EXPLOITING TARGET KNOWLEDGE FROM MLLMS FOR ROBUST FEW-SHOT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot segmentation (FSS) aims to segment unseen object categories with a few (*e.g.*, one or five) labeled examples, enabling efficient adaptation to novel classes. Conventional models typically rely on appearance-based visual matching between support and query images for segmentation. While straightforward, these methods often struggle to handle significant appearance discrepancies and occlusions in the query image due to insufficient target knowledge. To mitigate this, we introduce a novel framework that mines target knowledge using the strong reasoning capacity of Multimodal Large Language Models (MLLMs) and employs it to enhance FSS. Specifically, building on SAM 2, our method, named MK-FSS, exploits two forms of complementary knowledge derived from a query image by an MLLM for FSS, including spatial knowledge, which provides a spatial prior indicating the potential target location, and semantic knowledge, which describes the target using text. The spatial knowledge is first encoded into a memory representation, and then resulting memory is integrated with the support-guided memory feature from query image through a carefully designed dual-memory debate-fusion (DMDF) module, yielding a more robust target memory feature. In parallel, the semantic knowledge is encoded into the textual feature, which is fused with multi-scale query features via a progressive cross-modal prompt generator (PCPG), producing a target-aware multimodal prompt for segmentation. Working together, the dual-memory feature and the multimodal prompt provide a comprehensive representation of the target, enabling more robust segmentation under complicated scenarios. In extensive experiments on multiple datasets, MK-FSS shows state-of-the-art results and largely surpasses existing approaches. Our code will be released.

1 INTRODUCTION

Few-shot segmentation (FSS) aims at segmenting novel object classes using a few labeled examples. It is of great practical importance in scenarios where obtaining the pixel-level annotations is time-consuming and labor-intensive, *e.g.*, medical imaging, remote sensing, and autonomous driving. By enabling efficient adaptation to untrained categories, FSS provides a promising solution for scalable and data-efficient visual understanding and received extensive attention in the past decade (Xu et al., 2023; Fan et al., 2022; Qin et al., 2025; Zhang et al., 2022; Wang et al., 2024a; 2023).

Existing FSS approaches (Wen et al., 2025; Zou et al., 2025; Qin et al., 2025) often follow a support-guided segmentation paradigm, where the information extracted from a limited number of labeled support images is leveraged to identify and segment the target in a query image. Despite considerable progress, their performance remains limited in challenging scenarios, such as those involving large appearance variations, occlusions, or visually similar distractors. A key limitation is that using only a small number of support images is insufficient to describe the target across diverse conditions. To mitigate this, recent studies have therefore explored incorporating additional knowledge using visual foundation models (Xu et al., 2025; Zhang et al., 2024) or class-level textual descriptions derived from category names (Lüddecke & Ecker, 2022; Nandam et al., 2025; Chen et al., 2024) to complement the cues from support images. While these efforts demonstrate improvements, they still suffer from limited exploitation of target-specific information, resulting in suboptimal performance in complex scenarios. This naturally raises a question: *How can we obtain more discriminative and robust target-specific information to improve few-shot segmentation?*

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.

Recently, Multimodal Large Language Models (MLLMs) (Bai et al., 2025a; Comanici et al., 2025) have exhibited strong capabilities in visual understanding and reasoning, providing a promising way to extract rich target-specific knowledge directly from the given query image. Drawing inspiration from this, we propose a novel framework that mines target-specific knowledge from MLLMs and leverages such knowledge to enhance FSS. Our method, termed ***MK-FSS***, builds upon SAM 2 (Ravi et al., 2025) and exploits two complementary forms of MLLM-derived target-specific knowledge from the query image, including spatial knowledge and semantic knowledge, for FSS. Specifically, the spatial knowledge offers a spatial prior indicating the potential target location, while the semantic knowledge describes the target object through textual information. To obtain a more discriminative target representation, we first encode the spatial knowledge into a memory representation, and then integrate it with the support-guided memory feature from query image through a carefully designed dual-memory debate-fusion (DMDF) module. This module regards the two memory features as two complementary agents, which exchange, verify, and refine information through a structured debate-verification-integration process, thereby yielding a more robust target memory feature. Meanwhile, the semantic knowledge is first encoded into the textual feature, which is then fused with multi-scale query features via a progressive cross-modal prompt generator (PCPG). By explicitly injecting high-level linguistic semantics into visual space, PCPG progressively integrates textual and visual features to produce a target-aware multimodal prompt for segmentation. Working together, the dual-memory feature and the multimodal prompt provide a comprehensive target representation, enabling accurate and robust segmentation. To validate the effectiveness of our MK-FSS, we conduct experiments on PASCAL-5^i (Shaban et al., 2017) and COCO-20^i (Nguyen & Todorovic, 2019). MK-FSS achieves new state-of-the-art performance on both benchmarks, evidencing its efficacy.

We note several recent works (Zhu et al., 2024; Karimi & Poullis, 2025) have attempted to exploring LLMs for improving FSS. MK-FSS ***differs from*** these methods in two key aspects. First, the type of knowledge exploited for FSS is different. Existing methods either leverage LLM-generated polygon predictions (Zhu et al., 2024) or category-level textual descriptions (Karimi & Poullis, 2025), whereas our method extracts target-specific spatial and semantic knowledge directly from the query image, enabling the exploration of more discriminative information for segmentation. Second, the way in which LLM-derived knowledge is utilized is also different. Current approaches mainly use LLM-derived knowledge as direct segmentation guidance (Zhu et al., 2024) or combine it with visual prompts (Karimi & Poullis, 2025), while MK-FSS integrates MLLM-derived spatial and semantic knowledge for target representation learning, which achieves superior performance.

In summary, our contributions are summarized as follows: ♠ We propose MK-FSS, a novel framework that exploits complementary MLLM-derived spatial and semantic knowledge for FSS; ♥ We present a dual-memory debate-fusion (DMDF) module that integrates memory features derived from spatial knowledge and support information, yielding a more robust dual-memory representation. ♣ We propose a progressive cross-modal prompt generator (PCPG) that fuses textual feature encoded from semantic knowledge with multi-scale query features to produce a multimodal prompt for FSS. ♦ In extensive experiments, MK-FSS shows promising results, evidencing its effectiveness.

2 Related Work

Few-Shot Segmentation. Few-shot segmentation (FSS) (Shaban et al., 2017) aims to segment novel object categories using only a few annotated samples. Current approaches mainly follow prototype-based (Fan et al., 2022; Liu et al., 2022; Okazawa, 2022; Wang et al., 2024b; Zhang et al., 2021; Tian et al., 2020) or feature-matching-based (Hu et al., 2019; Zhang et al., 2022; Wang et al., 2020; Xie et al., 2021) paradigms. The former represents support information with prototypes to guide the segmentation of query images, while the latter exploits pixel-level support-query correspondences to achieve segmentation. Despite their effectiveness, both rely heavily on the limited support samples and may struggle with complex cases such as occlusions and visually similar distractors. Recent works incorporate foundation models (Zhang et al., 2024; Xu et al., 2025; Sun et al., 2024; Liu et al., 2024) or textual cues derived from category names (Nandam et al., 2025; Chen et al., 2024; Wang et al., 2024a; Yang et al., 2026) to provide additional target knowledge for FSS. Yet, such auxiliary information remains limited in resolving target ambiguity. More recently, Large Language Models (LLMs) have been explored to provide spatial localization cues (Zhu et al., 2024) or generate category-level textual descriptions (Karimi & Poullis, 2025) for FSS. Despite their improvements, these methods do not fully exploit query-specific target information. Different from them, MK-

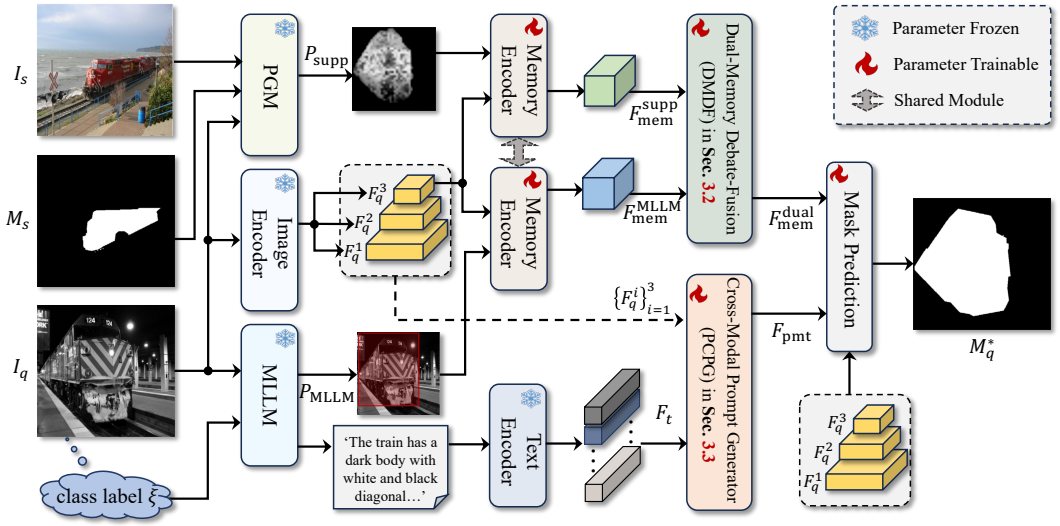


Figure 1: Overview of MK-FSS that exploits complementary spatial and semantic knowledge derived from a query image by an MLLM for few-shot segmentation.

