TOWARDS LANGUAGE-GUIDED UNIFIED VISUAL PROMPT LEARNING FOR CAMOUFLAGED OBJECT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Camouflaged object segmentation (COS) aims to segment objects that seamlessly blend into the background, making it a crucial task in visual perception. Most existing COS methods formulate this problem in a purely visual manner, and thus struggle to capture the semantic cues and contextual knowledge required for understanding camouflage. Recently, language-guided segmentation methods have offered a promising solution. To advance language-guided COS, we study Language-Guided Unified Visual Prompt Learning for COS, where camouflaged objects are segmented under text-prompt guidance with diverse visual-prompt conditions, including mask, point, scribble, and null-prompt settings. To support this setting, we re-annotate existing COS benchmarks by collecting 12,487 text descriptions and constructing corresponding visual prompts, including full-mask, point, and scribble annotations, together with a no-prompt protocol. Building on this, we propose a unified COS framework based on a multimodal large language model (MLLM), which incorporates different visual prompts under text-prompt guidance. Extensive experiments show that our framework achieves competitive performance across different visual prompt settings.

1 INTRODUCTION

Camouflage is a common defense strategy in the natural world. Organisms can adeptly blend into their surroundings by mimicking colors, textures, and shapes (Stevens, 2016; Reiter & Laurent, 2020; Han et al., 2025a; Millán et al., 2025), thereby avoiding detection by predators. Camouflaged object segmentation (COS) aims to accurately segment concealed objects that are visually embedded in complex backgrounds, posing a significant challenge for pixel-level visual understanding (Zhang et al., 2025a; Guo & Huang, 2025). Recently, COS has received increasing attention due to its broad applications in various domains, such as pest monitoring (Venkateswara & Padmanabhan, 2025; Dong et al., 2024), medical image segmentation (Gao et al., 2025; Chen et al., 2025b; Qayoom et al., 2025), industrial defect inspection (Dong et al., 2025; Han et al., 2025b), and ecological observation.

Recently, large foundation models have been explored to support camouflaged object segmentation. For example, Segment Anything Models (Chen et al., 2024a) can generate candidate masks as pseudo-labels, while diffusion models (Chen et al., 2024b) can synthesize labelled camouflaged images for data augmentation. Beyond these visual resources, multimodal large language models (MLLMs) (Zhang et al., 2025b) provide textual semantic cues for COS by describing camouflaged objects, surrounding contexts, and target-background relationships. However, as shown in Fig. 1, directly applying general MLLM-based segmentation methods to COS still leads to inaccurate localization and incomplete segmentation. Moreover, recent text-guided COD studies mainly introduce textual descriptions as auxiliary semantic cues, while general MLLM-based segmentation methods are primarily developed for natural images. Therefore, *they do not systematically study how text prompts can be combined with different visual prompts for unified camouflaged object segmentation*.

To this end, we study **Language-Guided Unified Visual Prompt Learning for Camouflaged Object Segmentation**. Different from conventional COS methods that usually follow a fixed mask-supervised paradigm, our goal is to build a unified visual-prompt learning framework guided by text prompts, as shown in Fig. 2. Specifically, text prompts provide semantic information about camouflaged

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

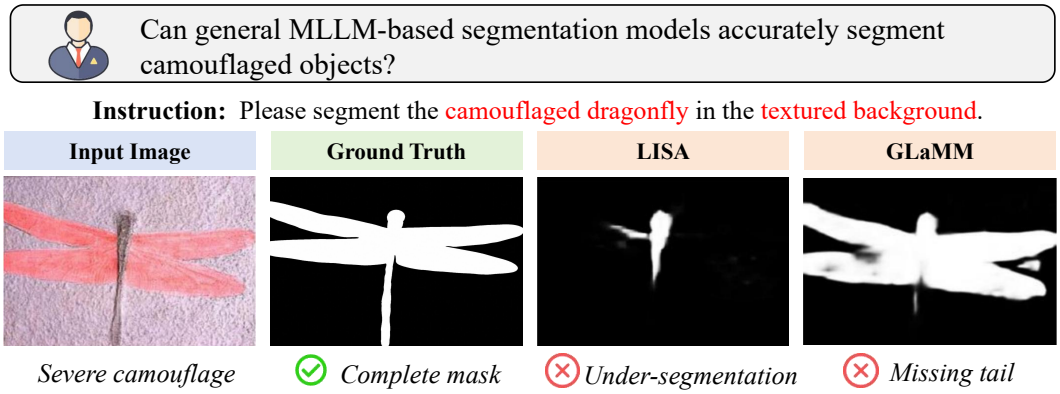

Figure 1: General MLLM-based segmentation results.

objects, background contexts, and camouflage relationships, helping the model understand how the target visually and semantically blends into its surroundings. Meanwhile, different visual annotation forms are organized into unified visual-prompt conditions, including full-mask, point (Chen et al., 2025a), scribble (He et al., 2023b), and no-prompt settings. Owing to such a flexible and cohesive design, the model can benefit from joint learning with text prompts and different visual prompts, facilitating the correlations among language semantics, camouflage contexts, visual cues, and pixel-level segmentation targets.

To support this study, we re-annotate existing COS benchmarks and collect 12,487 text descriptions as text prompts for camouflage-context understanding. We also organize visual prompts, including full-mask, point, scribble, and no-prompt settings. Based on these annotations, we propose an MLLM-based unified COS framework that bridges text-prompt-guided camouflage understanding and visual-prompt-driven mask prediction. The MLLM provides semantic understanding from image-text inputs, while the segmentation component incorporates visual prompts of different granularities to generate camouflaged object masks. Extensive experiments demonstrate that our framework achieves competitive performance under visual-prompt settings.These results indicate that text prompts provide useful semantic support for camouflage understanding, while unified visual-prompt learning offers a flexible formulation for COS under different annotation conditions.

Our main contributions are summarized as follows: **1)** We study Language-Guided Unified Visual Prompt Learning for COS, extending conventional COS to a language-guided promptable segmentation setting. **2)** We re-annotate existing COS benchmarks with 12,487 text descriptions and organize visual prompts, including full-mask, point, scribble, and no-prompt settings. **3)** We propose an MLLM-based unified COS framework that integrates text-prompt guidance with visual prompts of different granularities for camouflage-aware segmentation. **4)** Extensive experiments show that our framework achieves competitive performance under different visual-prompt settings in challenging camouflage scenes.

2 Related Work

Camouflaged Object Segmentation. COS has been extensively studied in recent years. Existing methods mainly design specialized architectures to capture visual cues of camouflaged objects, such as multi-scale feature fusion (Fan et al., 2020), boundary refinement (Pang et al., 2022), low-level feature enhancement (Huang et al., 2023), foreground-background separation, uncertainty modeling, and attention mechanisms (Fan et al., 2021). These methods improve COS performance by exploiting appearance, texture, structure, and boundary information. However, most of them formulate COS as a purely visual task and pay limited attention to textual semantic information about camouflaged objects, surrounding contexts, and target-background relationships. As a result, they may still struggle in complex scenes where objects are highly similar to their surroundings.

