UNIREFDET: UNIVERSAL REFERRING DETECTION OF SALIENT AND CAMOUFLAGED OBJECTS IN UNCONSTRAINED SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Salient object detection (SOD) and camouflaged object detection (COD) with language-driven understanding have attracted substantial attention in recent years. Existing studies predominantly focus on referential foreground object localization under individual task settings, while language-guided distinction between salient and camouflaged objects, particularly when these two visual attributes coexist in the same scene, remains underexplored. In this paper, we construct UniRefDet, a novel referring segmentation benchmark that unifies salient and camouflaged object detection tasks in unconstrained real-world scenes. The UniRefDet presents fine-grained referring expressions that explicitly describe object-environment relations. Built on this benchmark, we propose COREFNET, a SAM-based multimodal segmentation framework. We first propose a Language-Aware Semantic Adaptation (LASA), which inserts lightweight Text Adapters into the text encoder to adapt concept-level representations to fine-grained referring expressions. An Attribute Discrimination Module (ADM) is further designed, aiming to calibrate multiple proposals within each attribute and suppresses conflicting salient and camouflaged responses in coexistence scenes. Extensive experiments show that COREFNET achieves state-of-the-art performance across all scene configurations on UNIREFDET.

1 INTRODUCTION

Saliency and camouflage depend strongly on the relationship between an object and its surrounding environment. Salient object detection (SOD) and camouflaged object detection (COD) have achieved substantial progress (Pang et al., 2022; He et al., 2025), but most existing methods and benchmarks study the two settings separately. Representative SOD datasets, such as DUTS (Wang et al., 2017) and HKU-IS (Li & Yu, 2015), primarily contain salient objects, while COD datasets, such as COD10K (Fan et al., 2020) and CAMO (Le et al., 2019), mainly focus on camouflaged objects. Such settings are limited in unconstrained scenes, where salient and camouflaged objects may coexist and the same semantic category may exhibit different visual properties across environments. Recent methods such as VSCode (Luo et al., 2024) and USCNet (Zhou et al., 2025) have begun to jointly model the two settings, but remain primarily vision-driven.

Natural language provides fine-grained descriptions of target characteristics and object–environment relationships, offering explicit cues beyond visual appearance alone. Referring image segmentation (RIS) has achieved substantial progress in localizing and segmenting foreground objects by natural language expressions (Yu et al., 2016; Mao et al., 2016; Yang et al., 2022; Liu et al., 2023a; Huang et al., 2025; Jin et al., 2026; Cuttano et al., 2025), while recent multimodal benchmarks such as MMCSBench (Zhang et al., 2025) further demonstrate the value of language for describing object–environment relationships in camouflage scenes. Nevertheless, existing studies predominantly consider salient and camouflaged objects under separate settings, and leveraging language guidance to distinguish between them, particularly when both coexist in the same scene, remains underexplored.

To address this gap, we construct UNIREFDET, to the best of our knowledge, the first referring segmentation benchmark that jointly considers salient and camouflaged objects in unconstrained scenes. UNIREFDET covers four scene configurations: only salient objects, only camouflaged ob-

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.

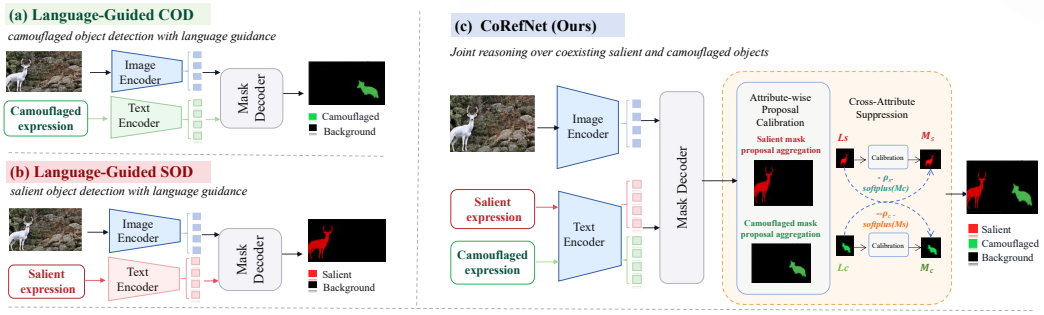

Figure 1: Comparison of task-specific language-guided segmentation and our unified paradigm. Existing approaches address salient or camouflaged objects separately, whereas CoRefNet jointly segments both types of objects and resolves their competing predictions in coexistence scenes.

jects, coexistence, and no-target scenes. It provides fine-grained referring expressions describing target identity, appearance, spatial location, and object–environment relationships. Beyond identifying the referred target, UniRefDet requires models to determine whether it is salient or camouflaged by jointly considering the language description and the surrounding scene.

Built upon this benchmark, we propose CoRefNet, a SAM-based multimodal segmentation framework to realize the unified setting illustrated in Fig. 1. CoRefNet introduces Language-Aware Semantic Adaptation (LASA) to adapt the pretrained text representation from concept-level prompts toward fine-grained referring expressions through lightweight Text Adapters. For coexistence scenes, the Attribute Discrimination Module (ADM) aggregates complementary proposals within each attribute and suppresses cross-attribute conflicts. Extensive experiments demonstrate that CoRefNet achieves state-of-the-art performance across all scene configurations on UniRefDet, reaching 82.53% mIoU and 82.54% oIoU, with particularly strong performance in scenes containing both salient and camouflaged objects.