FSS explicitly harnesses advanced reasoning capabilities of Multimodal Large Language Models (MLLMs) to derive complementary spatial and semantic knowledge from the query image, enabling more reliable target segmentation in complex scenes.

Multimodal Large Language Models. Multimodal Large Language Models (MLLMs) have shown remarkable capabilities in multimodal processing and reasoning in recent years (Brown et al., 2020; Liu et al., 2023; Bai et al., 2025a; Comanici et al., 2025). Given their extensive internal knowledge bases, numerous downstream visual tasks have leveraged these powerful models for auxiliary guidance (Kim et al., 2024; Dunlap et al., 2024; Lai et al., 2024; Jiang et al., 2025; Raza et al., 2025; Liu et al., 2025). For example, the work of SEAL (Wu & Xie, 2024) integrates an MLLM to provide informative visual cues for detection, while DeepEeyes (Zheng et al., 2026) leverages the inherent grounding capability of MLLMs to identify potential target regions. Different from these methods, our proposed MK-FSS specifically employ MLLM to generate spatial and semantic target-relevant knowledge from the query image and incorporate them into FSS through two dedicated modules.

3 Our Approach

Problem Definition FSS aims to segment targets in a query image I_q using only K (typically $K = 1$ or 5) annotated support samples $\{(I_s^i, M_s^i)\}_{i=1}^K$ from the same semantic category, where I_s^i and M_s^i denote the i-th support image and its corresponding mask. Following the standard episodic training protocol, each episode samples a target class ξ and contains a support set S and a query pair (I_q, M_q), where M_q is the groundtruth mask in I_q and used only for training supervision. The model is trained on base classes ξ_{train} and evaluated on disjoint novel classes ξ_{test}, with $\xi_{\text{train}} \cap \xi_{\text{test}} = \emptyset$. Following prior LLM/MLLM-based FSS methods (Karimi & Poullis, 2025; Zhu et al., 2024), we also use class label ξ to elicit target-related knowledge from the MLLM. Since the class label is readily available in standard FSS benchmarks, this process does not require any additional manual annotations.

Overview. In this work, we introduce MK-FSS, which exploits MLLM-derived target knowledge to improve FSS. As in Fig. 1, building upon SAM 2 (Ravi et al., 2025), MK-FSS first generates the target-specific knowledge from the query image using MLLM, and then leverages such knowledge to enhance target representation (§3.1). Specifically, we first encode the spatial knowledge into a memory representation, and then integrate it with the support-guided memory feature derived from the query image through our DMDF, yielding a more robust dual-memory feature (§3.2). Meanwhile, the semantic knowledge is encoded into a textual representation, which is progressively injected into query features by PCPG to produce a multimodal prompt (§3.3). The resulting dual-memory feature and the multimodal prompt are jointly used for mask prediction, enabling more robust segmentation.

3.1 Our MK-FSS Framework

In the following, we detail our MK-FSS by illustrating how target-specific knowledge derived from MLLM is leveraged to enhance few-shot segmentation.

3.1.1 Target Knowledge Generation

Our MK-FSS exploits two forms of target knowledge from an MLLM, including spatial knowledge and semantic knowledge, to improve FSS. The former provides spatial cues about the target location in the query image, while the latter captures high-level semantic information of the target. Together, they offer complementary information for robust few-shot segmentation.

Spatial Knowledge. Specifically, to generate spatial knowledge, we send the query image I_q and class label ξ to MLLM, and prompt it to identify all potential target regions and output their bounding box coordinates. The predicted boxes are then post-processed and mapped onto the query image to construct an MLLM-derived binary spatial prior $P_{\text{MLLM}} \in \{0,1\}^{H\times W}$, where H and W denote the height and width of the query image, highlighting the potential target regions in the query image. Let $\{b_i\}_{i=1}^{n}$ denote the n bounding boxes predicted by MLLM. Then, P_{MLLM} is defined as

$$P_{\text{MLLM}}(x,y) = \begin{cases} 1, & \text{if } (x,y) \in \bigcup_{i=1}^{n} b_n, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where (x, y) denotes a pixel location in the query image.

Semantic Knowledge. For semantic knowledge generation, we further prompt the MLLM with I_q and ξ to generate a textual description of the target, including its appearance and contextual cues. The generated description is encoded by a text encoder to produce a holistic sentence embedding $F_s \in \mathbb{R}^{1\times C}$ and word-level embedding $F_w \in \mathbb{R}^{L\times C}$, where L is the embedding sequence length and C denotes the feature dimension. These two representations are concatenated to form a unified textual feature $F_t = \texttt{Concat}(F_s, F_w) \in \mathbb{R}^{(L+1)\times C}$ for subsequent multimodal prompt generation.

Due to space limitations, please refer to the ***appendix*** for prompt templates used by the MLLM and representative examples of the generated spatial and semantic knowledge from the query image.

3.1.2 Few-Shot Segmentation with Target Knowledge

Given the target knowledge generated above, MK-FSS leverages it to improve FSS. Specifically, the spatial knowledge is encoded into the memory feature and integrated with a support-guided memory to construct a robust target memory, while the semantic knowledge is encoded into a textual feature and fused with multi-scale query features to generate a multimodal prompt for FSS.

Specifically, given a support set $\{(I_s^i, M_s^i)\}_{i=1}^{K}$ and the query image I_q, we first use a prior generation module (PGM), following (Xu et al., 2025), to generate the support-guided spatial prior. For clarity, we take the one-shot setting, *i.e.*, $K = 1$, as the example for illustration. Concretely, within PGM, a vision foundation model $\Phi_v(\cdot)$ (Oquab et al., 2024) is first used to extract query feature $\tilde{F}_q = \Phi_v(I_q)$ and support feature $\tilde{F}_s = \Phi_v(I_s)$. Afterwards, based on the support mask M_s, $\tilde{F}_s$ is compressed into foreground prototype F_{fg} and background prototype F_{bg} using global average pooling via $F_{\text{fg}} = \texttt{GAP}(\tilde{F}_s \odot M_s)$ and $F_{\text{bg}} = \texttt{GAP}(\tilde{F}_s \odot (1 - M_s))$, where $\texttt{GAP}(\cdot)$ denotes global average pooling and $\odot$ is element-wise multiplication. Finally, we compute the cosine similarity between the query feature $\tilde{F}_q$ and the two prototypes F_{fg} and F_{bg}, and obtain a non-negative support-guided spatial prior P_{supp} by subtracting the background similarity from the foreground similarity, as follows,

$$P_{\text{supp}} = \texttt{max}(\texttt{CosSim}(F_{\text{fg}}, \tilde{F}_q) - \texttt{CosSim}(F_{\text{bg}}, \tilde{F}_q), 0) \tag{2}$$

where $\texttt{CosSim}(\cdot,\cdot)$ is the cosine similarity function. Due to space limitation, the detailed architecture of PGM (Xu et al., 2025) is shown in the ***appendix***.

Once obtaining the support-guided spatial prior P_{supp}, we use it to generate the support-guided memory feature from the query image using memory encoder of SAM 2 (Ravi et al., 2025). Specifically, the query image I_q is first sent to the image encoder of SAM 2 to obtain multi-scale query features $\{F_q^1, F_q^2, F_q^3\}$, where F_q^1 to F_q^3 correspond to progressively deeper feature levels. Then, we use F_q^3

and P_{supp} to generate the support-guided memory $F_{\text{mem}}^{\text{supp}}$, as follows,

$$F_{\text{mem}}^{\text{supp}} = \texttt{MemEnc}(F_q^3, P_{\text{supp}}) \tag{3}$$

where $\texttt{MemEnc}(\cdot,\cdot)$ is the memory encoder of SAM 2. Because of large appearance variations, using $F_{\text{mem}}^{\text{supp}}$ alone may be insufficient to accurately segment the target due to lacking target knowledge. To mitigate this, we introduce an additional memory derived from the spatial knowledge provided by the MLLM and integrate it with F_{mem}^{s} for a more reliable memory. Specifically, given the MLLM-derived spatial prior P_{MLLM} and F_q^3, we generate the MLLM-derived memory $F_{\text{mem}}^{\text{MLLM}}$ as follows,

$$F_{\text{mem}}^{\text{MLLM}} = \texttt{MemEnc}(F_q^3, P_{\text{MLLM}}) \tag{4}$$

After obtaining $F_{\text{mem}}^{\text{supp}}$ and $F_{\text{mem}}^{\text{MLLM}}$, we integrate them with DMDF (explained in Sec. 3.2) via

$$F_{\text{mem}}^{\text{dual}} = \texttt{DMDF}(F_{\text{mem}}^{\text{MLLM}}, F_{\text{mem}}^{\text{supp}}) \tag{5}$$

where $F_{\text{mem}}^{\text{dual}}$ is the resulting dual-memory target feature.

