MulGeCa: ANCHORING SCATTERED SEMANTIC REPRESENTATIONS TO PHYSICAL ENTITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent open-vocabulary segmentation methods leverage pretrained vision-language models (VLMs) to recognize arbitrary text queries beyond the constraints of closed-set categories. Despite their promise, adapting image-level vision-language representations to dense pixel-level prediction remains fundamentally challenging. Semantic representations from VLMs are often scattered and weakly structured, making it difficult to bind open-vocabulary concepts to their corresponding physical entities. To solve this problem, our key insight is that category-agnostic depth cues provide spatial constraints that can bind semantic responses to depth-consistent physical support. To this end, we propose MulGeCa, which leverages estimated relative depth to hierarchically calibrate scattered VLM semantics at the scene, region, and instance levels. Specifically, MulGeCa incorporates a global-geometry text embedding calibrator to inject dynamic geometric context into text prompts, a geometric-contextual anchor aggregator to modulate attention mechanisms with geometric-consistency biases, and a centroid-purified mask refiner to suppress boundary noise via centripetal weighted aggregation. Extensive experiments on challenging benchmarks demonstrate that MulGeCa achieves superior performance, with particular improvements in cluttered scenarios. Notably, our method exhibits strong robustness in complex scenes, providing an effective pathway for anchoring scattered semantic representations to depth-consistent regions.

1 INTRODUCTION

Open-vocabulary segmentation (OVS) aims to recognize and segment arbitrary objects beyond the constraints of closed-set categories (Liang et al., 2023; Xu et al., 2023b; Shin et al., 2024). Vision-language models, or VLMs, such as the contrastive language-image pre-training model CLIP (Radford et al., 2021) and the large-scale image and noisy-text embedding model ALIGN (Jia et al., 2021), have significantly advanced this field through large-scale image–text pre-training. Consequently, current OVS research (Xu et al., 2023c; Jiang et al., 2024) primarily focuses on how to leverage the remarkable zero-shot generalizability of VLMs for dense pixel-level prediction (Zhou et al., 2022a; Jiao et al., 2023).

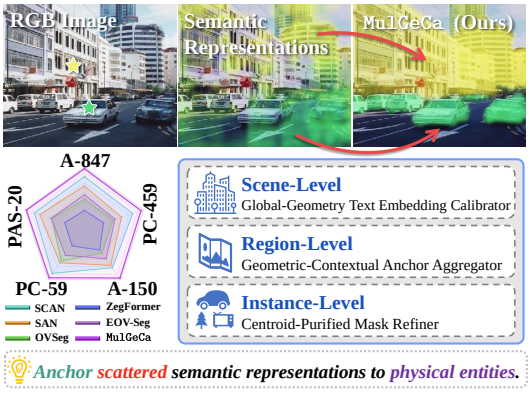

Figure 1: Overview, motivation, and performance comparison of MulGeCa.

Despite its promise, adapting VLMs to dense pixel-level prediction introduces a fundamental representational bottleneck. Semantic features from the VLM tend to be scattered and unstructured across the entire image, hindering precise localization (Vivanco Cepeda et al., 2023). This issue arises primarily from the image-level contrastive supervision used during VLM training, where the vision encoder captures global aligned context rather than local pixel-level details (Yu et al., 2023b; Jiang et al., 2023). To alleviate these issues, recent works (Yu et al., 2023a; Jiao et al., 2024; Ding et al., 2022; Li et al., 2025b; Wang et al., 2024c) introduce mask-proposal pipelines to reformulate the OVS

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.

task as region-level classification and matching. However, despite architectural advancements, efforts to bind semantic representations to physical entities have yet to yield significant results (Gao et al., 2025; Das et al., 2024). A common manifestation of this problem is the "Bag-of-Words" (Zhang et al., 2010; Yuksekgonul et al., 2022; Koishigarina et al., 2025) behavior: while the model can reliably detect high-level concepts, *e.g.*, "red" or "car", it struggles to bind these semantic representations to the corresponding physical entities. This issue becomes especially problematic in cluttered scenes with ambiguous boundaries or severe occlusions, further leading to classification hallucinations (Bai et al., 2024; Liu et al., 2024d). Additionally, directly fine-tuning the model to enforce such connections risks overfitting and catastrophic forgetting of the model's zero-shot generalizability (Zhai et al., 2024; Luo et al., 2025).

In this paper, we revisit this problem from a fundamentally different perspective: semantic concepts are not free-floating visual symbols, but should be anchored to physical entities with stable geometric support. In the physical world, objects typically exhibit surface continuity within their extent and geometric discontinuities at boundaries, which are largely independent of categories, textures, and colors (Xu et al., 2026). These properties make geometric priors a natural scaffold for anchoring open-vocabulary semantics (Xiao et al., 2017; Hou et al., 2019).

For image-based OVS, an effective geometric prior should capture physical structure while remaining compatible with dense 2D prediction. Among different geometric representations, depth maps provide a particularly suitable interface: they encode dense, pixel-aligned cues such as surface continuity and occlusion boundaries, while remaining directly compatible with masks, feature maps, and visual encoders (Chen et al., 2025). Therefore, depth offers a simple yet effective geometric prior for anchoring semantic representations to physical entities.

Despite this potential, integrating geometric priors into OVS remains largely unexplored, mainly due to the severe modality gap (Peng et al., 2023). In contrast to high-dimensional and abstract semantic features (Vesdapunt et al., 2024), geometric priors provide explicit but low-dimensional geometric constraints (Li et al., 2024). Consequently, indiscriminate concatenation or addition of geometric and contextual features is not only suboptimal for cross-modal alignment (Guo et al., 2026), but also disrupts the pre-trained semantic distribution, thereby compromising its zero-shot generalizability.

In light of this, we propose `MulGeCa`, a **Mul**ti-level **Ge**ometric **Ca**libration framework. We hierarchically reconstruct the model's geometric awareness at three levels: ❶ **Scene-level**: We introduce the global-geometry text embedding calibrator (G^2TEC). By injecting dynamic global geometric contexts into static prompts, it generates conditioned text embeddings with enhanced global scene perception. ❷ **Region-level**: We design the geometric-contextual anchor aggregator (GCA^2), which leverages a frozen vision foundation model (VFM), a geometric-consistency bias generator, and a feature interactor to enhance geometric consistency and locally aggregate geometric-contextual anchors. ❸ **Instance-level**: We present the centroid-purified mask refiner (CPMR), which replaces conventional equal-weight mask pooling with a geometric-guided centripetal weighted aggregation to suppress edge noise, thereby refining the mask queries.

In summary, we reveal the problem of "semantics anchoring failure" in current OVS models and propose `MulGeCa`, a geometry-aware framework that anchors semantic representations to physical entities at the scene, region, and instance levels, as illustrated in Fig. 1. Extensive experiments across multiple challenging OVS benchmarks (Zhou et al., 2019; Mottaghi et al., 2014; Everingham et al., 2010) demonstrate superior performance, with particular gains in complex occlusion and cluttered scenarios. We hope our study will inspire future research on anchoring semantic representations to physical entities for open-vocabulary segmentation.

2 Related Work

2.1 Open-Vocabulary Segmentation

Existing OVS methods (Wysoczańska et al., 2024; Lan et al., 2024; Hajimiri et al., 2025) can be broadly grouped into two categories: per-pixel dense matching and mask classification. Per-pixel methods, such as CAT-Seg (Cho et al., 2024) and SED (Xie et al., 2024), construct dense image-text or cost volumes and refine them via aggregation and decoding. In contrast, mask-classification approaches (Liang et al., 2023; Xu et al., 2023c; Han et al., 2023) first generate class-agnostic region proposals and then perform region–text alignment. Dense correspondence-based methods (Cho et al., 2024; Xie et al., 2024) facilitate global similarity modeling and long-range interactions, yet typically

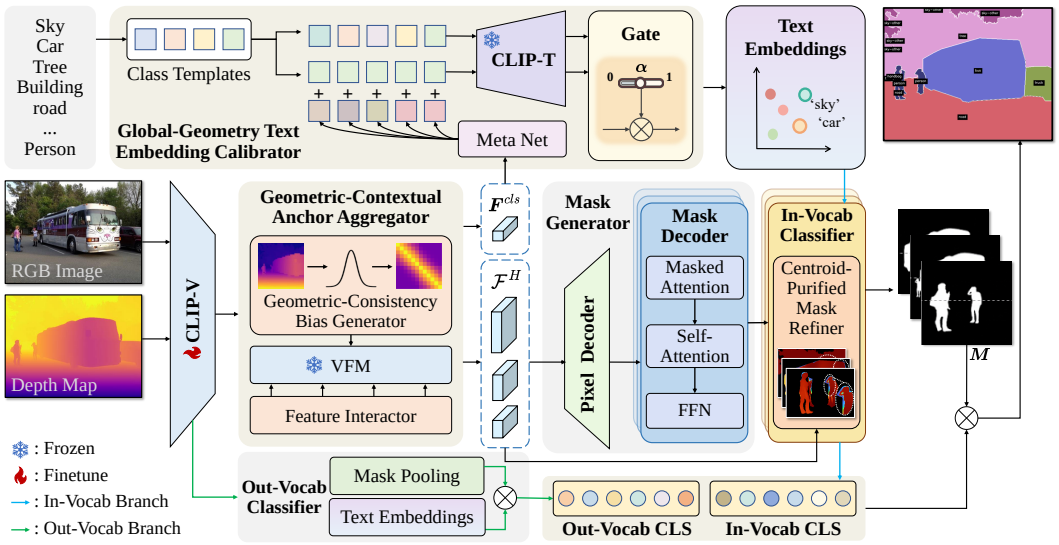

Figure 2: An illustration of our proposed `MulGeCa`. Given an RGB image, a depth map, and class labels, `MulGeCa` extracts and aggregates semantic features through vision-text interaction, grounding them in real-world physical entities.

introduce greater computational and architectural complexity. In contrast, mask-classification methods are more compatible with CLIP-based region–text alignment, but remain sensitive to the distribution gap between masked region features and CLIP's global image representations (Jiao et al., 2023; Li et al., 2025b). To alleviate this gap, we propose a novel framework that integrates geometric priors to anchor scattered semantic representations to physical entities.

2.2 Prompt Learning

To efficiently adapt pre-trained VLMs to downstream tasks, prompt learning has emerged as a more efficient alternative to full fine-tuning. Early static prompt learning methods, such as CoOp (Zhou et al., 2022b) and VPT (Jia et al., 2022), learn task-specific prompt tokens, enabling lightweight adaptation without updating the entire model. Subsequently, CoCoOp (Zhou et al., 2022c) introduces conditional prompt learning, generating instance-specific prompts conditioned on global image features. More recently, these dynamic prompting strategies have been extended to dense prediction tasks (Zhu et al., 2023; Liu et al., 2022; Zhao et al., 2025b), further guiding segmentation decoders through instance-aware dual-prompting mechanisms. Despite the efficacy of dynamic prompting, existing methods primarily rely on 2D appearance cues, which often fail to capture accurate physical boundaries in complex scenarios with severe occlusions or color ambiguities (Bai et al., 2024; Liu et al., 2024d). To bridge this gap, we incorporate scene-level geometric priors into text embedding generation, enabling physically consistent context aggregation.

3 Methodology

3.1 Overall Workflow

To anchor scattered semantic representations to physical entities, we propose `MulGeCa`, a multi-level geometric calibration framework. As illustrated in Fig. 2, our framework adopts a dual-branch design to bridge the modality gap across three levels. The vision branch comprises a CLIP-based vision encoder (CLIP-V), a GCA^2 module, and a mask generator. The text branch includes a G^{2}TEC module, along with in-vocabulary and out-of-vocabulary classifiers.

First, the RGB image $\boldsymbol{I}^R \in \mathbb{R}^{3\times H\times W}$ and the corresponding depth map $\boldsymbol{I}^D \in \mathbb{R}^{H\times W}$ are processed by a pre-trained CLIP-V to generate a hierarchical feature pyramid $\mathcal{F}^C$. Subsequently, the GCA^2 module is employed to aggregate a C-dimensional <CLS> feature $\boldsymbol{F}^{cls}$ and a heterogeneous feature pyramid $\mathcal{F}^H$. These multi-scale features are then fed into a mask generator based on Mask2Former (Cheng et al., 2022) to generate class-agnostic mask proposals $\boldsymbol{M}$. In parallel, a set of class names $\mathcal{C} = \{c_1, \ldots, c_K\}$ is mapped to natural language prompts $\boldsymbol{P} = \{\boldsymbol{P}_1, \ldots, \boldsymbol{P}_K\}$ and encoded into

text embeddings $\tilde{\boldsymbol{T}}$ via the G^{2}TEC module, where K denotes the number of classes. Once mask proposals are obtained, they are categorized with text embeddings $\tilde{\boldsymbol{T}}$ in a contrastive manner via the in-vocabulary classifier. Specifically, conventional equal-weight mask pooling is replaced by the CPMR module to suppress boundary noise. During inference, the out-vocab classifier performs mask pooling on CLIP-V features directly to derive object embeddings, enabling generalization to unseen classes. Ultimately, we aggregate the classification scores from both the in-vocab and out-vocab classifiers, and integrate them with mask proposals $\boldsymbol{M}$ to generate the final segmentation result.

The G^{2}TEC, GCA^2, and CPMR modules enforce geometric consistency while preserving the open-vocabulary generalizability. In the following sections, we detail each module and describe how they cooperate to enable robust geometry-aware representation learning.

3.2 Global-Geometry Text Embedding Calibrator

Standard VLMs typically rely on static, predefined templates that treat semantic concepts as abstract visual symbols, overlooking the concrete physical layout of the scene (Radford et al., 2021; Yuksekgonul et al., 2022). Although several prompt learning techniques (Zhou et al., 2022c) aim to generate instance-specific context, they rely solely on 2D RGB cues. In cluttered scenes, these methods tend to misinterpret local semantic features as independent objects, thereby generating misleading contextual information. To address this issue, the G^{2}TEC module injects scene-level geometric priors into the textual semantic space, producing text embeddings conditioned on the current geometric scene rather than fixed prompts or RGB-only cues.

First, we extract the <CLS> feature $\boldsymbol{F}^{cls} \in \mathbb{R}^{C\times\frac{H}{32}\times\frac{W}{32}}$ from the GCA^2 module and further generate a global geometric prior $\boldsymbol{G} \in \mathbb{R}^C$. In the text branch, each class label $c_i \in \mathcal{C}$ is prompted with n natural-language templates, forming the prompt token set $\boldsymbol{P}_i = \{\boldsymbol{p}_{i,1}, \ldots, \boldsymbol{p}_{i,n}\}$. Subsequently, we dynamically calibrate the text prompts by defining a learnable context based on static prompts $\boldsymbol{P}$. We employ a lightweight meta-Net $\mathcal{M}(\cdot)$ to map $\boldsymbol{G}$ into a scene-specific prompt shift $\Delta\mathcal{P} = \mathcal{M}(\boldsymbol{G})$, yielding dynamic prompt token $\mathcal{P} = \mathcal{P}_0 + \Delta\mathcal{P}$. For each class label c_i, we construct the k-th prompt sequence $\boldsymbol{s}_{i,k}$ by concatenating the prompt token $\boldsymbol{p}_{i,k}$ with the class name embedding $\boldsymbol{e}_i$, bounded by start-of-sequence (SOS) and end-of-sequence (EOS) tokens. These sequences are then encoded by the CLIP Text Encoder (CLIP-T) to obtain dynamic text embeddings $\boldsymbol{t}_{i,k} \in \mathbb{R}^C$ as follows:

$$\boldsymbol{t}_{i,k} = \text{CLIP-T}\left(\left[\boldsymbol{e}_{\texttt{SOS}}, \boldsymbol{p}_{i,k}, \boldsymbol{e}_i, \boldsymbol{e}_{\texttt{EOS}}\right]\right). \tag{1}$$

Directly replacing the original text embeddings with $\boldsymbol{t}_{i,k}$ may disrupt CLIP's pre-aligned semantic space, leading to catastrophic forgetting and degraded zero-shot generalizability (Jia et al., 2022; Zhou et al., 2022c). To mitigate this issue, we introduce a gating mechanism that adaptively fuses the dynamic geometric context with the original text embeddings. Let $\boldsymbol{t}^0_{i,k}$ denote the static semantic text embeddings for class c_i, derived from standard prompt templates, *e.g.*, *"a photo of a {c_i}"*. The calibrated text embedding $\tilde{\boldsymbol{t}}_{i,k}$ is computed via an adaptive gating mechanism:

$$\tilde{\boldsymbol{t}}_{i,k} = \text{Norm}\left((1-\boldsymbol{\alpha}) \odot \boldsymbol{t}^0_{i,k} + \boldsymbol{\alpha} \odot \boldsymbol{t}_{i,k}\right), \tag{2}$$

where $\boldsymbol{j} \in \mathbb{R}^C$ is a learnable channel-wise gating vector, $\boldsymbol{\alpha} = \text{Sigmoid}(\boldsymbol{j}) \in \mathbb{R}^C$ gives the per-channel fusion weights in the range $(0, 1)$. $\text{Sigmoid}(\cdot)$ denotes the Sigmoid function, $\odot$ is element-wise multiplication, and $\text{Norm}(\cdot)$ denotes a normalization operation. After appending the void class, the calibrated text embeddings $\tilde{\boldsymbol{T}}$ are further expanded to $\mathbb{R}^{(N+1)\times C}$ for downstream dense matching in the classifier, where N denotes the total number of prompts across all classes.

