the three-dimensional representation. These methods exhibit strong semantic capacity, but the resulting object representations are tightly coupled to the original scene training process; they also tend to be computationally slow and to retain a large number of noisy primitives.

The third category performs segmentation **after scene-level 3DGS reconstruction**. FlashSplat Shen et al. (2024) formulates scene-conditioned Gaussian label assignment as a linear program with a closed-form solution and introduces a background bias to control foreground–background assignment. Trace3D Shen et al. (2025) traces Gaussian contributions across views to identify ambiguity, at substantially higher cost and with limited handling of fully occluded errors. SAGA Cen et al. (2023) clusters contrastive Gaussian affinities but remains computationally expensive. COB-GS Zhang et al. (2025) improves boundaries through adaptive Gaussian splitting, while NG-GS He et al. (2026b) introduces a continuous implicit field to mitigate discrete boundary artifacts. VCAR Cao et al. (2026) leverages spherical spiral sampling and axis-aware boundary refinement to alleviate boundary artifacts caused by insufficient view coverage and anisotropic Gaussian overflow. Despite different objectives, these methods infer labels under full-scene visibility/transmittance or refine object boundaries, rather than explicitly verifying candidate validity after scene-context removal.

2.2 Rendering Contribution and Boundary Artifacts

With fixed Gaussian geometry and opacity, label rendering is linear in Gaussian attributes, enabling contribution-based label lifting Shen et al. (2024). However, rendering contributions are inherently scene-conditioned: they measure foreground compatibility under the original occlusion structure, not validity after context removal. RCA therefore uses rendering contributions to establish a fixed

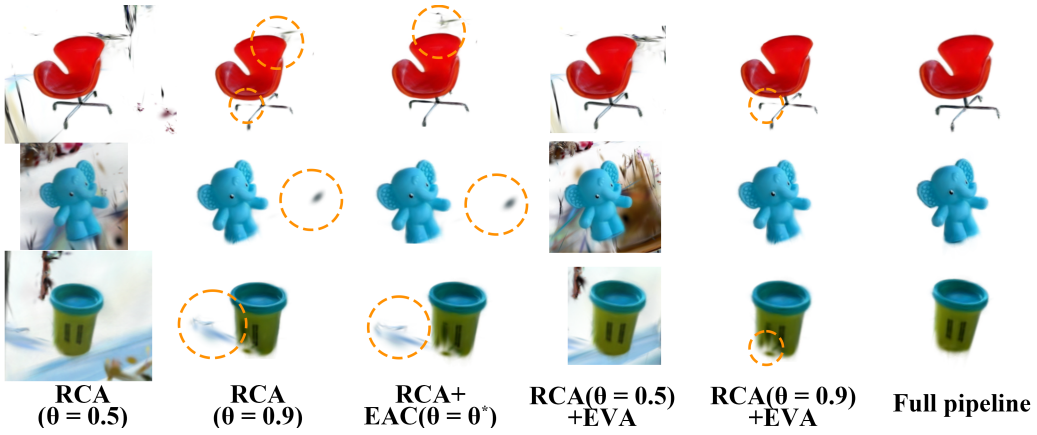


Figure 3: Qualitative ablation of our pipeline on the LERF-Mask's figurines as described in Sec 4.4.

observational association, not a final binary assignment. EVI-GS subsequently determines the selection boundary through independent-object rendering and checks candidate admissibility using evidence that a normalized association score alone cannot express.

COB-GS and NG-GS primarily address boundary ambiguity and the spatial discretization of Gaussian primitives. Their objectives improve object-boundary representation but are not explicitly designed to verify whether low-visibility Gaussians remain valid after non-target scene context is removed. EVI-GS instead treats the pretrained scene as a frozen representation and formulates extraction as evidence-based selection: questionable Gaussians are neither moved nor split; the method determines whether the available observations provide sufficient support for retaining them in the exported asset. Consequently, our method does not eliminate errors by moving, splitting, or retraining Gaussians; instead, it judges, from existing observational evidence, whether each Gaussian has sufficient grounds to be retained.

