GR-GS : GEOMETRIC REGULARIZATION 3D GAUSSIAN SPLATTING SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in 3D Gaussian Splatting (3DGS) have achieved remarkable success in real-time scene rendering. However, interactive 3DGS scene segmentation remains challenging. Current MLLM-based agents often treat spatial grounding as a black box, propagating multi-instance ambiguities into the 3D domain where an absence of structural constraints allows boundary Gaussians to erroneously elongate along viewing rays. In this paper, we introduce GR-GS, a unified framework that tightly couples explicit topological reasoning with physical covariance regularization to achieve high-fidelity 3D segmentation. Rather than relying on unstable multi-view voting to resolve semantic ambiguities, our framework introduces an *LLM-Driven Relational Extraction* (LDRE) module that enforces rigorous, pre-lifting topological constraints, isolating the exact target directly in 2D before spatial corruption can occur. To ensure these precise semantic boundaries are preserved during 3D optimization, we propose a *Ray-Aligned Covariance Penalty* (RACP). Instead of generic regularization, RACP leverages cross-view Jacobian projection to artificially inflate the screen footprint of ill-posed, elongated Gaussians. This forced boundary spillover triggers steep positional gradients that reactivate the splitting mechanism, organically compressing artifacts without additional parameters. Extensive experiments demonstrate that GR-GS achieves state-of-the-art results on NVOS and LERF-OVS. By treating semantic grounding and geometric fidelity as a coupled optimization problem, our method delivers superior relational reasoning and flawlessly delineated object boundaries.

1 INTRODUCTION

Interactive 3D scene segmentation serves as a fundamental pillar for advanced spatial understanding, facilitating critical applications ranging from autonomous robotic navigation (Jatavallabhula et al., 2023; Gu et al., 2024) to immersive mixed reality interactions (Xu et al., 2026; Wang et al., 2024). Recently, 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) has emerged as a dominant representation paradigm, distinguished by its real-time rendering efficiency and explicit geometric manipulability. To facilitate interactive segmentation within this framework, recent methodologies increasingly rely on natural language guidance. However, accurately grounding open-vocabulary instructions (Zhai et al., 2025), particularly complex or implicit spatial queries, into these unconstrained discrete representations remains a formidable challenge. The primary bottleneck stems from the inherent geometric ambiguity of language: while textual prompts (e.g., "The toy beside the plush pikachu!") provide rich relational and semantic contexts, they lack explicit physical coordinates. This discrepancy creates a profound mapping gap between 2D vision-language relational disambiguation and 3D physical space.

Contemporary segmentation pipelines attempt to bridge this semantic-physical gap by unprojecting 2D generic masks (e.g., SAM (Kirillov et al., 2023)) onto spatial representations. However, recent MLLM-agent frameworks treat spatial reasoning as a black box, relying on unstructured prompt mapping that is highly susceptible to multi-instance ambiguity. Consequently, they are forced to employ expensive, multi-view ID voting to resolve 3D inconsistencies. Furthermore, even with accurate 2D masks, lifting planar features into volumetric space remains fundamentally ill-posed along the projection axis. Lacking mathematically rigorous geometric constraints, boundary Gaussians erroneously deform and elongate along the camera viewing ray to minimize 2D projection error (Fig. 1). This structural degradation triggers prominent aliasing artifacts and unconstrained boundary bleeding when observed from novel viewpoints.

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.

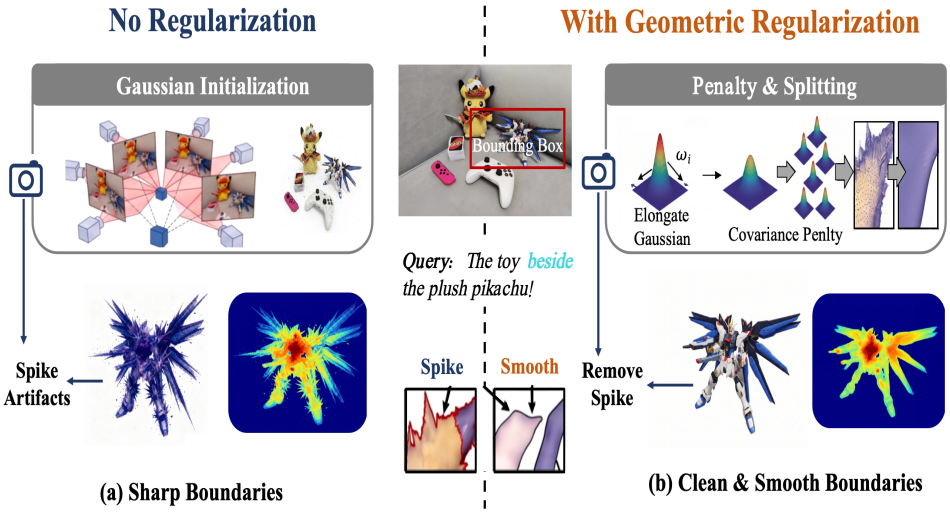

Figure 1: Effect of Geometric Regularization on 3D Gaussian segmentation. Left (a): without regularization, Gaussian splatting optimization is plagued by spike artifacts on boundaries. Right (b): our proposed geometric regularization resolves these issues by maintaining geometric consistency, yielding defined surfaces with clean and smooth boundaries.

To fundamentally rectify these intrinsic limitations, we propose ***G**eometric-**R**egularization 3D**GS*** (GR-GS), a unified framework that tightly couples explicit topological reasoning with physical covariance regularization. To eliminate semantic ambiguity at its source, we introduce an *LLM-driven relational extraction* (LDRE) module. Unlike methods that rely on unstable multi-view voting heuristics, LDRE acts as a deterministic pre-lifting gatekeeper. It decomposes textual queries into structured semantic tuples to physically validate topological constraints directly in the 2D domain, isolating the exact target before spatial corruption can occur. To ensure these precise semantic priors safely guide the 3D optimization, a robust mask filtering (RMF) mechanism systematically purifies view-dependent inconsistencies. Finally, to eradicate geometric degradation, we devise a *ray-aligned covariance penalty* (RACP). Rather than applying generic boundary heuristics, RACP leverages cross-view Jacobian projections to artificially inflate the screen footprint of ill-posed, elongated Gaussians. This forced boundary spillover strictly guarantees a positive-definite gradient update that physically compresses the geometric artifacts and reactivates the native 3DGS split-and-prune cycle, rectifying boundary structures without requiring additional learnable parameters.