3.3 Geometric-Contextual Anchor Aggregator

While G^{2}TEC improves scene-level perception through global geometric priors, region-level visual modeling remains limited without explicit geometric constraints (Wang et al., 2024a; Tao et al., 2026). Such a limitation often results in feature bleeding (Saxena, 2025; Khosla et al., 2025) or semantic leakage (Chen et al., 2026; Gonen et al., 2025) across occlusion boundaries and between visually similar yet distinct objects. To tackle these challenges, we introduce the GCA^2 module, which enhances geometric consistency without updating the frozen VFM. Specifically, it first modulates the VFM's self-attention mechanism via geometric-consistency bias and then reinforces geometric constraints through heterogeneous feature interaction.

Geometric-Consistency Bias Generator. As shown in Fig. 3, to construct geometric-consistency bias, we first flatten the depth map $\boldsymbol{I}^D$ into a depth sequence $\boldsymbol{D} \in \mathbb{R}^{\frac{HW}{16\times16}}$. We append a summary token $\boldsymbol{d}$ encoding the mean scene depth to the depth sequence, yielding the augmented geometric sequence $\hat{\boldsymbol{D}} = [\boldsymbol{D}; \boldsymbol{d}] \in \mathbb{R}^{(\frac{HW}{16\times16}+1)}$. We further construct a pairwise depth-relation matrix $\boldsymbol{\Delta} \in \mathbb{R}^{(\frac{HW}{16\times16}+1)\times(\frac{HW}{16\times16}+1)}$, by measuring absolute depth differences between augmented geometric sequence: $\Delta_{m,n} = \left|\hat{d}_m - \hat{d}_n\right|$, where $\Delta_{m,n}$ denotes the pairwise depth difference between the m-th and n-th tokens, and $\hat{d}_m$ is the corresponding value of the m-th token in $\hat{\boldsymbol{D}}$. Subsequently, we convert the pairwise depth differences into geometric-consistency bias using a Gaussian kernel. The affinity matrix $\boldsymbol{A}$ and the bias $\boldsymbol{B}$ are computed as:

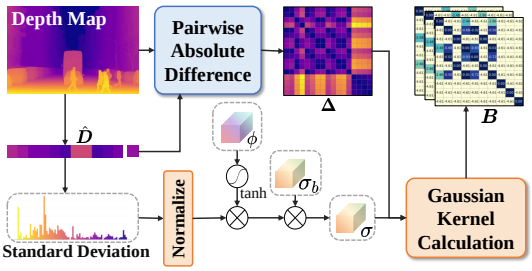

Figure 3: Details of our proposed geometric-consistency bias generator.

$$\boldsymbol{A} = \exp\left(-\frac{\Delta^2}{2\sigma^2}\right), \tag{3}$$

$$\boldsymbol{B} = \mathrm{clamp}\big(\log\left(\mathrm{clamp}(\boldsymbol{A}, 0.05, 1)\right), -8, 0\big), \tag{4}$$

where clamp operation bounds $\boldsymbol{A}$ and $\boldsymbol{B}$ within $[0.05, 1]$ and $[-8, 0]$, respectively, to stabilize the logarithm and prevent the bias from destabilizing or dominating attention. To account for regionally specific geometric variations, we introduce an adjustment to a learnable parameter σ_b, defining the Gaussian bandwidth σ as $\sigma = \sigma_b \cdot (1 + w \cdot \tanh(\phi) \cdot (\mathrm{Norm}(\mathrm{Std}(\hat{\boldsymbol{D}})) - 1))$, where learnable parameter ϕ modulates the sensitivity to the geometric sequence, and w is a fixed scaling factor.

Formally, let $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{n_h\times(\frac{HW}{16^2}+1)\times d_k}$ denote the query, key, and value embeddings in the VFM attention layers, where n_h is the number of attention heads and d_k is the per-head feature dimension. We incorporate the geometric-consistency bias $\boldsymbol{B}$ into the self-attention computation $\mathrm{Attn}(\cdot)$ as:

$$\mathrm{Attn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}} + \boldsymbol{B}\right)\boldsymbol{V}. \tag{5}$$

As a plug-and-play module, this formulation explicitly injects geometric-consistency constraints into the attention process and can be readily integrated into existing Transformer architectures, thereby improving the robustness of geometric semantic alignment.

Heterogeneous Feature Interactor. While the geometric-consistency bias constrains VFM attention, its effectiveness is limited when the backbone features lack explicit geometric context. Therefore, we further introduce a heterogeneous feature interactor to inject multi-scale spatial priors into the VFM backbone. We first flatten and concatenate the hierarchical CLIP feature pyramid $\mathcal{F}^C = \{\boldsymbol{F}^C_{\frac{1}{4}}, \boldsymbol{F}^C_{\frac{1}{8}}, \boldsymbol{F}^C_{\frac{1}{16}}\}$ across scales, forming a unified spatial token sequence $\boldsymbol{T} \in \mathbb{R}^{N_H\times C}$, where $N_H = \frac{HW}{4^2} + \frac{HW}{8^2} + \frac{HW}{16^2}$. In the i-th interaction block, we use the VFM token sequence $\boldsymbol{F}^i_{\mathrm{vfm}} \in \mathbb{R}^{\frac{HW}{16^2}\times C}$ as queries and the unified spatial token sequence $\boldsymbol{T}$ as keys and values. The VFM token sequence in the i-th interaction block is updated as:

$$\boldsymbol{F}^{i+1}_{\mathrm{vfm}} = \boldsymbol{F}^i_{\mathrm{vfm}} + \boldsymbol{\gamma}^i \odot \mathrm{Attn}\left(\mathrm{Norm}\left(\boldsymbol{F}^i_{\mathrm{vfm}}\right), \mathrm{Norm}\left(\boldsymbol{T}\right), \mathrm{Norm}\left(\boldsymbol{T}\right)\right), \tag{6}$$

where $\boldsymbol{\gamma}^i \in \mathbb{R}^C$ is a learnable channel-wise scaling vector. The interactor and the bias play complementary roles, jointly enhancing region-level geometric consistency and robustness. Finally, GCA^2 fuses the spatial token sequence $\boldsymbol{T}$ with the VFM tokens $\boldsymbol{F}^i_{\mathrm{vfm}}$ to produce a <CLS> representation $\boldsymbol{F}^{cls}$ and a heterogeneous feature pyramid $\mathcal{F}^H$ for downstream mask generation.

3.4 Centroid-Purified Mask Refiner

As shown in Fig. 4, mask embeddings are typically extracted using mask cropping (Ding et al., 2022; Liang et al., 2023; Xu et al., 2022) or mask pooling (Ghiasi et al., 2022; Jiao et al., 2024;

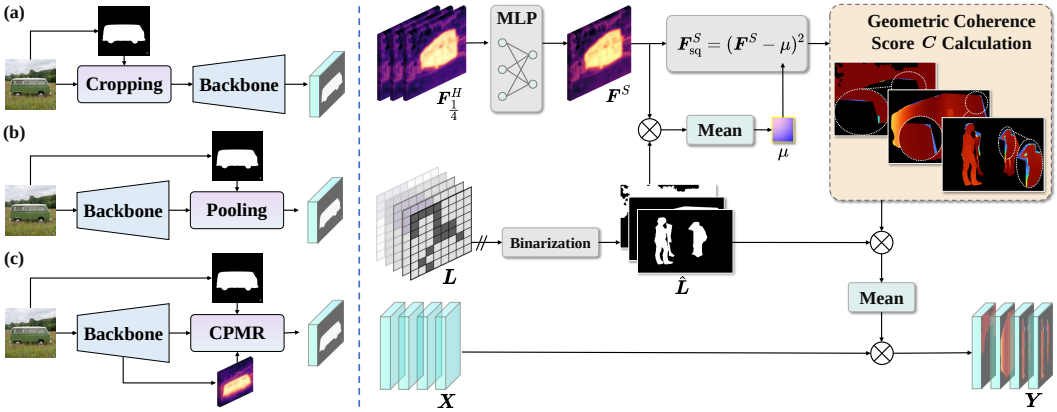

Figure 4: Comparison of mask embedding extraction methods and details of centroid-purified mask refiner. (a) Mask Cropping: directly feeding image crops into the vision backbone. (b) Mask Pooling: aggregating contextual features within the predicted mask regions. (c) CPMR: deriving purified mask embeddings by filtering out geometrically and semantically inconsistent regions.

Shan et al., 2024; Xu et al., 2025; Yu et al., 2023a). Although mask pooling is widely adopted for its efficiency, it is susceptible to outlier noise from imperfect masks (Li et al., 2025b). To address this limitation, we propose the CPMR module, which refines mask embeddings through centripetal weighted aggregation.

Raw depth map provides limited semantic cues, whereas mask classification requires instance-level geometric–semantic alignment. To bridge this gap, the CPMR module leverages the high-level heterogeneous features $\boldsymbol{F}^H_{\frac{1}{4}}$ from the GCA^2 module to guide centroid-purified mask aggregation. First, a lightweight MLP projects $\boldsymbol{F}^H_{\frac{1}{4}}$ into a single-channel response map $\boldsymbol{F}^S \in \mathbb{R}^{1\times\frac{H}{4}\times\frac{W}{4}}$. Let $\boldsymbol{L} \in \mathbb{R}^{Q\times\frac{H}{4}\times\frac{W}{4}}$ denote the predicted mask logits, where Q is the number of mask queries. Then, we detach and binarize it to obtain the coarse query masks $\hat{\boldsymbol{L}}$. We further compute the mean of the single-channel response map $\boldsymbol{F}^S$ within each rough query mask, obtaining the instance-level feature-response centroid μ. Subsequently, we derive a squared deviation map $\boldsymbol{F}^S_{\text{sq}} = (\boldsymbol{F}^S - \mu)^2$, which encodes the response deviation of each location from the instance centroid. The squared deviation map is then transformed into a non-negative coherence score $\boldsymbol{C}$ using a Gaussian kernel:

$$\boldsymbol{C} = \exp\left(-\frac{\boldsymbol{F}^S_{\text{sq}}}{2\sigma_p^2}\right), \tag{7}$$

where σ_p is the Gaussian bandwidth. This formulation assigns higher weights to pixels geometrically closer to the instance centroid and suppresses centroid-inconsistent regions. The final aggregation weight $\boldsymbol{W}$ is computed by masking the score $\boldsymbol{C}$ and normalizing it over the spatial domain:

$$\boldsymbol{W} = \frac{\hat{\boldsymbol{L}} \odot \boldsymbol{C}}{\sum_{h,w}\left(\hat{\boldsymbol{L}} \odot \boldsymbol{C}\right)}, \tag{8}$$

where $\sum_{h,w}$ denotes summation over the spatial dimensions. This step preserves the geometrically consistent core of each mask while suppressing boundary noise and outliers introduced by imperfect segmentation. The final pooled context feature $\boldsymbol{Y}$ is obtained by $\boldsymbol{Y} = \boldsymbol{W} \odot \boldsymbol{X}$, where the mask embeddings $\boldsymbol{X} \in \mathbb{R}^{C\times\frac{H}{4}\times\frac{W}{4}}$ comes from the mask generator. By replacing standard mask pooling, the lightweight plug-and-play CPMR module can be seamlessly integrated into existing architectures (Shan et al., 2024; Xu et al., 2025; Yu et al., 2023a) to anchor scattered instance-level mask embeddings to physical entities.

4 Experiments

4.1 Experimental Setting

Datasets. Following standard zero-shot evaluation protocols in open-vocabulary segmentation (Ghiasi et al., 2022; Liang et al., 2023; Yu et al., 2023a), we train our model on the COCO-Stuff

Table 1: Comparisons with SoTA methods on various datasets. "LN" denotes the Localized Narratives dataset (Pont-Tuset et al., 2020). The best results are highlighted in **bold**.

Model	VLM	Training Dataset	A-847	PC-459	A-150	PC-59	PAS-20
LSeg (Li et al., 2022)	ViT-B	PASCAL VOC-15	-	-	-	-	47.4
LSeg+ (Ghiasi et al., 2022)	ALIGN	COCO-Stuff	2.5	5.2	13.0	36.0	-
ZegFormer (Ding et al., 2022)	ViT-B	COCO-Stuff-156	4.9	9.1	16.9	42.8	86.2
ZegFormer† (Ding et al., 2022)	ViT-B	COCO-Stuff-156	5.6	10.4	18.0	45.5	89.5
OpenSeg (Ghiasi et al., 2022)	ALIGN	COCO Panoptic+LN	4.4	7.9	17.5	40.1	-
OVSeg (Liang et al., 2023)	ViT-B	COCO-Stuff+COCO Captions	7.1	11.0	24.8	53.3	92.6
ZegCLIP (Zhou et al., 2023)	ViT-B	COCO-Stuff-156	-	-	-	41.2	93.6
SAN (Xu et al., 2023c)	ViT-B	COCO-Stuff	10.1	12.6	27.5	53.8	94.0
FC-CLIP (Yu et al., 2023a)	ConvNeXt-L	COCO Panoptic	-	-	-	-	-
MAFT+ (Jiao et al., 2024)	ConvNeXt-B	COCO-Stuff	13.2	14.2	33.6	55.9	93.9
EOV-Seg (S) (Niu et al., 2025)	ResNet50	COCO Panoptic	6.6	11.5	21.9	46.0	87.2
EOV-Seg (M) (Niu et al., 2025)	ResNet50x4	COCO Panoptic	7.8	12.2	25.5	51.8	91.2
ELSE-Net (Zhao et al., 2025a)	ViT-B	COCO-Stuff	10.3	-	27.9	54.3	95.1
`MulGeCa` (Ours)	ConvNeXt-B	COCO-Stuff	**13.5**	**19.0**	**35.3**	**58.9**	**95.9**
LSeg (Li et al., 2022)	ViT-L	PASCAL VOC-15	-	-	-	-	52.3
OpenSeg (Ghiasi et al., 2022)	ALIGN	COCO Panoptic + LN	8.1	11.5	26.4	44.8	-
OVSeg (Liang et al., 2023)	ViT-L	COCO-Stuff + COCO Captions	9.0	12.4	29.6	55.7	94.5
ODISE (Xu et al., 2023a)	ViT-L	COCO-Stuff	11.1	14.5	29.9	57.3	-
FC-CLIP (Yu et al., 2023a)	ConvNeXt-L	COCO Panoptic	14.8	18.2	34.1	58.4	95.4
SAN (Xu et al., 2023c)	ViT-L	COCO-Stuff	12.4	15.7	32.1	57.7	94.6
MAFT+ (Jiao et al., 2024)	ConvNeXt-L	COCO-Stuff	15.1	**21.6**	36.1	59.4	**96.5**
EOV-Seg (L) (Niu et al., 2025)	ConvNeXt-L	COCO Panoptic	12.8	16.8	32.1	56.9	94.8
`MulGeCa` (Ours)	ConvNeXt-L	COCO-Stuff	**15.5**	21.2	**36.2**	**59.6**	95.4

(Caesar et al., 2018) dataset, which provides 118K images annotated with 171 classes. We then perform zero-shot evaluation on standard benchmarks: ADE20K (Zhou et al., 2019) with the A-847 and A-150 splits, PASCAL-Context (Mottaghi et al., 2014) with the PC-459 and PC-59 splits, and PASCAL VOC (Everingham et al., 2010) with the PAS-20 split. We employ mean Intersection over Union, or mIoU, to evaluate OVS performance.