3 METHOD

3.1 PROBLEM FORMULATION

Given a fully trained 3D Gaussian scene $\mathcal{G}^0 = \{g_i\}_{i=1}^N$, each Gaussian is represented as $g_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \omega_i, \mathbf{c}_i)$, corresponding to the Gaussian center, covariance, opacity, and color or spherical-harmonic appearance parameters, respectively. We hold out one view every eight frames in the original view sequence for testing and use the remaining views for extraction. Using only the training views and the two-dimensional masks M_v of the same target object in these views. M_v denotes the visible target region in the original complete-scene image. The goal is to obtain, without modifying the parameters of $\mathcal{G}^0$ Kerbl et al. (2023),

$$\mathcal{G}_{\text{obj}} = \{g_i \in \mathcal{G}^0 : h_i = 1\}, \qquad h_i \in \{0, 1\}, \tag{1}$$

where h_i indicates whether the i-th Gaussian is ultimately retained in the independently renderable object asset. EVI-GS decomposes this process in Figure 1 into three stages: 1. Rendering-Contribution Attribution (RCA); 2. Extraction-Aligned Calibration (EAC); 3. Evidence-Validated Admission (EVA). This decomposition reflects a deliberate separation of three decisions that are often conflated in post-hoc 3DGS segmentation. RCA establishes a fixed association score from scene-conditioned observations; EAC uses this score to calibrate independent-object selection; and EVA performs a separate admission check without modifying the score. These stages distinguish how strongly a Gaussian is associated with the observed target from whether it should be retained in the exported asset. In particular, neither a high membership score nor a good scene-conditioned projection is treated as sufficient evidence that $\mathcal{G}_i$ should survive independent object extraction.

3.2 RENDERING-CONTRIBUTION ATTRIBUTION

RCA prepares observational evidence for extraction rather than directly assigning final object labels. It attributes each Gaussian's rendered contribution in the frozen scene to target and non-target observations, yielding the absolute foreground and background supports

$$S_i^F = \sum_{v,u} q_v(u) M_v(u) w_{vi}(u), S_i^B = \sum_{v,u} q_v(u)(1 - M_v(u)) w_{vi}(u), \tag{2}$$

Table 1: Standard-view quantitative comparison on 3D-OVS, LERF-Mask, and LLFF. We report independently rendered object masks using mIoU, mAcc, and mBIoU, visible-region appearance using SSIM(×100) and LPIPS(×100), and per-object extraction time. For methods that perform scene-level processing, runtime is reported as scene-level runtime divided by the number of evaluated objects.

Method	mIoU ↑	mAcc ↑	mBIoU ↑	SSIM ↑	LPIPS ↓	Obj-Time (s)
EVI-GS	**90.29**	**99.67**	**77.44**	<u>81.87</u>	**1.07**	<u>30.7</u>
VCAR	82.34	99.14	57.54	78.75	1.31	**4.7**
NG-GS	<u>89.86</u>	<u>99.64</u>	76.12	81.86	**1.07**	40.7
COB-GS	89.85	<u>99.64</u>	<u>76.17</u>	81.84	**1.07**	36.9
Trace3D	89.48	99.60	73.65	**82.72**	1.20	390.2
FlashSplat	87.54	99.42	70.86	76.82	1.79	3.2
SAGA	74.90	98.56	53.30	72.87	1.67	98.7

where $q_v(u) \in \{0, 1\}$ excludes pixels with insufficient complete-scene coverage. We recover these contribution statistics through a single auxiliary-color backward pass, avoiding explicit construction of a pixel–Gaussian contribution matrix. The auxiliary colors are used only for attribution; the pretrained Gaussian parameters remain unchanged. From the resulting evidence, RCA defines the fixed association score

$$p_i = \begin{cases} \dfrac{S_i^F}{S_i^F + S_i^B}, & S_i^F + S_i^B > \epsilon, \\ 0.5, & \text{otherwise.} \end{cases} \tag{3}$$

Insufficiently observed Gaussians receive the neutral score $p_i = 0.5$, rather than an unsupported background assignment. The score measures relative foreground association among observed contributions; it is neither a calibrated probability nor a certificate of object validity. In particular, weakly and strongly supported Gaussians can share the same normalized score. This distinction motivates the subsequent stages: EAC calibrates selection using independent-object rendering, while EVA evaluates support sufficiency and geometric consistency without rewriting p_i. RCA is therefore an attribution stage whose output remains fixed during subsequent calibration and admission.