For evaluation, we compare our method against state-of-the-art baselines on three standard interactive segmentation benchmarks: NVOS (Jambon et al., 2023), LERF-OVS (Kerr et al., 2023), and ScanNet (Dai et al., 2017). Specifically, our method conclusively surpasses the strongest baseline, REALM (Shi et al., 2026), on all three datasets, achieving average gains of 5.3% in mAcc, 3.2% in mIoU, and 6.2% in B-mIoU. These results underscore the superior accuracy and robustness of our coupled optimization strategy in resolving intricate geometric boundaries. We will publicly release the code upon acceptance.

In summary, our main contributions are as follows:

- We propose GR-GS, a unified 3DGS segmentation framework that tightly couples explicit 2D semantic reasoning with 3D physical geometry regularization, overcoming the fundamental misalignment inherent in cross-dimensional feature lifting.
- We introduce the LDRE module paired with RMF, which explicitly resolves multi-instance ambiguities through rigorous pre-lifting topological constraint validation, completely bypassing the need for unstable post-hoc multi-view voting.
- We formulate the RACP module, a mathematically rigorous regularizer that leverages cross-view Jacobian projections to guarantee positive-definite gradient updates, physically compressing ray-aligned artifacts to achieve flawlessly smooth boundaries.
- Extensive experiments demonstrate that GR-GS achieves state-of-the-art performance across competitive benchmarks, delivering significant quantitative gains in complex relational reasoning and boundary-aware metrics.

2 Related Work

2.1 3D Scene Representations

Although 3D Gaussian Splatting (3DGS) has revolutionized novel view synthesis with its real-time rendering efficiency and explicit geometric manipulability, seamlessly adapting this discrete representation for interactive scene understanding remains a formidable challenge. The assignment of highly accurate, instance-level masks to target objects based on user prompts is a key task for downstream applications, which span autonomous robotic navigation (Jatavallabhula et al., 2023), intelligent vehicle perception (Behley et al., 2019; Lai et al., 2022), and immersive mixed reality (Xu et al., 2026). However, a significant barrier persists in accurately decoupling fine-grained object geometries from the underlying 3D representations without introducing severe artifacts. To overcome this inherent limitation and propel human-machine collaborative perception, it is imperative to explore advanced text-guided 3DGS segmentation paradigms that inherently enforce geometric consistency.

2.2 Interactive 3D Scene Segmentation

Transitioning from static reconstruction to interactive scene understanding (Chen et al., 2024; Roessle et al., 2024; Zuo et al., 2025) requires elevating image-level semantic priors into the spatial domain. Contemporary open-world segmentation methodologies operate primarily through feature distillation or mask-driven guidance. Adapting this to explicit representations, LangSplat (Qin et al., 2024) and Feature3DGS (Zhou et al., 2024) project 2D semantic features onto 3D Gaussians, allowing real-time spatial interaction. Alternatively, mask-driven methodologies directly process the output from 2D foundation models to preserve instance integrity. Gaussian Grouping (Ye et al., 2024) introduces 2D spatial consistency constraints across rendered masks to improve per-point discrimination, while mask-guided optimization frameworks such as SAGA (Cen et al., 2025a) and COB-GS (Zhang et al., 2025) explicitly utilize boundary-adaptive splitting heuristics based on spatial gradient statistics. Furthermore, REALM (Shi et al., 2026) explores the multimodal reasoning capabilities of MLLMs in the context of 3D visual grounding.

Crucially, a fundamental bottleneck across these methodologies remains the absence of explicit 3D geometric constraints for boundary structures. This unconstrained spatial ambiguity directly triggers structural degradation and pervasive elongated artifacts at object margins, underscoring the urgent need for explicit geometric regularization within the 3D representation. However, rather than relying purely on feature-level interactions, our framework introduces explicit topological constraint validation and robust mask filtering prior to 3D optimization, strictly eliminating multi-instance ambiguities.

3 Method

3.1 Framework Overview

As illustrated in Figure 2, our GR-GS framework operates within an explicit neural representation paradigm, introducing a semantic-spatial constraint mechanism to combat geometric degradation and semantic ambiguity prevalent in interactive 3D segmentation.

Given a multi-view image sequence $\mathcal{I} = \{\mathbf{I}_1, \ldots, \mathbf{I}_V\}$ and a complex relational text query $\mathcal{Q}$, GR-GS reconstructs a semantically decoupled scene through three stages:

- *LLM-Driven Relational Extraction (LDRE)*: This component parses raw user queries into structured semantic tuples, translating abstract natural language into actionable topological priors that guide spatial alignment and mask consistency.
- *Robust Mask Filtering (RMF)*: It maintains training efficiency and accuracy by identifying and discarding low-confidence supervisory signals, such as occlusion-induced artifacts or incomplete boundary masks.
- *Ray-Aligned Covariance Penalty (RACP)*: Serving as our primary geometric regularizer, this module enforces physical plausibility by penalizing the unconstrained radial elongation of Gaussian primitives. It functions within a closed-loop optimization framework where boundary-level gradient signals trigger adaptive splitting and pruning, ensuring the learned 3D representation is tightly anchored to the target surface geometry.