Implementation Details. Our framework extends the FC-CLIP architecture (Yu et al., 2023a), using MAFT+ (Jiao et al., 2024) as the baseline model. Following prior open-vocabulary segmentation settings (Yu et al., 2023a; Jiao et al., 2024), we train the model for 80K iterations with the AdamW (Loshchilov et al., 2017) optimizer, using a learning rate of 1×10^{-4} and a weight decay of 2×10^{-5}. Additional implementation details regarding the batch size, learning rate scheduler, and data augmentations are provided in the technical appendix.

4.2 Comparisons with State-of-the-art Methods

As presented in Table 1, we compare our proposed `MulGeCa` with a representative set of state-of-the-art (SoTA) methods in terms of open-vocabulary segmentation performance. For a fair comparison, we categorize methods by VLM backbone scale. It can be observed that `MulGeCa` achieves SoTA performance across the majority of datasets with both base and large VLM backbones. Notably, `MulGeCa` achieves the strongest results on the A-150 (Zhou et al., 2019), A-847 (Zhou et al., 2019), and PC-59 (Mottaghi et al., 2014) datasets under both Base and Large configurations, further demonstrating the effectiveness of our method. Nevertheless, despite its significant gains on ADE20K, its improvement on PC-459 exhibits different trends across backbone scales. Specifically, while `MulGeCa` improves mIoU by 4.8% over the baseline with the base backbone, it experiences a performance regression with the large-scale VLM. We attribute this to the fact that the PC-459 dataset contains fine-grained labels belonging to the same geometric entity, *e.g.*, various components of a building, that often lack clear physical boundaries. For a strong VFM, imposing rigid geometry from noisy geometric priors may suppress its ability to distinguish such categories. These challenges

Table 2: Component-wise ablation study across different datasets.

G²TEC	GCA²	CPMR	A-847	A-150	PAS-20
			13.2	33.6	93.9
✓			13.3	34.9	95.4
	✓		13.2	34.8	95.0
		✓	13.4	35.0	95.5
✓	✓		13.3	35.0	95.5
✓		✓	**13.5**	35.1	95.7
	✓	✓	13.4	35.0	95.6
✓	✓	✓	**13.5**	**35.3**	**95.9**

Table 3: Ablation of G²TEC module design and inputs.

G²TEC Design	Input	A-150
Frozen CLIP, static context	–	31.6
+ $\mathcal{M}(\cdot)$, dynamic context	–	30.2
+ $\mathcal{M}(\cdot) \rightarrow$ dynamic Bias	RGB	33.4
+ $\mathcal{M}(\cdot) \rightarrow$ dynamic Bias	RGB-D	34.7
+ $\mathcal{M}(\cdot) \rightarrow$ dynamic token	RGB	34.1
+ $\mathcal{M}(\cdot) \rightarrow$ **dynamic token**	**RGB-D**	**34.9**

Table 4: Comparison among various design of GCA² module on the A-150 (Zhou et al., 2019) dataset.

GCA² Architecture	A-150	A-847
W/O Geometric-Consistency Bias	33.6	13.2
frozen VFM	33.1	12.9
frozen VFM + Bias	34.2	13.4
frozen VFM + Interactor	34.0	13.4
frozen VFM + Bias + Interactor	**34.8**	**13.5**

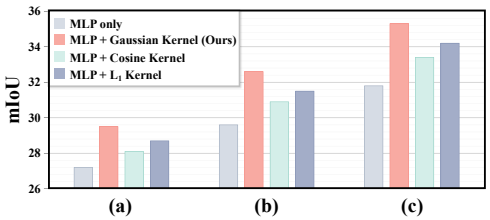

Figure 5: Ablation study of the CPMR module across different designs and geometric feature inputs: (a) raw relative depth map, (b) middle-level geometric features from GCA² module, and (c) high-level geometric-contextual features from the mask generator.

motivate future work on improving the adaptability of geometric anchoring in massive semantic spaces with high category density.

4.3 Ablation Study

Component-wise ablations As presented in Table 2, we first conduct comprehensive ablation experiments across different datasets to evaluate the individual and collective effectiveness of G²TEC, GCA², and the CPMR module within `MulGeCa`. Compared with the baseline model, the integration of any individual component consistently improves performance, with each component contributing an increase of over 1.2% in mIoU on the A-150 (Zhou et al., 2019) dataset. The optimal result is achieved when all three components are jointly employed. Notably, the exclusion of any proposed module results in a noticeable performance degradation, indicating that each component contributes complementary benefits.

Effect of Global-Geometry Text Embedding Calibrator. To investigate the most effective design of the G²TEC module, we evaluate several structural variants and feature inputs. As shown in Table 3, a naive MLP without auxiliary features decreases the mIoU of the static-context baseline from 31.6% to 30.2%, indicating that unconstrained prompt learning may disturb the original text embedding space. In contrast, generating dynamic tokens from auxiliary features consistently outperforms the variant that predicts only a dynamic bias, with detailed architectural comparisons provided in the technical appendix. This improvement is mainly because the bias-based design bypasses CLIP-T and directly modifies the output embeddings, whereas our token-based design preserves the original text encoder pathway, leading to more stable cross-modal alignment. Moreover, using RGB-D features in the G²TEC module improves mIoU by 0.8-1.3% over using RGB context features alone. This further supports our hypothesis in Sect. 1 that global geometric priors can bridge scattered semantic representations and mask proposals, thereby anchoring them to physical entities.

Effect of Geometric-Contextual Anchor Aggregator. We further investigate the impact of different designs of the GCA² module. As shown in Table 4, directly using frozen VFM leads to worse performance than the baseline model, because the frozen VFM cannot effectively enhance the fine-grained perception without proper task-specific adaptation. As anticipated in Sect. 3.3, equipping the frozen VFM with both the interactor and geometric-consistency bias leads to accurate segmentation results, highlighting the necessity of both geometric context injection and geometric-consistency bias. Further ablations on the GCA² designs are provided in the technical appendix.

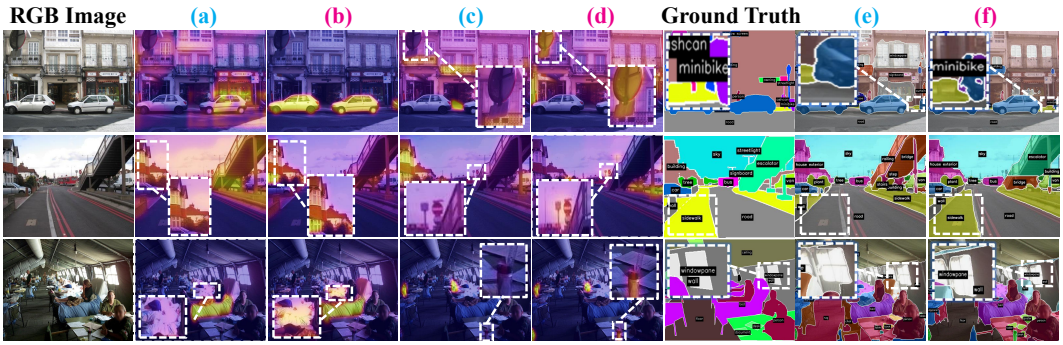

Figure 7: Qualitative comparisons of prediction probabilities and segmentation results. (a), (c), and (e) are the prediction probabilities and final prediction results of the baseline model. (b), (d), and (f) are the corresponding outputs of `MulGeCa`.

Effect of Centroid-Purified Mask Refiner. As shown in Fig. 5, we evaluate the CPMR module across various kernel designs and inputs, with different kernel formulations provided in the technical appendix. Notably, the MLP + Gaussian kernel design demonstrates superior performance. This indicates that a simple MLP alone may be insufficient for modeling complex spatial relationships, whereas the Gaussian kernel further suppresses noisy responses and refines the predicted masks. Moreover, performance steadily improves as the input representation evolves from raw depth maps to semantically enriched features, suggesting that these high-level features provide richer context and geometric cues for centroid purification.

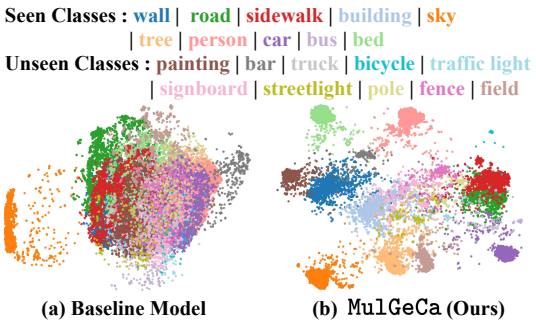

Figure 6: t-SNE embeddings of the feature space from: (a) baseline model, and (b) `MulGeCa`. Each point is colored by its semantic class.

4.4 Qualitative Study

Qualitative Comparisons. As shown in Figs. 7(a) and (c), the baseline model frequently suffers from scattered feature activations, assigning high probabilities to irrelevant backgrounds or failing to separate from neighboring objects. In contrast, `MulGeCa` produces more localized and accurate probability maps, as illustrated in Figs. 7(b) and (d). Additionally, the baseline model exhibits over-segmented or hallucinated prediction results in cluttered scenes, whereas `MulGeCa` produces sharper object boundaries and better segmentation results (see Figs. 7(e) and (f)). This further highlights the effectiveness of `MulGeCa` to anchor scattered semantic representations to physical entities. More comparisons across indoor and outdoor scenes are provided in the technical appendix.

t-SNE Visualization of Dense Embeddings. The t-SNE visualizations of feature embeddings produced by the baseline model and our proposed `MulGeCa` on the A-150 (Zhou et al., 2019) dataset are shown in Fig. 6. The baseline model produces highly entangled feature embeddings, failing to form distinct decision boundaries. In contrast, `MulGeCa` yields well-separated semantic clusters across both seen and unseen categories.

5 Conclusion and Future Work

In this paper, we proposed `MulGeCa`, a multi-level geometric calibration framework that addresses the "Bag-of-Words" behavior in OVS by anchoring scattered semantic representations to physical entities at the scene, region, and instance levels through the G^{2}TEC, GCA^2, and CPMR modules. Extensive experiments validated the effectiveness of our novel contributions and the superior performance of `MulGeCa`. One limitation is that the framework relies on the quality of geometric priors, which may be unreliable under extreme lighting or transparent surfaces. Future work will explore jointly learning geometric and semantic representations and extending `MulGeCa` to video open-vocabulary segmentation with temporal geometric consistency.

AI USE STATEMENT

Generative AI tools were used only to assist with limited language editing and LaTeX formatting. They were not involved in formulating the research problem, developing the proposed method, designing or conducting the experiments, or interpreting the results. Every AI-assisted revision was checked and, where necessary, corrected by the authors. The authors made all final decisions and assume full responsibility for the manuscript, code, claims, and reported results.

REPRODUCIBILITY STATEMENT

We support reproducibility by documenting the model architecture, datasets, training and evaluation protocols, implementation details, and hyperparameter settings in the main paper and technical appendix. The appendix further reports controlled ablations, computational-cost measurements, robustness analyses, and qualitative results to facilitate independent verification and direct comparison. The complete source code and experiment configurations will be released upon publication.

REFERENCES

Zechen Bai et al. Hallucination of multimodal large language models: A survey. *Computing Research Repository (CoRR)*, 2024. URL https://arxiv.org/abs/2404.18930.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

Alexey Bochkovskiy et al. Depth Pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations (ICLR)*, 2025.

Maxime Bucher et al. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Holger Caesar et al. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

Siyu Chen et al. Leveraging depth and language for open-vocabulary domain-generalized semantic segmentation. *Computing Research Repository (CoRR)*, abs/2506.09881, 2025. URL https://arxiv.org/abs/2506.09881.

Yiyi Chen et al. Semantic leakage from image embeddings. *Computing Research Repository (CoRR)*, abs/2601.22929, 2026. URL https://arxiv.org/abs/2601.22929.

Bowen Cheng et al. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 1290–1299, 2022.

Seokju Cho et al. CAT-Seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4113–4123, 2024.

Anurag Das et al. MTA-CLIP: Language-guided semantic segmentation with mask-text alignment. In *European Conference on Computer Vision (ECCV)*, pp. 39–56. Springer, 2024.

Jian Ding et al. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11583–11592, 2022.

Mark Everingham et al. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88:303–338, 2010.

Guangyu Gao et al. PRFormer: Matching proposal and reference masks by semantic and spatial similarity for few-shot semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Golnaz Ghiasi et al. Scaling open-vocabulary image segmentation with image-level labels. In *European conference on computer vision (ECCV)*, pp. 540–557. Springer, 2022.

Hila Gonen et al. Does liking yellow imply driving a school bus? semantic leakage in language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 785–798, 2025.

Sicen Guo et al. Fully exploiting vision foundation model's profound prior knowledge for generalizable RGB-Depth driving scene parsing. *IEEE Transactions on Intelligent Vehicles*, 2026.

Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3099–3108, 2025.

Cong Han et al. Open-vocabulary semantic segmentation with decoupled one-pass network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1086–1096, 2023.

Ji Hou et al. 3D-SIS: 3D semantic instance segmentation of RGB-D scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4421–4430, 2019.

Chao Jia et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, pp. 4904–4916. PMLR, 2021.

Menglin Jia et al. Visual prompt tuning. In *European conference on computer vision (ECCV)*, pp. 709–727. Springer, 2022.

Dongsheng Jiang et al. From CLIP to DINO: Visual encoders shout in multi-modal large language models. *Computing Research Repository (CoRR)*, abs/2310.08825, 2023. URL https://arxiv.org/abs/2310.08825.

Li Jiang et al. Open-vocabulary 3D semantic segmentation with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21284–21294, 2024.

Siyu Jiao et al. Learning mask-aware CLIP representations for zero-shot segmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:35631–35653, 2023.

Siyu Jiao et al. Collaborative vision-text representation optimizing for open-vocabulary segmentation. In *European conference on computer vision (ECCV)*, pp. 399–416. Springer, 2024.

Savya Khosla et al. Relocate: A simple training-free baseline for visual query localization using region-based representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 3697–3706, 2025.

Darina Koishigarina et al. CLIP behaves like a bag-of-words model cross-modally but not unimodally. *Computing Research Repository (CoRR)*, 2025. URL https://arxiv.org/abs/2502.03566.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ProxyCLIP: Proxy attention improves CLIP for open-vocabulary segmentation. In *European conference on computer vision (ECCV)*, pp. 70–88. Springer, 2024.

Boyi Li et al. Language-driven semantic segmentation. *Computing Research Repository (CoRR)*, 2022. URL https://arxiv.org/abs/2201.03546.

Fan Li, Xuanbin Wang, Xuan Wang, Zhaoxiang Zhang, and Yuelei Xu. Images as noisy labels: Unleashing the potential of the diffusion model for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025a.

Peng Li et al. Geodc: Geometry-constrained depth completion with depth distribution modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

Yongkang Li et al. Mask-Adapter: The devil is in the masks for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14998–15008, 2025b.

Feng Liang et al. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 7061–7070, 2023.

Jinfeng Liu et al. Mono-ViFI: A unified learning framework for self-supervised single and multi-frame monocular depth estimation. In *European Conference on Computer Vision (ECCV)*, pp. 90–107, 2024a.

Lingbo Liu et al. Prompt-matched semantic segmentation. *Computing Research Repository (CoRR)*, 2022. URL https://arxiv.org/abs/2208.10159.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024b.

Yong Liu et al. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3491–3500, 2024c.

Yufang Liu et al. Investigating and mitigating object hallucinations in pretrained vision-language (CLIP) models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 18288–18301, 2024d.

Ilya Loshchilov et al. Decoupled weight decay regularization. *Computing Research Repository (CoRR)*, abs/1711.05101, 2017. URL https://arxiv.org/abs/1711.05101.

Yun Luo et al. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Roozbeh Mottaghi et al. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 891–898, 2014.

Hongwei Niu et al. Eov-seg: Efficient open-vocabulary panoptic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 6254–6262, 2025.

Maxime Oquab et al. DINOv2: Learning Robust Visual Features without Supervision. *Computing Research Repository (CoRR)*, abs/2304.07193, 2023. URL https://arxiv.org/abs/2304.07193.