3.3 Extraction-Aligned Threshold Calibration

Given RCA's fixed association scores, EAC calibrates a shared selection boundary for independent-object rendering. The key distinction is the rendering process through which selection is optimized, rather than the use of a learnable scalar alone. Removing non-target Gaussians changes transmittance and the visibility of retained primitives, whereas scene-conditioned label rendering preserves the original occlusion structure. EAC therefore gates Gaussian opacities and recomputes transmittance to optimize a differentiable relaxation of independent-object selection.

$$\theta = \sigma(\xi), \qquad \gamma_i(\theta) = \sigma\left(\frac{p_i - \theta}{T}\right), \tag{4}$$

which induce the soft-extracted opacities

$$\widetilde{\omega}_i = \gamma_i(\theta)\omega_i. \tag{5}$$

Under this soft extraction, we directly render the independent-object accumulated opacity $\widehat{A}_v(u; \theta)$. The threshold is optimized with a Soft-IoU Rahman & Wang (2016) objective:

$$\mathcal{L}_{\text{ext}} = \frac{1}{K} \sum_{v \in \mathcal{V}_K} \left(1 - \frac{\sum_u \widehat{A}_v(u; \theta) M_v(u)}{\sum_u \left[\widehat{A}_v(u; \theta) + M_v(u) - \widehat{A}_v(u; \theta) M_v(u)\right]}\right), \qquad K = 8. \tag{6}$$

During EAC, all Gaussian parameters and the memberships p_i remain fixed; only the scalar parameter ξ, and consequently $\theta = \sigma(\xi)$, is optimized. By default, θ is initialized to 0.9. At each optimization step, $K = 8$ views are randomly sampled from V_{fit}. We optimize for $P = \max(|V_{\text{fit}}|, 64)$ steps with a learning rate of 0.01 and retain the threshold $\theta^\star$ that attains the lowest fitting loss during optimization. Importantly, EAC does not introduce a new Gaussian representation or modify the pretrained scene. It changes where the selection decision is optimized: the scalar threshold is fitted against the rendering behavior of the standalone object itself. We therefore interpret $\theta^\star$ as an

Table 2: Segmentation quality evaluated via **SNV-Eval**. The best result is bolded, and the second-best is underlined.

Method	mIoU ↑	mAcc ↑	mBIoU ↑	SSIM ↑	LPIPS ↓
EVI-GS	**90.46**	**99.64**	**78.87**	**87.70**	**0.59**
VCAR	84.97	99.03	62.46	74.45	1.19
NG-GS	89.01	<u>99.03</u>	<u>77.36</u>	84.80	0.65
COB-GS	<u>89.29</u>	<u>99.03</u>	77.32	<u>85.30</u>	<u>0.64</u>
Trace3D	82.56	98.24	54.16	79.18	1.32
FlashSplat	87.47	98.45	73.11	83.12	1.27
SAGA	75.20	97.55	53.29	75.29	1.89

export-time decision boundary, rather than a threshold for an auxiliary scene-conditioned label map. The initial object candidate set is

$$\mathcal{C} = \{i : p_i \geq \theta^\star\}. \tag{7}$$

EAC optimizes a continuous relaxation of independent-object selection. The final asset is obtained by hard thresholding followed by EVA validation, with the RCA association scores kept fixed.

3.4 Evidence-Validated Admission

EAC thresholds the fixed RCA scores to obtain $C = \{i : p_i \geq \theta^\star\}$, but candidates with similar association scores can have substantially different observational support. EVA therefore performs a separate, read-only admission check on the thresholded proposal. It combines contribution purity and absolute support with cross-view contribution–surface consistency and bounded connectivity to reliable seeds. Unlike RCA, whose output is the association score p_i, EVA outputs an admission indicator η_i. Neither p_i nor the calibrated threshold $\theta^\star$ is updated during this stage.

The rasterizer also produces a weighted inverse depth $\sum_i w_{vi}(u)/d_{vi}(u)$. When $A_v^0(u) \geq a_{\min}$, this quantity is converted into a harmonic surface depth comparable to Gaussian camera depths:

$$D_v^0(u) = \frac{A_v^0(u)}{\sum_i w_{vi}(u)/d_{vi}(u)}. \tag{8}$$

A_v^0 and D_v^0 serve as read-only geometric evidence and do not enter optimization. During candidate validation, EVA re-reads foreground and background contribution statistics on the evidence views through auxiliary-color backpropagation. These reads support admission tests rather than a second membership-estimation stage: their normalized ratio is used only as a purity diagnostic and is never written back to the RCA score p_i. We define this diagnostic as