Besides the spatial knowledge, we further employ the semantic knowledge derived from the MLLM to provide high-level target features for FSS, which is beneficial for disambiguating visually similar regions in the query image. Specifically, given multi-scale query features $\{F_q^1, F_q^2, F_q^3\}$ and textual feature F_t encoded from the semantic knowledge (see Sec. 3.1.1), we fuse them via PCPG (explained in Sec. 3.3) to generate a multimodal prompt for segmentation as follows,

$$F_{\text{pmt}} = \texttt{PCPG}(F_q^1, F_q^2, F_q^3, F_t) \tag{6}$$

where F_{pmt} is the learned target-aware multimodal prompt for segmentation.

Finally, we combine the dual-memory feature $F_{\text{mem}}^{\text{dual}}$ and multimodal prompt F_{pmt} to jointly guide the mask decoder for final target segmentation, as follows,

$$M_q^* = \texttt{MaskDec}(F_q^1, F_q^2, \hat{F}_q^3) \quad \hat{F}_q^3 = \texttt{MemAtt}(F_q^3 + F_{\text{pmt}}, F_{\text{mem}}^{\text{dual}}) \tag{7}$$

where $\texttt{MemAtt}(\cdot,\cdot)$ and $\texttt{MaskDec}(\cdot,\cdot,\cdot)$ are the memory attention and mask decoder of SAM 2, and M_q^* is the predicted segmentation mask in the query image I_q.

3.2 Dual-Memory Debate-Fusion

The recent LLM systems have applied the Multi-Agent Debate (MAD) mechanism (Chen et al., 2025; Chan et al., 2024) to boost the reasoning reliability. In these systems, multiple LLM agents discuss the shared question and update their answers based on peer responses, yielding more robust reasoning and improved outcomes. Inspired by this, we introduce a novel dual-memory debate-fusion (DMDF) module that mutually enhances the support-guided and MLLM-derived memories, producing a more robust dual-memory feature. Analogous to the MAD process in LLM systems, DMDF operates in three stages, including debate, verification and integration, as described in the following.

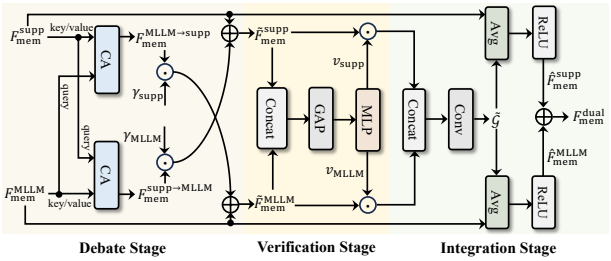


Figure 2: Overview of DMDF.

Debate Stage. In debate stage, each memory feature attends to the other to extract complementary information, enabling bilateral information exchange analogous to the debate between two agents. Specifically, given the support-guided memory feature $F_{\text{mem}}^{\text{supp}}$ and the MLLM-derived memory feature $F_{\text{mem}}^{\text{MLLM}}$, we apply a bidirectional cross-attention mechanism to facilitate mutual feature refinement. As illustrated in Fig. 2, this process can be expressed as follows,

$$F_{\text{mem}}^{\text{MLLM}\to\text{supp}} = \texttt{CA}(F_{\text{mem}}^{\text{supp}}, F_{\text{mem}}^{\text{MLLM}}), \quad F_{\text{mem}}^{\text{supp}\to\text{MLLM}} = \texttt{CA}(F_{\text{mem}}^{\text{MLLM}}, F_{\text{mem}}^{\text{supp}}) \tag{8}$$

where $F_{\text{mem}}^{\text{MLLM}\to\text{supp}}$ and $F_{\text{mem}}^{\text{supp}\to\text{MLLM}}$ are refined memory features, and $\texttt{CA}(\mathbf{z}, \mathbf{u})$ represents the cross-attention operation with $\mathbf{z}$ generating query and $\mathbf{u}$ key/value. The resulting refined features are then fused with the original memory features to incorporate complementary information from each other,

yielding the updated memory features as follows,

$$\tilde{F}_{\text{mem}}^{\text{supp}} = F_{\text{mem}}^{\text{supp}} + \gamma_{\text{supp}} \cdot F_{\text{mem}}^{\text{MLLM}\to\text{supp}} \quad \tilde{F}_{\text{mem}}^{\text{MLLM}} = F_{\text{mem}}^{\text{MLLM}} + \gamma_{\text{MLLM}} \cdot F_{\text{mem}}^{\text{supp}\to\text{MLLM}} \tag{9}$$

where $\tilde{F}_{\text{mem}}^{\text{supp}}$ and $\tilde{F}_{\text{mem}}^{\text{MLLM}}$ denotes updated memory features, and γ_{supp} and γ_{MLLM} are learnable scaling factors that control the memory fusion in the debate stage.

Verification Stage. After obtaining the updated memories, we estimate channel-wise importance weights for two memories, enabling subsequent fusion to emphasize more informative cues. Specifically, the two memory features are globally pooled and concatenated along the channel dimension. Then, the concatenated feature is fed into a multilayer perceptron (MLP) to generate the reliability weights for the two memories, as follows,

$$v_{\text{supp}}, v_{\text{MLLM}} = \texttt{MLP}(\texttt{Concat}(\texttt{GAP}(\tilde{F}_{\text{mem}}^{\text{supp}}), \texttt{GAP}(\tilde{F}_{\text{mem}}^{\text{MLLM}}))) \tag{10}$$

where v_{supp} and v_{MLLM} are reliability weights of the support-guided and MLLM-derived memories.

Integration Stage. In the integration stage, we fuse the updated memories according to their reliability weights to produce the final dual-memory feature. Specifically, each updated memory is first weighted by its corresponding reliability weight, and the weighted memories are concatenated and passed through a convolution layer to obtain a combined representation $\tilde{\mathcal{G}}$ via

$$\tilde{\mathcal{G}} = \texttt{Conv}(\texttt{Concat}(\tilde{F}_{\text{mem}}^{\text{supp}} \odot v_{\text{supp}}, \tilde{F}_{\text{mem}}^{\text{MLLM}} \odot v_{\text{MLLM}})) \tag{11}$$

where $\texttt{Conv}(\cdot)$ denotes the convolution operation. After this, the combined representation $\tilde{\mathcal{G}}$ is fused with each original memory feature by averaging, followed by a ReLU activation to obtain the final memory features, as follows,

$$\hat{F}_{\text{mem}}^{\text{supp}} = \texttt{ReLU}(\texttt{Avg}(\tilde{\mathcal{G}}, F_{\text{mem}}^{\text{supp}})) \quad \hat{F}_{\text{mem}}^{\text{MLLM}} = \texttt{ReLU}(\texttt{Avg}(\tilde{\mathcal{G}}, F_{\text{mem}}^{\text{MLLM}})) \tag{12}$$

where $\texttt{Avg}(\cdot)$ is the averaging operation and $\texttt{ReLU}(\cdot)$ denotes the ReLU activation. The final dual-memory target feature $F_{\text{mem}}^{\text{dual}}$ is then obtained via

$$F_{\text{mem}}^{\text{dual}} = \hat{F}_{\text{mem}}^{\text{supp}} + \hat{F}_{\text{mem}}^{\text{MLLM}} \tag{13}$$

By effectively incorporating support-guided and MLLM-derived memories, $F_{\text{mem}}^{\text{dual}}$ provides a more reliable target representation for target segmentation.

3.3 Progressive Cross-modal Prompt Generator

To further exploit the semantic knowledge, we propose a progressive cross-modal prompt generator (PCPG) that fuses multi-scale query features with the textual feature to produce a multimodal prompt for segmentation. As illustrated in Fig. 3, PCPG first applies a cross-modal fusion (CMF) module (as described later) to each query feature F_q^i, where $i \in \{1, 2, 3\}$, together with the textual feature F_t, to generate cross-modal features F_{vl}^i as follows,

$$F_{vl}^i = \texttt{CMF}(F_q^i, F_t) \quad i \in \{1, 2, 3\} \tag{14}$$

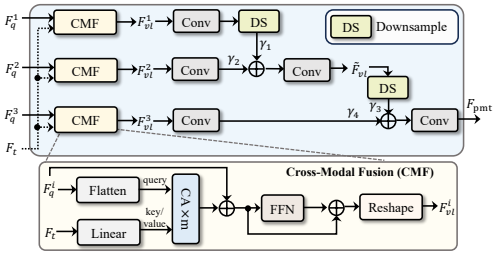


Figure 3: Overview of PCPG.