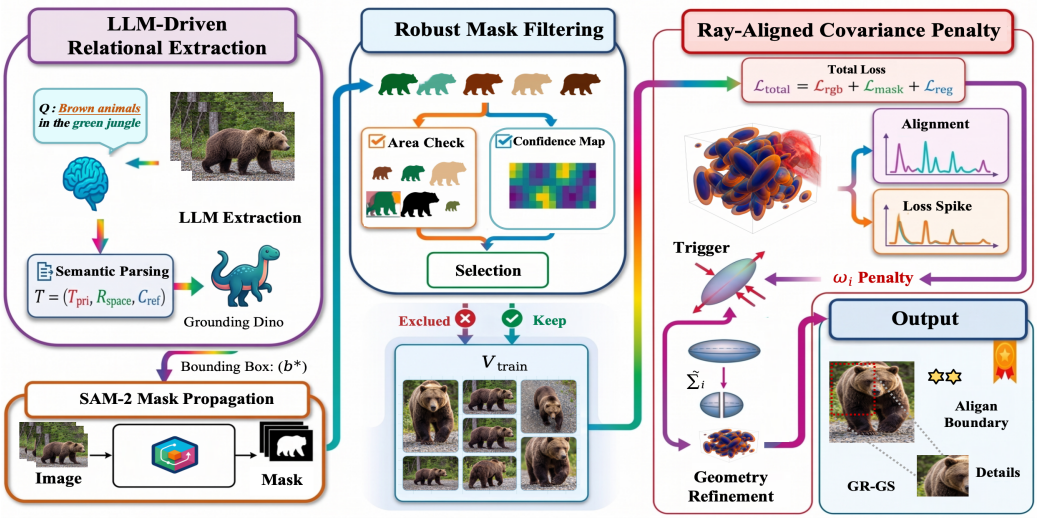

Figure 2: Overview of the GR-GS framework. The framework proceeds through three stages: (1) LDRE parses the query into a structured semantic tuple, and grounds the target object via joint visual-semantic similarity and topological constraint validation; (2) RMF evaluates mask area ratios and pixel-wise confidence scores, retaining only reliable supervision signals; (3) RACP explicitly penalizes elongation of boundary Gaussians along the viewing ray, artificially inflating their screen footprint to trigger gradient-driven splitting and pruning.

3.2 LLM-Driven Relational Extraction

While recent frameworks like REALM (Shi et al., 2026) treat spatial reasoning as a black box, their unstructured prompt mapping causes multi-instance ambiguity and necessitates expensive post-hoc multi-view ID voting. To prevent these semantic errors from propagating into 3D space, our LDRE module achieves precise grounding through explicit, pre-lifting topological constraint validation.

Initially, rather than directly predicting bounding boxes from raw text, an LLM parses the complex query $\mathcal{Q}$ into a structured semantic tuple $\mathcal{T} = \langle T_{\text{pri}}, R_{\text{space}}, C_{\text{ref}} \rangle$. Specifically, T_{pri} designates the core target object (e.g., "Brown animals"), R_{space} defines the explicit topological or positional relationship (e.g., "in the"), and C_{ref} specifies the contextual anchor providing a spatial reference (e.g., "green jungle").

Given view $\mathbf{I}_i \in \mathcal{I}$, we utilize Grounding DINO (Liu et al., 2024) to generate a comprehensive set of candidate bounding box proposals $\mathcal{B}_i$ for the primary target T_{pri}. To evaluate visual-semantic alignment, the cross-modal cosine similarity between a candidate box $b \in \mathcal{B}_i$ and the target text is computed as:

$$\text{Sim}_{\text{VL}}(b, T_{\text{pri}}) = \frac{\Phi_{\text{img}}(b)^\top \Phi_{\text{text}}(T_{\text{pri}})}{\|\Phi_{\text{img}}(b)\|_2 \|\Phi_{\text{text}}(T_{\text{pri}})\|_2}, \tag{1}$$

where Φ_{img} and Φ_{text} represent the pre-trained vision and language encoders, respectively.

Crucially, relying solely on isolated feature similarity inevitably fails when multiple visually identical instances exist in the scene. To strictly resolve this pre-projection, we integrate a relation-sensitive Boolean indicator $\Omega(b, R_{\text{space}}, C_{\text{ref}}) \in \{0, 1\}$. Specifically, $\Omega = 1$ if the spatial arrangement between the candidate box b and the detected reference anchor C_{ref} strictly satisfies the topological condition defined by R_{space} (implemented via depth-aware bounding box heuristics, e.g., relative center offsets for "left/right" and depth-ordering for "front/behind"), and $\Omega = 0$ otherwise. The optimal bounding box b^* for the reference view $\mathbf{I}_i$ is determined by jointly maximizing visual similarity subject to this hard topological constraint:

$$b^* = \arg\max_{b \in \mathcal{B}_i} \Big(\text{Sim}_{\text{VL}}(b, T_{\text{pri}}) \cdot \Omega(b, R_{\text{space}}, C_{\text{ref}}) \Big). \tag{2}$$

By enforcing R_{space} relative to the contextual anchor C_{ref}, Ω acts as a deterministic spatial gatekeeper that masks out visually identical distractors. This allows our framework to successfully isolate the unique target and resolve multi-instance ambiguities directly in the 2D domain. After acquiring this rigorously validated bounding box b^*, we utilize it as an initialization prompt for SAM-2 (Ravi et al., 2025) to extract an uncontaminated target mask, completely bypassing the need for unstable multi-view ID voting heuristics.

3.3 ROBUST MASK FILTERING

Unconstrained perspectives inevitably introduce view-dependent mask inconsistencies and occlusion artifacts. Rather than blindly trusting 2D foundation models, we introduce a dual-criteria robust mask filtering (RMF) mechanism as a strict gatekeeper for 3DGS optimization:

Geometric Plausibility: To detect catastrophic tracking failures (e.g., mask explosion or disappearance), we compute the mask area ratio A_i for each reference view $i \in \{1, \dots, V\}$ given the binary mask $\mathbf{M}_i \in [0,1]^{H \times W}$:

$$A_i = \frac{1}{HW} \sum_{\mathbf{p} \in \{1..W\} \times \{1..H\}} \mathbb{I}\left(\mathbf{M}_i(\mathbf{p}) > 0.5\right). \tag{3}$$

Probabilistic Certainty: To eliminate fuzzy boundaries, we integrate the SAM-2 pixel-wise confidence map $\mathbf{C}_i \in [0,1]^{H \times W}$.