$$\rho_i = \frac{S_i^F}{S_i^F + S_i^B}. \tag{9}$$

The absolute foreground support is also retained for the seed-support test in Eq. (10). Thus, EVA uses contribution statistics as evidence for admission while leaving the association used by EAC unchanged. We further project each Gaussian center and local probes along its principal axis onto each view, and, on sufficiently covered pixels, compare the camera depth d_{vi} with D_v^0. A match with $|d_{vi} - D_v^0| \leq \theta_d D_v^0$ that falls on the foreground mask is counted as a foreground surface match; the same near-surface condition on the background mask is counted as a background surface conflict; and $d_{vi} < D_v^0 - \theta_d D_v^0$ is counted as a free-space conflict. Accumulating these events across views yields the number of foreground surface matches n_i^{fg}, the number of background and free-space conflicts n_i^-, and the number of joint supports n_i^{joint} that simultaneously satisfy foreground contribution and foreground surface matching. Reliable seeds require not only membership above the threshold but also sufficient absolute foreground quality and cross-view surface support. Using high-purity, multi-view jointly supported candidates as a reference, we define a support lower bound $\underline{S} = \max(S_{\min}, \kappa \cdot \mathrm{median}(S^F))$ and

$$\mathcal{S} = \left\{ i \in \mathcal{C} : S_i^F \geq \underline{S},\ n_i^{\mathrm{joint}} \geq n_{\min},\ n_i^{\mathrm{fg}} \geq n_{\min},\ \rho_i \geq \rho_{\min},\ n_i^- \leq n_i^{\mathrm{fg}} + 1 \right\}. \tag{10}$$

Candidates that repeatedly exhibit background-surface or free-space conflicts while possessing almost no foreground surface support are marked as hard negatives; they may serve neither as seeds nor as subsequent propagation nodes. The remaining non-hard-negative candidates form a local adjacency graph whose edge length is at most a bounded multiple of the median seed scale, and the

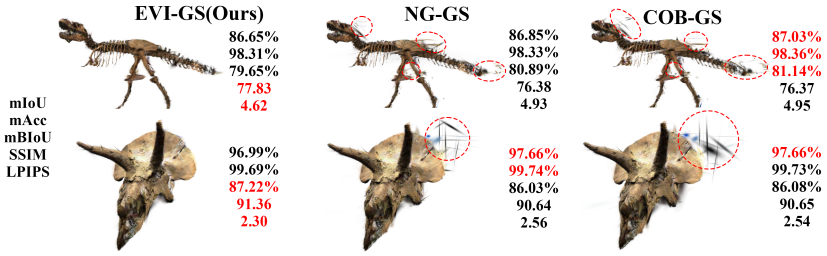


Figure 4: Sparse-view projection metrics may fail to expose view-dependent extraction artifacts. Although the compared methods obtain similar scores under the evaluated cameras, additional viewpoints reveal detached Gaussians and elongated residuals that are not visible in those projections. Red denotes the best value for each reported metric.

number of propagation hops is likewise limited. We set $\eta_i = 1$ if and only if the candidate can be connected to $\mathcal{S}$ along this local structure. The rule retains multiple spatially separated object parts that each possess reliable observational evidence, rather than collapsing to the globally largest connected component. To avoid disconnecting genuine fine structure whose membership is slightly below $\theta^\star$, near-surface Gaussians with membership marginally below the threshold are allowed as bridging nodes in the propagation; they themselves do not enter the exported set. For non-hard-negative candidates near an already admitted body that have weaker evidence, we allow only a single local attachment, recovering fine structure without introducing new disconnected components.

To prevent evidence validation from over-pruning genuine object structure, we further validate the EVA proposal on calibration views that are disjoint from the evidence views. We independently render both the original candidate set C and the post-admission proposal under V_{cal}. If the proposal causes the IoU or foreground recall to decrease beyond a predefined tolerance, retains an insufficient fraction of the original candidates, or becomes empty, EVA is rejected and the original candidate set C is retained. We likewise fall back to C when no reliable seed can be established from the evidence views. This safeguard allows EVA to suppress unsupported residuals while avoiding aggressive deletion under insufficient evidence. EVA changes admission, not attribution: re-reading S_i^F and S_i^B does not update p_i or recalibrate $\theta^\star$. Combining the EAC candidate mask with the EVA admission decision gives

$$h_i = [p_i \geq \theta^\star]\, \eta_i, \tag{11}$$

where a rejected EVA proposal corresponds to $\eta_i \equiv 1$, i.e., falling back to the extraction-aligned candidate set C. This design suppresses low-evidence residuals exposed after scene-context removal without modifying the pretrained Gaussian representation.