Foundation Models for COS. Large foundation models have recently been explored for COS. For example, SAM-based methods can generate prompt-conditioned masks and use them as pseudo masks for camouflaged images (Kirillov et al., 2023; He et al., 2023a). Diffusion-based methods can

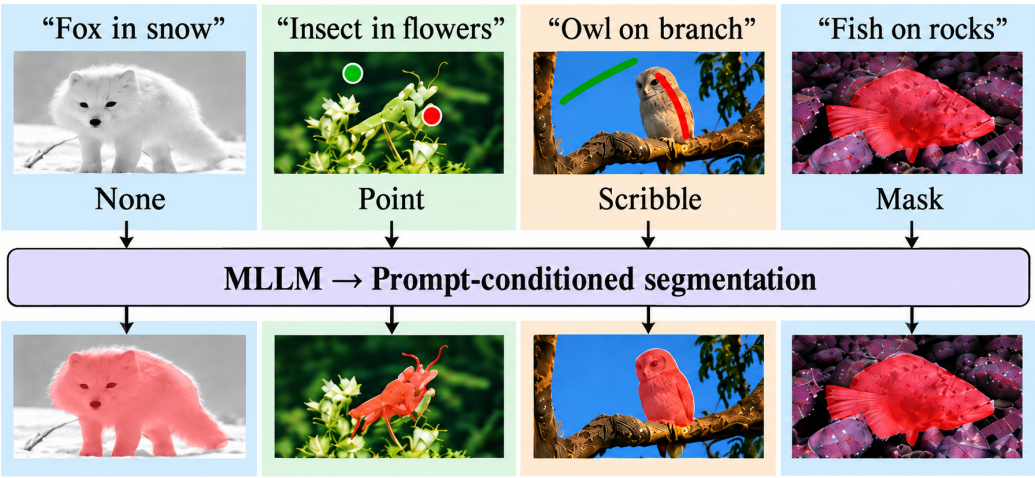

Figure 2: Illustration of the proposed language-guided unified visual-prompt setting for COS. Each sample has a text prompt for camouflage-context understanding, while the visual prompt varies among null, point, scribble, and mask prompts. The null prompt means no explicit visual guidance.

synthesize camouflaged samples or reformulate COD as conditional mask generation (Luo et al., 2023; Chen et al., 2024b). These methods mainly exploit foundation models from the visual side. Beyond visual cues, recent text-guided COD methods introduce textual descriptions to provide semantic guidance for camouflaged object understanding (Chen et al., 2026). Meanwhile, general MLLM-based segmentation methods have shown strong potential in language-guided mask generation and pixel-level visual grounding (Lai et al., 2024; Rasheed et al., 2024; Ren et al., 2024). However, existing text-guided COD methods mainly use text as auxiliary semantic information, while general MLLM-based segmentation methods are developed for natural images; neither systematically studies how text prompts can be combined with different visual prompts for unified COS. Different from them, our work builds text prompts for camouflage-context understanding and organizes corresponding visual prompts for text-guided unified visual prompt learning in COS.

Visual Prompt Learning for Segmentation. Visual prompts provide flexible spatial guidance for segmentation models under different levels of annotation granularity. In COS, dense masks are widely used in fully supervised methods (Fan et al., 2020; Mei et al., 2021; Fan et al., 2021), while point and scribble annotations provide sparse spatial cues for annotation-efficient learning (Chen et al., 2025a; He et al., 2023b; Han et al., 2025a). Null prompts correspond to the no-visual-prompt setting, which is closely related to unsupervised COS methods that segment camouflaged objects without human-provided spatial annotations (Zhang & Wu, 2023; Yan et al., 2025). However, existing COS studies usually investigate these settings separately, and their effectiveness under text-prompt guidance remains underexplored. In this work, we organize mask, point, scribble, and null prompts into a visual-prompt formulation, and study how they can be combined with text prompts for COS.

3 BENCHMARK: MULTICAMO

Existing COS datasets usually contain only images and pixel-level masks, which are mainly used for visual segmentation training and evaluation. However, they lack textual descriptions and cannot explicitly provide semantic information about camouflaged objects, surrounding environments, and target-background relationships. To support text-guided unified visual-prompt learning, we construct MultiCamo by adding text prompts to existing COS datasets and organizing different visual prompts following established COS settings.

MultiCamo Construction. For each source image, we preserve the original pixel-level mask and construct text prompts based on object category, location, image content, and background context. These prompts describe the camouflaged target and its surrounding scene, helping the model capture target-background relationships. Meanwhile, we organize four types of visual prompts: mask, point, scribble, and null prompts. Mask prompts are derived from pixel-level

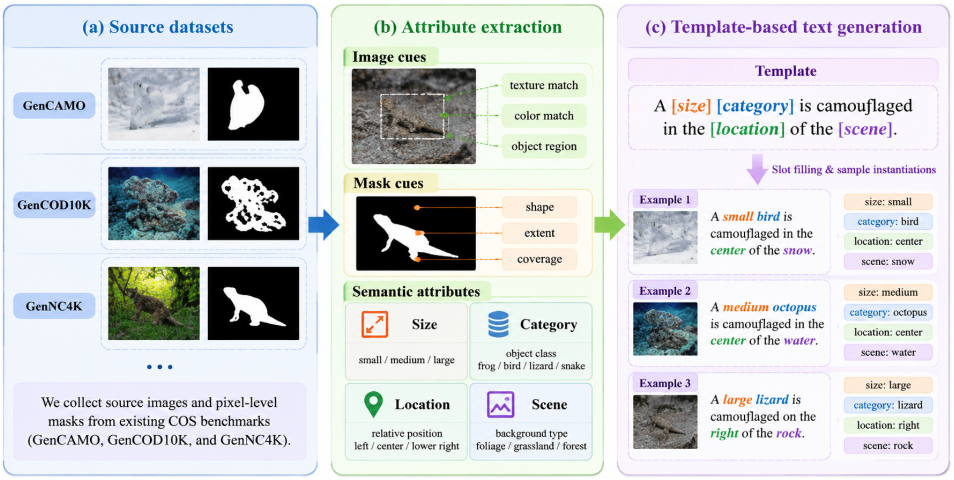

Figure 3: Overview of the text prompt construction process in MultiCamo. We collect source images and pixel-level masks from existing COS benchmarks, extract target-related attributes including size, category, location, and scene information, and use GPT-4o to generate text prompts for camouflage-context understanding.

annotations, point and scribble prompts follow point- and scribble-supervised COS settings, and null prompts indicate the no-visual-prompt setting. Detailed dataset statistics are reported in Table 1.

Data Sources. MultiCamo is built upon three widely used COS benchmarks, including CAMO, COD10K, and NC4K. We preserve their original images and masks, and reorganize them into GenCAMO, GenCOD10K, and GenNC4K. MultiCamo follows the official or commonly used training/testing splits of these benchmarks to ensure fair comparison with existing methods.

Table 1: Comparison of the original COS benchmarks and the proposed MultiCamo. M/P/S/N/T denote mask, point, scribble, null, and text prompts, respectively. Train and Test indicate the numbers of images in each split, while Text indicates the number of text prompts. The number of text prompts may exceed the number of images because some images have multiple camouflage-context descriptions.