The resulting multi-scale cross-modal features F_{vl}^i are then progressively aggregated from the shallow to deep levels. Specifically, F_{vl}^1 and F_{vl}^2 are first fed into separate convolutional layers, after which the former is downsampled and combined with the latter using learnable weights. The fused feature is further processed by a convolutional layer to obtain the intermediate representation $\tilde{F}_{vl}$. Afterwards, $\tilde{F}_{vl}$ is downsampled and combined with the convolutionally processed F_{vl}^3 using learnable weights, followed by another convolutional layer to generate the multimodal prompt. Mathematically, this process can be expressed as follows,

$$\begin{aligned} \tilde{F}_{vl} &= \texttt{Conv}(\texttt{DS}\left(\gamma_1 \cdot \texttt{Conv}(F_{vl}^1)\right) + \gamma_2 \cdot \texttt{Conv}(F_{vl}^2)) \\ F_{\text{pmt}} &= \texttt{Conv}(\gamma_3 \cdot \texttt{DS}(\tilde{F}_{vl}) + \gamma_4 \cdot \texttt{Conv}(F_{vl}^3)) \end{aligned} \tag{15}$$

where F_{pmt} is the generated multimodal prompt feature, $\{\gamma_i\}_{i=1}^{4}$ are learnable parameters, and $\texttt{DS}(\cdot)$ represents the downsampling operation.

Cross-Modal Fusion. The cross-modal fusion (CMF) module aims to fuse the textual feature F_t with the query feature F_q^i. Specifically, F_q^i is first flattened into visual tokens, while F_t is linearly projected to match its channel dimension. We then apply m (m is empirically set to 2) consecutive cross-attention layers to enable interaction between F_q^i and F_t, followed by a residual connect. The resulting fused feature is further processed by an FFN with another residual connection and reshaped to obtain the cross-modal feature F_{vl}^i. Fig. 3 shows the architecture of CMF.

3.4 K-SHOT EXTENSION AND OPTIMIZATION

Extending to K-shot Setting. In K-shot FSS, each support sample is applied to generate a support-guided prior mask. We average all these masks to obtain the final support-guided spatial prior P_{supp}. Once P_{supp} is obtained, the subsequent FSS procedure is identical to that in the one-shot setting.

Optimization. Given a query image and a support set, MK-FSS predicts a foreground segmentation mask M_q^*. During the training process, the model is optimized using the standard Dice loss (Milletari et al., 2016) between M_q^* and the target groundtruth mask.

4 EXPERIMENTS

Implementation. MK-FSS is implemented in Python using PyTorch (Paszke et al., 2019) and built upon the pretrained SAM 2-small model (Ravi et al., 2025). The memory attention and mask decoder modules are fine-tuned, while the newly introduced modules are trained from scratch. In MK-FSS, we employ Qwen3-VL-8B-Instruct (Bai et al., 2025a) as the MLLM. To preserve localization cues, we post-process the bounding boxes predicted by the MLLM. Specifically, the invalid outputs, such as empty or malformed boxes, are replaced with a box that covers the entire image; otherwise, the predicted boxes are utilized as the spatial prior. We employ RoBERTa (Liu et al., 2019) as the text encoder. For PGM, DINOv2-B (Oquab et al., 2024) is applied for visual feature extraction. During training, images are randomly cropped to 512×512, and predictions are resized to their original resolution for evaluation. The model is trained using AdamW (Loshchilov & Hutter, 2019) with an initial learning rate of 0.001 for 150 and 100 epochs on PASCAL-5^i and COCO-20^i, respectively. All models are trained on three NVIDIA A100 (40G) GPUs and tested on a single A100 (40G) GPU.

4.1 DATASETS AND EVALUATION METRICS.

Datasets. We evaluate MK-FSS on PASCAL-5^i (Shaban et al., 2017) and COCO-20^i (Nguyen & Todorovic, 2019). PASCAL-5^i includes 20 classes, built upon the PASCAL VOC 2012 (Everingham et al., 2010) with additional annotations from SDS (Hariharan et al., 2014). COCO-20^i (Lin et al., 2014) is larger in scale and comprises 80 classes. Both datasets are divided into 4 disjoint folds for cross-validation, with each fold containing 5 classes for PASCAL-5^i and 20 classes for COCO-20^i. In each iteration, 3 folds are used for training, and the remaining fold is used for testing. Following existing works, we randomly sample 1,000 episodes for testing by default. Due to limited space, the results for error bars evaluation and different number of testing episodes are included in ***appendix***.

Evaluation Metrics. Following prior works, we employ mean intersection-over-union (mIoU) and foreground-background IoU (FB-IoU) as evaluation metrics. Specifically, mIoU averages IoU across foreground classes within each fold, while FB-IoU treats all foreground classes as a single category and measures binary foreground-background segmentation performance.

4.2 STATE-OF-THE-ART COMPARISON

Tables 1 and 2 summarize the comparisons on PASCAL-5^i and COCO-20^i, respectively. We evaluate our proposed MK-FSS against recent state-of-the-art non-LLM/MLLM-based and LLM/MLLM-based FSS methods under standard 1-shot and 5-shot settings. As demonstrated, MK-FSS establishes new state-of-the-art performance. Specifically, on PASCAL-5^i, it achieves mIoU scores of 87.1% (1-shot) and 87.7% (5-shot), surpassing all prior methods. On the more challenging COCO-20^i benchmark, MK-FSS maintains its superiority, delivering highly competitive mIoU scores of

Table 1: Comparisons with state-of-the-arts on PASCAL-5^i. Best results are highlighted in **bold**.

Method	1-shot						5-shot					
	5^0	5^1	5^2	5^3	Mean	FB-IoU	5^0	5^1	5^2	5^3	Mean	FB-IoU
Non-LLM/MLLM-Based Few-Shot Segmentation												
MSI (Moon et al., 2023) [ICCV'23]	73.1	73.9	64.7	68.8	70.1	82.3	73.6	76.1	68.0	71.3	72.2	82.3
PI-CLIP (Wang et al., 2024a) [CVPR'24]	76.4	83.5	74.7	72.8	76.8	-	76.7	83.8	75.2	73.2	77.2	-
AENet (Xu et al., 2024) [ECCV'24]	72.2	75.5	68.5	63.1	69.8	80.8	74.2	76.5	74.8	70.6	74.1	84.5
GF-SAM (Zhang et al., 2024) [NeurIPS'24]	71.1	75.7	69.2	73.3	72.1	-	81.5	86.3	79.7	82.9	82.6	-
OCNet (Wen et al., 2025) [ICCV'25]	73.5	75.9	71.1	64.9	71.4	82.2	75.9	77.1	74.1	70.9	74.5	84.7
PAHNet (Zou et al., 2025) [ICCV'25]	72.1	76.4	71.0	66.6	71.5	82.4	75.8	78.6	76.9	73.2	76.1	85.7
DeFSS (Qin et al., 2025) [ICCV'25]	73.7	76.6	70.9	65.8	71.7	-	75.0	78.6	74.7	72.2	75.1	-
FSSAM (Xu et al., 2025) [ICML'25]	81.6	84.9	81.6	76.0	81.0	89.4	84.1	88.5	83.8	85.0	85.4	91.9
LLM/MLLM-based Few-Shot Segmentation												
LLaFS (Zhu et al., 2024) [CVPR'24]	74.2	78.8	72.3	68.5	73.5	-	75.9	80.1	75.8	70.7	75.6	-
DSV-LFS (Karimi & Poullis, 2025) [CVPR'25]	71.7	82.0	71.2	75.0	75.0	-	72.0	82.0	71.3	75.5	75.2	-
MLLM+SAM 2 (our baseline)	74.7	79.2	73.5	75.8	75.8	86.9	74.7	79.3	72.7	75.3	75.5	86.1
MK-FSS (ours)	**85.0**	**89.3**	**86.1**	**87.8**	**87.1**	**93.0**	**85.2**	**90.8**	**85.8**	**88.9**	**87.7**	**93.3**

Table 2: Comparisons with state-of-the-arts on COCO-20^i. Best results are highlighted in **bold**.