Implementation details. We use $\theta_d = 0.05$, $S_{\min} = 1e-3$, $\kappa = 0.05$, $n_{\min} = 2$, and $\rho_{\min} = 0.6$. Hard-negative, graph-radius, propagation-hop, bridging-margin, and attachment thresholds are 2, $8\tilde{s}$, 4, 0.1, and $12\tilde{s}$, respectively. EVA is rejected when the IoU/recall drop exceeds $0.01/0.02$ or the retained-candidate ratio falls below 0.5.

4 Experiments

4.1 Experimental Setup

Data and splits. The experimental scenes are drawn from 3D-OVS, LERF-Mask, and LLFF, comprising 18 scenes and 71 target objects. For each scene, one view in every eight is held out for testing in the original view sequence, and all remaining views form the training set. The extraction stage uses target masks from training views only; ground-truth masks of test views do not participate in Gaussian membership computation, threshold learning, candidate admission, hyperparameter selection, or early stopping. Within EVI-GS, the training views are further partitioned into disjoint fitting and calibration sets with a default ratio $|V_{\text{fit}}| : |V_{\text{cal}}| = 8 : 2$. The fitting views are used for RCA attribution, EAC optimization, and EVA evidence collection; the calibration views are used only to accept or reject the EVA proposal. All compared methods take identical dataset target masks as two-dimensional input; differences among SAM Kirillov et al. (2023), GroundingDINO Liu et al. (2023b), DEVACheng et al. (2023), or other two-dimensional front-ends are not included in the comparison.

Standard test metrics. Each method first exports an independently renderable target Gaussian set $\mathcal{G}_{\text{obj}}$ and independently renders its accumulated α under the test cameras, A_v^{obj}. Using a fixed

Table 3: Ablation studies on 3D-OVS, LERF-Mask and LLFF datasets.

Method	mIoU ↑	mAcc ↑	mBIoU ↑	SSIM ↑	LPIPS ↓	Obj-Time(s)
RCA(θ =0.5)	47.79	80.82	22.28	73.93	1.22	**12.8**
RCA(θ =0.9)	77.64	95.41	68.17	78.05	1.20	12.3
RCA+EAC	88.33	99.56	72.19	81.02	1.09	22.9
FlashSplat+LB	87.86	99.51	71.22	78.53	1.47	15.6
RCA(θ =0.5)+EVA	51.94	81.29	31.65	87.81	0.77	35.3
RCA(θ =0.9)+EVA	89.00	99.62	74.00	79.96	1.15	27.5
Ours	**90.29**	**99.67**	**77.44**	81.87	1.07	30.7
Init-0.5	88.35	99.53	73.01	**84.81**	**0.88**	27.2
Init-0.7	89.32	99.62	74.88	83.06	0.99	30.4
V4	89.44	99.64	75.01	80.73	1.12	24.9
Vall	89.54	99.64	75.21	80.90	1.11	98.6
C91	89.30	99.61	74.76	80.85	1.10	29.3
C73	89.41	99.64	75.02	80.79	1.11	28.4
S4	88.56	99.57	72.81	79.88	1.13	28.7
S8	89.42	99.63	74.96	80.71	1.12	24.7

threshold 0.5, the standard evaluation comprises mIoU, mAcc, mBIoU Cheng et al. (2021), foreground SSIM, masked LPIPS , and extraction time. The comparative results are presented in Figure 2 and Table 1.

4.2 VIEW-COVERAGE FAILURE MODES IN OBJECT-EXTRACTION EVALUATION

Conventional projection metrics remain our primary quantitative benchmark. Our concern is not the validity of these metrics themselves, but the viewpoint coverage under which they are measured. For independently extracted 3D assets, errors such as detached Gaussians or elongated structures may remain completely occluded from a small set of test cameras and therefore incur little or no projection penalty. Consequently, sparse-view evaluation measures the quality of the observed projections, but does not necessarily reveal every failure mode of the underlying extracted representation. Figure 4 illustrates this coverage effect: methods with comparable projection scores can exhibit substantially different behavior when the same extracted object is inspected from additional viewpoints. We therefore treat dense-view inspection as a complementary stress test rather than a replacement for conventional metrics.