Dataset	M	P	S	N	T	Train	Test	Text
Original COS								
CAMO (Le et al., 2019)	✓	✗	✗	✗	✗	1000	250	–
COD10K (Fan et al., 2020)	✓	✗	✗	✗	✗	3040	2026	–
NC4K (Lv et al., 2021)	✓	✗	✗	✗	✗	0	4121	–
MultiCamo								
GenCAMO	✓	✓	✓	✓	✓	1000	250	1496
GenCOD10K	✓	✓	✓	✓	✓	3040	2026	6061
GenNC4K	✓	✓	✓	✓	✓	0	4121	4930
TG-COD (Chen et al., 2026)	✓	✗	✗	✗	✓	4040	6397	10437
MultiCamo	✓	✓	✓	✓	✓	**4040**	**6397**	**12487**

Text Prompt Construction. To construct reliable text prompts, we first extract target category, scale, spatial location, and scene information from the original annotations, masks, and image content. These attributes provide stable target grounding and scene constraints. Then, we use GPT-4o to rewrite the initial structured descriptions into more natural text prompts by considering target appearance, background environment, and camouflage relationships, as illustrated in Fig. 3. The rewritten prompts further describe camouflage cues such as color similarity, texture consistency, partial occlusion, and ambiguous boundaries.

Text Verification. After automatic generation, we manually check and revise the text prompts. Annotators compare each prompt with the original image and mask to ensure that it describes the correct target and is consistent with the scene. Prompts with category errors, target shifts, background confusion, redundant expressions, or ambiguous descriptions are corrected or removed. This process improves the reliability of text prompts for learning the correspondence between language descriptions and camouflaged targets.

Visual Prompt Organization. Besides text prompts, MultiCamo organizes four visual-prompt conditions. The mask prompt provides dense region guidance from the original pixel-level mask. The point prompt provides sparse location cues following point-supervised COS annotations. The scribble prompt provides sparse structural cues following scribble-supervised COS annotations. The null prompt provides no additional spatial cue and represents the no-visual-prompt condition. By

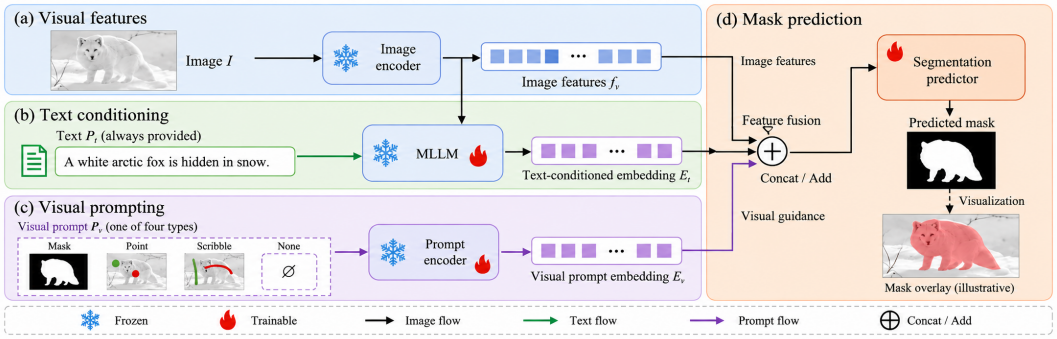

Figure 4: Overview of the proposed language-guided unified visual-prompt framework for COS. Each image is paired with a text prompt for camouflage-context understanding, while the visual prompt varies among mask, point, scribble, and null prompts. The framework encodes image features, text-conditioned camouflage semantics, and visual-prompt embeddings, and integrates them to predict the final camouflaged object mask.

organizing these prompts in a unified format, MultiCamo supports the evaluation of text-guided COS under different visual-prompt settings.

4 METHOD

Overview. The overall architecture of our framework is illustrated in Fig. 4. Given an input image, our framework performs COS under text-prompt guidance and different visual-prompt conditions. The framework consists of an image encoder, an MLLM, a visual prompt encoder, and a segmentation predictor. Specifically, the input image is first encoded into visual features by the image encoder. The text prompt is combined with the image representation and fed into the MLLM to obtain text-conditioned semantic embeddings for camouflage-context understanding. Meanwhile, different visual prompts are encoded into unified prompt embeddings by the visual prompt encoder. The segmentation predictor then integrates image features, text-conditioned camouflage semantics, and visual prompt embeddings to generate the final camouflaged object mask. In this way, our framework unifies text-guided camouflage understanding and visual-prompt-driven mask prediction within a consistent COS formulation.

Text-Guided Semantic Encoding. We adopt an MLLM to extract text-conditioned semantic embeddings for camouflaged object segmentation. Given an input image I and a text prompt P_t, the image is first encoded by a visual encoder F_v to obtain visual features f_v. These visual features are projected into the language space through a projection function G_v and then combined with the text prompt in the MLLM. This process can be written as $f_v = F_v(I)$ and $E_t = F_{\text{MLLM}}(G_v(f_v), P_t)$, where E_t denotes the text-conditioned output embeddings.

The text prompt describes the camouflaged object, the surrounding context, and the relationship between the target and the background. Therefore, E_t provides camouflage-aware semantic cues that are difficult to obtain from visual appearance alone, especially when the target shares similar color, texture, or structure with its surroundings. These semantic embeddings are then used as language guidance for pixel-level mask prediction, helping the model distinguish the camouflaged target from visually similar background regions.

Unified Prompt Formulation. To support different COS conditions, we design a unified prompt formulation that contains two parts: a text prompt P_t and a visual prompt P_v. The text prompt provides semantic information about the camouflaged scene, while the visual prompt provides spatial guidance with different annotation granularities.

1) Text prompt. The text prompt describes the target object, background context, and target-background relationship. A typical instruction is formulated as:

> "Please segment the camouflaged object according to the following description: [text description]."

The text prompt helps the model capture camouflage-specific semantic cues related to the target category, scene context, and the visual similarity between the target and background.

2) Visual prompt. The visual prompt provides spatial or structural guidance for mask prediction. We consider four visual-prompt conditions: mask prompt, point prompt, scribble prompt, and null-prompt condition. The mask prompt provides dense region-level reference cues, the point prompt provides sparse location cues, and the scribble prompt provides sparse structural cues. For the null-prompt condition, no additional visual prompt is provided, and the model predicts the camouflaged object based on the input image and text prompt. In this way, different annotation forms are organized into a unified visual-prompt formulation for COS, enabling the model to handle different levels of spatial ambiguity in camouflaged scenes.

Visual Prompt Encoder. Given a visual prompt P_v, the visual prompt encoder F_p converts it into a unified prompt embedding E_v. Different prompt types are represented in the same embedding space, allowing the model to handle different visual-prompt conditions within one framework. For point prompts, the encoder embeds point coordinates and point labels to provide sparse target localization. For scribble prompts, sparse foreground traces are encoded as structural guidance for weakly visible object regions. For mask prompts, dense visual reference cues are encoded to guide mask prediction. For the null-prompt condition, we use a learnable null prompt embedding to represent the absence of visual guidance. This unified design allows the segmentation predictor to receive consistent prompt representations across different annotation conditions and adapt to the ambiguity of camouflaged target boundaries.

Segmentation Predictor. The segmentation predictor generates the final camouflaged object mask by jointly using image features, text-conditioned embeddings, and visual prompt embeddings. Given the image feature f_v, the text-conditioned embedding E_t, and the visual prompt embedding E_v, the predicted mask is obtained as $\hat{M} = F_{\text{seg}}(f_v, E_t, E_v)$.