By coupling these metrics, a discrete filtering policy $\mathcal{K}_i \in \{0,1\}$ strictly retains only frames ($\mathcal{K}_i = 1$) that satisfy both confidence thresholds and spatial bounds:

$$\frac{1}{\max(|\mathcal{P}_i|, 1)} \sum_{\mathbf{p} \in \mathcal{P}_i} \mathbf{C}_i(\mathbf{p}) > \tau_{\text{conf}} \quad \text{and} \quad A_i \in [\tau_{\min}, \tau_{\max}], \tag{4}$$

where $\mathcal{P}_i = \{\mathbf{p} \mid \mathbf{M}_i(\mathbf{p}) > 0.5\}$ denotes foreground pixels. Restricting the 3DGS training subset to $\mathcal{V}_{\text{train}} = \{v \mid \mathcal{K}_v = 1\}$ aggressively prunes noisy viewpoints, preventing 2D hallucinations from accumulating into permanent 3D projection errors.

3.4 RAY-ALIGNED COVARIANCE PENALTY

Given a reliable mask sequence $\mathcal{V}_{\text{train}}$ we optimize boundary geometry. Since minimizing projection loss often forces boundary Gaussians to erroneously elongate along the viewing ray into aliasing spikes, we introduce explicit structural regularization directly into the covariance parameterization.

The geometry of each Gaussian is parameterized by a *positive semi-definite* (PSD) covariance matrix Σ_i and a mean vector $\boldsymbol{\mu}_i$, and its unnormalized spatial influence (or opacity field) is defined as:

$$\mathcal{G}(\mathbf{x}) = \exp\Big(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^\top \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)\Big). \tag{5}$$

To maintain the PSD property during optimization, Σ_i is explicitly decomposed into a scaling matrix $\mathbf{S}_i$ and a rotation matrix $\mathbf{R}_i$, such that $\Sigma_i = \mathbf{R}_i \mathbf{S}_i \mathbf{S}_i^\top \mathbf{R}_i^\top$.

Standard α-blending lacks depth constraints along the viewing ray. To regularize this, we define the normalized viewing ray $\mathbf{v}_i = \frac{\boldsymbol{\mu}_i - \mathbf{c}}{\|\boldsymbol{\mu}_i - \mathbf{c}\|_2}$ for a Gaussian at $\boldsymbol{\mu}_i$ observed from camera c. To ensure gradient propagation during affine projection, we apply a geometric penalty directly along $\mathbf{v}_i^*$, the Gaussian principal axis most aligned with $\mathbf{v}_i$.

To avoid distortion of interior object volumes, this ray-aligned penalty is selectively activated along object edges by evaluating the spatial gradient of the rendered mask, represented by $\|\nabla \mathbf{M}_{\text{pred}}\|_2$, which is computed via a Sobel operator (Seger, 2012). For the i-th Gaussian primitive, we extract the magnitude of the gradient at its projected screen center $p_i = \pi(\boldsymbol{\mu}_i)$. If the projection resides in a high-gradient transition zone ($\|\nabla \mathbf{M}_{\text{pred}}(p_i)\|_2 > \delta_{\text{edge}}$), we map this magnitude to a continuous dynamic scalar penalty weight ω_i:

$$\omega_i = \epsilon \cdot \Big(1 - \exp\Big(-\frac{\|\nabla \mathbf{M}_{\text{pred}}(p_i)\|_2^2}{2\sigma_{\text{edge}}^2}\Big)\Big), \tag{6}$$

where ϵ modulates the maximum penalty scale and σ_{edge} is a smoothing factor. This dynamic weight governs our geometric regularizer, which is injected directly into the original covariance of the i-th Gaussian to yield the constrained ray-aligned covariance $\widetilde{\Sigma}_i$:

$$\widetilde{\Sigma}_i = \Sigma_i + \omega_i \cdot \mathbf{v}_i^* \mathbf{v}_i^{*\top}. \tag{7}$$

Crucially, the rank-1 PSD penalty term $\mathbf{v}^* \mathbf{v}^{*\top}$ ensures a strict quadratic accumulation across views, monotonically regularizing ray directions without mutual nullification. To mathematically elucidate how this 3D penalty guarantees screen-space inflation and subsequent compression, we must derive its effect through the camera projection. The 2D covariance in an arbitrary view with view matrix $\mathbf{W}_c$ and affine Jacobian $\mathbf{J}_c$ is computed as $\Sigma_{\text{2D}} = \mathbf{J}_c \mathbf{W}_c \Sigma \mathbf{W}_c^\top \mathbf{J}_c^\top$. Substituting our penalized

covariance $\widetilde{\Sigma}_i$ yields:

$$\widetilde{\Sigma}_{\text{2D},c} = \mathbf{J}_c\mathbf{W}_c\left(\Sigma_i + \omega_i\mathbf{v}_i^*\mathbf{v}_i^{*\top}\right)\mathbf{W}_c^\top\mathbf{J}_c^\top = \Sigma_{\text{2D},c} + \omega_i(\mathbf{J}_c\mathbf{W}_c\mathbf{v}_i^*)(\mathbf{J}_c\mathbf{W}_c\mathbf{v}_i^*)^\top. \tag{8}$$

Let $\mathbf{u}_c = \mathbf{J}_c\mathbf{W}_c\mathbf{v}_i^*$ be the 2D projection of the penalized principal axis. In the collinear reference view where $\mathbf{v}_i^*$ aligns perfectly with the viewing ray, $\mathbf{v}_i^*$ resides in the null space of the projection Jacobian. Thus, the penalty remains invisible in the specific view where the elongation originated.

However, when this Gaussian is observed from an adjacent, non-coincident cross-view c', the viewing ray shifts. Consequently, $\mathbf{v}_i^*$ escapes the null space, yielding a non-zero projected vector $\mathbf{u}_{c'}$. The updated 2D covariance in this cross-view becomes:

$$\widetilde{\Sigma}_{\text{2D},c'} = \Sigma_{\text{2D},c'} + \omega_i\mathbf{u}_{c'}\mathbf{u}_{c'}^\top. \tag{9}$$

Because $\omega_i > 0$ and $\mathbf{u}_{c'} \neq \mathbf{0}$, the term $\omega_i\mathbf{u}_{c'}\mathbf{u}_{c'}^\top$ is a positive semi-definite matrix with a strictly positive eigenvalue along the direction of $\mathbf{u}_{c'}$. This explicitly derives the cross-view inflation: the 2D screen footprint strictly expands by a positive-definite increment ($\Delta\Sigma_{\text{2D},c'} \succ 0$).