Method	1-shot						5-shot					
	20^0	20^1	20^2	20^3	Mean	FB-IoU	20^0	20^1	20^2	20^3	Mean	FB-IoU
Non-LLM/MLLM-Based Few-Shot Segmentation												
MSI (Moon et al., 2023) [ICCV'23]	44.8	54.2	52.3	48.0	49.8	-	49.3	58.0	56.1	52.7	54.0	-
PI-CLIP (Wang et al., 2024a) [CVPR'24]	49.3	65.7	55.8	56.3	56.8	-	56.4	66.2	55.9	58.0	59.1	-
AENet (Xu et al., 2024) [ECCV'24]	43.1	56.0	50.3	48.4	49.4	73.6	51.7	61.9	57.9	55.3	56.7	76.5
GF-SAM (Zhang et al., 2024) [NeurIPS'24]	56.6	61.4	59.6	57.1	58.7	-	67.1	69.4	66.0	64.8	66.8	-
OCNet (Wen et al., 2025) [ICCV'25]	45.9	56.9	52.9	50.4	51.5	73.7	52.7	63.1	57.4	54.8	57.0	76.8
PAHNet (Zou et al., 2025) [ICCV'25]	45.2	58.2	54.4	51.2	52.3	75.2	53.0	64.1	60.8	58.7	59.2	78.1
DeFSS (Qin et al., 2025) [ICCV'25]	46.5	59.7	54.8	53.2	53.7	-	55.1	66.0	61.3	57.5	60.0	-
FSSAM (Xu et al., 2025) [ICML'25]	59.9	65.6	62.1	61.6	62.3	77.3	68.6	74.0	64.5	69.9	69.3	82.9
LLM/MLLM-based Few-Shot Segmentation												
LLaFS (Zhu et al., 2024) [CVPR'24]	47.5	58.8	56.2	53.0	53.9	-	53.2	63.8	63.1	60.0	60.0	-
DSV-LFS (Karimi & Poullis, 2025) [CVPR'25]	70.0	73.3	70.7	71.3	71.3	-	71.0	73.8	71.3	71.4	71.9	-
MLLM+SAM 2 (our baseline)	64.3	64.2	68.3	61.1	64.5	78.7	65.9	63.3	65.8	61.6	64.2	79.0
MK-FSS (ours)	**73.2**	**76.6**	**75.9**	**71.6**	**74.3**	**84.8**	**75.8**	**77.1**	**76.4**	**74.3**	**75.9**	**86.1**

74.3% and 75.9% in the 1-shot and 5-shot settings, significantly outperforming existing models. Furthermore, we construct a naive baseline, denoted as MLLM+SAM 2, which directly feeds MLLM-predicted bounding boxes as spatial prompts to SAM 2. The sub-optimal performance of this variant reveals that while MLLMs possess basic detection capabilities, their raw bounding boxes lack the precision strictly required for pixel-level segmentation. This performance gap strongly validates the necessity and effectiveness of our proposed framework.

4.3 Ablation Study

To better understand MK-FSS, we conduct ablation studies in the following. Unless explicitly specified, the ablation experiments are conducted on PASCAL-5^i under the 1-shot setting.

Table 3: Ablation on DMDF and PCPG.

	DMDF	PCPG	5^0	5^1	5^2	5^3	Mean
❶	-	-	79.3	81.1	75.9	72.7	77.3
❷	✓	-	84.8	88.2	84.1	87.0	86.0
❸	-	✓	80.3	82.7	77.9	77.2	80.0
❹	✓	✓	**85.0**	**89.3**	**86.1**	**87.8**	**87.1**

Ablation study on DMDF and PCPG in MK-FSS. DMDF and PCPG are two crucial modules in MK-FSS for incorporating MLLM-derived knowledge. To better understand them, we conduct ablation experiments in Tab. 3. We begin with a model that uses only the support-guided memory feature for FSS (❶), achieving a mean score of 77.3%. Using DMDF alone to incorporate MLLM-derived spatial knowledge improves the mean score to 86.0% (❷), while PCPG alone for semantic knowledge raises it to 80.0% (❸). These results demonstrate the effectiveness of both spatial and semantic knowledge for FSS. When combining DMDF and PCPG, the model achieves the best mean score of 87.1% (❹), showing that the two types of knowledge provide complementary benefits and jointly improve FSS.

Ablation on different memories. In MK-FSS, we integrate support-guided memory and MLLM-derived memories through DMDF for segmentation. To analyze the impact of different memories on MK-FSS, we conduct an ablation in Tab. 4. As shown in the table, using either $F_{\text{mem}}^{\text{supp}}$ (❶) or $F_{\text{mem}}^{\text{MLLM}}$ (❷) alone leads to suboptimal performance compared to their combination via DMDF, indicating

Table 4: Ablation on different memories.

	$F_{\text{mem}}^{\text{supp}}$	$F_{\text{mem}}^{\text{MLLM}}$	5^0	5^1	5^2	5^3	Mean
❶	✓	-	80.3	82.7	77.9	77.2	80.0
❷	-	✓	84.1	87.8	85.9	85.2	85.8
❸	✓	✓	**85.0**	**89.3**	**86.1**	**87.8**	**87.1**

Table 5: Ablation on stages in DMDF.

	Deb.	Ver.& Int.	5^0	5^1	5^2	5^3	Mean
❶	✓	-	84.7	88.8	85.7	87.0	86.6
❷	-	✓	84.7	88.9	85.3	87.3	86.6
❸	✓	✓	**85.0**	**89.3**	**86.1**	**87.8**	**87.1**

Table 6: Ablation on the comparison of DMDF with other fusion strategies.

	Memory fusion	5^0	5^1	5^2	5^3	Mean
❶	Learnable addition	84.9	88.6	85.6	86.9	86.5
❷	Concatenation	84.5	88.4	85.1	86.0	86.0
❸	DMDF	**85.0**	**89.3**	**86.1**	**87.8**	**87.1**

Table 7: Ablation on fusion method of the multi-scale cross-modal features of PCPG.

	Multi-scale fusion	5^0	5^1	5^2	5^3	Mean
❶	w/o fusion	84.8	88.8	85.7	86.7	86.5
❷	Learnable addition	**85.0**	88.8	85.4	87.7	86.7
❸	PCPG	**85.0**	**89.3**	**86.1**	**87.8**	**87.1**

that the two memories provide complementary target guidance. By integrating both memories, MK-FSS effectively leverages their respective strengths and achieves superior performance (❸).

Ablation on stages in DMDF. DMDF contains three stages of debate, verification, and integration for memory fusion. To analyze the contribution of each stage in DMDF, we compare the full model (❸) with partial stages in Tab. 5, including a setting using only the debate (Deb.) stage with simple addition for fusion (❶) and another setting using only the verification and integration (Ver.&Int.) stages (❷). The results show that the three stages collaborate effectively, and combining them is crucial for improving segmentation accuracy.

Comparison of DMDF with other fusion strategies. To evaluation the effectiveness of our DMDF, we compare it with two alternative strategies in Tab. 6. The first strategy, named learnable addition, combines $F_{\text{mem}}^{\text{supp}}$ and $F_{\text{mem}}^{\text{MLLM}}$ through addition with learnable memory weights (❶), while the second concatenates two memory features along the channel dimension and fuses them using an MLP (❷). Compared with these alternatives, our DMDF (❸) achieves the best performance across all evaluation splits, demonstrating its ability to more effectively exploit the complementary information between support-guided memory and MLLM-derived memory.

Effectiveness of the PCPG Design. To evaluate the effectiveness of PCPG, we compare it with two alternative strategies in Tab. 7, one relying only on the deepest textual feature F_q^3 (w/o fusion) (❶) and the other aggregating multi-scale features $\{F_q^i\}_{i=1}^3$ through addition with learnable weights (❷). Our progressive fusion strategy (❸) consistently outperforms both the single-level and direct fusion variants, demonstrating that progressively integrating hierarchical features is crucial for constructing a robust multimodal prompt that captures both local details and global semantic cues.

Ablation on text embedding. In MK-FSS, the extracted semantic knowledge is encoded to features that is composed of both sentence embedding F_s with word-level embedding F_w. We compare different text embedding strategies in Tab. 8. The results show that the combination of F_s and F_w achieves superior performance (❸), validating the efficacy of our design.

Table 8: Ablation on text embedding.

	F_s	F_w	5^0	5^1	5^2	5^3	Mean
❶	✓	-	84.8	88.2	85.4	87.4	86.4
❷	-	✓	84.8	88.5	85.4	87.3	86.5
❸	✓	✓	**85.0**	**89.3**	**86.1**	**87.8**	**87.1**

Due to limited space, we show more results and analysis in the ***appendix***.

5 Conclusion

In this paper, we present MK-FSS, a novel framework that exploits both spatial and semantic knowledge derived by MLLMs to enhance FSS through carefully designed DMDF and PCPG modules. Specifically, DMDF incorporates MLLM-derived spatial knowledge into the memory representation by effectively fusing it with the support-guided memory, leading to a more robust dual-memory feature. Meanwhile, PCPG leverages semantic knowledge to enrich visual representations and generate a multimodal prompt that provides additional guidance for target segmentation. By jointly exploiting these two forms of knowledge, MK-FSS captures complementary target cues beyond those available from the limited support samples alone. Extensive experiments on multiple FSS benchmarks demonstrate that MK-FSS consistently achieves strong performance and generalization, validating the effectiveness of incorporating MLLM-derived spatial and semantic knowledge for FSS.

AI USE STATEMENT

In this work, the generative AI tools are used *solely* for language polishing and readability improvement. We do *not* use the generative AI to generate datasets, produce experimental results, formulate mathematical claims, design experiments, implement methods, analyze results, or conduct the final interpretation of the findings. All AI-assisted text has been carefully reviewed, revised, and verified by the authors, who take full responsibility for the final content of this work.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv*, 2025b.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *ICLR*, 2024.