4.3 SYNTHETIC NOVEL-VIEW AUXILIARY EVALUATION

To specifically test the view-dependent failure mode identified above, we introduce **Synthetic Novel-View Auxiliary Evaluation** as an auxiliary stress test. SNV-Eval does not define a new source of GT supervision and is not used for training, model selection, threshold calibration, or post-processing. Instead, it increases the density of observation directions under which the final extracted representation is inspected, reducing the chance that erroneous Gaussians remain hidden simply because of sparse test-camera coverage. The corresponding results are reported in Table 2.

Each novel view lies on the SE(3) geodesic between two consecutive held-out test cameras whose masks contain the target object. With camera-to-world poses T_a and T_b, the inserted pose is the midpoint

$$T_{1/2} = T_a \exp\left(\frac{1}{2}\log\left(T_a^{-1}T_b\right)\right) \tag{12}$$

and the resulting rotation is projected back to $\mathrm{SO}(3)$. Applying this midpoint-interpolation rule yields the auxiliary camera set $\widetilde{V}_{\mathrm{aux}} = \{\widetilde{C}_j\}_{j=1}^{N_{\mathrm{aux}}}$. The generation rule is fixed across scenes, and all methods are evaluated using the same generated cameras and intrinsics. The auxiliary cameras are determined without reference to the outputs of any extraction method. Test masks are used here only to identify eligible camera pairs for auxiliary evaluation, not for extraction or hyperparameter selection.

For each $\widetilde{C}_j$, RGB rendering uses only the original frozen complete scene $\mathcal{G}^0$:

$$\widetilde{I}_j^{\mathrm{scene}} = \mathcal{R}_{\mathrm{RGB}}\left(\mathcal{G}^0, \widetilde{C}_j\right). \tag{13}$$

Synthetic image generation is therefore independent of any object extraction method. We then obtain a pseudo-reference mask $\widetilde{M}_j$ of the corresponding target with SAM 3. We emphasize that this mask is not treated as a ground-truth annotation, but as a pseudo-reference used to increase evaluation viewpoint coverage. Under an auxiliary camera $\widetilde{C}_j$, independent rendering yields

$$\widetilde{A}_j^{(m)} = \mathcal{R}_\alpha\big(\mathcal{G}_{\text{obj}}^{(m)}, \widetilde{C}_j\big). \tag{14}$$

Using the same fixed α threshold as in the standard test,

$$\widetilde{M}_j^{(m)} = \mathbf{1}\big[\widetilde{A}_j^{(m)} \geq 0.5\big]. \tag{15}$$

We then compute mIoU, mAcc, and mBIoU, applying the same macro-averaging rule as the standard metrics. The pseudo-reference appearance is constructed as

$$\widetilde{I}_j^{\text{ref}} = \widetilde{I}_j^{\text{scene}} \odot \widetilde{M}_j. \tag{16}$$

The object extraction is independently rendered to $\widetilde{I}_j^{(m)}$. SSIM and LPIPS are then evaluated only in the pseudo-reference foreground. These metrics still assess visible appearance preservation of the extracted target, rather than directly detecting floating Gaussians outside the object region. The design goal of SNV-Eval is not to replace genuine test ground truth with pseudo-labels, but to reduce, via denser camera directions, the probability that erroneous Gaussians happen to remain occluded under a small number of test viewpoints.

4.4 Ablation Studies

We first examine the roles of attribution, calibration, and admission in Table 3. Directly thresholding the RCA scores yields 47.79 mIoU at a fixed threshold of 0.5 and 77.64 at 0.9, showing that observational association alone does not determine a reliable extraction boundary. RCA+EAC improves the latter result to 88.33 mIoU and increases mBIoU from 68.17 to 72.19. Applying EVA to the fixed-0.9 proposal yields 89.00 mIoU and 74.00 mBIoU, whereas the full pipeline reaches 90.29 and 77.44, respectively. These results support combining selection calibration with evidence-based admission. Relative to RCA+EAC, EVA adds 7.8 s per object. Figure 3 shows the visualization results of ablation studies.