This forced expansion explicitly causes the 2D Gaussian footprint to spill outside the ground-truth mask boundaries. The segmentation loss $\mathcal{L}_{\text{total}}$ strictly penalizes these false-positive foreground predictions. Because any increase in the 2D footprint directly increases this penalty, the gradient matrix with respect to the 2D covariance, $\mathbf{G}_{\text{2D}} = \frac{\partial\mathcal{L}_{\text{total}}}{\partial\widetilde{\Sigma}_{\text{2D},c'}}$, is guaranteed to be positive-definite.

With this positive-definite gradient established, we apply the chain rule to the scale matrix $\mathbf{S}_i$ to observe the corrective gradient flow (see Appendix for the detailed mathematical derivation):

$$\frac{\partial\mathcal{L}_{\text{total}}}{\partial\mathbf{S}_i} = 2\mathbf{R}_i^\top\left(\mathbf{W}_{c'}^\top\mathbf{J}_{c'}^\top\mathbf{G}_{\text{2D}}\mathbf{J}_{c'}\mathbf{W}_{c'}\right)\mathbf{R}_i\mathbf{S}_i. \tag{10}$$

This dynamic physically squeezes the Gaussian, compressing the elongated scales along the viewing direction. Simultaneously, the footprint inflation breaks the native 3DGS optimization deadlock. Elongated artifacts typically become trapped in local optima because their positional gradients $\|\nabla_{\boldsymbol{\mu}_i}\mathcal{L}_{\text{total}}\|$ are too small to trigger densification. The positional gradient depends on the projected effective area Ω_i:

$$\nabla_{\boldsymbol{\mu}_i}\mathcal{L}_{\text{total}} = \mathbf{W}_{c'}^\top\mathbf{J}_{c'}^\top\sum_{\mathbf{p}\in\Omega_i}\frac{\partial\mathcal{L}_{\text{total}}}{\partial\alpha_{\mathbf{p}}}\nabla_{\boldsymbol{\mu}_{\text{2D},i}}\alpha_{\mathbf{p}}. \tag{11}$$

By inflating Ω_i to explicitly intersect the mask boundary, the derivative $\frac{\partial\mathcal{L}_{\text{total}}}{\partial\alpha_{\mathbf{p}}}$ for spillover pixels ($\mathbf{p} \in \Omega_{\text{out}}$) scales up massively, drastically sensitizing the integral over the boundary:

$$\|\nabla_{\boldsymbol{\mu}_i}\mathcal{L}_{\text{total}}\| \approx \left\|\mathbf{W}_{c'}^\top\mathbf{J}_{c'}^\top\sum_{\mathbf{p}\in\Omega_{\text{out}}}\lambda_1\nabla_{\boldsymbol{\mu}_{\text{2D},i}}\alpha_{\mathbf{p}}\right\| \gg \tau_{\text{pos}}. \tag{12}$$

The artificially inflated footprint thereby sensitizes the positional gradient, encouraging it to surge past the native densification threshold (τ_{pos}) and reactivating the splitting mechanism for the deadlocked primitives. Once split, the new sub-Gaussians possess highly compacted scales ($\mathbf{S}_k \ll \mathbf{S}_i$), bounding their localized projections entirely within the valid object mask ($\Omega_{\text{out}} \to \emptyset$). Consequently, the anomalous gradients diminish ($\|\nabla_{\boldsymbol{\mu}_k}\mathcal{L}_{\text{total}}\| < \tau_{\text{pos}}$), halting the densification cycle and achieving a compact, surface-aligned geometry without requiring additional learnable parameters.

3.5 Loss Functions

The total loss $\mathcal{L}_{\text{total}}$ is formulated as a weighted sum:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{color}} + \lambda_1\mathcal{L}_{\text{mask}} + \lambda_2\mathcal{L}_{\text{reg}}, \tag{13}$$

where λ_1, λ_2 balance the optimization objectives. Specifically, we utilize the standard photometric loss $\mathcal{L}_{\text{color}}$ (Kerbl et al., 2023) for high-fidelity rendering. To mitigate boundary class imbalance, $\mathcal{L}_{\text{mask}}$ amalgamates *binary cross-entropy* (BCE) (Li et al., 2024b) and Dice loss (Li et al., 2020):

$$\mathcal{L}_{\text{mask}} = \text{BCE}(\mathbf{M}_{\text{pred}}, \mathbf{M}_{\text{gt}}) + \text{Dice}(\mathbf{M}_{\text{pred}}, \mathbf{M}_{\text{gt}}), \tag{14}$$

where $\mathbf{M}_{\text{pred}}$ and $\mathbf{M}_{\text{gt}}$ are the differentiable mask rendered from 3D Gaussian semantic logits via α-blending and the SAM-2 pseudo-ground-truth, respectively. Finally, we employ a spatial regularization term $\mathcal{L}_{\text{reg}}$ to enforce smoothness among K-nearest neighbors:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N_{\text{g}}}\sum_i\sum_{j\in\text{KNN}(i)}\exp\Big(-\frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2}{2\sigma_s^2}\Big)\|m_i - m_j\|_2^2, \tag{15}$$

where N_g is the total Gaussian count, σ_s controls the distance sensitivity, and m_i (m_j) denote the learnable semantic mask logits of the i-th (j-th) 3D Gaussian primitives.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets. We evaluated our framework on three datasets: NVOS (Jambon et al., 2023) (eight scenes with reference graffiti and 2D target masks), LERF-OVS (Kerr et al., 2023) (three scenes with manual annotations for interactive 3D segmentation), and ScanNet (Dai et al., 2017) (posed RGB sequences with reconstructed point clouds and 3D semantic labels). Across all datasets, we adopt two categories of text prompts for quantitative evaluation:

- *Explicit noun prompts*: Utilizing the standard explicit object names provided by the original dataset annotations to evaluate direct semantic mapping.
- *Implicit reasoning prompts*: We rephrase all explicit queries using Qwen2.5-VL (Bai et al., 2023) and manual refinement, covering spatial relation and sense reasoning to rigorously evaluate the semantic-geometric decoupling capability of our framework. The complete prompt set, rewriting instructions, and relation-wise performance breakdowns are detailed in the Supplementary Material.