Shuai Chen, Fanman Meng, Runtong Zhang, Heqian Qiu, Hongliang Li, Qingbo Wu, and Linfeng Xu. Visual and textual prior guided mask assemble for few-shot segmentation and beyond. *IEEE TMM*, 26:7197–7209, 2024.

Xi Chen, Mao Mao, Shuo Li, and Haotian Shangguan. Debate-feedback: A multi-agent framework for efficient legal judgment prediction. In *NAACL*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv*, 2025.

Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In *CVPR*, 2024.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. In *IJCV*, 2010.

Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. Self-support few-shot semantic segmentation. In *ECCV*, 2022.

Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *ECCV*, 2014.

Tao Hu, Pengwan Yang, Chiliang Zhang, Gang Yu, Yadong Mu, and Cees GM Snoek. Attention-based multi-context guiding for few-shot semantic segmentation. In *AAAI*, 2019.

Wen Jiang, Boshu Lei, Katrina Ashton, and Kostas Daniilidis. Multimodal llm guided exploration and active mapping using fisher information. In *ICCV*, 2025.

Amin Karimi and Charalambos Poullis. Dsv-lfs: Unifying llm-driven semantic cues with visual features for robust few-shot segmentation. In *CVPR*, 2025.

Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, and Chanyoung Park. Llm4sgg: Large language models for weakly supervised scene graph generation. In *CVPR*, 2024.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Fang Liu, Yuhao Liu, Ke Xu, Shuquan Ye, Gerhard Petrus Hancke, and Rynson WH Lau. Language-guided salient object ranking. In *CVPR*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. In *ICLR*, 2024.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv*, 2019.

Yuanwei Liu, Nian Liu, Qinglong Cao, Xiwen Yao, Junwei Han, and Ling Shao. Learning non-target knowledge for few-shot semantic segmentation. In *CVPR*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *CVPR*, 2022.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016.

Seonghyeon Moon, Samuel S Sohn, Honglu Zhou, Sejong Yoon, Vladimir Pavlovic, Muhammad Haris Khan, and Mubbasir Kapadia. Msi: Maximize support-set information for few-shot segmentation. In *ICCV*, 2023.

Srinivasa Rao Nandam, Sara Atito, Zhenhua Feng, Josef Kittler, and Muhammad Awais. Text augmented correlation transformer for few-shot classification & segmentation. In *CVPR*, 2025.

Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *ICCV*, 2019.

Atsuro Okazawa. Interclass prototype relation for few-shot segmentation. In *ECCV*, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. In *TMLR*, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Zishu Qin, Junhao Xu, and Weifeng Ge. Defss: Image-to-mask denoising learning for few-shot segmentation. In *ICCV*, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *ICLR*, 2025.

Mubashar Raza, Zarmina Jahangir, Muhammad Bilal Riaz, Muhammad Jasim Saeed, and Muhammad Awais Sattar. Industrial applications of large language models. *Scientific Reports*, 15(1): 13755, 2025.

Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *BMVC*, 2017.

Yanpeng Sun, Jiahui Chen, Shan Zhang, Xinyu Zhang, Qiang Chen, Gang Zhang, Errui Ding, Jingdong Wang, and Zechao Li. Vrp-sam: Sam with visual reference prompt. In *CVPR*, 2024.

Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE TPAMI*, 44(2):1050–1065, 2020.

Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. In *ECCV*, 2020.

Jin Wang, Bingfeng Zhang, Jian Pang, Honglong Chen, and Weifeng Liu. Rethinking prior information generation with clip for few-shot segmentation. In *CVPR*, 2024a.

Jing Wang, Jiangyun Li, Chen Chen, Yisi Zhang, Haoran Shen, and Tianxiang Zhang. Adaptive fss: a novel few-shot segmentation framework via prototype enhancement. In *AAAI*, 2024b.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv*, 2025.

Yuan Wang, Rui Sun, and Tianzhu Zhang. Rethinking the correlation in few-shot segmentation: A buoys view. In *CVPR*, 2023.

Chunlin Wen, Yu Zhang, Jie Fan, Hongyuan Zhu, Xiu-Shen Wei, Yijun Wang, Zhiqiang Kou, and Shuzhou Sun. Object-level correlation for few-shot segmentation. In *ICCV*, 2025.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *CVPR*, 2024.

Guo-Sen Xie, Jie Liu, Huan Xiong, and Ling Shao. Scale-aware graph neural network for few-shot semantic segmentation. In *CVPR*, 2021.

Qianxiong Xu, Wenting Zhao, Guosheng Lin, and Cheng Long. Self-calibrated cross attention network for few-shot segmentation. In *ICCV*, 2023.

Qianxiong Xu, Guosheng Lin, Chen Change Loy, Cheng Long, Ziyue Li, and Rui Zhao. Eliminating feature ambiguity for few-shot segmentation. In *ECCV*, 2024.

Qianxiong Xu, Lanyun Zhu, Xuanyi Liu, Guosheng Lin, Cheng Long, Ziyue Li, and Rui Zhao. Unlocking the power of sam 2 for few-shot segmentation. In *ICML*, 2025.

Yi Yang, Qiang Jiao, Mengrui Shi, and Qiang Zhang. Towards complete activation: Foreground-background multi-perspective guided cross-support for few-shot segmentation. In *CVPR Findings*, 2026.

Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, et al. Mimo-vl technical report. *arXiv*, 2025.

Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi Liu, and Yunchao Wei. Bridge the points: Graph-based few-shot segment anything semantically. In *NeurIPS*, 2024.

Bingfeng Zhang, Jimin Xiao, and Terry Qin. Self-guided and cross-guided learning for few-shot segmentation. In *CVPR*, 2021.

Gengwei Zhang, Shant Navasardyan, Ling Chen, Yao Zhao, Yunchao Wei, Humphrey Shi, et al. Mask matching transformer for few-shot segmentation. In *NeurIPS*, 2022.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *ICLR*, 2026.

Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. Llafs: When large language models meet few-shot segmentation. In *CVPR*, 2024.

Tianyu Zou, Shengwu Xiong, Ruilin Yao, and Yi Rong. Balancing conservatism and aggressiveness: Prototype-affinity hybrid network for few-shot segmentation. In *ICCV*, 2025.

A APPENDIX

The appendix is structured as follows:

- **Section A.1**: We detail the exact prompt templates designed to instruct the MLLMs within our framework and show several examples of the generated spatial and semantic knowledge.
- **Section A.2**: We present the detailed architecture of the PGM.
- **Section A.3**: Visualization examples of P_{MLLM} and P_{supp} are provided.
- **Section A.4**: We report the error bar evaluation results under different random seeds to verify the stability and robustness of our model.
- **Section A.5**: We analyze the influence of different numbers of testing episodes on model performance on the COCO-20^i dataset.
- **Section A.6**: We present additional qualitative segmentation results and comparisons with other representative models.
- **Section A.7**: We then provide qualitative examples illustrating the effect of linguistic cues, showing how MLLM-derived semantic knowledge enhances visual understanding and disambiguates confusing backgrounds.
- **Section A.8**: We further examine the effect of cross-attention depth in the PCPG module.
- **Section A.9**: Additional results evaluated on alternative MLLMs are presented.
- **Section A.10**: We further evaluate the effectiveness of our target-specific language descriptions by comparing them with class-specific descriptions, validating the advantage of our fine-grained semantic information.

A.1 PROMPT DESIGN FOR MLLMS AND EXAMPLES OF GENERATED KNOWLEDGE

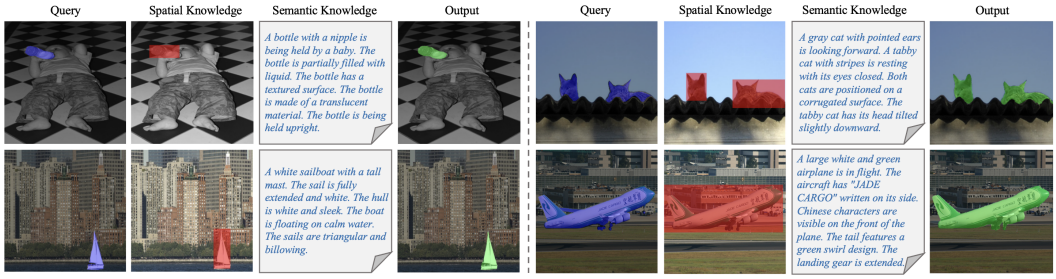


Figure 4: Visualization of MLLM-derived target knowledge and segmentation results on PASCAL-5^i under 1-shot setting. For each example, blue masks indicate the ground-truth target regions in query images, red masks indicate the MLLM-predicted spatial regions, and green masks indicate the final segmentation results of our proposed MK-FSS.