Yuwen Pan et al. Purify then guide: A bi-directional bridge network for open-vocabulary semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(1):343–356, 2025.

Songyou Peng et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–824, 2023.

Jordi Pont-Tuset et al. Connecting vision and language with Localized Narratives. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 647–664, 2020.

Alec Radford et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PmLR, 2021.

Pranav Saxena. Gen-LangSplat: Generalized language gaussian splatting with pre-trained feature compression. *Computing Research Repository (CoRR)*, abs/2510.22930, 2025. URL https://arxiv.org/abs/2510.22930.

Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient RGB-D semantic segmentation for indoor scene analysis. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

Xiangheng Shan et al. Open-vocabulary semantic segmentation with image embedding balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28412–28421, 2024.

Heeseong Shin et al. Towards open-vocabulary semantic segmentation without semantic labels. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:9153–9177, 2024.

Libo Sun et al. SC-DepthV3: Robust self-supervised monocular depth estimation for dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):497–508, 2024.

Xujing Tao et al. GeoGuide: Hierarchical geometric guidance for open-vocabulary 3D semantic segmentation. *Computing Research Repository (CoRR)*, abs/2603.26260, 2026. URL https://arxiv.org/abs/2603.26260.

Noranart Vesdapunt et al. Hvclip: High-dimensional vector in clip for unsupervised domain adaptation. In *European Conference on Computer Vision (ECCV)*, pp. 36–54. Springer, 2024.

Vicente Vivanco Cepeda et al. GeoCLIP: CLIP-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:8690–8701, 2023.

Pengfei Wang et al. Open vocabulary 3D scene understanding via geometry guided self-distillation. In *European Conference on Computer Vision (ECCV)*, pp. 442–460. Springer, 2024a.

Xuehui Wang et al. Opmapper: Enhancing open-vocabulary semantic segmentation with multi-guidance information. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Youhong Wang et al. SQLdepth: Generalizable self-supervised fine-structured monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 5713–5721, 2024b.

Ziyi Wang et al. XMask3D: Cross-modal mask reasoning for open vocabulary 3d semantic segmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:74922–74944, 2024c.

Yuxin Wu et al. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

Monika Wysoczańska et al. CLIP-DINOiser: Teaching CLIP a few DINO tricks for open-vocabulary semantic segmentation. In *European conference on computer vision (ECCV)*, pp. 320–337. Springer, 2024.

Yongqin Xian et al. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8256–8265, 2019.

Jingjing Xiao et al. Robust fusion of color and depth data for RGB-D target tracking using adaptive range-invariant depth models and spatio-temporal consistency constraints. *IEEE transactions on cybernetics*, 48(8):2485–2499, 2017.

Bin Xie et al. SED: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3426–3436, 2024.

Jiarui Xu et al. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2955–2966, 2023a.

Jilan Xu et al. Learning open-vocabulary semantic segmentation models from natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2935–2944, 2023b.

Lintao Xu et al. Occlusion boundary and depth: Mutual enhancement via multi-task learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4427–4437, 2026.

Mengde Xu et al. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European conference on computer vision (ECCV)*, pp. 736–753. Springer, 2022.

Mengde Xu et al. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2945–2954, 2023c.

Wenhao Xu et al. Generalization boosted adapter for open-vocabulary segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(1):520–532, 2025. doi: 10.1109/TCSVT.2024.3454227.

Lihe Yang et al. Depth anything v2. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:21875–21911, 2024.

Bowen Yin, Xuying Zhang, Zhongyu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. DFormer: Rethinking RGBD representation learning for semantic segmentation. In *International Conference on Learning Representations (ICLR)*, 2024.

Qihang Yu et al. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:32215–32234, 2023a.

Seonghoon Yu et al. Zero-shot referring image segmentation with global-local context features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19456–19465, 2023b.

Mert Yuksekgonul et al. When and why vision-language models behave like bags-of-words, and what to do about it? *Computing Research Repository (CoRR)*, 2022. URL https://arxiv.org/abs/2210.01936.

Yuexiang Zhai et al. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning*, pp. 202–227. PMLR, 2024.

Dengke Zhang, Fagui Liu, and Quan Tang. CorrCLIP: Reconstructing patch correlations in CLIP for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

Yin Zhang et al. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1(1):43–52, 2010.

Shiyuan Zhao et al. Excluding the impossible for open vocabulary semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 10457–10464, 2025a.

Ziyu Zhao et al. DPSeg: dual-prompt cost volume learning for open-vocabulary semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 25346–25356, 2025b.

Bolei Zhou et al. Semantic understanding of scenes through the ade20k dataset. *International journal of computer vision*, 127(3):302–321, 2019.

Chong Zhou et al. Extract free dense labels from clip. In *European conference on computer vision (ECCV)*, pp. 696–712. Springer, 2022a.

Kaiyang Zhou et al. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, July 2022b. ISSN 1573-1405. doi: 10.1007/s11263-022-01653-1. URL http://dx.doi.org/10.1007/s11263-022-01653-1.

Kaiyang Zhou et al. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 16816–16825, 2022c.

Ziqin Zhou et al. ZegClIP: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11175–11185, 2023.

Muzhi Zhu et al. SegPrompt: Boosting open-world segmentation via category-level prompt learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 999–1008, 2023.

MulGeCa: Anchoring Scattered Semantic Representations to Physical Entities

-Supplementary Material-

Anonymous submission

CONTENTS

A Symbol Definitions **16**

B Discussions **18**
B.1 Detailed Architectural Comparisons of G^{2}TEC Variants 18
B.2 Mathematical Formulations of Centroid-Purified Kernels 18

C Datasets Information **19**

D Additional Experimental Results **19**
D.1 Implementation Details 19
D.2 Computational Cost and End-to-End Efficiency 20
D.3 Depth Sources and Controlled Perturbations 21
D.4 VLM Backbones, VFM Initialization, and Input Resolution 22
D.5 Isolating Geometric Cues from Additional Feature Capacity 23
D.6 Module-Level Sensitivity and Statistical Stability 23
D.7 Architectural Generalization and State-of-the-Art Comparisons 24
D.8 Category-Level Quantitative Analysis 27
D.9 Detailed Results of the Ablation Study on Meta-Net Architecture 28
D.10 Hyperparameter Sensitivity and Optimization Complexity 29
D.11 Grounding-Based Comparison and Generalized Zero-Shot Evaluation 29
D.12 Additional Qualitative Results on Prediction Probabilities 30
D.13 Additional Qualitative Results of Semantic Segmentation 32

E Additional Clarifications **33**

F Broader Applications and Future Directions **34**
F.1 Failure Mechanisms and Scope of the Geometric Assumption 34

In the following technical appendix, we provide a comprehensive extension of the main text to further support the effectiveness and robustness of MulGeCa. First, we present detailed architectural comparisons of different G^{2}TEC variants, illustrating why dynamically generated context tokens better preserve the pre-trained vision-language alignment than dynamic bias fusion. We then provide the mathematical formulations of the centroid-purified kernels used in the CPMR module, including Gaussian, L_1, and cosine kernels, to clarify the design choices behind our geometric purification strategy. In addition, we describe the datasets, implementation details, training settings, evaluation protocol, and compared open-vocabulary segmentation methods to ensure reproducibility and fair comparison. We further include extensive quantitative analyses, such as full state-of-the-art, or SoTA, comparisons across multiple benchmarks, per-category performance discussions, depth-source ablations, and ablation studies on key geometric parameters. Finally, we present additional qualitative results on prediction probabilities and segmentation results in both indoor and outdoor scenarios, demonstrating that MulGeCa effectively suppresses background noise, improves boundary localization, and enhances dense open-vocabulary scene understanding under complex scenarios.

A SYMBOL DEFINITIONS

In this section, Table 5 provides a comprehensive summary of the notations and their corresponding physical or semantic meanings used in the MulGeCa framework.

Table 5: Key notations used throughout the `MulGeCa` framework.

Notations	Descriptions
General and Dimensions	
H, W	Height and width of the input RGB image and depth map.
C	Dimensionality of the joint vision-language embedding space.
K	Total number of target semantic categories in the label set.
Q	Number of mask queries initialized in the mask generator.
N	Total number of text prompt instances across all classes.
Input and Basic Features	
$\boldsymbol{I}^R, \boldsymbol{I}^D$	Paired input RGB image and its corresponding estimated depth map.
$\mathcal{F}^C, \mathcal{F}^H$	Feature pyramids from the CLIP vision encoder and the GCA^2 module.
$\boldsymbol{F}^{cls}$	Global visual feature used for extracting scene-level geometric context.
$\boldsymbol{M}$	Set of class-agnostic mask proposals generated by the mask decoder.
Global-Geometry Text Embedding Calibrator (G^{2}TEC)	
$\mathcal{C}$	Set of class labels used for open-vocabulary recognition.
$\boldsymbol{P}$	Collection of natural language prompt templates.
$\mathcal{P}_0, \Delta\mathcal{P}, \mathcal{P}$	Prompt contexts; $\mathcal{P}_0$ is the static prompt context, $\Delta\mathcal{P}$ is the scene-specific prompt shift, and $\mathcal{P}$ is the calibrated dynamic prompt context.
$\boldsymbol{G}$	Global geometric prior vector derived from the visual/geometric context.
$\mathcal{M}(\cdot)$	Meta-Net used to map geometric priors into dynamic prompt shifts.
$\boldsymbol{p}$	Base prompt token; $\boldsymbol{p}_{i,k}$ denotes the k-th template token for class c_i.
$\boldsymbol{e}$	Class-name embedding; $\boldsymbol{e}_i$ denotes the embedding of class c_i.
$\boldsymbol{s}$	Complete prompt sequence; $\boldsymbol{s}_{i,k}$ denotes the sequence for class c_i under template k.
$\boldsymbol{t}, \boldsymbol{t}^0$	Text embeddings; $\boldsymbol{t}_{i,k}$ denotes the conditioned dynamic embedding for class c_i under template k, and $\boldsymbol{t}^0_{i,k}$ denotes the original static embedding under standard prompts.
$\tilde{\boldsymbol{T}}$	Final calibrated and fused text embeddings for dense contrastive matching.
$\boldsymbol{j}$	Learnable channel-wise gating vector for context fusion.
$\boldsymbol{\alpha}$	Derived channel-wise weights computed as $\mathrm{Sigmoid}(\boldsymbol{j})$.
Geometric-Contextual Anchor Aggregator (GCA^2)	
$\boldsymbol{D}, \hat{\boldsymbol{D}}$	Geometric sequences; $\boldsymbol{D}$ is the flattened depth sequence and $\hat{\boldsymbol{D}}$ is the augmented geometric sequence with the appended summary token.
$\boldsymbol{d}, \hat{d}$	Depth tokens; $\boldsymbol{d}$ is the mean scene-depth summary token and $\hat{d}$ is an individual token value in the augmented geometric sequence used in pairwise depth relations.
$\boldsymbol{\Delta}$	Pairwise depth-relation matrix used to measure surface continuity.
$\boldsymbol{A}, \boldsymbol{B}$	Pairwise geometric affinity matrix and the resulting geometry-aware attention bias.
σ, σ_b	Gaussian bandwidth terms; σ is dynamic and σ_b is the learnable base scaling parameter.
ϕ	Learnable parameter for modulating geometric sensitivity.
w	Fixed scaling factor used for Gaussian bandwidth adaptation.
$\boldsymbol{T}$	Unified multi-scale spatial token sequence for heterogeneous interaction.
$\boldsymbol{\gamma}$	Learnable channel-wise scaling vector used in the feature interactor.
Centroid-Purified Mask Refiner (CPMR)	
$\boldsymbol{L}, \hat{\boldsymbol{L}}$	Mask representations; $\boldsymbol{L}$ is the logit map and $\hat{\boldsymbol{L}}_q$ is the q-th binarized mask.
$\boldsymbol{F}^S, \boldsymbol{F}^S_{\mathrm{sq}}$	Response maps; $\boldsymbol{F}^S$ is for alignment and $\boldsymbol{F}^S_{\mathrm{sq}}$ is the squared deviation map measuring the geometric distance from the centroid.
μ	Instance-level geometric centroid computed for each query; μ_q denotes the centroid for the q-th query.

Continued on the next page

Table 5 continued

Notations	Descriptions
$\boldsymbol{C}$	Geometric coherence score assigned to each pixel within the mask.
σ_p	Specific Gaussian bandwidth for centripetal mask purification.
$\boldsymbol{W}$	Final normalized aggregation weights for mask query refinement.
$\boldsymbol{X}$	Input mask-level visual features before refinement.
$\boldsymbol{Y}$	Final refined and purified context feature embeddings.

B Discussions

B.1 Detailed Architectural Comparisons of G^{2}TEC Variants

In this section, we provide detailed architectural overviews of the Global-Geometry Text Embedding Calibrator, denoted as G^{2}TEC, and its structural variants discussed in the main text. As quantitatively demonstrated in the corresponding ablation table in the main text, integrating auxiliary geometric features via dynamic context tokens significantly and consistently outperforms the dynamic bias addition variant.

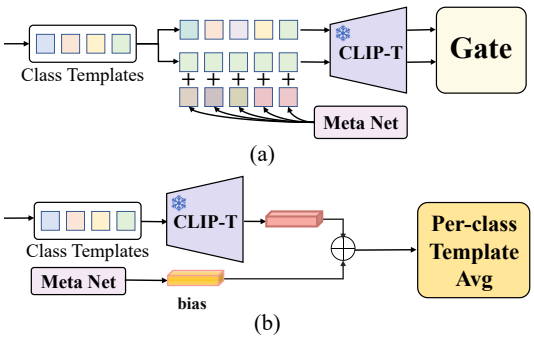

Figure 8: Comparison of G^{2}TEC design variants. (a) utilizing meta-Net to produce dynamic bias, and (b) our proposed design that generates dynamic context tokens to explicitly condition the input prompt sequence.

To comprehensively understand this performance gap, we illustrate the distinct designs of these two prompt learning strategies in Fig. 8. Specifically, Fig. 8(a) depicts the dynamic bias variant, which employs a meta-Net to generate an additive bias vector. This bias is applied to the output representations in a post-hoc manner and bypasses the CLIP Text Encoder, denoted as CLIP-T. Consequently, such unconstrained late-fusion severely disrupts the pre-trained cross-modal distribution, leading to suboptimal feature alignment.

In contrast, Fig. 8(b) showcases our proposed G^{2}TEC module, which dynamically generates context tokens to explicitly condition the input prompt sequence. By injecting these geometric priors directly into the input space, our token-based design allows the G^{2}TEC module to deeply interact with semantic concepts through the self-attention layers of CLIP-T. This strategy effectively preserves the well-aligned zero-shot representation space, yielding superior open-vocabulary segmentation performance.

B.2 Mathematical Formulations of Centroid-Purified Kernels

In the CPMR ablation study in the main text, we evaluate various kernel designs and geometric inputs. This section provides the mathematical definitions of the three kernel designs involved in the ablation experiments.

All kernels are based on the deviation between the single-channel response map $\boldsymbol{F}^S$ and its calculated geometric centroid μ within the rough query mask $\hat{\boldsymbol{L}}$. The mathematical definitions of the three kernel designs are expressed as follows:

Gaussian Kernel, Ours: This is the default configuration proposed in our `MulGeCa`. It utilizes the squared deviation feature $\boldsymbol{F}^S_{\text{sq}} = (\boldsymbol{F}^S - \mu)^2$ to generate smooth, distance-aware weighting:

$$\boldsymbol{C}_{\text{Gaussian}} = \exp\left(-\frac{(\boldsymbol{F}^S - \mu)^2}{2\sigma_p^2}\right), \tag{9}$$

where $\sigma_p > 0$ is the Gaussian bandwidth that controls the softness of the centripetal weighting. The non-linear decay of the Gaussian kernel ensures that pixels closer to the geometric core receive higher weights, effectively suppressing boundary noise.

L_1 Kernel: The L_1 kernel employs absolute distance rather than squared distance. Compared to the Gaussian kernel, it exhibits linear sensitivity to deviations:

$$\boldsymbol{C}_{L_1} = \exp\left(-\frac{|\boldsymbol{F}^S - \mu|}{\sigma_p}\right). \tag{10}$$

Experimental results demonstrate that it is less robust than the Gaussian kernel when handling complex boundaries due to its slower decay rate.