Baselines. To rigorously evaluate our framework, we compare against two primary 3D segmentation methods: feedforward and mask-guided architectures. Feedforward methods (e.g., LSeg (Li et al., 2022), LangSplat (Qin et al., 2024), LangSurf (Li et al., 2024a), LSM (Fan et al., 2024), and LanScene-X (Liu et al., 2025)) rely on rapid end-to-end feature distillation, but frequently compromise fine-grained geometric fidelity. The mask-guided methods include optimization-based approaches driven by two-dimensional priors, ranging from standard mask-projection models (SA3D-GS (Cen et al., 2025b), SAGA (Cen et al., 2025a), FlashSplat (Shen et al., 2024), COB-GS (Zhang et al., 2025)) to recent state-of-the-art language-reasoning and advanced segmentation frameworks (GS-Group (Ye et al., 2024), Gaga (Lyu et al., 2024), GAGS (Peng et al., 2026), NG-GS (He et al., 2026), and REALM (Shi et al., 2026)).

Implementation details. Quantitative experiments are conducted on an NVIDIA RTX 3090 with PyTorch, focusing on metrics including mIoU, mAcc, and *boundary mIoU* (B-mIoU). The 3DGS optimizers employ Adam, with an initial learning rate of $1.6e^{-4}$.

Evaluation Protocol. Unless otherwise specified, all reported metrics (mIoU, mAcc, B-mIoU) are evaluated using 2D rendered masks on held-out novel views. This protocol rigorously tests 3D geometric consistency, ensuring that methods cannot artificially inflate scores by merely overfitting to the training viewpoints. For B-mIoU, we utilize a standard 3-pixel boundary width tolerance.

4.2 QUANTITATIVE COMPARISONS

We comprehensively evaluate our framework against state-of-the-art baselines across the NVOS, LERF-OVS, and ScanNet datasets (Tables 1–5). To clarify the scope of our evaluation, Tables 1–3 report performance strictly under implicit reasoning prompts, serving as a rigorous stress test for semantic-geometric decoupling capability, while Table 5 demonstrates that our method remains highly competitive even under explicit noun prompts, achieving near state-of-the-art volumetric performance on ScanNet and superior boundary precision throughout.

While recent reasoning baselines (e.g., REALM) often compromise geometric fidelity, our synergy of LLM reasoning and ray-aligned constraints eliminates cross-instance interference to achieve unprecedented boundary precision. Compared to the leading baseline (REALM), our method yields significant gains across all datasets: 4.2%–8.0% in B-mIoU and 2.5%–3.8% in mIoU. These results confirm that our approach effectively mitigates depth projection artifacts, ensuring accurate implicit target grounding while preserving strict 3D structural integrity.

4.3 QUALITATIVE COMPARISONS

Baselines relying on direct semantic similarity are often susceptible to geometric ambiguities during 2D localization and 3D projection. As shown in Figure 3, our framework mitigates these limitations through geometric constraints:

Table 1: Performance (%) under **implicit** reasoning on NVOS.

Method	B-mIoU ↑	mAcc ↑	mIoU ↑
SA3D-GS (2023)	66.5	92.1	81.4
SAGA (2024)	67.2	92.4	81.9
FlashSplat (2024)	69.5	93.5	83.2
Gaga (2024)	58.2	85.4	68.5
GAGS (2024)	45.1	55.2	52.4
GS-Group (2024)	55.4	83.1	66.2
COB-GS (2025)	73.4	94.8	85.1
NG-GS (2026)	77.6	96.2	87.4
REALM (2026)	82.6	97.6	88.2
Ours	**89.1**	**99.8**	**90.7**

Table 2: Performance (%) under **implicit** reasoning on ScanNet.

Method	B-mIoU ↑	mAcc ↑	mIoU ↑
Gaga (2024)	30.5	55.2	38.4
GAGS (2024)	22.4	38.6	25.2
GS-Group (2024)	28.6	52.1	35.4
COB-GS (2025)	44.5	72.1	52.4
NG-GS (2026)	48.5	74.6	54.8
REALM (2026)	51.2	76.5	56.8
Ours	**55.4**	**79.2**	**60.1**

Table 3: Performance (%) under **implicit** reasoning on LERF-OVS.

Method	B-mIoU ↑	mAcc ↑	mIoU ↑
LSeg (2022)	21.5	51.2	28.4
LangSplat (2024)	17.2	34.1	20.5
LangSurf (2024)	18.5	37.6	23.2
LSM (2024)	19.8	45.2	26.1
LanScene-X (2024)	42.1	68.5	49.6
SA3D-GS (2023)	56.4	75.1	68.5
SAGA (2024)	57.2	75.4	69.1
FlashSplat (2024)	58.6	76.8	70.4
Gaga (2024)	38.5	58.6	41.2
GAGS (2024)	30.2	42.3	32.4
GS-Group (2024)	36.4	55.1	38.9
COB-GS (2025)	64.5	78.4	73.8
NG-GS (2026)	73.8	83.2	79.6
REALM (2026)	74.3	88.4	80.2
Ours	**82.3**	**99.5**	**84.0**

Table 4: Time consumption on LERF-OVS.

Method	Train (m) ↓	Inf (s) ↓
LSeg (2022)	-	1.8
LangSplat (2024)	-	1.5
LangSurf (2024)	-	1.3
LSM (2024)	-	1.4
LanScene-X (2024)	-	1.2
SA3D-GS (2023)	17	1.4
SAGA (2024)	14	1.2
FlashSplat (2024)	15	1.1
Gaga (2024)	15	1.2
GAGS (2024)	16	1.3
GS-Group (2024)	14	1.1
COB-GS (2025)	13	1.0
NG-GS (2026)	13	1.0
REALM (2026)	12	0.9
Ours	**11**	**0.9**

Table 5: Performance (%) under **explicit noun prompts**.