We carefully design structured prompts to guide the MLLM in generating precise and task-aligned spatial and semantic knowledge. For spatial knowledge generation, the goal is to obtain structured coordinates that can be directly parsed by downstream components. Specifically, the MLLM is prompted with: `Locate all the [class_name] in this tile accurately and completely. Respond ONLY with the bounding box locations in valid JSON format as follows: ["bbox_2d": [10, 20, 990, 980], "label": "[class_name]"]`. This prompt encourages the MLLM to localize all potential target instances and return valid bounding boxes. For semantic knowledge generation, we use an explicit restriction prompt to reduce generic semantic noise and fine-grained hallucinations: `Describe the key visible attributes of all [class_name] in the image, focusing only on their appearance and state. Use several short factual SENTENCES. Do not mention background objects or irrelevant details. Do not include any guesses about age, condition, history, usage, or purpose.` These explicit instructions constrain the MLLM to focus exclusively on instance-level visual evidence present in the query scene.

Several examples of the generated spatial and semantic knowledge are shown in Fig. 4. We can observe that the two types of knowledge provide complementary target cues. The spatial knowledge roughly localizes potential target regions, while the semantic knowledge describes discriminative visual attributes of the target. By integrating these complementary cues, MK-FSS is able to produce more accurate segmentation results in complicated and challenging scenes.

A.2 Detailed Architecture of PGM

We illustrate the prior generation module (PGM) in Fig. 5. PGM generates a support-guided prior map P_{supp} by measuring the visual correspondence between query features and support prototypes. Following (Xu et al., 2025), given a support image I_s, a query image I_q, and the corresponding support mask M_s, we first utilize a frozen DINOv2 encoder (Oquab et al., 2024) to extract the support feature $\tilde{F}_s$ and query feature $\tilde{F}_q$. The foreground and background prototypes are then obtained through the mask-guided global average pooling,

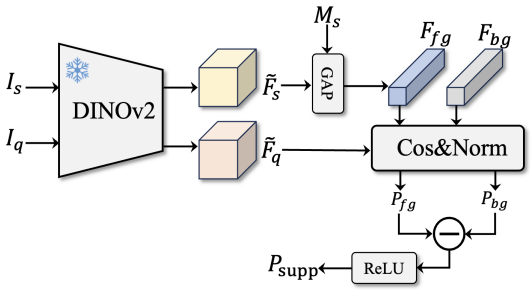


Figure 5: Architecture of the PGM.

$$F_{fg} = \texttt{GAP}(\tilde{F}_s \odot M_s), \quad F_{bg} = \texttt{GAP}(\tilde{F}_s \odot (1 - M_s)) \tag{16}$$

We then compute the normalized cosine similarity between the query feature and the two prototypes to obtain the foreground and background similarity maps,

$$P_{fg} = \texttt{Norm}(\texttt{CosSim}(\tilde{F}_q, F_{fg})), \quad P_{bg} = \texttt{Norm}(\texttt{CosSim}(\tilde{F}_q, F_{bg})) \tag{17}$$

where $\texttt{Norm}(\cdot)$ denotes min-max normalization. The final support-guided prior is obtained by contrasting the foreground and background similarity maps. Since this operation may introduce negative responses, we apply a ReLU activation to suppress them. This process can be expressed as follows,

$$P_{\text{supp}} = \texttt{ReLU}(P_{fg} - P_{bg}). \tag{18}$$

This contrastive design emphasizes query regions that are more similar to the support foreground while suppressing background-like responses. In the K-shot setting, PGM generates one prior map for each support image and averages them to obtain the final P_{supp}.

A.3 Visualization Examples of P_{MLLM} and P_{SUPP}

Fig. 6 visualizes the support-guided prior P_{supp} and the MLLM-derived prior P_{MLLM}. The support-guided prior tends to capture fine-grained visual correspondence, producing detailed activations around the target object. Nevertheless, it may be distracted by visually similar background regions and fail to highlight complete target regions. The MLLM-derived prior provides a more semantic localization cue, which helps identify the target object at a coarse level and suppress irrelevant regions. By combining these two priors, our method benefits from both detailed visual matching and high-level semantic guidance, leading to more complete and accurate segmentation results. These examples qualitatively validate the complementary roles of P_{supp} and P_{MLLM}.

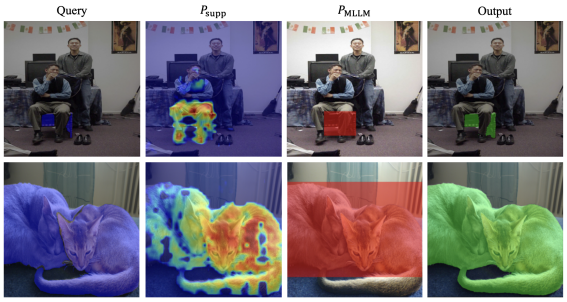


Figure 6: Visualization of P_{supp}, P_{MLLM}, and segmentation results on PASCAL-5^i. From left to right, each row shows the query image, support-guided prior rendered as a heatmap, MLLM-derived prior, and final prediction. Blue, red, and green masks indicate ground truth, MLLM-derived prior regions, and predicted masks, respectively.

A.4 Error Bars Evaluation

Tab. 9 shows the error bar evaluation results of our model on the PASCAL-5^i and COCO-20^i datasets under five different random seeds, each tested across four folds with 1,000 episodes. The results

Table 9: Error bars evaluation on PASCAL-5^i and COCO-20^i datasets under 1-shot setting. The number of testing episodes is 1,000. The selected random seeds are {0, 1, 2, 3, 321 (default)}. "Mean" is the averaged mIoU score of 4 folds.

Seed	PASCAL-5^i					COCO-20^i				
	5^0	5^1	5^2	5^3	Mean	20^0	20^1	20^2	20^3	Mean
0	84.8	89.0	85.3	87.2	86.6	71.9	77.5	73.8	69.4	73.2
1	85.0	89.2	86.2	87.7	87.0	69.1	78.4	73.0	73.3	73.5
2	84.3	88.5	84.4	87.8	86.3	66.6	76.3	73.9	72.7	72.4
3	84.3	89.1	86.3	87.5	86.8	71.9	75.5	70.9	73.9	73.1
321	85.0	89.3	86.1	87.8	87.1	73.2	76.6	75.9	71.6	74.3
Mean	84.7	89.0	85.7	87.6	86.8	70.5	76.8	73.5	72.2	73.3
Std	0.4	0.3	0.8	0.3	0.3	2.7	1.1	1.8	1.8	0.7

demonstrate that our model maintains high stability and robustness across different seeds. Specifically, on the PASCAL-5^i dataset, the mean mIoU remains consistently around 86.8% with a very small standard deviation of 0.3. And the fold-wise standard deviations are all around 0.5, confirming the model's consistent generalization ability across different folds. For the more challenging COCO-20^i dataset, although the mean mIoU is relatively lower (73.3%), the standard deviation remains small, showing stable performance despite higher data complexity. The fold-wise variations are also limited, with Fold-1 achieving the best results and Fold-0 showing slightly larger variance, which stems from higher intra-class diversity of COCO-20^i. Overall, the low standard deviations across both datasets validate the reliability of our MK-FSS and its robustness to random sampling.

A.5 Different Number of Testing Episodes

Since COCO-20^i is a challenging dataset, the evaluation results based on only 1,000 randomly sampled testing episodes may not be sufficiently convincing. Following (Xu et al., 2025), we further investigate the influence of different testing episode numbers on model performance. Specifically, we vary the number of testing episodes among 1,000, 4,000, 10,000, 20,000 under 1-shot setting and report the corresponding quantitative results in Tab. 10.

Table 10: Performance on COCO-20^i with different number of testing episodes in {1,000, 4,000, 10,000, 20,000}. The random seed is fixed as 321. "Mean" is the averaged mIoU score of 4 folds, "20^i" denotes the mIoU score of the i-th fold.

Testing Episodes	20^0	20^1	20^2	20^3	Mean
1,000	73.2	76.6	75.9	71.6	74.3
4,000	72.7	75.5	72.2	72.5	73.2
10,000	71.2	75.4	73.4	72.7	73.2
20,000	70.6	75.7	73.2	72.6	73.0
Mean	71.9	75.8	73.7	72.4	73.5

From the results, it can be observed that varying the number of testing episodes has a slight impact on the performance. For example, the mIoU achieved with 1,000 testing episodes is 74.3%, while the average mIoU across all four episode settings is 73.5%, yielding a marginal gap of only 0.8%. This demonstrates that our model achieves stable and reliable performance even with a relatively small number of testing episodes. Furthermore, the fold-wise variations remain consistent across different episode settings, indicating that the model generalizes well regardless of the number of sampled episodes. Overall, these results validate the robustness and consistency of our approach when evaluated under different testing conditions.