The image features preserve fine-grained visual details, the text-conditioned embeddings provide camouflage-aware semantic cues from the text prompt, and the visual prompt embeddings provide spatial guidance. By integrating these three types of information, the model can better distinguish camouflaged targets from visually similar backgrounds under different visual-prompt conditions while maintaining a consistent text-guided segmentation formulation.

Training Objectives. Our framework is trained under different visual-prompt conditions in a unified manner. For each training image, we use different visual prompts, including mask, point, scribble, and null-prompt conditions. The mask prompt provides dense region-level guidance, the point and scribble prompts provide sparse spatial guidance, and the null-prompt condition provides no additional visual cue. For all prompt conditions, the predicted mask is supervised by the ground-truth mask M, encouraging the model to recover complete camouflaged object regions under different levels of visual guidance. We use binary cross-entropy loss and Dice loss (Milletari et al., 2016) for mask optimization:

$$\mathcal{L}_{seg}(\hat{M}_{P_v}, M) = \mathcal{L}_{bce}(\hat{M}_{P_v}, M) + \mathcal{L}_{dice}(\hat{M}_{P_v}, M), \tag{1}$$

where $\hat{M}_{P_v}$ denotes the predicted mask under the visual-prompt condition P_v. When multiple visual-prompt conditions are jointly used during training, we calculate the segmentation loss for each condition and combine them as:

$$\mathcal{L} = \sum_{P_v \in \mathcal{P}} \lambda_{P_v} \mathcal{L}_{seg}(\hat{M}_{P_v}, M), \tag{2}$$

where $\mathcal{P} = \{\text{mask}, \text{point}, \text{scribble}, \text{null}\}$ denotes the set of visual-prompt conditions. By default, all prompt conditions are assigned equal weights.

5 Experiments

5.1 Experimental Setup

Datasets. We conduct experiments on MultiCamo, the re-annotated COS benchmark introduced earlier in this paper. MultiCamo provides text prompts and four visual-prompt settings, including mask, point, scribble, and null prompts. Following common COS evaluation protocols (Chen et al., 2026), we train the model on the training split and evaluate it on multiple test sets.

Table 2: Quantitative comparison on MultiCamo. Methods include visual COS, FM/Text, MLLM, and our unified framework. **Ours (Mask)** uses mask supervision, while **Ours (All)** jointly uses mask, point, and scribble supervision with text guidance. Best results are in **bold**.

Methods	Base model	GenCAMO (250)				GenCOD10K (2026)				GenNC4K (4121)			
		$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	MAE↓	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	MAE↓	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	MAE↓
Conventional visual COS methods													
FSPNet (Huang et al., 2023)	ViT-B	0.857	0.901	0.803	0.051	0.849	0.895	0.733	0.027	0.878	0.916	0.820	0.035
ZoomNeXt (Pang et al., 2024)	PVTv2-B5	0.889	0.945	0.857	0.041	0.898	0.956	0.827	0.018	0.903	0.951	0.863	0.028
ARM (Chen et al., 2025c)	SAM-H	0.887	0.935	0.883	0.046	0.909	0.955	0.885	0.020	0.906	0.947	0.897	0.033
SDRNet (Zhang & Jiang, 2026)	PVTv2-B4	0.872	0.924	0.826	0.049	0.871	0.924	0.785	0.023	0.889	0.934	0.842	0.032
Foundation-model-based and text-guided COS methods													
EVP (Liu et al., 2023)	SegFormer-B4	0.847	0.897	0.781	0.058	0.842	0.906	0.742	0.029	0.872	0.917	0.810	0.039
SAM-Adapter (Chen et al., 2023a)	SAM-L	0.851	0.901	0.781	0.068	0.884	0.917	0.805	0.025	0.887	0.912	0.821	0.040
diffCOD (Chen et al., 2023b)	Diffusion	0.795	0.852	0.704	0.082	0.812	0.892	0.684	0.036	0.837	0.891	0.761	0.051
CamoDiffusion (Chen et al., 2024b)	Diffusion	0.879	0.940	0.854	0.042	0.880	0.943	0.815	0.020	0.892	0.941	0.858	0.029
TG-COD (Chen et al., 2026)	SAM-B	0.876	0.916	0.830	0.048	0.895	0.940	0.834	0.021	0.892	0.927	0.849	0.034
General MLLM-based segmentation methods adapted to COS													
LISA (Lai et al., 2024)	LLaVA-7B	0.801	0.846	0.714	0.081	0.808	0.872	0.672	0.039	0.832	0.884	0.752	0.051
PixelLM (Ren et al., 2024)	LLaVA-7B	0.819	0.867	0.742	0.073	0.823	0.889	0.704	0.034	0.846	0.899	0.781	0.046
GLaMM (Rasheed et al., 2024)	LLaVA-7B	0.835	0.884	0.766	0.066	0.838	0.901	0.724	0.031	0.859	0.912	0.803	0.041
Our COS-specific text-guided unified visual-prompt framework													
Ours (All)	LLaVA-7B	0.892	0.947	0.886	0.039	0.911	0.958	0.888	0.017	0.908	0.953	0.900	0.027
Ours (Mask)	LLaVA-7B	**0.895**	**0.950**	**0.890**	**0.037**	**0.914**	**0.961**	**0.892**	**0.015**	**0.911**	**0.956**	**0.904**	**0.025**

Evaluation metrics. We adopt widely used COS evaluation metrics, including structure-measure S_α, enhanced-alignment measure E_ϕ, weighted F-measure F_β^w, and mean absolute error (MAE). For all metrics except MAE, higher values indicate better performance.

Implementation details. For text prompt construction, we use GPT-4o to generate textual descriptions, followed by manual checking and refinement. For our framework, we adopt LLaVA-1.5-7B (Liu et al., 2024) as the MLLM for image-text understanding and SAM-H (Kirillov et al., 2023) as the visual segmentation backbone. All visual-prompt conditions are trained with the same mask prediction objective. Input images are resized to 1024×1024. Experiments are conducted on NVIDIA RTX 4090 GPUs with 24GB memory. We use AdamW (Loshchilov & Hutter, 2019) with a learning rate of 3×10^{-4}, weight decay of 0, and cosine learning rate decay.

5.2 Comparison with SOTA Methods

Quantitative results. We compare our method with three groups of representative methods under the mask-prompt setting, as shown in Table 2. The first group includes conventional visual COS methods, which are image-only models and do not use text prompts. The second group contains foundation model-based and text-guided COS methods, which introduce foundation models or textual information into COS. The third group includes general MLLM-based segmentation methods adapted to COS, where the same text prompts and mask prompts are used for fair comparison. Ours (Mask) achieves the best overall performance. Ours (All) obtains the second-best results among our settings, indicating that multiple visual prompts can be effectively unified. Overall, our framework better integrates language-guided camouflage understanding with visual-prompt-driven mask prediction, demonstrating the effectiveness of unified visual prompt learning for COS.

Qualitative results. We further provide qualitative comparisons under the mask-prompt setting. As shown in Fig. 5, existing COS methods often suffer from incomplete localization, or false responses to background distractors, especially when camouflaged objects share similar colors and textures with their surroundings. In contrast, our method effectively leverages mask prompts as visual guidance and produces more accurate and complete masks in challenging camouflage scenarios.