Dataset	Method	B-mIoU	mAcc	mIoU
NVOS	NG-GS	84.7	99.2	92.6
	REALM	86.3	99.4	92.8
	Ours	**90.2**	**99.8**	**93.1**
LERF-OVS	NG-GS	72.8	85.5	82.9
	REALM	76.5	86.6	84.6
	Ours	**83.5**	**89.6**	**85.2**
ScanNet	NG-GS	59.6	84.1	64.3
	REALM	60.8	82.2	**64.6**
	Ours	**62.1**	**87.8**	64.0

Table 6: Ablation study of different components.

Scene	Components			Performance		
	LDRE	RMF	RACP	B-mIoU ↑	Acc ↑	mIoU ↑
NVOS	✓	✓	✓	**89.1**	99.8	**90.7**
	✓	✓	-	85.3	**99.9**	86.7
	✓	-	✓	87.2	99.1	88.4
	-	✓	✓	71.5	96.2	78.3
	-	-	-	53.3	94.7	74.7
LERF-OVS	✓	✓	✓	**82.3**	**99.5**	**84.0**
	✓	✓	-	80.2	99.5	82.0
	✓	-	✓	79.4	99.2	82.6
	-	✓	✓	72.1	99.1	74.5
	-	-	-	70.4	99.1	72.8

2D Semantic-Topological Grounding. Lacking spatial awareness, traditional methods suffer from multi-instance ambiguity. For example, given the query "The drink held the teddy" (Figure 3, top), baselines misidentify targets or produce fragmented masks, whereas our approach elegantly isolates the exact target. We achieve this by extracting relational tuples to enforce strict topological constraints. Validating spatial relationships prior to 3D lifting accurately anchors semantics to physical bounding boxes.

3D Ray-Aligned Structural Regularization. Unconstrained ray directions projection forces boundary Gaussians to erroneously elongate along the viewing ray, causing structural collapse in novel views. As seen with the "The toy hooking on the bookshelf" query (Figure 3, bottom), competing baselines exhibit stretching spike artifacts. Our explicit covariance penalizes this elongation; this explicit penalty triggers targeted Gaussian splitting and pruning, thereby yielding sharply delineated boundaries.

Our method leverages accurate geometric spatial grounding for high-fidelity scene editing, outperforming REALM (refer to the Appendix).

4.4 Computational Efficiency Analysis

Table 4 compares computational efficiency in LERF-OVS, where all runtime metrics are benchmarked on identical hardware for fairness. Although feedforward models completely bypass the

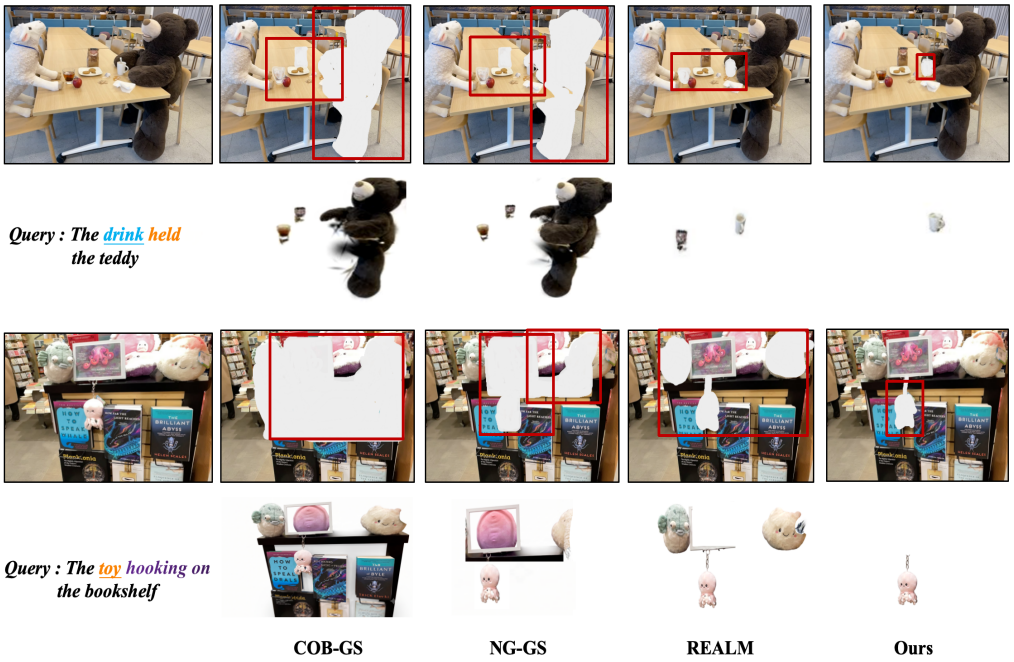

Figure 3: Qualitative Results on LERF and LERF-OVS.

per-scene training phase for rapid deployment, this shortcut inevitably compromises segmentation accuracy. Our method accelerates training convergence by dynamically pruning redundant and noisy viewpoints through robust mask filtering. Similarly, instead of heavy auxiliary boundary refinement networks, our geometric covariance regularization maintains the native high-speed inference of 3DGS.

4.5 Ablation Study

Table 6 presents the ablation results on NVOS and LERF-OVS. Separating LDRE and RMF reveals their distinct contributions: omitting LDRE causes the most severe degradation (e.g., a 17.6 percentage-point B-mIoU drop on NVOS), as the model fails to resolve multi-instance ambiguities. While the combination of LDRE and RMF establishes a strong baseline by filtering 2D semantic noise (achieving 80.2% B-mIoU on LERF-OVS and surpassing REALM), the RACP module provides the critical 3D geometric refinement. By explicitly suppressing unconstrained boundary bleeding in novel views, RACP contributes the final performance margin (increasing B-mIoU to 82.3% on LERF-OVS and 89.1% on NVOS).