Cosine Kernel: This kernel evaluates the directional alignment between the response values and the centroid. For the single-channel response map $\boldsymbol{F}^S$, it is defined as follows:

$$\boldsymbol{C}_{\text{Cosine}} = \frac{\boldsymbol{F}^S \cdot \mu}{\|\boldsymbol{F}^S\|\|\mu\|}. \tag{11}$$

Notably, for single-channel scalars, this formulation essentially degenerates into a constant value of 1 when both $\boldsymbol{F}^S$ and μ are positive. Unlike the Gaussian or L_1 kernels that leverage absolute spatial distances, the Cosine kernel fails to generate effective spatial weight gradients in a scalar feature space. This mathematical limitation prevents the module from performing meaningful geometric purification, explaining its significantly inferior performance in our ablation studies.

C Datasets Information

We use the COCO-Stuff (Caesar et al., 2018) dataset as the training dataset and evaluate generalizability of our proposed `MulGeCa` in a cross-dataset setting on the ADE20K (Zhou et al., 2019), Pascal-Context (Mottaghi et al., 2014), and Pascal-VOC (Everingham et al., 2010) datasets.

COCO-Stuff 164K (Caesar et al., 2018) is a large-scale dense semantic segmentation dataset with 171 categories, including 80 thing classes and 91 stuff classes. Following standard open-vocabulary protocols, we use its training split as the dataset domain for model training.

ADE20K (Zhou et al., 2019) is a scene parsing benchmark containing over 25K images. We report results on both A-150 which contains 150 classes and A-847 which consists of 847 classes, providing standard and fine-grained open-vocabulary evaluation settings respectively.

Pascal-Context (Mottaghi et al., 2014) is built on the PASCAL benchmark and provides pixel-level annotations at different granularities. We evaluate on PC-59 and PC-459 to test robustness across medium- and large-vocabulary label spaces.

Pascal-VOC (Everingham et al., 2010) is a classic semantic segmentation benchmark with 20 foreground object categories. We use the PAS-20 split to assess generalizability on common object-centric categories.

This multi-benchmark protocol enables a comprehensive evaluation of our geometric-aware representations across different semantic granularities and dataset domains.

D Additional Experimental Results

Unless otherwise stated, the controlled studies fix the RGB input, segmentation network, optimizer, augmentation, and evaluation protocol. Depth-control variants are trained for 10K iterations and evaluated on A-150. The following analyses complement the full-training results in the main paper.

D.1 Implementation Details

Framework Architecture. All experiments are conducted on a single NVIDIA RTX 3090 GPU. Our model is implemented based on the Detectron2 (Wu et al., 2019) library in PyTorch. We utilize

OpenCLIP ConvNeXt-Base and the ViT-B/16 variant of DINOv2 (Oquab et al., 2023) as the base VLM and VFM, respectively. The input depth maps are generated by Depth Anything V2 (Yang et al., 2024), which uses a DINOv2 backbone; this estimator is distinct from the frozen DINOv2 VFM used inside GCA^2. The VFM remains frozen apart from the task-relevant adaptation modules and a small number of learnable scalar parameters.

Data Augmentation. During the training phase on COCO-Stuff, we apply synchronized augmentations to both RGB and depth images. This includes random horizontal flipping with a 0.5 probability, random rescaling with a ratio between 0.1 and 2.0, and color jittering. Finally, the inputs are randomly cropped to a fixed resolution of 384 $\times$ 384 pixels. This extensive augmentation strategy facilitates robust cross-modal alignment in diverse driving and indoor scenes.

Training Strategy. We optimize the model for 80k iterations using the AdamW optimizer with a base learning rate of 1e-4 and a weight decay of 2e-5. We employ a Poly learning rate decay policy with a power factor of 0.9, and the total batch size is set to 4. To accelerate training and reduce VRAM consumption, automatic mixed precision, abbreviated as AMP, is enabled throughout the optimization process.

Evaluation Metrics. We report mean Intersection over Union, abbreviated as mIoU, as the primary metric. For inference, we adopt sliding-window inference at a base scale of 1,024 $\times$ 1,024. All results are obtained using the standard Detectron2 evaluation pipeline to ensure fair comparison and reproducibility.

D.2 Computational Cost and End-to-End Efficiency

To provide a comprehensive evaluation of computational efficiency, we compare **MulGeCa** with representative open-vocabulary segmentation methods in terms of inference throughput, trainable parameters, and FLOPs. All models are evaluated using synthetic inputs on a single NVIDIA RTX PRO 6000 GPU with a batch size of 1 and an input resolution of 384 $\times$ 384 pixels. RGB and estimated depth are represented as separate three-channel tensors. During evaluation, all models are set to evaluation mode with gradient computation disabled. The reported runtime excludes data loading, disk I/O, visualization, and metric evaluation. After warm-up, FPS is averaged over 100 forward passes with CUDA synchronization, and FLOPs are estimated using `fvcore.nn.FlopCountAnalysis`.

As shown in Table 6, different methods exhibit distinct trade-offs between throughput and model complexity. Under this isolated segmentation-core protocol, **MulGeCa** achieves 19.7683 FPS with 207.14M trainable parameters and 193.188 GFLOPs. The additional cost mainly arises from processing RGB-D input and performing geometry-aware cross-modal feature interaction. Complete online measurements that additionally include depth estimation are reported in Sect. D.2.

Table 8 separates the original MAFT+ components from the additional **MulGeCa** components. Under the actual single-text-encoder configuration, the complete model contains 385.56M parameters and 207.14M trainable parameters, excluding the teacher model. The additional trainable parameters mainly arise from the Depth SPM and GCA^2; the deployment cost of GCA^2 is driven by the VFM and interaction between VFM tokens and multi-scale CLIP features.

For a fixed image collection, such as a training set, benchmark, repeated evaluation set, or archival batch, depth can be generated once and cached. This setting measures the segmentation core when RGB-D input is already available. For a previously unseen online image, however, depth estimation is indispensable. We therefore measure the complete online pipeline using Depth Anything V2 Small, Base, and Large, as reported in Table 9. The end-to-end timing includes depth inference and **MulGeCa** segmentation.

The isolated large-estimator measurement used during the initial efficiency audit is 13.38 FPS, 1.87 GB peak inference memory, and 74.58 ms latency. Because it was measured separately from the complete online protocol above, we report it only as an isolated depth-generation measurement and do not combine it arithmetically with the end-to-end table. For online deployment, the Small or Base estimator and a lower segmentation resolution provide more practical latency-memory trade-offs.

Table 6: Model-complexity comparison on an NVIDIA RTX PRO 6000 at 384×384.

Method	Total (M) ↓	Trainable (M) ↓
CAT-Seg [CVPR'24] (Cho et al., 2024)	**154.29**	25.12
EBSeg [CVPR'24] (Shan et al., 2024)	261.83	26.62
MAFT+ [ECCV'24] (Jiao et al., 2024)	381.19	109.98
Mask-Adapter [CVPR'25] (Li et al., 2025b)	196.25	14.81
`MulGeCa`, Ours	385.56	207.14

Table 7: Inference-efficiency comparison under the same protocol.

Method	FLOPs (G) ↓	FPS ↑
CAT-Seg [CVPR'24] (Cho et al., 2024)	1073.41	9.31
EBSeg [CVPR'24] (Shan et al., 2024)	6401.92	25.57
MAFT+ [ECCV'24] (Jiao et al., 2024)	87.28	13.18
Mask-Adapter [CVPR'25] (Li et al., 2025b)	**80.81**	11.67
`MulGeCa`, Ours	193.188	19.7683

Table 8: Parameter decomposition of the complete model.

Component group	Module	Functional role	Total (M)	Trainable (M)
MAFT+ core	CLIP-V	RGB visual encoding	88.22	88.22
	CLIP-T	Class-text encoding	91.16	0.00
	Mask Generator	Pixel and mask decoding	20.66	20.66
	OV Classifiers + CDT	Open-vocabulary classification	2.05	2.05
`MulGeCa` addition	Depth SPM	Depth visual encoding	88.22	87.56
	GCA^2 VFM	Geometric feature backbone	85.94	0.59
	G^{2}TEC	Scene-level text calibration	0.55	0.55
	GCA^2 Generator and Interactor	Region-level geometric interaction	9.25	9.25
	CPMR	Instance-level mask refinement	0.05	0.05
Complete `MulGeCa`, excluding teacher			**385.56**	**207.14**

D.3 Depth Sources and Controlled Perturbations

Since `MulGeCa` explicitly relies on estimated depth maps as geometric priors, we study estimator capacity, alternative depth models, and controlled perturbations. The comparison includes a *No depth* baseline, the Small, Base, and Large variants of Depth Anything V2, four additional monocular depth estimators, Gaussian noise, and shuffled depth.

For all settings, we keep the RGB images, semantic annotations, model configuration, optimizer, data augmentation, and evaluation protocol unchanged. Only the depth input is varied. Each model is trained for 10K iterations and evaluated on A-150, enabling a controlled comparison of depth capacity and noise perturbation.

Table 10 changes only depth quality. Valid aligned depth performs best, mild noise causes a small drop, strong noise causes a larger degradation, and shuffled depth performs worst. The result confirms that useful gains require spatial correspondence between RGB and depth rather than an arbitrary auxiliary signal.

We further replace the depth estimator while keeping depth normalization, input resolution, and the remaining training protocol unchanged. Table 10 shows that all valid estimators outperform removing

Table 9: Complete online pipeline for unseen images, including depth latency and inference memory.

Estimator	Depth (ms)	E2E (ms)	P95 (ms)	Alloc. (GB)	Reserv. (GB)
DA-V2-Small	10.65	166.09	194.39	4.97	5.58
DA-V2-Base	18.17	168.03	183.44	5.24	6.16
DA-V2-Large	40.53	192.34	229.26	6.12	6.79

Table 10: Depth-source and controlled-corruption study on A-150. All models are trained for 10K iterations under the same protocol.

Group	Depth source	Variant/venue	Corruption	A-150
Control	No depth	–	Removed	32.7
Control	Aligned depth	–	Mild Gaussian noise	34.3
Control	Aligned depth	–	Strong Gaussian noise	34.0
Control	Aligned depth	–	Shuffled	32.7
Estimator	Depth Anything V2 (Yang et al., 2024)	Small	None	34.2
Estimator	Depth Anything V2 (Yang et al., 2024)	Base	None	**34.5**
Estimator	Depth Anything V2 (Yang et al., 2024)	Large	None	**34.5**
Estimator	Depth Anything V2 (Yang et al., 2024)	Small	Gaussian noise	33.5
Estimator	Depth Anything V2 (Yang et al., 2024)	Base	Gaussian noise	33.9
Estimator	Depth Anything V2 (Yang et al., 2024)	Large	Gaussian noise	34.0
Estimator	SQLdepth [AAAI'24] (Wang et al., 2024b)	–	None	34.3
Estimator	Mono-ViFI [ECCV'24] (Liu et al., 2024a)	–	None	34.1
Estimator	Depth Pro [ICLR'25] (Bochkovskiy et al., 2025)	–	None	34.4
Estimator	SC-DepthV3 [TPAMI'24] (Sun et al., 2024)	–	None	33.9

depth. Depth Anything V2 variants achieve similar results, while SQLdepth, Mono-ViFI, Depth Pro, and SC-DepthV3 also provide consistent gains. Gaussian noise degrades every Depth Anything V2 variant, but all noisy variants remain above the no-depth setting.

D.4 VLM Backbones, VFM Initialization, and Input Resolution

To investigate the influence of VFM initialization, we compare DINOv2 (Oquab et al., 2023) and BEiT (Bao et al., 2022) pre-trained weights while keeping the ViT-B/16 VFM architecture, model configuration, training schedule, and evaluation protocol unchanged. As reported in Table 11, the DINOv2 initialization achieves 35.3% mIoU on A-150 (Zhou et al., 2019), while the BEiT initialization obtains 35.2% mIoU. The marginal difference of 0.1% demonstrates that `MulGeCa` is insensitive to the choice of VFM pre-trained weights and that

Table 13: Input resolution ablation.

Resolution	A-150	Memory (GB)	FPS
320 × 320	34.9	3.19	13.29
384 × 384	35.0	3.46	19.7683
512 × 512	35.1	4.15	7.63
640 × 640	**35.3**	5.04	4.70

Table 11: VFM initialization ablation on A-150.

VFM pre-trained weights	A-150
BEiT (Bao et al., 2022)	35.2
DINOv2 (Oquab et al., 2023)	**35.3**

Table 12: VLM scale ablation across five benchmarks.

VLM	A-847	PC-459	A-150	PC-59	PAS-20
ConvNeXt-B	13.5	19.0	35.3	58.9	**95.9**
ConvNeXt-L	**15.5**	**21.2**	**36.2**	**59.6**	95.4

Table 14: Text-side controls for G^{2}TEC on A-150.

Control	Input	A-150
Static context, frozen CLIP	–	31.6
Dynamic bias	RGB	33.4
Dynamic token	RGB	34.1
Dynamic bias	RGB-D	34.7
Dynamic token, G^{2}TEC	RGB-D	**34.9**

Table 15: Visual-side controls for GCA^2 on A-150.

Control	Input	A-150
No geometric-consistency bias	RGB	33.6
Frozen VFM	RGB	33.1
VFM + interactor	RGB	34.0
VFM + bias	RGB-D	34.2
VFM + bias + interactor, GCA^2	RGB-D	**34.8**

its performance gains do not depend on a specific VFM initialization. We therefore adopt DINOv2 pre-trained weights as the default setting in the main experiments.

As shown in Table 12, the complete model remains effective with both ConvNeXt-B and ConvNeXt-L VLMs. Table 11 further shows that replacing DINOv2 initialization with BEiT changes A-150 by only 0.1 mIoU, indicating limited dependence on a particular VFM initialization.

Table 13 shows that when only input resolution changes, accuracy increases gradually while memory and latency increase more sharply. This provides practical operating points for resource-constrained deployment.

D.5 Isolating Geometric Cues from Additional Feature Capacity

We first isolate the effect of aligned depth from the additional feature capacity introduced by the VFM stream. The text-side and visual-side controls in Tables 14 and 15 progressively add RGB features and aligned depth while preserving the corresponding architecture.

On the text side, adding the RGB feature stream improves the static-context result from 31.6 to 34.1, and aligned depth further improves it to 34.9. On the visual side, the frozen VFM alone obtains only 33.1, while RGB feature interaction reaches 34.0 and the complete depth-consistency design reaches 34.8. Thus, additional pretrained features provide some benefit, but feature capacity alone is insufficient; the strongest results require aligned depth and its geometric-consistency modeling.

D.6 Module-Level Sensitivity and Statistical Stability

To localize error propagation, we retain clean depth for two modules and inject Gaussian noise only into the depth-feature input of the remaining module. Table 17 shows that perturbing any module reduces performance. The numerical differences among G^{2}TEC, GCA^2, and CPMR are small, so this experiment supports a shared sensitivity to depth errors but does not establish a statistically significant robustness ranking among the three modules.

The component contributions themselves are evaluated with five independent random seeds in Table 16. G^{2}TEC, GCA^2, and CPMR individually improve over the baseline by 1.31, 1.19, and 1.38 mIoU, respectively, each substantially larger than the run-to-run standard deviation. The complete model improves A-150 from 33.63 ± 0.15 to 35.27 ± 0.10, an absolute gain of 1.64 mIoU, and obtains the highest mean with the lowest standard deviation.

To provide a more comprehensive evaluation of our contributions and fulfill the requirements for experimental statistical significance, we present an extended component-wise ablation study in Table 16. Specifically, to account for the performance variability, the reported main metric for each configuration represents the *mean* mIoU, while the blue subscripts denote the *standard deviation*. These statistics are calculated across five independent training runs utilizing different random seeds, which inherently capture the variability introduced by network initialization and data loader shuffling.

Table 17: Module-level sensitivity to depth perturbation on A-150.

Perturbed module	A-150
None; clean depth for all modules	**34.5**
G^{2}TEC	34.3
GCA^2	34.0
CPMR	34.1

As shown in Table 16, the performance improvement yielded by the progressive integration of the G^{2}TEC, GCA^2, and CPMR modules significantly exceeds the standard deviations, which range from ± 0.06 to ± 0.18. For instance, the full implementation establishes an absolute advantage over the baseline and single-module variants. This empirical evidence firmly confirms that the improvements brought by our `MulGeCa` framework are statistically significant and highly robust, rather than being the artifact of random variance.