Comparison with learnable background bias. We additionally evaluate **FlashSplat+LB**, a variant that learns the background bias using scene-conditioned label rendering. Although this makes foreground–background assignment adaptive, the fitting process retains the transmittance of the complete scene. RCA+EAC instead optimizes selection through opacity gating and recomputed transmittance. The lower extraction quality of FlashSplat+LB is consistent with this difference: residual Gaussians suppressed by the original occlusion structure may remain weakly penalized during scene-conditioned fitting but become visible after extraction. This comparison suggests that making the decision parameter learnable alone is insufficient; the rendering process used for calibration also matters.

We further examine four implementation choices. Init-0.5 and Init-0.7 initialize the learnable EAC threshold at 0.5 and 0.7, respectively, whereas Ours uses 0.9. V4 and Vall use four and all fitting views per EAC optimization step, respectively, compared with eight in the default setting. C91 and C73 set the view-count ratio $|V_{\text{fit}}| : |V_{\text{cal}}|$ to $9:1$ and $7:3$, respectively; the default is $8:2$. S4 and S8 use four and eight views selected from V_{fit} for RCA attribution, whereas the default uses all fitting views. Among the tested initializations, 0.9 achieves the highest mIoU and mBIoU, while lower initializations better preserve visible-region appearance. Eight views per EAC step provide a favorable quality–efficiency balance: using all fitting views increases runtime from 30.7 s to 98.6 s without improving segmentation quality. The default fit/calibration ratio and full-view RCA attribution also achieve the best mIoU and mBIoU within their respective ablation groups.

5 Conclusion

We presented EVI-GS for independently renderable object extraction from pretrained 3DGS scenes. The key insight is that full-scene membership does not guarantee validity after scene-context removal. EVI-GS uses RCA to establish fixed observational association, EAC to calibrate

independent-object selection, and EVA to validate candidate admission without revising that association or modifying the pretrained representation. SNV-Eval further exposes view-dependent errors missed by sparse test cameras. These results frame post-hoc 3DGS object extraction as export-time representation selection rather than scene-conditioned label propagation.

AI USE STATEMENT

In this work, we used generative AI tools solely to assist with organizing and summarizing experimental results, preparing preliminary sketches for scientific figures, and translating the manuscript into English. Generative AI was not used for research hypothesis formulation, method design, algorithm development, code implementation, experimental design, data generation, or the production of experimental results.

REPRODUCIBILITY STATEMENT

We place particular emphasis on the reproducibility of this work. A fully anonymized repository containing the complete implementation of our method has already been prepared and is available for review. The repository includes the full main pipeline, optimization and inference code, as well as the evaluation scripts used to compute all reported quantitative metrics. We also provide the datasets used in our experiments together with the corresponding data organization, preprocessing, and evaluation protocols. Detailed hyperparameter settings, experimental configurations, and implementation details required to reproduce the reported results are documented in the paper, supplementary materials, and anonymous repository. The repository is maintained in an anonymized form during the double-blind review process and will be converted to a permanent public repository after the review process.

REFERENCES

Kun Cao, Di Wang, Haibin Zhu, Haozhi Huang, Xu Wang, Zheng Shi, and Guanghua Yang. Vcar: Training-free 3dgs segmentation via view completeness and axis-aware boundary refinement, 2026. URL https://arxiv.org/abs/2608.30870.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3d gaussians. *arXiv preprint arXiv:2312.00860*, 2023.

Bowen Cheng, Ross Girshick, Piotr Dollar, Alexander C. Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15334–15342, June 2021.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *ICCV*, 2023.

Shuting He, Peilin Ji, Yitong Yang, Changshuo Wang, Jiayi Ji, Yinglin Wang, and Henghui Ding. A survey on 3d gaussian splatting applications: Segmentation, editing, and generation, 2026a. URL https://arxiv.org/abs/2508.09977.

Yi He, Tao Wang, Yi Jin, Congyan Lang, Yidong Li, and Haibin Ling. Ng-gs: Nerf-guided 3d gaussian splatting segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42061–42070, 2026b.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric Xing, and Shijian Lu. Weakly supervised 3d open-vocabulary segmentation. *arXiv preprint arXiv:2305.14093*, 2023a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023b.

Yiren Lu, Yunlai Zhou, Yiran Qiao, Chaoda Song, Tuo Liang, Jing Ma, Huan Wang, and Yu Yin. Segment then splat: Unified 3d open-vocabulary segmentation via gaussian splatting. *Advances in Neural Information Processing Systems*, 2025.