5 Conclusion

In this paper, we present GR-GS, a novel text-guided 3DGS segmentation framework that bridges semantic reasoning with explicit geometric regularization. By integrating an *LLM-driven relational reasoning* module, our framework successfully decodes complex implicit spatial instructions that traditional baselines struggle to parse. Crucially, we introduced a *ray-aligned covariance penalty* module to mathematically penalize the unconstrained elongation of Gaussians along the viewing ray. This explicit geometric regularization effectively suppresses view-dependent aliasing spikes and ensures exceptional boundary precision. Comprehensive evaluations confirm that GR-GS outperforms existing methods in both reasoning-based target extraction and boundary geometric fidelity.

Limitations and Future Work. Although our framework effectively mitigates boundary artifacts, it still faces challenges in modeling highly complex dynamic scenes and handling extreme occlusions. Addressing these specific limitations, our future work will focus on extending GR-GS to general dynamic scenes and optimizing the reasoning-to-rendering pipeline for highly interactive applications.

Ethics Statement

This research focuses on the segmentation of explicitly defined objects in open-vocabulary 3D environments. We do not use human subjects in this study and no personally identifiable information or proprietary spatial data has been released. The models and methodologies presented are designed to improve spatial reasoning and 3D perception with potential applications in robotics and automated systems. We encourage responsible deployment of this technology, being mindful of privacy and ethical considerations when operating these sensors and representations in real-world environments.

Reproducibility Statement

To ensure reproducibility, we have explicitly laid out the derivations for the positional and scale gradients induced by our regularizer in the methodology section, while further specifics are deferred to the supplementary material. All datasets used (NVOS, LERF-OVS, ScanNet) are publicly available standard benchmarks. Furthermore, we commit to releasing our source code, including pre-trained weights, training scripts, and custom CUDA kernels for the ray-aligned covariance penalty, upon acceptance of the paper.

AI Use Statement

In this work, we used generative AI tools, specifically Qwen2.5-VL and ChatGPT, to assist in refining and editing the implicit reasoning prompts used for evaluation datasets to ensure complex semantic variety, as well as to assist with the language and grammar editing of the manuscript text. We have reviewed all AI-assisted work. All scientific claims, models, architectural derivations, and analyses are the original work of the authors. We take responsibility for the final content of this work.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Juergen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9296–9306. IEEE, 2019.

Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3d gaussians. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 1971–1979, 2025a.

Jiazhong Cen, Jiemin Fang, Zanwei Zhou, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment anything in 3d with radiance fields. *International Journal of Computer Vision (IJCV)*, 133(8):5138–5160, 2025b.

Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhongang Cai, Lei Yang, Huaping Liu, and Guosheng Lin. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21476–21485, 2024.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5828–5839, 2017.

Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, et al. Large spatial model: End-to-end unposed images to semantic 3d. *Conference on Neural Information Processing Systems (NeurIPS)*, 37:40212–40229, 2024.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028. IEEE, 2024.

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

Clément Jambon, Bernhard Kerbl, Georgios Kopanas, Stavros Diolatzis, Thomas Leimkühler, and George Drettakis. Nerfshop: Interactive editing of neural radiance fields. *Proceedings of the ACM on Computer Graphics and Interactive Techniques (ACM CGIT)*, 6(1), 2023.

Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*, 2023.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):139–1, 2023.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19729–19739, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8500–8509, 2022.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022.

Hao Li, Minghan Qin, Zhengyu Zou, Diqi He, Xinhao Ji, Bohan Li, Bingquan Dai, Dingewn Zhang, and Junwei Han. Langsurf: Language-embedded surface gaussians for 3d scene understanding. *arXiv preprint arXiv:2412.17635*, 2024a.

Qiufu Li, Xi Jia, Jiancan Zhou, Linlin Shen, and Jinming Duan. Rediscovering bce loss for uniform classification. *arXiv preprint arXiv:2403.07289*, 2024b.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 465–476, 2020.

Fangfu Liu, Hao Li, Jiawei Chi, Hanyang Wang, Minghui Yang, Fudong Wang, and Yueqi Duan. Langscene-x: Reconstruct generalizable 3d language-embedded scenes with trimap video diffusion. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 29010–29020, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, pp. 38–55. Springer, 2024.

Weijie Lyu, Xueting Li, Abhijit Kundu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Gaga: Group any gaussians via 3d-aware memory bank. *arXiv preprint arXiv:2404.07977*, 2024.

Yuning Peng, Haiping Wang, Yuan Liu, Chenglu Wen, Zhen Dong, and Bisheng Yang. Gags: Granularity-aware feature distillation for language gaussian splatting. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 8376–8384, 2026.

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

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 28085–28128, 2025.

Barbara Roessle, Norman Müller, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, Angela Dai, and Matthias Nießner. L3dg: Latent 3d gaussian diffusion. In *ACM SIGGRAPH*, pp. 1–11, 2024.

Olle Seger. Generalized and separable sobel operators. *Machine vision for three-dimensional scenes*, pp. 347, 2012.

Qiuhong Shen, Xingyi Yang, and Xinchao Wang. Flashsplat: 2d to 3d gaussian splatting segmentation solved optimally. In *European Conference on Computer Vision (ECCV)*, pp. 456–472. Springer, 2024.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16779–16788, 2026.

Junjie Wang, Jiemin Fang, Xiaopeng Zhang, Lingxi Xie, and Qi Tian. Gaussianeditor: Editing 3d gaussians delicately with text instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20902–20911, 2024.

Boshen Xu, Yuting Mei, Sipeng Zheng, Qin Jin, et al. Egodtm: Towards 3d-aware egocentric video-language pretraining. *Conference on Neural Information Processing Systems (NeurIPS)*, 38:122549–122576, 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. Springer, 2024.

Hongjia Zhai, Hai Li, Zhenzhe Li, Xiaokun Pan, Yijia He, and Guofeng Zhang. Panogs: Gaussian-based panoptic segmentation for 3d open vocabulary scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14114–14124, 2025.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19335–19344, 2025.

Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21676–21685, 2024.

Xingxing Zuo, Pouya Samangouei, Yunwen Zhou, Yan Di, and Mingyang Li. Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding. *International Journal of Computer Vision (IJCV)*, 133(2):611–627, 2025.