D.7 Architectural Generalization and State-of-the-Art Comparisons

Table 18 summarizes the architectural distinction from representative depth-aware and RGB-X segmentation frameworks. Existing methods primarily fuse visual modalities for task-specific or fixed-label prediction. `MulGeCa` instead preserves pretrained image-text alignment and applies depth at the text-context, regional-interaction, and query-mask-pooling levels for zero-shot transfer.

Table 19 compares recent representative methods under Base and Large configurations. The Base `MulGeCa` provides the strongest A-150 and PAS-20 results in this comparison while using 18.0% of CAT-Seg's FLOPs. Under the Large configuration, `MulGeCa` trails CAT-Seg and DEDOS, showing that the advantage is not uniform across model scales.

To test compatibility beyond the MAFT+/FC-CLIP pipeline, we insert the architecturally independent G^{2}TEC module into CAT-Seg. The unified table shows consistent gains on all three datasets, demonstrating that scene-level text calibration can complement cost aggregation.

Traditional depth-guided segmentation mainly fuses RGB-D visual features within a fixed label space. In contrast, `MulGeCa` addresses open-vocabulary segmentation and preserves pretrained image-text alignment while applying geometry to text context, regional interaction, and query-mask pooling. Its novelty is therefore multi-level depth-guided semantic anchoring for OVS rather than the use of depth alone.

Constrained by the page limit of the main manuscript, we provide a more exhaustive quantitative comparison with recent state-of-the-art, or SoTA, open-vocabulary segmentation methods in Table 21. Methods not included in the compact main-text benchmark table, including SCAN, CAT-Seg, CorrCLIP, and DEDOS, are reported either in this complete table or in the focused recent-method comparison in Table 19. This extended evaluation incorporates a broader spectrum of baselines across different VLMs, *i.e.*, Base and Large configurations. As quantitatively demonstrated in the table, our proposed `MulGeCa` with a Base-level backbone establishes leading results across all five challenging benchmarks. Under the Large configuration, `MulGeCa` improves over its direct MAFT+ baseline but trails dense cost aggregation and diffusion-distillation methods on several datasets. These findings support strong generalization while avoiding a uniform state-of-the-art claim across model scales.

In the following section, we provide a detailed summary of the open-vocabulary segmentation methods compared in our experiments.

SPNet (Xian et al., 2019) addresses the challenging task of zero-shot and few-shot semantic segmentation by incorporating class-level semantic information into any framework designed for

Table 16: Component-wise ablation study across different datasets.

[1pt] G^{2}TEC	GCA^2	CPMR	A-847	A-150	PAS-20
			13.20$_{\pm 0.12}$	33.63$_{\pm 0.15}$	93.93$_{\pm 0.18}$
✓			13.33$_{\pm 0.09}$	34.94$_{\pm 0.14}$	95.39$_{\pm 0.11}$
	✓		13.24$_{\pm 0.18}$	34.82$_{\pm 0.12}$	95.03$_{\pm 0.15}$
		✓	13.36$_{\pm 0.11}$	35.01$_{\pm 0.16}$	95.52$_{\pm 0.08}$
✓	✓		13.29$_{\pm 0.14}$	34.98$_{\pm 0.10}$	95.45$_{\pm 0.13}$
✓		✓	13.47$_{\pm 0.09}$	35.09$_{\pm 0.12}$	95.68$_{\pm 0.07}$
	✓	✓	13.38$_{\pm 0.16}$	35.03$_{\pm 0.14}$	95.57$_{\pm 0.11}$
✓	✓	✓	**13.54**$_{\pm 0.08}$	**35.27**$_{\pm 0.10}$	**95.92**$_{\pm 0.06}$

Table 18: Architectural comparison with representative depth-aware segmentation methods.

Method	Task	Input	Fusion target	Text treatment	Generalization
ESANet [ICRA'21] (Seichter et al., 2021)	Closed-set segmentation	RGB-D	Encoder/decoder features	None	Fixed labels
TokenFusion [CVPR'22] (Wang et al., 2022)	Multimodal dense prediction	RGB-X	Transformer tokens	None	Task-specific
CMX [T-ITS'23] (Zhang et al., 2023)	Closed-set segmentation	RGB-X	Cross-modal exchange	None	Fixed labels
DFormer [ICLR'24] (Yin et al., 2024)	Closed-set segmentation	RGB-D	Depth-aware visual features	None	Fixed labels
`MulGeCa`	Open-vocabulary segmentation	RGB-D + text	Text, regions, mask pooling	Gated CLIP calibration	Zero-shot transfer

Table 19: Accuracy, computational cost, and plug-and-play comparison with recent OVS methods.

Scale	Method	A-847	PC-459	A-150	PC-59	PAS-20	FLOPs (G)
Base	CAT-Seg [CVPR'24] (Cho et al., 2024)	12.0	19.0	31.8	57.5	94.6	1073.41
Base	CorrCLIP [ICCV'25] (Zhang et al., 2025)	–	–	24.6	–	87.8	366.35
Base	DEDOS [ICCV'25] (Li et al., 2025a)	–	–	33.7	–	95.7	**163.59**
Base	`MulGeCa`	**13.5**	**19.0**	**35.3**	**58.9**	**95.9**	193.188
Large	CAT-Seg [CVPR'24] (Cho et al., 2024)	16.0	23.8	37.9	63.3	97.0	1976.40
Large	CorrCLIP [ICCV'25] (Zhang et al., 2025)	–	–	26.9	–	90.8	564.89
Large	DEDOS [ICCV'25] (Li et al., 2025a)	–	–	**39.4**	–	**97.6**	**487.33**
Large	`MulGeCa`	15.5	21.2	36.2	59.6	95.4	851.37

Table 20: Plug-and-play G^2TEC experiment with CAT-Seg.

Setting	A-150	PC-59	PAS-20
CAT-Seg (Cho et al., 2024) baseline	31.8	57.5	94.6
CAT-Seg + G^2TEC	**34.7**	**58.2**	**95.4**

semantic segmentation in an end-to-end manner, thereby transferring knowledge from previously seen classes to novel classes.

ZS3Net (Bucher et al., 2019) introduces a new task of zero-shot semantic segmentation and proposes a novel architecture that combines a deep visual segmentation model with an approach to generate visual representations from semantic word embeddings to address both generalized and standard zero-shot classification at pixel level.

LSeg (Li et al., 2022) proposes a novel language-driven semantic image segmentation model that uses a text encoder to compute embeddings of descriptive input labels and contrastively aligns pixel-wise

Table 21: Complete comparisons with state-of-the-art methods across multiple benchmarks. "LN" denotes the Localized Narratives dataset (Pont-Tuset et al., 2020).

[1pt] Model	VLM	Training Dataset	A-847	PC-459	A-150	PC-59	PAS-20
SPNet [CVPR'19] (Xian et al., 2019)	-	PASCAL VOC	-	-	-	24.3	18.3
ZS3Net [NeurIPS'19] (Bucher et al., 2019)	-	PASCAL VOC	-	-	-	19.4	38.3
LSeg [ICLR'22] (Li et al., 2022)	ViT-B	PASCAL VOC-15	-	-	-	-	47.4
LSeg+ [ECCV'22] (Ghiasi et al., 2022)	ALIGN	COCO-Stuff	2.5	5.2	13.0	36.0	-
ZegFormer [CVPR'22] (Ding et al., 2022)	ViT-B	COCO-Stuff-156	4.9	9.1	16.9	42.8	86.2
ZegFormer† [CVPR'22] (Ding et al., 2022)	ViT-B	COCO-Stuff-156	5.6	10.4	18.0	45.5	89.5
ZSseg [ECCV'22] (Xu et al., 2022)	ViT-B	COCO-Stuff	7.0	-	20.5	47.7	88.4
OpenSeg [ECCV'22] (Ghiasi et al., 2022)	ALIGN	COCO Panoptic+LN	4.4	7.9	17.5	40.1	-
OVSeg [CVPR'23] (Liang et al., 2023)	ViT-B	COCO-Stuff+COCO Captions	7.1	11.0	24.8	53.3	92.6
ZegCLIP [CVPR'23] (Zhou et al., 2023)	ViT-B	COCO-Stuff-156	-	-	-	41.2	93.6
SAN [CVPR'23] (Xu et al., 2023c)	ViT-B	COCO-Stuff	10.1	12.6	27.5	53.8	94.0
DeOP [ICCV'23] (Han et al., 2023)	ViT-B	COCO-Stuff	7.1	9.4	22.9	48.8	91.7
SCAN [CVPR'24] (Liu et al., 2024c)	ViT-B	COCO-Stuff	10.8	13.2	30.8	58.4	**97.0**
SED [CVPR'24] (Xie et al., 2024)	ViT-B	COCO-Stuff	11.4	18.6	31.6	57.3	94.4
EBSeg [CVPR'24] (Shan et al., 2024)	ViT-B	COCO-Stuff	11.1	17.3	30.0	56.7	94.6
MAFT+ [ECCV'24](Jiao et al., 2024)	ConvNeXt-B	COCO-Stuff	13.2	14.2	33.6	55.9	93.9
EOV-Seg (S) [AAAI'25] (Niu et al., 2025)	ResNet50	COCO Panoptic	6.6	11.5	21.9	46.0	87.2
EOV-Seg (M) [AAAI'25] (Niu et al., 2025)	ResNet50x4	COCO Panoptic	7.8	12.2	25.5	51.8	91.2
ELSE-Net [AAAI'25] (Zhao et al., 2025a)	ViT-B	COCO-Stuff	10.3	-	27.9	54.3	95.1
OPMapper [NeurIPS'25] (Wang et al., 2025)	ViT-B	COCO-Stuff	-	-	31.3	56.2	94.0
`MulGeCa`, Ours	ConvNeXt-B	COCO-Stuff	**13.5**	**19.0**	**35.3**	**58.9**	**95.9**
LSeg [ICLR'22] (Li et al., 2022)	ViT-L	PASCAL VOC-15	-	-	-	-	52.3
OpenSeg [ECCV'22] (Ghiasi et al., 2022)	ALIGN	COCO Panoptic+LN	8.1	11.5	26.4	44.8	-
OVSeg [CVPR'23] (Liang et al., 2023)	ViT-L	COCO-Stuff+COCO Captions	9.0	12.4	29.6	55.7	94.5
ODISE [CVPR'23] (Xu et al., 2023a)	ViT-L	COCO-Stuff	11.1	14.5	29.9	57.3	-
FC-CLIP [NeurIPS'23] (Yu et al., 2023a)	ConvNeXt-L	COCO Panoptic	14.8	18.2	34.1	58.4	95.4
SAN [CVPR'23] (Xu et al., 2023c)	ViT-L	COCO-Stuff	12.4	15.7	32.1	57.7	94.6
MAFT [NeurIPS'23] (Jiao et al., 2023)	ViT-L	COCO-Stuff	12.7	16.2	33.0	59.0	92.1
MAFT+ [ECCV'24] (Jiao et al., 2024)	ConvNeXt-L	COCO-Stuff	15.1	**21.6**	36.1	59.4	**96.5**
EOV-Seg (L) [AAAI'25] (Niu et al., 2025)	ConvNeXt-L	COCO Panoptic	12.8	16.8	32.1	56.9	94.8
`MulGeCa`, Ours	ConvNeXt-L	COCO-Stuff	**15.5**	21.2	**36.2**	**59.6**	95.4

embeddings to the corresponding text embeddings, which allows generalization to previously unseen categories at test time without requiring additional training samples.

LSeg+ (Ghiasi et al., 2022) designs an open-vocabulary image segmentation model that organizes an image into meaningful regions indicated by arbitrary texts. It focuses on visual grouping by learning to propose segmentation masks and then learns visual-semantic alignments by aligning each word in a caption to one or a few predicted masks, which significantly boosts performance and scalability.

ZegFormer (Ding et al., 2022) proposes to decouple the zero-shot semantic segmentation task into two sub-tasks: a class-agnostic grouping task to group pixels into segments and a segment-level zero-shot classification task, which provides a natural way to leverage large-scale vision-language models like CLIP.

ZSseg (Xu et al., 2022) advocates a two-stage open-vocabulary semantic segmentation framework that first extracts generalizable mask proposals and then utilizes a pre-trained CLIP model to perform open-vocabulary classification on the masked image crops.

OpenSeg (Ghiasi et al., 2022) emphasizes the importance of visual grouping before learning visual-semantic alignments by learning to propose segmentation masks and aligning them with words in captions, enabling scalable open-vocabulary image segmentation.

OVSeg (Liang et al., 2023) identifies the performance bottleneck of CLIP on masked images and proposes to fine-tune CLIP on a collection of masked image regions and their text descriptions, while introducing mask prompt tuning to further improve generalizability.

ZegCLIP (Zhou et al., 2023) proposes an efficient one-stage solution that extends CLIP's zero-shot capability directly from image to pixel level, utilizing three simple designs to retain CLIP's inherent capacity and significantly accelerating inference speed compared to two-stage methods.

SAN (Xu et al., 2023c) introduces a side adapter network attached to a frozen CLIP model to predict mask proposals and attention biases, enabling fast and accurate open-vocabulary semantic segmentation in a lightweight and end-to-end manner.

SED (Xie et al., 2024) employs a hierarchical encoder-decoder architecture to generate pixel-level image-text cost maps and introduces a category early rejection scheme in the decoder to significantly accelerate inference without accuracy degradation.

EBSeg (Shan et al., 2024) incorporates an Adaptively Balanced Decoder and a Semantic Structure Consistency loss to balance embeddings for seen and novel classes, while leveraging a frozen SAM image encoder to complement the spatial information missing in CLIP features.

CAT-Seg (Cho et al., 2024) introduces a cost-based approach that adapts vision-language models by aggregating the cost volume between image and text embeddings, enabling effective fine-tuning of CLIP encoders for both seen and unseen class segmentation.

EOV-Seg (Niu et al., 2025) presents a single-stage, efficient framework for open-vocabulary panoptic segmentation, utilizing a Vocabulary-Aware Selection module and Two-way Dynamic Embedding Experts to achieve high inference speeds without sacrificing accuracy.

ELSE-Net (Zhao et al., 2025a) adopts reverse thinking by excluding improbable categories to narrow the selection range. It combines a General Processing Block for inclusion probabilities with an EXcluding the ImPossible Block that leverages CLIPN to refine predictions and reduce misclassification.

BBN (Pan et al., 2025) proposes a bi-directional bridge network that purifies noisy text embeddings via an optimal transport algorithm and integrates reliable textual information into vision embeddings to bridge the gap between pre-trained models and downstream segmentation tasks.

ODISE (Xu et al., 2023a) presents a unified framework for open-vocabulary panoptic segmentation by leveraging the frozen internal representations of both pre-trained text-image diffusion models and discriminative models like CLIP to perform segmentation of any category.

FC-CLIP (Yu et al., 2023a) proposes a single-stage framework using a shared Frozen Convolutional CLIP backbone that serves as both a mask generator and an open-vocabulary classifier, significantly simplifying the existing two-stage pipeline and improving efficiency.

D.8 Category-Level Quantitative Analysis

We fix four category groups in advance and select 10 images containing the corresponding categories for each group, yielding 40 images in total. MAFT+ and `MulGeCa` use the same images, category mapping, resolution, and evaluation code. We calculate mIoU over the categories included in each group, with the quantitative results reported in Table 22.

Object-like categories benefit most because depth ordering and occlusion boundaries often align with semantic boundaries. Surface continuity also helps stuff-like classes. Thin or distant boundaries can be smoothed by low-resolution depth, while coplanar parts and materials can change semantics without a corresponding depth change. The latter case explains the decrease for part/material-like categories and the relatively limited improvement on fine-grained PC-459 labels.

Table 22: Category-group mIoU on the fixed 40-image subset.