A.6 Qualitative Analysis

To further qualitatively validate the effectiveness of our proposed method, we provide visual segmentation results and compare them with representative methods (Zou et al., 2025; Zhang et al., 2024; Xu et al., 2025; Ravi et al., 2025; Karimi & Poullis, 2025), as shown in Fig. 7. It can be observed that existing methods often produce incomplete masks or include irrelevant background regions, especially in challenging cases with occlusions and visually similar distractors, e.g., the bottles in Fig. 7. In contrast, our method generates more accurate and complete target masks with clearer object boundaries. These results demonstrate that MK-FSS effectively exploits MLLM-derived spatial and semantic knowledge, which provides complementary localization and semantic cues to enhance the model's segmentation capability in complex scenarios.

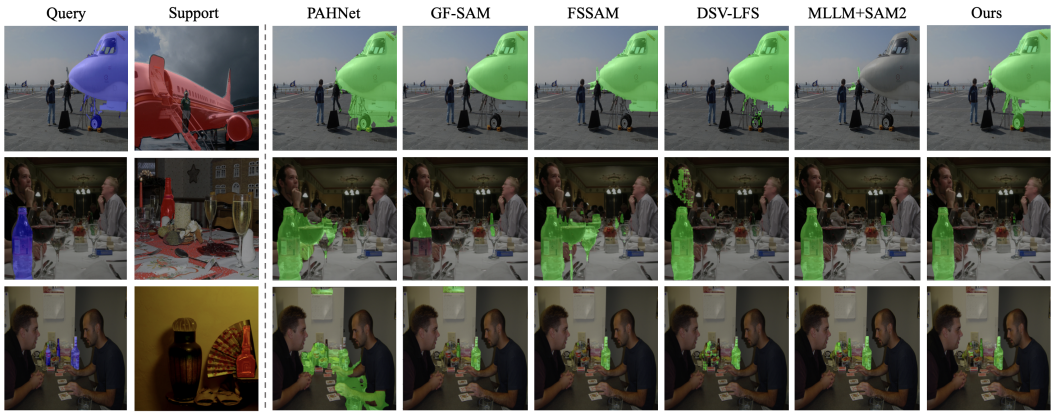


Figure 7: Visual comparison on PASCAL-5^i dataset under 1-shot setting. The ground-truth objects in query image are highlighted in blue, the referred objects in support image are highlighted in red, and the predicted objects in query image are highlighted in green.

A.7 Visualization of the Effect of Linguistic Cues

In Fig. 8, we present several examples illustrating the impact of incorporating high-level linguistic cues into purely visual features. From these examples, we can see that models relying solely on visual matching often struggle to handle complex scenarios, where the same object category shows significant variations in appearance across different images (e.g., the people and plant in Fig. 8). However, the integration of linguistic information provides additional semantic priors that offer explicit and discriminative guidance for the visual features. Such language-guided cues help the model focus on the semantically relevant regions and disambiguate visually similar background objects, leading to more accurate and robust segmentation results.

A.8 Impact of Cross-Attention Depth in PCPG

To examine the effect of cross-attention depth in the PCPG module, we vary the number of cross-attention layers from 1 to 3. The quantitative results, presented in Tab. 11, indicate that the 2-layer design (❷) yields the optimal balance and achieves the highest overall performance.

Table 11: Ablation on CA number in PCPG.

	CA layer number	5^0	5^1	5^2	5^3	Mean
❶	1	84.8	88.8	85.8	87.7	86.8
❷	2	**85.0**	**89.3**	**86.1**	**87.8**	**87.1**
❸	3	84.9	88.8	85.5	87.5	86.7

A.9 Analysis of Impact across MLLMs

To show that the performance gains are brought by the proposed architecture rather than tied to a specific MLLM, we evaluate several recent MLLMs within our framework. As shown in Tab. 12, our method consistently achieves strong performance across different MLLM choices. Notably, using a more advanced model, Qwen3-VL-8B (❺), further improves the mean mIoU to 87.1%. These results indicate that our framework can effectively leverage target knowledge from different MLLMs, and its performance can further benefit from the improved reasoning capabilities of newer models.

Table 12: Performance comparison under different MLLM configurations on PASCAL-5^i.

	MLLM	5^0	5^1	5^2	5^3	Mean
❶	Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	83.7	89.2	81.7	85.0	84.9
❷	Qwen2.5-VL-32B-Instruct (Bai et al., 2025b)	82.8	89.7	83.8	87.5	86.0
❸	MiMo-VL-7B-RL (Yue et al., 2025)	**85.9**	**90.2**	85.2	86.3	86.9
❹	InternVL3.5-8B (Wang et al., 2025)	83.0	87.8	84.8	85.8	85.4
❺	Qwen3-VL-8B-Instruct (Bai et al., 2025a)	85.0	89.3	**86.1**	**87.8**	**87.1**

A.10 Effectiveness of Our Target-specific Language Description

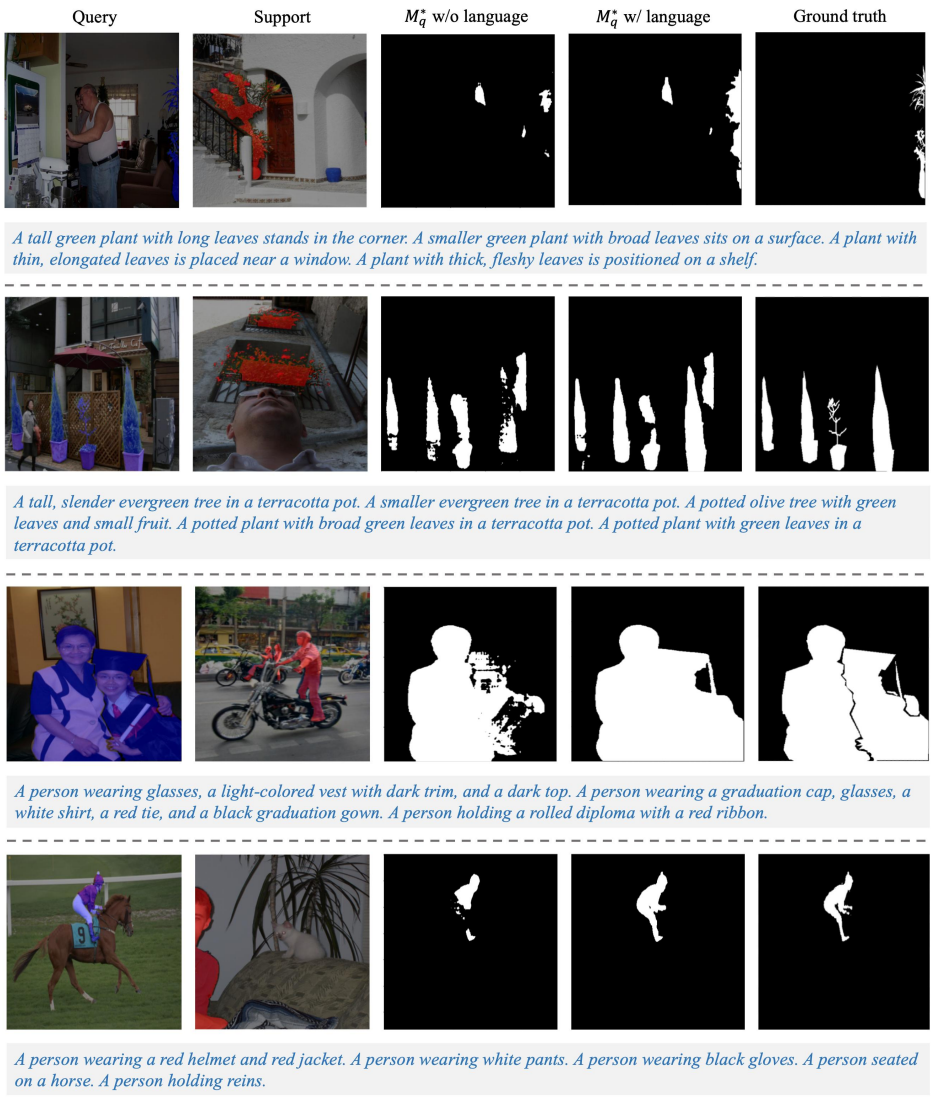


Figure 8: Qualitative examples showing the effect of linguistic cues on visual understanding.

To further verify the effectiveness of our generated target-specific language descriptions, we conduct a comparative experiment between using two types of textual descriptions: ❶ the class-specific descriptions generated by ChatGPT-4.0 as adopted in DSV-LFS (Karimi & Poullis, 2025), and ❷ our proposed target-specific descriptions, which are generated individually for each query image. As shown in Tab. 13, our image-specific descriptions achieve consistently higher performance across all folds on PASCAL-5^i. This improvement demonstrates that general class-level language descriptions are insufficient to capture fine-grained and context-dependent visual details, while our specific descriptions provide more precise semantic cues that better align with the specific visual content of each image.

Table 13: Performance comparison under different language description configurations.

	Language Description	5^0	5^1	5^2	5^3	Mean
❶	Class-specific	84.7	89.2	84.4	87.2	86.4
❷	Target-specific	**85.0**	**89.3**	**86.1**	**87.8**	**87.1**