Juliette Marrie, Romain Menegaux, Michael Arbel, Diane Larlus, and Julien Mairal. Ludvig: Learning-free uplifting of 2d visual features to gaussian splatting scenes, 2025. URL https://arxiv.org/abs/2410.14462.

Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019.

Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Fatih Porikli, Sandra Skaff, Alireza Entezari, Jianyuan Min, Daisuke Iwai, Amela Sadagic, Carlos Scheidegger, and Tobias Isenberg (eds.), *Advances in Visual Computing*, pp. 234–244, Cham, 2016. Springer International Publishing. ISBN 978-3-319-50835-1.

Hongyu Shen, Junfeng Ni, Yixin Chen, Weishuo Li, Mingtao Pei, and Siyuan Huang. Trace3d: Consistent segmentation lifting via gaussian instance tracing. 2025.

Qiuhong Shen, Xingyi Yang, and Xinchao Wang. Flashsplat: 2d to 3d gaussian splatting segmentation solved optimally. *European Conference of Computer Vision*, 2024.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

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 *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pp. 19114–19138, 2024.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *ECCV*, 2024.

Jiaxin Zhang, Junjun Jiang, Youyu Chen, Kui Jiang, and Xianming Liu. Cob-gs: Clear object boundaries in 3dgs segmentation based on boundary-adaptive gaussian splitting. In *CVPR*, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Ruijie Zhu, Mulin Yu, Linning Xu, Lihan Jiang, Yixuan Li, Tianzhu Zhang, Jiangmiao Pang, and Bo Dai. Objectgs: Object-aware scene reconstruction and scene understanding via gaussian splatting. In *International Conference on Computer Vision (ICCV), 2025*, 2025a.

Runsong Zhu, Shi Qiu, Zhengzhe Liu, Ka-Hei Hui, Qianyi Wu, Pheng-Ann Heng, and Chi-Wing Fu. Rethinking end-to-end 2d to 3d scene segmentation in gaussian splatting. *arXiv preprint arXiv:2503.14029*, 2025b.

A APPENDIX

Table 4: Segmentation quality evaluated on **3D-OVS**. The best result is bolded, and the second-best is underlined.

Method	mIoU ↑	mAcc ↑	mBIoU ↑	SSIM ↑	LPIPS ↓
EVI-GS	**92.91**	**99.77**	**80.17**	<u>84.53</u>	**0.77**
VCAR	82.62	99.24	54.44	78.61	0.99
NG-GS	<u>92.38</u>	<u>99.75</u>	<u>78.67</u>	83.70	<u>0.79</u>
COB-GS	92.35	<u>99.75</u>	78.73	83.69	<u>0.79</u>
Trace3D	91.96	99.69	76.01	**84.67**	0.88
FlashSplat	89.69	99.55	71.83	75.25	1.25
SAGA	76.26	98.69	53.27	71.03	1.38

Table 5: Segmentation quality evaluated on **LERF-Mask**. The best result is bolded, and the second-best is underlined.

Method	mIoU ↑	mAcc ↑	mBIoU ↑	SSIM ↑	LPIPS ↓
EVI-GS	**87.58**	**99.66**	**73.44**	78.69	1.43
VCAR	81.61	99.29	62.34	77.55	1.65
NG-GS	<u>87.06</u>	99.56	<u>71.98</u>	**80.18**	**1.33**
COB-GS	87.00	99.56	71.75	<u>80.09</u>	<u>1.34</u>
Trace3D	86.98	<u>99.62</u>	70.64	79.37	1.83
FlashSplat	85.81	99.52	69.29	73.12	2.17
SAGA	73.60	98.40	53.25	71.08	2.35

Table 6: Segmentation quality evaluated on **LLFF**. The best result is bolded, and the second-best is underlined.

Method	mIoU ↑	mAcc ↑	mBIoU ↑	SSIM ↑	LPIPS ↓
EVI-GS	**93.73**	<u>98.71</u>	<u>74.29</u>	<u>85.25</u>	<u>3.89</u>
VCAR	86.84	96.89	56.14	79.74	4.44
NG-GS	93.21	**98.87**	**74.85**	84.84	3.99
COB-GS	93.24	**98.87**	74.02	84.85	3.98
Trace3D	<u>93.43</u>	<u>98.71</u>	73.03	**87.20**	**3.66**
FlashSplat	91.85	97.85	72.50	79.56	4.26
SAGA	78.97	98.44	52.68	75.70	4.67