Category group	MAFT+	`MulGeCa`	Δ mIoU
Object-like	31.74	39.58	**+7.84**
Stuff-like	68.81	70.34	+1.54
Part/material-like	41.28	40.23	-1.06
Thin/distant objects	27.82	28.69	+0.87

To further analyze the effectiveness of `MulGeCa` at a per-category level, we visualize the normalized confusion matrices in Fig. 9. The baseline model suffers from severe background interference, where a substantial portion of foreground pixels are misclassified into the "*oth*" category denoting others. This results

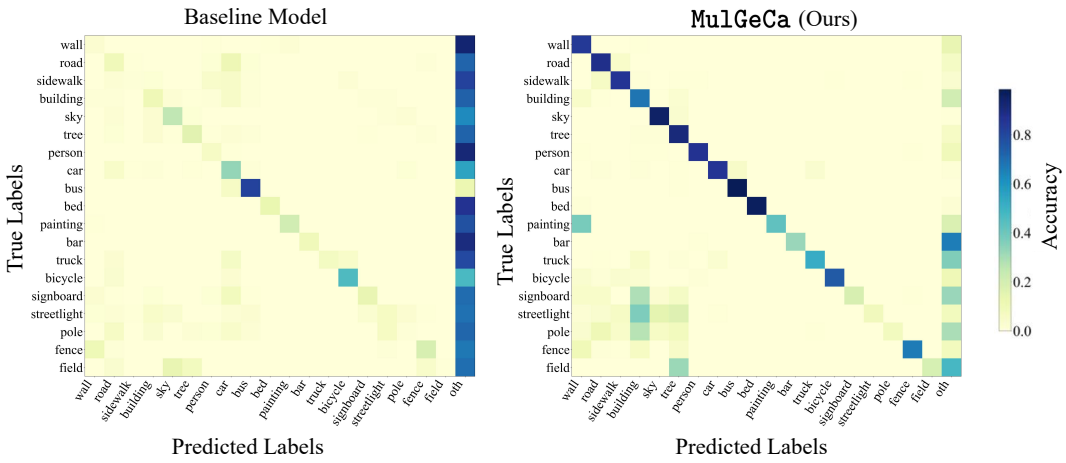

Figure 9: Normalized confusion matrices for the baseline and `MulGeCa` on A-150 (Zhou et al., 2019).

in a faded diagonal response, particularly for challenging categories such as "*pole*", "*fence*", and "*signboard*".

In contrast, our proposed `MulGeCa` demonstrates a significantly sharper diagonal, indicating a substantial improvement in per-category precision. By effectively infusing fine-grained spatial geometric knowledge, `MulGeCa` suppresses the background noise that previously dominated the baseline's predictions. Notably, challenging categories that were almost entirely neglected by the baseline, such as "*fence*" and "*painting*", now exhibit distinct and accurate diagonal squares. This improvement confirms that anchoring semantic features within a geometric context prevents the model from over-relying on the background class and enhances its ability to distinguish small or visually similar entities.

D.9 Detailed Results of the Ablation Study on Meta-Net Architecture

In this section, we present an in-depth exploration of the architectural design of the Meta-Net within the G^{2}TEC module. The Meta-Net is responsible for projecting the global geometric prior from the vision backbone into a sample-specific continuous prompt shift. To identify the optimal configuration, we conduct ablation experiments on the A-150 (Zhou et al., 2019) dataset, varying the number of multilayer perceptron layers, abbreviated as MLP layers, from one to four and investigating the impact of dropout regularization.

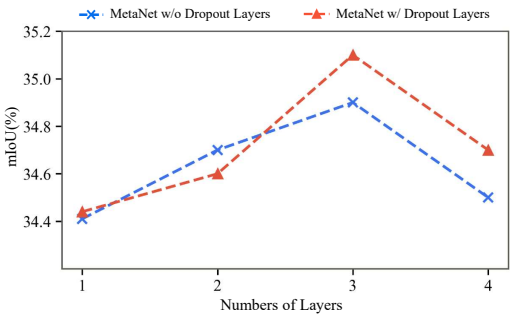

Figure 10: Meta-Net architecture ablation on A-150 (Zhou et al., 2019). The three-layer MLP with dropout yields the best result.

As illustrated in Fig. 10, increasing the network depth from one to three layers yields a steady improvement in mIoU. This trend indicates that a sufficiently deep architecture is required to capture the complex non-linear mapping between low-dimensional geometric statistics and the high-dimensional textual semantic space. However, further extending the network to four layers leads to a noticeable performance degradation, with a mIoU dropping to 34.5% without Dropout and 34.7% with dropout, which can be attributed to overfitting and optimization challenges associated with excessive parameters in the prompt generation pathway.

Furthermore, the integration of dropout layers proves to be highly effective in mitigating this overfitting tendency, particularly in deeper configurations. While the performance gap is marginal at shallower depths, the model with a three-layer MLP equipped with dropout achieves the superior performance of 35.1% in terms of mIoU, outperforming its non-dropout designs. Consequently, we

Table 23: Affinity-range ablation for geometric constraints. The upper bound A_{max} is fixed at 1.0.

[1pt] A_{min}	PC-459	A-150	PC-59
1e-4	14.61	34.91	58.65
1e-3	14.45	35.04	58.70
0.01	14.58	34.87	58.75
0.05	**14.96**	35.05	**58.77**
0.1	14.55	**35.17**	58.73

Table 24: Coupled depth-bias log-space-bound ablation. For each lower bound, the affinity floor is $\exp(\log_{\text{min}})$ and the upper bound is 0.0.

[1pt] $\log_{\text{min}}$	PC-459	A-150	PAS-20
-6	14.50	34.91	95.80
-8	**14.67**	**35.03**	**95.83**
-10	14.63	34.98	95.68
-12	14.53	34.86	95.57

select the three-layer MLP with dropout as our default Meta-Net architecture to achieve the optimal balance between representation capacity and generalization robustness.

D.10 Hyperparameter Sensitivity and Optimization Complexity

In this section, we provide an in-depth exploration of the key hyper-parameters of the GCA^2 module that control the intensity of the geometric-consistency bias. Specifically, the affinity matrix $\boldsymbol{A}$ represents the spatial-geometric similarity derived from depth discrepancies, and the resulting geometric-consistency bias $\boldsymbol{B}$ that constrains the attention mechanism. To stabilize the log operation and prevent the geometric-consistency bias from inducing numerical instability or dominating the weights within the attention mechanism, we constrain the elements of the affinity matrix $\boldsymbol{A}$ and the values of $\boldsymbol{B}$ within the ranges of $[0.05, 1.0]$ and $[-8.0, 0.0]$, respectively.

We validate the effectiveness of these boundary values through comprehensive ablation experiments. As presented in Table 23, we fix the upper bound A_{max} at 1.0 and evaluate various lower bounds for A_{min}. The results indicate that $A_{\text{min}} = 0.05$ effectively maintains the validity of the geometric constraints while avoiding the computational fluctuations associated with extremely small weights.

Furthermore, Table 24 evaluates coupled log-space bounds. For each $\log_{\text{min}}$, the affinity floor is set consistently to $\exp(\log_{\text{min}})$ before applying the outer clamp $[\log_{\text{min}}, 0]$. This coupling is necessary because, if the affinity floor were fixed at 0.05, lower outer bounds below $\log(0.05)$ would be inactive. The coupled sweep shows that overly broad suppression, *e.g.*, -12.0, may over-attenuate semantic correlations in distant regions, whereas -6.0 does not sufficiently eliminate geometric mismatches. We therefore use -8.0 as the coupled log-space lower bound in this sensitivity study.

The GCA^2 parameters σ_b, ϕ, and γ are learned jointly through backpropagation and do not require manual per-dataset adjustment. The principal fixed hyperparameters are the scale factor w, affinity clamp range $[0.05, 1.0]$, bias clamp range $[-8.0, 0.0]$, and the depth and dropout of the G^{2}TEC Meta-Net. The same settings are shared across all datasets without per-dataset search. Meta-Net depth and dropout are ablated in Fig. 10, while the affinity and bias clamp ranges are analyzed in Tables 23 and 24. All three modules are trained jointly and require neither multi-stage pretraining nor independent optimization.

D.11 Grounding-Based Comparison and Generalized Zero-Shot Evaluation

The extended state-of-the-art table already includes ODISE (Xu et al., 2023a), a representative diffusion-based OVS method that uses frozen text-to-image diffusion representations. We additionally consider grounding-based features below.

Grounding DINO (Liu et al., 2024b) is an object detector and phrase-grounding model rather than a semantic segmentation model. For a comparable experiment, we integrate the open-vocabulary classifiers and segmentation head while preserving the remaining training and evaluation settings. Table 25 reports the resulting A-150 comparison. Under this common segmentation setting,

Table 25: Grounding DINO (Liu et al., 2024b) comparison on A-150.

Method	A-150
Grounding DINO + segmentation head	33.2
`MulGeCa`	**35.3**

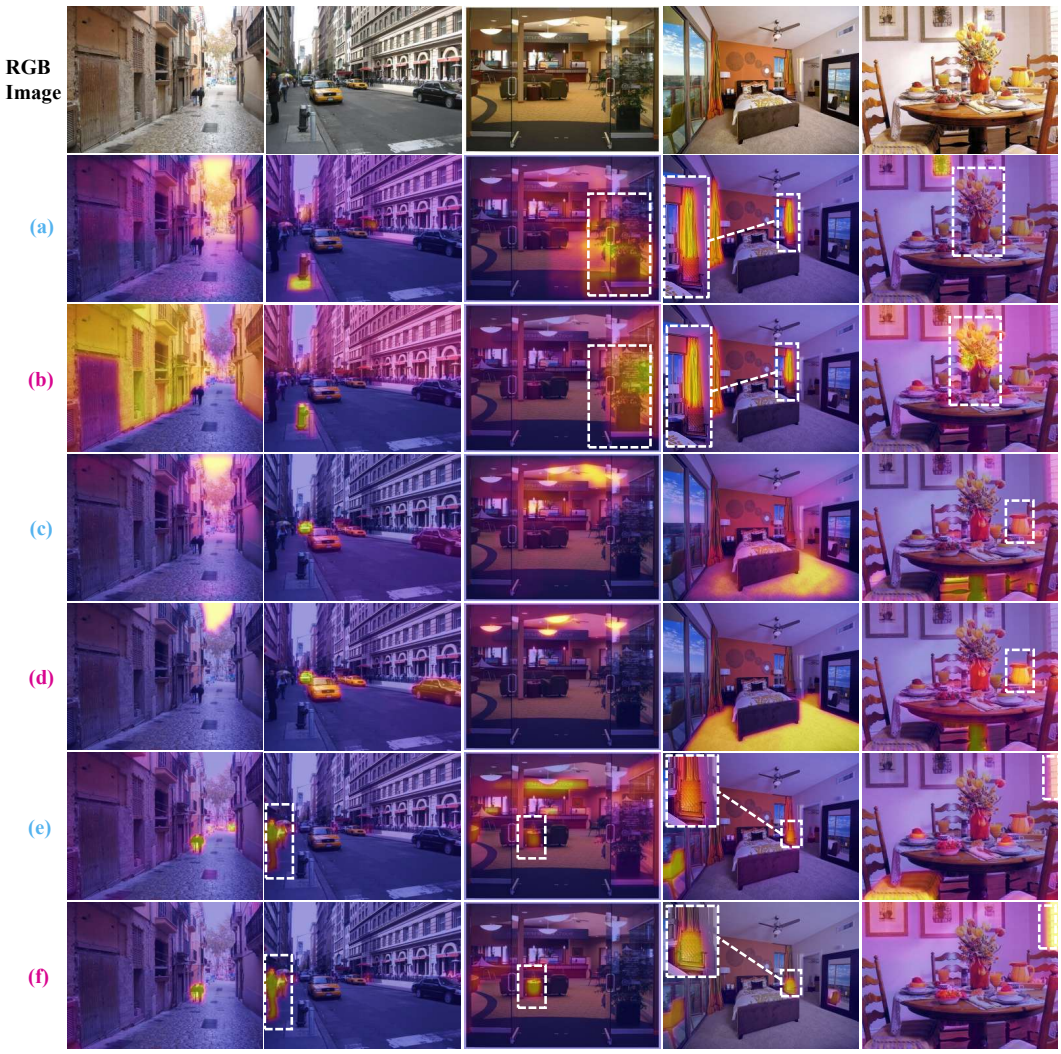

Figure 11: Additional qualitative comparisons of prediction probabilities in indoor and outdoor scenes. Baseline results are shown in cyan-labeled rows, while the corresponding `MulGeCa` results are shown in magenta-labeled rows. White dashed boxes highlight representative improvements.

`MulGeCa` obtains 35.3 mIoU, compared with 33.2 mIoU for the Grounding DINO prior.

We additionally evaluate the generalized zero-shot setting in a unified label space. Unseen mIoU directly measures zero-shot ability, while seen and overall mIoU diagnose bias toward training categories and aggregate performance. The model reaches 45.5 seen mIoU, 26.8 unseen mIoU, and 35.2 overall mIoU, providing a direct measurement of transfer to unseen categories alongside aggregate performance.

Table 26: Generalized zero-shot evaluation.

Metric	Result
Seen mIoU	45.5
Unseen mIoU	26.8
Overall mIoU	35.2

D.12 Additional Qualitative Results on Prediction Probabilities

To further validate the effectiveness of our proposed approach, we provide additional qualitative comparisons of prediction probabilities under diverse and cluttered scenarios. As shown in Fig. 11, significantly improved areas are highlighted with white dashed boxes. It reveals that the baseline

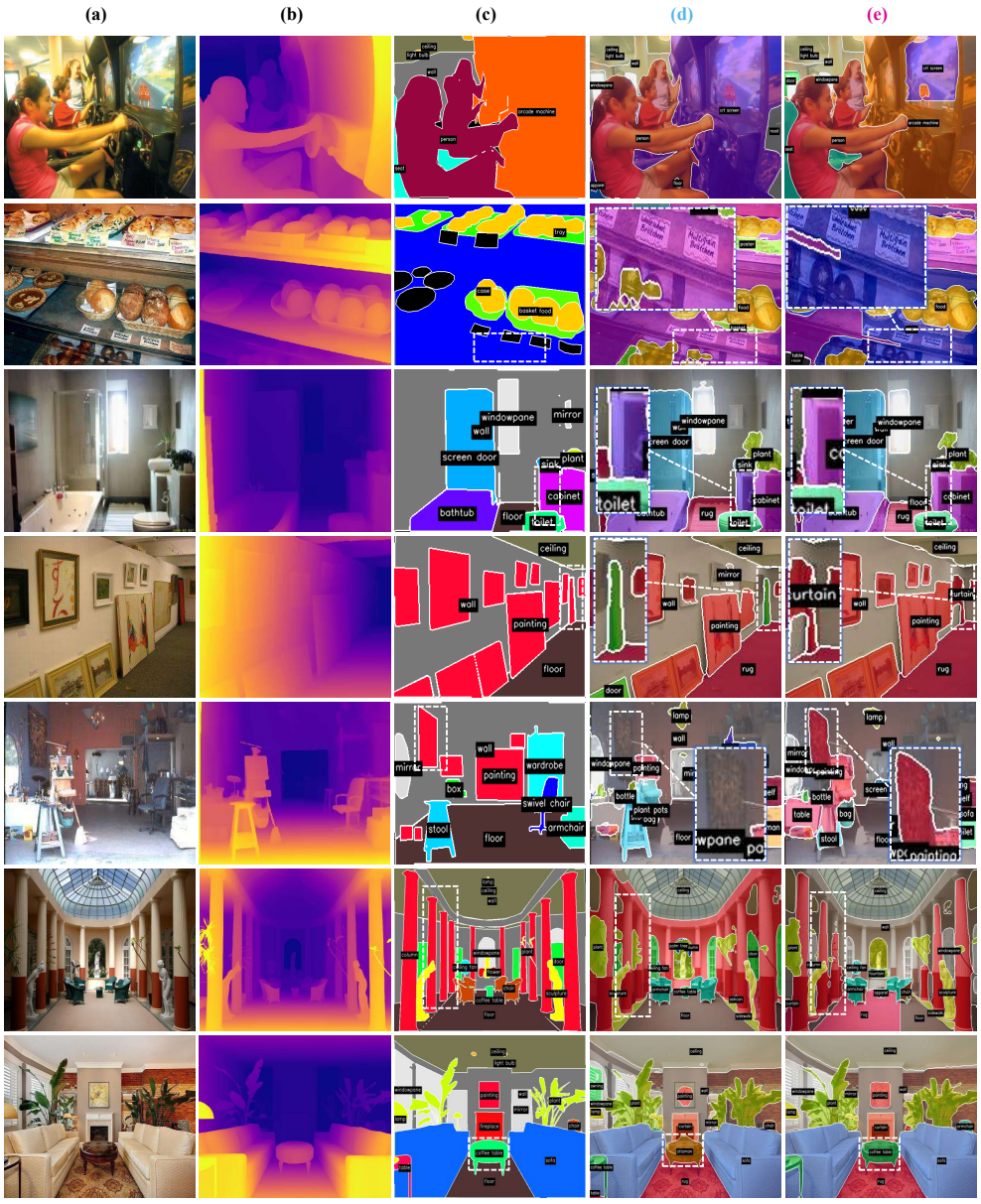

Figure 12: Additional qualitative comparisons of semantic segmentation results in **indoor scenarios**. (a), (b), and (c) denote the input RGB images, estimated depth maps, and Ground Truth annotations, respectively. (d) shows the segmentation results produced by the baseline model. (e) displays the final predictions generated by our proposed `MulGeCa`. Challenging areas with significant improvements are highlighted with white dashed boxes for a clearer comparison.

model frequently struggles with severe feature diffusion. Particularly in visually ambiguous regions, such as the distant objects in complex street views or the intricate boundaries of indoor furniture, the baseline's prediction probabilities in Figs. 11(a), (c) and (e) are loosely distributed and often bleed into irrelevant backgrounds. This phenomenon reflects the inherent limitation of standard VLMs that lack explicit spatial constraints, leading to suboptimal cross-modal alignment. In contrast, by effectively injecting global geometric priors, our `MulGeCa` produces significantly more confident and localized probability maps. As observed in Figs. 11(b), (d) and (f), the high-probability regions, indicated by warmer colors, are tightly anchored to the actual physical boundaries of the target objects. This demonstrates that our method effectively suppresses background noise and disambiguates spatial overlaps, thereby yielding highly discriminative semantic representations that are crucial for precise dense prediction tasks.