5.3 Visual-Prompt Analysis

Table 3 reports the results under different visual-prompt conditions. All settings use the same text prompt, and only the visual prompt is changed. The *Text + Null visual prompt* setting provides no explicit visual guidance, so the model relies on image content and text semantics. Point and scribble prompts offer sparse spatial cues, while mask prompts provide dense and accurate region guidance.

Table 3: Results under different visual-prompt conditions on our benchmark. All settings use the text prompt. Null visual prompt denotes the setting in which no explicit visual prompt is provided. All visual prompts denotes the setting in which mask, point, and scribble prompts are jointly used. The best results are highlighted in **bold**.

Input Setting	GenCAMO				GenCOD10K				GenNC4K			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
Text + Null Visual Prompt	0.815	0.865	0.735	0.075	0.825	0.890	0.700	0.035	0.848	0.900	0.780	0.046
Text + Point	0.845	0.891	0.780	0.062	0.855	0.920	0.755	0.028	0.873	0.925	0.820	0.038
Text + Scribble	0.865	0.918	0.815	0.052	0.872	0.935	0.795	0.023	0.885	0.935	0.840	0.032
Text + Mask	**0.895**	**0.950**	**0.890**	**0.037**	**0.914**	**0.961**	**0.892**	**0.015**	**0.911**	**0.956**	**0.904**	**0.025**
Text + All Visual Prompts	0.892	0.947	0.886	0.039	0.911	0.958	0.888	0.017	0.908	0.953	0.900	0.027

Table 4: Ablation studies of the proposed framework on GenCOD10K under the All Visual Prompts setting. Mask, point, and scribble prompts are jointly used unless otherwise specified. The default configuration in each study is highlighted in light gray.

Text Prompt	S_α	E_ϕ	F_β^w	M
w/o Text	0.858	0.918	0.778	0.030
Category	0.874	0.935	0.804	0.023
Obj. Desc.	0.878	0.939	0.811	0.021
Full Desc.	**0.911**	**0.958**	**0.888**	**0.017**

(a) Text prompt form.

Semantic Encoder	S_α	E_ϕ	F_β^w	M
w/o Text	0.858	0.918	0.778	0.030
CLIP Text	0.866	0.926	0.791	0.027
LLaVA Text-only	0.873	0.934	0.803	0.024
LLaVA Image-Text	**0.911**	**0.958**	**0.888**	**0.017**

(b) Semantic encoder.

Text Construction	S_α	E_ϕ	F_β^w	M
Template	0.853	0.914	0.782	0.034
Template + Check	0.856	0.927	0.797	0.032
GPT-4o	0.879	0.935	0.809	0.025
GPT-4o + Check	**0.911**	**0.958**	**0.888**	**0.017**

(c) Text prompt construction.

The results show that *Text + Mask* achieves the best performance, as the mask prompt gives the most complete target information. *Text + All Visual Prompts* obtains the second-best results. This setting combines mask, point, and scribble prompts, but does not include the null prompt because it represents the absence of visual guidance. Although multiple explicit prompts improve robustness, point and scribble cues may be less precise or redundant compared with the dense mask prompt. Overall, these results show that our framework can handle visual prompts of different granularities, and stronger spatial guidance generally leads to better COS performance.

5.4 Ablation Studies

We conduct ablation studies on the GenCOD10K test set to analyze three key factors of our framework: text prompt form, semantic encoder, and text prompt construction strategy. The results are shown in Table 4. All results are reported under the all visual-prompt setting, where mask, point, and scribble prompts are jointly used.

Effect of text prompt form. We first study the influence of different text prompt forms. The w/o Text setting removes textual semantic guidance and relies on image features and visual prompts only. Category prompts provide basic object-level semantics, while object descriptions further introduce target-related attributes. Full descriptions contain richer information about the camouflaged object, surrounding context, and target-background relationship. The comparison shows that more complete textual descriptions provide stronger semantic support for camouflage-context understanding.

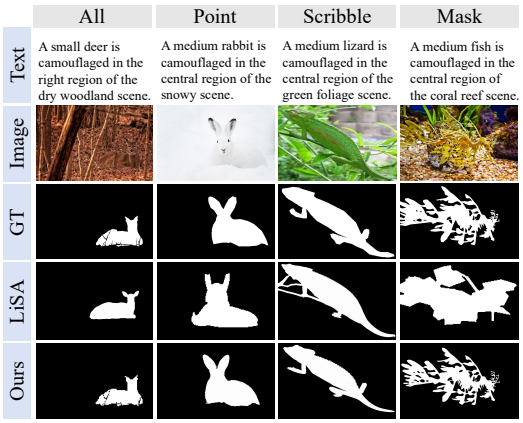

Figure 5: Qualitative comparison of COS under different visual prompt conditions. Columns denote **(a) all**, **(b) point**, **(c) scribble**, and **(d) mask** prompts, while rows show the text prompt, input image, ground truth, LISA prediction, and our prediction. Our method produces more accurate and complete masks across different prompt conditions.

Effect of semantic encoder. We then compare different semantic encoding strategies. The w/o Text setting removes textual semantic encoding, while CLIP Text encodes the text prompt independently without image-conditioned reasoning. LLaVA Text-only uses the language branch to encode

textual information without visual context. In contrast, LLaVA Image-Text jointly models the input image and text prompt, enabling better alignment between textual semantics and camouflaged regions. This indicates that image-text semantic encoding is important for text-guided COS.

Effect of text prompt construction. Finally, we evaluate different text prompt construction strategies. Template-based prompts provide structured but relatively simple descriptions. Manual checking improves their correctness and consistency with the target region. Compared with template prompts, GPT-4o-generated prompts offer more natural and context-aware descriptions. After manual checking, GPT-4o + Check further improves prompt reliability by reducing target mismatch expressions.

Effect of unified visual-prompt encoding. Mask, point, and scribble prompts are converted into SAM-compatible representations and encoded by a shared SAM prompt encoder, while a learnable embedding represents the null prompt. As shown in Table 5, sharing encoder parameters improves overall performance, and prompt-type embeddings provide further consistent gains by distinguishing different prompt forms.

Table 5: Ablation study of the unified visual-prompt encoder on GenCOD10K under the All Visual Prompts setting. Different encoder variants are evaluated, and the best results are highlighted in **bold**.

SAM Encoder	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
Type-specific	0.876	0.936	0.808	0.023
Shared w/o type embedding	0.879	0.940	0.812	0.021
Shared + type embedding	**0.911**	**0.958**	**0.888**	**0.017**

5.5 Failure Case Analysis

We present representative failure cases in Fig. 6. Although our framework supports diverse prompt conditions, failures remain in challenging camouflage scenarios. Under the null-prompt setting, the model relies on image content and text guidance, and may confuse targets with surrounding regions when foreground-background contrast is extremely low, as shown in the human and flatfish examples. With point prompts, sparse spatial cues can localize targets but may be insufficient to recover complete structures for thin or irregular objects, leading to background inclusion in the caterpillar example. Even with mask prompts, fine-grained details such as boundaries and small structures may be inaccurately predicted when objects are partially occluded or highly similar to the background, as observed in the butterfly example. These cases indicate that improving prompt-target alignment and boundary-aware prediction remains important for complex camouflage scenes.