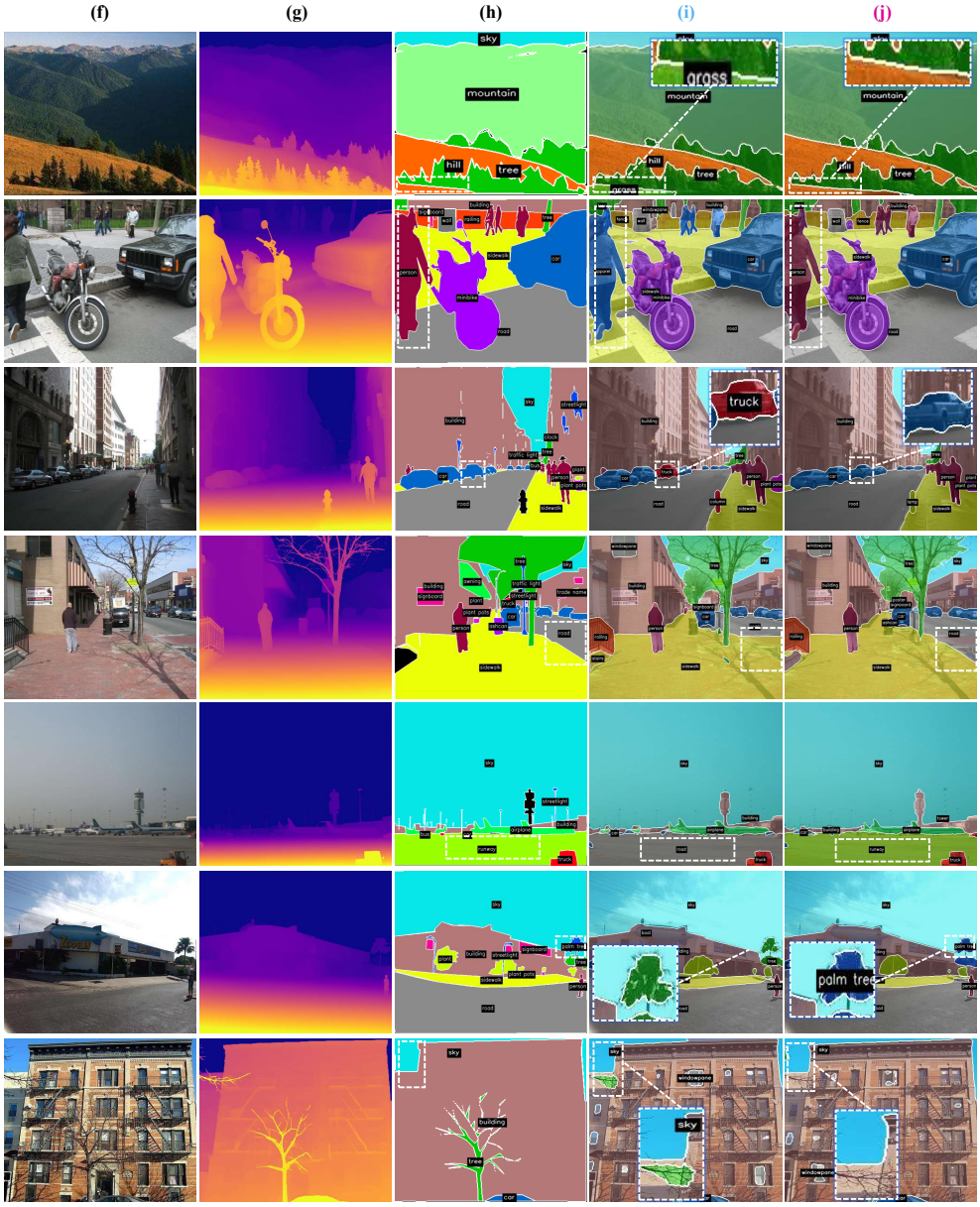

Figure 13: Additional qualitative comparisons of semantic segmentation results in **outdoor scenarios**. (f), (g), and (h) denote the input RGB images, estimated depth maps, and Ground Truth annotations, respectively. (i) shows the segmentation results produced by the baseline model. (j) displays the final predictions generated by our proposed `MulGeCa`. Challenging areas with significant improvements are highlighted with white dashed boxes for a clearer comparison.

D.13 Additional Qualitative Results of Semantic Segmentation

To further validate the robustness and superior generalizability of our approach, we present extensive qualitative comparisons of semantic segmentation results across diverse and cluttered scenes. Indoor results are shown in Fig. 12, and outdoor results are shown in Fig. 13. As illustrated in Fig. 12(d) and Fig. 13(i), the baseline model frequently struggles with visual ambiguities, such as objects with similar textures or severe spatial occlusions. Relying exclusively on 2D RGB cues, the baseline inevitably suffers from semantic diffusion, yielding over-segmented masks, hallucinated categories, and coarse object boundaries. In stark contrast, our proposed `MulGeCa` in Fig. 12(e) and Fig. 13(j) effectively overcomes these limitations by integrating explicit global geometric priors derived from the depth maps in Fig. 12(b) and Fig. 13(g). As highlighted by the white dashed boxes, our method

successfully disambiguates complex spatial overlaps and delineates sharp, physical-aware boundaries. By anchoring scattered abstract semantics to concrete 3D layouts, `MulGeCa` produces highly precise segmentation masks that closely align with the Ground Truth in Fig. 12(c) and Fig. 13(h), further confirming its exceptional capability in dense open-vocabulary scene parsing.

E Additional Clarifications

Are the geometric priors sensor-measured depth or complete 3D geometry? No. The official ADE20K, PASCAL-Context, PASCAL VOC, and COCO-Stuff settings used here do not provide synchronized sensor depth ground truth. We use monocular relative depth estimated from RGB images by Depth Anything V2 (Yang et al., 2024). Its useful signals are relative boundaries, surface continuity, and foreground-background ordering rather than metric scale, sensed geometry, or a complete 3D reconstruction. Accordingly, "depth-consistent physical support" is the intended scope of the geometric prior. Estimated depth is a realistic and scalable setting for RGB-only OVS benchmarks, while evaluation with synchronized sensor depth remains a valuable direction for separating physical boundary alignment from monocular-estimation artifacts.

Do the gains come only from additional foundation-model capacity? The controlled text-side and visual-side studies in Tables 14 and 15 progressively introduce auxiliary RGB features and aligned depth while preserving the corresponding architecture. Auxiliary RGB features improve performance, but the best result in both controls requires aligned depth and geometric-consistency modeling. Thus, additional pretrained capacity contributes to the result but does not by itself explain the complete gain. Comparisons with appearance-based foundation features and mask adapters in Tables 6 and 7 further place this gain against contemporary alternatives with different capacity-efficiency trade-offs.

When can depth be precomputed, and what is the cost for unseen online images? Depth can be generated once and cached when the image collection is known in advance, including training sets, fixed benchmarks, repeated evaluations, archival collections, and offline batch processing. This setting measures the segmentation core with RGB-D input already available. For a previously unseen image or a real-time stream, depth estimation is indispensable and must be included in end-to-end cost. Table 9 therefore reports the complete online pipeline with Depth Anything V2 Small, Base, and Large, including depth latency, end-to-end latency, P95 latency, and peak memory. These end-to-end measurements are distinct from the synchronized 384×384 single-crop segmentation-core benchmark in Table 7; they should not be combined arithmetically.

What does the CPMR centroid represent? CPMR does not compute a centroid from raw depth values, 3D coordinates, or complete object geometry, and it is not a simple mean of a learned response map. It projects the depth-conditioned high-level features produced by GCA^2 and the mask generator into a response map, computes the first moment of the response distribution inside each query mask, and applies Gaussian centripetal weights to suppress boundary and outlier responses. We therefore use the term *feature-response centroid*. The resulting weighted contraction maps the query-specific spatial weights and visual features to the pooled query embedding, preserving the embedding dimension required by the open-vocabulary classifier.

How are the affinity and log-bias bounds related? Under a fixed affinity floor of 0.05, the smallest attainable log affinity is $\log(0.05)$; consequently, an outer lower clamp below that value acts only as a numerical safeguard. The sensitivity experiment in Table 24 instead uses coupled bounds: the affinity floor is $\exp(\log_{\min})$ for each tested log-space lower bound. This clarification distinguishes the default stable formulation from the controlled bound sweep and prevents interpreting the outer clamp as independently active under a fixed 0.05 floor.

What is the scope of the Grounding DINO and generalized zero-shot controls? Grounding DINO (Liu et al., 2024b) is a phrase-grounding detector rather than a semantic segmentation model. Table 25 is therefore a controlled auxiliary-prior comparison: we attach the same open-vocabulary classifiers and segmentation head and preserve the remaining training and evaluation protocol. It is not presented as a comparison with a native Grounding DINO segmentation system. Likewise, Table 26 is a diagnostic of seen-category bias and unseen-category transfer in a unified label space,

rather than a claim of comparability with GZSS methods that use different datasets or class partitions. We report it separately from the standard zero-shot benchmark tables and do not use it to support the principal state-of-the-art claim.

Why can coplanar semantic boundaries remain difficult? When different materials, functions, or parts occupy the same geometric surface, a flat and confident depth prediction can still disagree with the semantic boundary. In that case, geometric continuity is not neutral: GCA^2 may preserve attention across the semantic transition, and CPMR may downweight peripheral fine-grained responses, producing an over-smoothing tendency. This mechanism is consistent with the part/material-group decrease and the limited PC-459 gain. A suitable future remedy is cross-modal boundary-discordance detection, which identifies strong RGB or semantic gradients coinciding with weak depth gradients, rather than relying only on depth uncertainty.

F Broader Applications and Future Directions

Our framework improves the spatial localization of open-vocabulary semantics in cluttered and occluded scenes, which can support open-world perception in autonomous systems, robotic navigation, assistive perception, and large-scale scene analysis. Because `MulGeCa` uses estimated relative depth rather than specialized sensing hardware, it can be applied to existing RGB image collections and deployed with different monocular depth estimators according to the available latency and memory budget. At the same time, safety-critical use requires explicit treatment of unreliable geometry, especially under extreme lighting, reflective or transparent surfaces, small distant objects, and coplanar fine-grained categories. Future work will investigate cross-modal boundary-discordance detection, adaptive geometric weighting, temporally consistent calibration for video, and evaluation on RGB-D benchmarks with synchronized sensor depth.

F.1 Failure Mechanisms and Scope of the Geometric Assumption

Geometric calibration assumes that estimated depth preserves RGB-aligned relative boundaries, surface continuity, and foreground-background ordering, and that semantic boundaries have some correspondence with depth discontinuities or surface organization. Figure 14 shows violations of these assumptions: overexposure and reflection distort depth in Figs. 14(a)-(b), small or distant objects are omitted in Fig. 14(c), and occlusion weakens foreground-background ordering in Fig. 14(d). Figs. 14(e)-(f) show coplanar semantic distinctions without depth discontinuities. Table 27 traces these violations through the three modules.

Table 27: Failure conditions, violated assumptions, and affected modules.

Failure condition	Violated geometric information	Module-level effect
Extreme lighting, transparency, or reflection	Flat, fragmented, or false depth boundaries; RGB-depth misalignment	G^{2}TEC receives an incorrect scene summary; GCA^2 connects or separates regions using incorrect pairwise depth differences; CPMR responses consequently shift.
Small or distant objects	True boundaries are smoothed or missed	GCA^2 lacks effective discontinuities, while CPMR may weaken true small-object or boundary responses.
Crowding and severe occlusion	Incorrect foreground-background order or merged instances	GCA^2 is guided by incorrect occlusion relations, and errors propagate to instance-level pooling in CPMR.
Coplanar materials, functions, or parts	Semantics change without a depth change	Continuous depth can preserve attention across the semantic transition and over-smooth fine-grained responses in GCA^2 and CPMR.

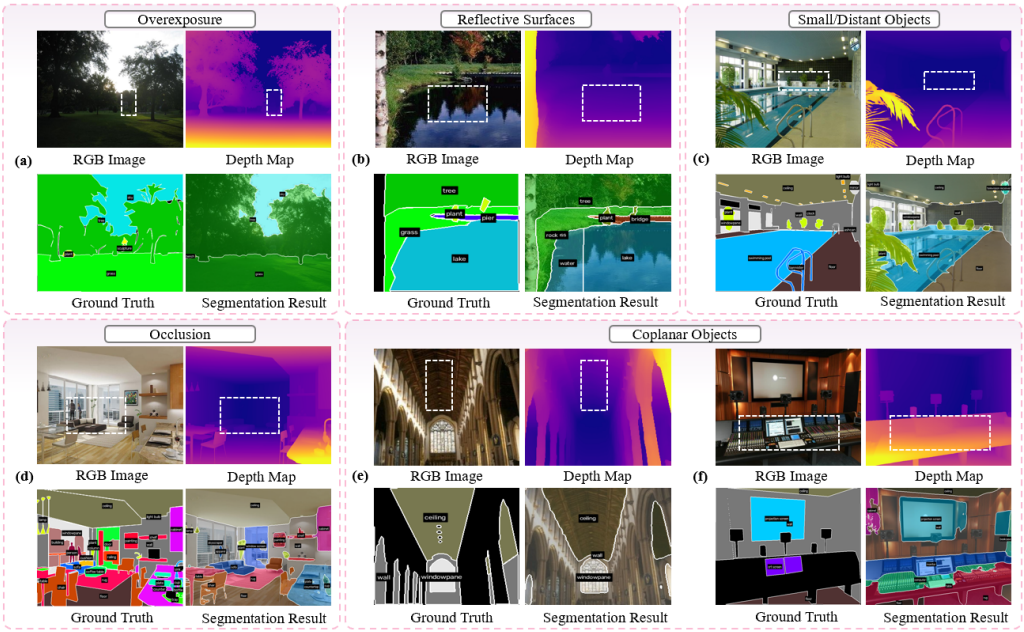

Figure 14: Representative failure cases under unreliable or insufficient geometric cues. The examples cover overexposure, reflective surfaces, small or distant objects, occlusion, and coplanar objects. White dashed boxes highlight regions where estimated depth loses a valid boundary or where semantic distinctions have no corresponding depth discontinuity.

These cases do not imply that predictions are determined entirely by depth. The G^{2}TEC gate retains static CLIP text embeddings, and the visual pathway preserves RGB/VLM semantics. For uncertain or false depth boundaries, future work can use reliability-aware weighting to weaken GCA^2 bias and CPMR weighting. Coplanar transitions require a different signal: cross-modal boundary-discordance detection can identify strong RGB or semantic gradients that coincide with weak depth gradients and prevent continuous geometry from suppressing useful semantic distinctions.