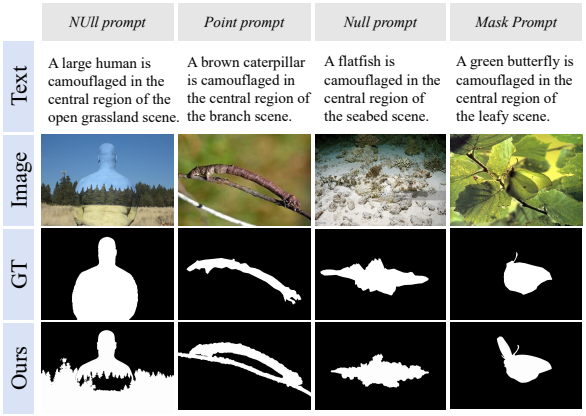

Figure 6: Representative failure cases of the proposed framework. Failures mainly occur in challenging camouflage scenarios with low foreground-background contrast, ambiguous semantic cues, and complex background patterns, where accurate target localization and boundary recovery remain difficult.

6 Conclusion

In this paper, we proposed language-guided unified visual prompt learning for COS, aiming to establish a unified formulation that integrates textual guidance with diverse visual prompts. To facilitate systematic study of promptable COS, we introduced MultiCamo, a benchmark extending existing COS datasets with 12,487 text descriptions generated and refined through a GPT-4o-assisted annotation pipeline, together with unified mask, point, scribble, and null visual-prompt settings. Based on this benchmark, we developed an MLLM-based framework that jointly models semantic information from text prompts and spatial guidance from visual prompts for pixel-level segmentation. Extensive experiments and ablation studies demonstrate that our framework effectively exploits complementary text and visual cues, highlighting the importance of prompt design, semantic encoding, and unified visual-prompt representations. Comprehensive evaluations establish a new benchmark and provide insights into multimodal prompt learning for camouflage understanding. Despite strong performance across various settings, challenges remain in scenarios with severe target-background similarity, ambiguous semantic cues, and fine-grained boundary uncertainty, motivating future research on robust cross-modal reasoning and adaptive prompt optimization.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language translation, improve the readability and clarity of the manuscript, and provide limited feedback on research methodology and experimental design. AI-generated suggestions were used only as references and did not replace the authors' independent research decisions. All AI-assisted translations and revisions were manually reviewed for technical accuracy, terminology, and consistency, and all methodological and experimental decisions were made and verified by the authors. Generative AI was not used to generate experimental results, formulate mathematical claims, implement the proposed method, conduct data analysis, or interpret the reported results. The authors take full responsibility for the final content, claims, and conclusions of this work.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of this work. The experimental section provides details of the datasets, evaluation metrics, model configurations, training settings, and implementation protocols used in our experiments. Additional implementation details, parameter analyses, and supplementary experiments are provided in the Appendix. The anonymized source code is also included in the supplementary material for double-blind review. Together, these materials are intended to facilitate independent reproduction and verification of the reported results.

REFERENCES

Huafeng Chen, Pengxu Wei, Guangqian Guo, and Shan Gao. Sam-cod: Sam-guided unified framework for weakly-supervised camouflaged object detection. In *European Conference on Computer Vision*, pp. 315–331, 2024a.

Huafeng Chen, Dian Shao, Guangqian Guo, and Shan Gao. Just a hint: Point-supervised camouflaged object detection. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 332–348, Cham, 2025a. Springer Nature Switzerland. ISBN 978-3-031-72761-0.

Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3367–3375, 2023a.

Tianrun Chen, Runlong Cao, Xinda Yu, Lanyun Zhu, Chaotao Ding, Deyi Ji, Cheng Chen, Qi Zhu, Chunyan Xu, Papa Mao, et al. Sam3-adapter: Efficient adaptation of segment anything 3 for camouflage object segmentation, shadow detection, and medical image segmentation. *arXiv preprint arXiv:2511.19425*, 2025b.

Xuehan Chen, Guangyu Ren, Tianhong Dai, Tania Stathaki, and Hengyan Liu. Enhancing prompt generation with adaptive refinement for camouflaged object detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20672–20682, 2025c.

Zefeng Chen, Yunqi Xue, Zhijiang Li, Philip Torr, and Jindong Gu. Text-guided camouflaged object detection. *Pattern Recognition*, 170:112058, 2026.

Zhennan Chen, Rongrong Gao, Tian-Zhu Xiang, and Fan Lin. Diffusion model for camouflaged object detection, 2023b. URL https://arxiv.org/abs/2308.00303.

Zhongxi Chen, Ke Sun, and Xianming Lin. Camodiffusion: Camouflaged object detection via conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1272–1280, 2024b.

Qing Dong, Lina Sun, Tianxin Han, Minqi Cai, and Ce Gao. Pestlite: A novel yolo-based deep learning technique for crop pest detection. *Agriculture*, 14(2):228, 2024.

Qing Dong, Tianxin Han, Gang Wu, Lina Sun, Min Huang, and Fu Zhang. Industrial device-aided data collection for real-time rail defect detection via a lightweight network. *Engineering Applications of Artificial Intelligence*, 161:112102, 2025.

Ji Du, Xin Wang, Fangwei Hao, Mingyang Yu, Chunyuan Chen, Jiesheng Wu, Bin Wang, Jing Xu, and Ping Li. Beyond single images: Retrieval self-augmented unsupervised camouflaged object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22131–22142, 2025.

Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4548–4557, 2017. doi: 10.1109/ICCV.2017.487.

Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 698–704, 2018.

Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2777–2787, 2020.

Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6024–6042, 2021.

Yuxiao Gao, Yang Jiang, Yanhong Peng, Fujiang Yuan, Xinyue Zhang, and Jianfeng Wang. Medical image segmentation: A comprehensive review of deep learning-based methods. *Tomography*, 11 (5):52, 2025.

Cunhan Guo and Heyan Huang. Cofinet: Unveiling camouflaged objects with multi-scale finesse. *Neurocomputing*, 614:128763, 2025.

Tianxin Han, Xingwei Wang, Qing Dong, Min Huang, Jie Jia, and Fu Zhang. Weakly supervised camouflaged object detection as progressive perception learning. *Knowledge-Based Systems*, 325: 113993, 2025a.

Tianxin Han, Xingwei Wang, Qing Dong, Jie Jia, Fu Zhang, and Min Huan. Cdnet: Efficient real-time chip defect detection for consumer electronics. *IEEE Transactions on Consumer Electronics*, 2025b.

Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. In *Advances in Neural Information Processing Systems*, volume 36, pp. 30726–30737, 2023a.

Ruozhen He, Qihua Dong, Jiaying Lin, and Rynson WH Lau. Weakly-supervised camouflaged object detection with scribble annotations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 781–789, 2023b.

Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, Huai-Xin Chen, Jie Qin, and Huan Xiong. Feature shrinkage pyramid for camouflaged object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5557–5566, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589, 2024.

Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer vision and image understanding*, 184:45–56, 2019.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024.

Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19434–19445, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Xue-Jing Luo, Shuo Wang, Zongwei Wu, Christos Sakaridis, Yun Cheng, Deng-Ping Fan, and Luc Van Gool. Camdiff: Camouflage image augmentation via diffusion. *CAAI Artificial Intelligence Research*, 2:9150021, 2023.

Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11591–11601, 2021.

Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2014.

Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8772–8781, 2021.

Ana P Millán, Hanlin Sun, Lorenzo Giambagli, Riccardo Muolo, Timoteo Carletti, Joaquín J Torres, Filippo Radicchi, Jürgen Kurths, and Ginestra Bianconi. Topology shapes dynamics of higher-order networks. *Nature Physics*, 21(3):353–361, 2025.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the Fourth International Conference on 3D Vision*, pp. 565–571, 2016.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 2160–2170, 2022.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9205–9220, 2024.

Abdul Qayoom, Juanying Xie, and Haider Ali. Polyp segmentation in medical imaging: challenges, approaches and future directions. *Artificial Intelligence Review*, 58(6):169, 2025.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad Shahbaz Khan. GLaMM: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13009–13018, 2024.

Sam Reiter and Gilles Laurent. Visual perception and cuttlefish camouflage. *Current Opinion in Neurobiology*, 60:47–54, 2020.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. PixelLM: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26374–26383, 2024.

Martin Stevens. Color change, phenotypic plasticity, and camouflage. *Frontiers in Ecology and Evolution*, 4:51, 2016.

Stella Mary Venkateswara and Jayashree Padmanabhan. Deep learning based agricultural pest monitoring and classification. *Scientific Reports*, 15(1):8684, 2025.

Ranwan Wu, Tian-Zhu Xiang, Guo-Sen Xie, Rongrong Gao, Xiangbo Shu, Fang Zhao, and Ling Shao. Uncertainty-aware transformer for referring camouflaged object detection. *IEEE Transactions on Image Processing*, 2025.

Weiqi Yan, Lvhai Chen, Huaijia Kou, Shengchuan Zhang, Yan Zhang, and Liujuan Cao. UCOD-DPL: Unsupervised camouflaged object detection via dynamic pseudo-label learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30365–30375, 2025.

Guangjian Zhang, Zhengming Yang, Yong Wang, Yuliang Chen, and Duoqian Miao. Camouflaged object detection with boundary localization in complex backgrounds. *Engineering Applications of Artificial Intelligence*, 161:112047, 2025a.

Hongliang Zhang and Sanxin Jiang. Sdnet: A two-stage segmentation-diffusion network for camouflaged object detection. *Applied Soft Computing*, 188:114390, 2026. ISSN 1568-4946.

Jin Zhang, Ruiheng Zhang, Zhe Cao, and Kaizheng Chen. Mmcsbench: A fine-grained benchmark for large vision-language models in camouflage scenes. In *Advances in Neural Information Processing Systems*, 2025b.

Yi Zhang and Chengyi Wu. Unsupervised camouflaged object segmentation as domain adaptation. *arXiv preprint arXiv:2308.04528*, 2023.

A APPENDIX

A.1 EVALUATION METRICS

Following common practice in camouflaged object segmentation (COS) (Wu et al., 2025; Guo & Huang, 2025; Zhang et al., 2025a), we adopt four widely used evaluation metrics, including structure-measure (S_α) (Fan et al., 2017), enhanced-alignment measure (E_ϕ) (Fan et al., 2018), weighted F-measure (F_β^w) (Margolin et al., 2014), and mean absolute error (MAE).

These metrics evaluate the predicted masks from different perspectives, including structural consistency, pixel-level alignment, precision-recall quality, and prediction error.

Structure-measure. S_α evaluates the structural similarity between the predicted mask and the ground-truth mask by considering both object-aware and region-aware structural information:

$$S_\alpha = \alpha S_o + (1-\alpha) S_r, \tag{3}$$

where S_o denotes object-aware structural similarity, S_r denotes region-aware structural similarity, and α is empirically set to 0.5 following previous COS works. Higher S_α indicates that the prediction better preserves the overall object structure.

Enhanced-alignment measure. E_ϕ measures the alignment quality between the prediction and ground truth by jointly considering local pixel-level matching and global image-level statistics. Compared with pixel-wise accuracy, E_ϕ better reflects whether the predicted mask is locally accurate and globally consistent with the ground truth. Higher E_ϕ indicates better foreground-background separation.

Weighted F-measure. F_β^w evaluates the weighted precision-recall quality of the predicted mask. Different from conventional F-measure, it assigns different weights to pixels and considers spatially inconsistent errors. It is defined as:

$$F_\beta^w = \frac{(1+\beta^2) \cdot Precision^w \cdot Recall^w}{\beta^2 \cdot Precision^w + Recall^w}, \tag{4}$$

where $Precision^w$ and $Recall^w$ denote weighted precision and weighted recall, respectively. Following standard COS evaluation protocols, β^2 is set to 0.3.

Mean absolute error. MAE measures the average pixel-wise absolute difference between the predicted mask and the binary ground-truth mask:

$$MAE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} |P(i,j) - G(i,j)|, \tag{5}$$

where $P(i, j)$ and $G(i, j)$ represent the predicted value and ground-truth value at pixel (i, j), respectively. H and W denote the height and width of the mask.

For S_α, E_ϕ, and F_β^w, higher values indicate better performance, while lower MAE is preferred.

A.2 Implementation Details

For text prompt construction, we use GPT-4o to generate candidate textual descriptions, followed by manual checking and refinement. For the segmentation framework, we adopt LLaVA-1.5-7B (Liu et al., 2024) as the MLLM for image-text understanding and SAM-H (Kirillov et al., 2023) as the visual segmentation backbone. The MLLM provides text-conditioned semantic embeddings, while the segmentation module generates camouflaged object masks under different visual-prompt conditions.

A.2.1 Training Details

Input images are resized to 1024×1024. All visual-prompt conditions are trained with the same mask prediction objective. We use AdamW (Loshchilov & Hutter, 2019) as the optimizer, with a learning rate of 3×10^{-4}, weight decay of 0, and cosine learning rate decay.

All experiments are conducted on NVIDIA RTX 4090 GPUs with 24GB memory.

A.2.2 Prompt Settings

During training, different visual prompts are provided according to the corresponding prompt condition.

The mask prompt provides dense visual region references, the point prompt provides sparse location cues, the scribble prompt provides sparse structural cues, and the null prompt provides no explicit spatial guidance.

Under the all visual-prompt setting, mask, point, and scribble prompts are jointly used, while the null prompt is excluded because it represents the absence of visual guidance.

A.3 Details of MultiCamo Construction

MultiCamo is constructed to support text-guided unified visual-prompt learning for camouflaged object segmentation (COS). Existing COS benchmarks mainly provide images and pixel-level masks, but lack explicit textual descriptions and unified organization of different visual-prompt conditions.

To address this limitation, we enrich existing COS datasets with text prompts and organize multiple visual-prompt settings, including mask, point, scribble, and null prompts. This design enables controlled evaluation of language-guided COS under different levels of visual guidance.

A.3.1 Prompt Organization

For each image, MultiCamo provides one or more text prompts together with four visual-prompt conditions.

The text prompt describes the camouflaged object, its spatial location, the surrounding scene, and the target-background relationship. The mask prompt is formulated as a dense visual region reference. The point prompt provides sparse location guidance, while the scribble prompt provides sparse structural information. The null prompt represents the no-visual-prompt condition, where no explicit spatial reference is provided.

A.3.2 Comparison with TG-COD

TG-COD provides text descriptions for camouflaged object detection, but does not systematically organize different visual-prompt conditions.

In contrast, MultiCamo is designed for text-guided unified visual-prompt learning, where text prompts are consistently paired with mask, point, scribble, and null visual prompts. This enables systematic comparison under different visual guidance levels and facilitates the study of promptable COS.

A.4 Text Prompt Generation and Quality Control

For each image, we construct text prompts according to the camouflaged object category, object size, spatial location, surrounding scene, and target-background relationship.

Different from pixel-level masks that provide explicit spatial supervision, text prompts aim to provide semantic context, describing what the target is, where it appears, and how it blends into the surrounding environment.

A.4.1 Prompt Template

To maintain consistency across different datasets, we use a unified prompt template:

> A [size] [category] is camouflaged in the [location] of the [scene].

The placeholders correspond to four semantic attributes: object size, object category, spatial location, and scene context.

For example, the template can be instantiated as:

> A small insect is camouflaged in the central region of the flowering vegetation scene.

This template keeps the descriptions concise while preserving the camouflage-related semantic information required for text-guided COS.

A.4.2 Candidate Prompt Generation

We first extract target-related attributes from the image and annotation, including object category, approximate scale, spatial location, and background scene.

GPT-4o is then used to transform these structured attributes into natural language descriptions. During generation, the model is encouraged to describe both the camouflaged target and its surrounding context rather than only providing the object category.

This process enables the generated text prompts to provide additional semantic cues beyond pixel-level annotations.

A.4.3 Manual Verification

After automatic generation, all text prompts are manually checked and refined.

Annotators compare each prompt with the original image and mask to ensure that it describes the correct target and remains consistent with the visible scene. Prompts containing incorrect object categories, inaccurate spatial descriptions, ambiguous scene information, or hallucinated content are corrected or removed.

We also avoid descriptions that directly reveal exact mask boundaries or object shapes, since text prompts are intended to provide semantic guidance rather than pixel-level supervision.

A.4.4 Quality Control Principles

We follow three principles during manual verification.

First, the prompt should faithfully describe visible image content without introducing absent objects or scene elements.

Second, the prompt should provide discriminative camouflage-context cues that help distinguish the target from background distractors.

Third, the prompt should remain concise and natural, avoiding redundant or overly complex descriptions that may introduce linguistic noise.

Following these principles, the final MultiCamo text prompts provide reliable semantic guidance while maintaining consistency across different COS datasets.

A.5 Visual Prompt Construction

MultiCamo organizes four visual-prompt conditions: mask, point, scribble, and null prompts. These prompts correspond to different levels of spatial guidance and annotation granularity.

A.5.1 Mask Prompt

The mask prompt is formulated as a dense visual region reference. It provides explicit object-level spatial guidance and represents the strongest visual-prompt condition in our unified formulation.

In practice, mask prompts are constructed from available pixel-level annotations and converted into the input format required by the visual prompt encoder.

A.5.2 Point Prompt

The point prompt provides sparse location guidance. It indicates representative foreground locations of the camouflaged object and follows point-supervised COS settings (Chen et al., 2025a).

Compared with mask prompts, point prompts require less annotation effort but provide weaker spatial constraints.

A.5.3 Scribble Prompt

The scribble prompt provides sparse structural guidance. Following scribble-supervised COS settings (He et al., 2023b; Han et al., 2025a), coarse foreground traces are used to indicate the approximate structure of the camouflaged object.

Scribble prompts provide richer spatial information than point prompts, while remaining much cheaper than dense mask annotations.

A.5.4 Null Prompt

The null prompt represents the no-visual-prompt condition. No explicit spatial reference is provided, and the model relies on the input image and text prompt to infer the camouflaged target.

A.5.5 Unified Visual-Prompt Representation

All visual prompts are converted into a unified representation before being fused with image features and text-conditioned semantic embeddings.

This design allows the framework to process different visual-prompt conditions within the same architecture and enables controlled comparison among mask, point, scribble, and null prompts.

A.6 Text Prompt Examples

Table 6 presents representative examples of text prompts in MultiCamo. The prompts describe semantic attributes including object category, spatial location, scene context, and target-background relationships.

A.7 Additional Experimental Analysis

A.7.1 Comparison under Single Visual-Prompt Settings

We further compare our framework with annotation-efficient COS methods under corresponding single visual-prompt settings, including no-prompt, point-prompt, and scribble-prompt conditions.

Table 6: Examples of text prompts in MultiCamo.

Attribute Type	**Example Text Prompt**
Size + Category + Location + Scene	A small insect is camouflaged in the central region of the flowering vegetation scene.
Category + Background	A medium owl is camouflaged in the lower-central region of the tree-branch scene.
Target-background relation	A large arctic fox is camouflaged in the central region of the snowy scene.
Underwater scene	A medium fish is camouflaged in the central region of the rocky underwater scene.

Table 7: Comparison with annotation-efficient COS methods under single visual-prompt settings. Conventional methods are evaluated under their original settings, while our method uses the same text prompt with a single visual-prompt condition. Here, N, P, and S denote null, point, and scribble prompts, respectively, and T+N, T+P, and T+S indicate text guidance combined with the corresponding visual prompts. ↑ (↓) denotes larger (smaller) is better.

Method	Prompt	GenCAMO				GenCOD10K				GenNC4K			
		$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
No-prompt / Unsupervised COS methods													
UCOS-DA (Zhang & Wu, 2023)	N	0.701	0.784	0.606	0.127	0.689	0.740	0.513	0.086	0.755	0.819	0.656	0.085
UCOD-DPL (Yan et al., 2025)	N	0.793	0.862	0.747	0.077	0.834	0.916	0.763	0.031	0.850	0.923	0.818	0.043
RISE (Du et al., 2025)	N	0.734	0.787	0.610	0.109	0.763	0.840	0.600	0.049	0.805	0.868	0.705	0.061
Ours (Text + Null)	T+N	0.815	0.865	0.735	0.075	0.825	0.890	0.700	0.035	0.848	0.900	0.780	0.046
Point-supervised COS methods													
Point-COD (Chen et al., 2025a)	P	0.798	0.872	0.727	0.074	0.784	0.859	0.650	0.042	0.822	0.889	0.748	0.051
Ours (Text + Point)	T+P	0.845	0.891	0.780	0.062	0.855	0.920	0.755	0.028	0.873	0.925	0.820	0.038
Scribble-supervised COS methods													
Scribble-COD (He et al., 2023b)	S	0.735	0.815	0.641	0.092	0.733	0.832	0.576	0.049	–	–	–	–
PPL (Han et al., 2025a)	S	0.755	0.843	0.697	0.090	0.751	0.849	0.656	0.047	0.794	0.874	0.754	0.060
Ours (Text + Scribble)	T+S	0.865	0.918	0.815	0.052	0.872	0.935	0.795	0.023	0.885	0.935	0.840	0.032

As shown in Table 7, Ours (Text + Null) is compared with unsupervised COS methods, Ours (Text + Point) is compared with point-supervised COS methods, and Ours (Text + Scribble) is compared with scribble-supervised COS methods.

Compared with existing methods, our framework introduces text prompts to provide additional semantic cues about object category, spatial location, surrounding scene, and target-background relationships.

The results demonstrate that language guidance can complement different levels of visual supervision and improve segmentation under diverse visual-prompt conditions.