Our contributions in this paper are as follows:

- We construct UniRefDet, the first unified referring segmentation benchmark for salient and camouflaged objects in unconstrained scenes, with fine-grained expressions that explicitly describe object details and their environmental context.
- We propose CoRefNet, a SAM-based multimodal segmentation framework for unified salient and camouflaged object segmentation. It introduces Language-Aware Semantic Adaptation (LASA) to adapt concept-level text representations to fine-grained referring expressions, and an Attribute Discrimination Module (ADM) to aggregate complementary proposals and suppress conflicting salient and camouflaged responses.
- We conduct extensive experiments and analyses on UniRefDet. CoRefNet achieves state-of-the-art performance across all scene configurations, demonstrating the effectiveness of fine-grained language guidance and cross-attribute discrimination.

2 Related Work

Salient and camouflaged object detection. Salient object detection (SOD) aims to identify visually distinctive foreground objects, with representative methods improving multi-scale aggregation, boundary refinement, and long-range modeling, such as PoolNet (Liu et al., 2019), BASNet (Qin et al., 2019), and Samba (He et al., 2025). In contrast, camouflaged object detection (COD) focuses on targets that blend into their surroundings and exploits subtle object–background discrepancies. Representative methods including SINet (Fan et al., 2020), ZoomNet (Pang et al., 2022), FSPNet (Huang et al., 2023), and ESCNet (Ye et al., 2025) improve camouflage perception through search-and-identification, multi-scale feature modeling, and edge–semantic cues. More recent studies have begun to unify the two settings. VSCode (Luo et al., 2024) jointly models SOD and COD with task prompts, FOCUS (You et al., 2025) incorporates them into universal foreground segmentation, and USCNet (Zhou et al., 2025) investigates the misdetection of salient and camouflaged

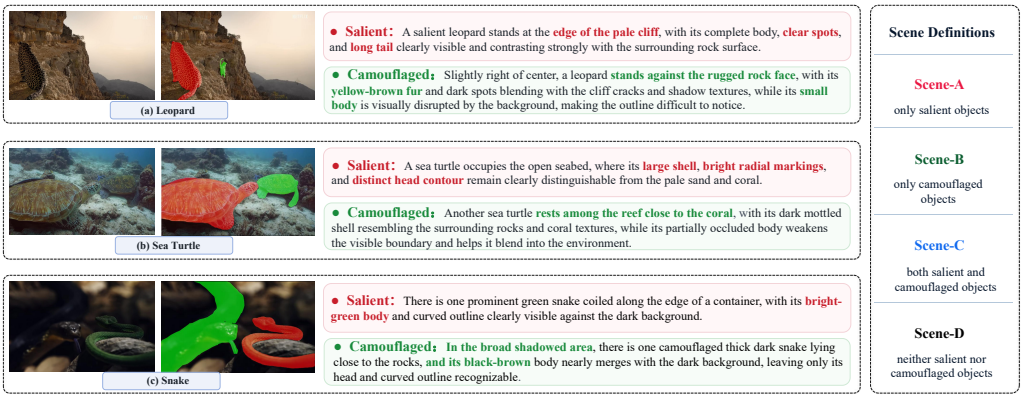

Figure 2: Scene configurations and representative examples in UNIREFDET. The benchmark considers four scene configurations according to the presence of salient and camouflaged objects, while examples show that the same object category can exhibit opposite contextual attributes.

objects and models their attribute relationships in unconstrained scenes. Existing methods remain primarily vision-driven, while language-guided distinction between salient and camouflaged objects remains underexplored. Our work addresses this gap with fine-grained referring expressions that provide explicit relational cues between targets and their surroundings.

Referring image segmentation. Referring image segmentation (RIS) predicts masks for targets specified by natural-language expressions through vision–language grounding. LAVT (Yang et al., 2022) enhances cross-modal interaction, PolyFormer (Liu et al., 2023b) formulates RIS as sequential polygon prediction, and GRES (Liu et al., 2023a) extends conventional RIS to multiple and absent referents. With the development of vision–language pretraining and segmentation foundation models, Prompt-RIS (Shang et al., 2024) connects CLIP and SAM through cross-modal prompting, while DETRIS (Huang et al., 2025) and TALENT (Jin et al., 2026) investigate parameter-efficient adaptation and suppression of non-target responses. Language-driven segmentation has also been extended to more general foundation models: SAMWISE (Cuttano et al., 2025) injects natural-language cues into SAM2 for referring video segmentation, and SAM3 (Carion et al., 2026) further introduces promptable concept segmentation with textual prompts. More recently, InstructSAM (Yuan et al., 2026) extends SAM3 to instruction-driven instance segmentation through language-conditioned instance queries. However, existing RIS mainly addresses *what* to segment, while the object–environment context explaining *why* a target is salient or camouflaged remains largely unexplored.

3 UNIREFDET BENCHMARK

3.1 BENCHMARK CONSTRUCTION

Dataset organization. Existing SOD and COD datasets predominantly focus on a single visual attribute and provide annotations for either salient or camouflaged objects (Wang et al., 2017; Li & Yu, 2015; Fan et al., 2020; Le et al., 2019). Existing referring segmentation benchmarks (Yu et al., 2016; Mao et al., 2016) use natural-language expressions to specify target objects, while MMCSBench (Zhang et al., 2025) provides fine-grained descriptions for camouflage-scene understanding. However, these benchmarks do not support joint language-guided segmentation of salient and camouflaged objects across unconstrained scene configurations. We derive UNIREFDET from USC12K (Zhou et al., 2025) by randomly sampling 6,000 images from its training and test splits while retaining their original split assignments. The sampled images are evenly distributed across the four scene configurations shown in Fig. 2, with 1,500 images per configuration, resulting in 4,400 training images and 1,600 test images. Scene membership follows the object-presence annotations inherited from USC12K rather than the identity of the source dataset. Each image is further paired with manually verified fine-grained referring expressions describing object details and their environmental context.

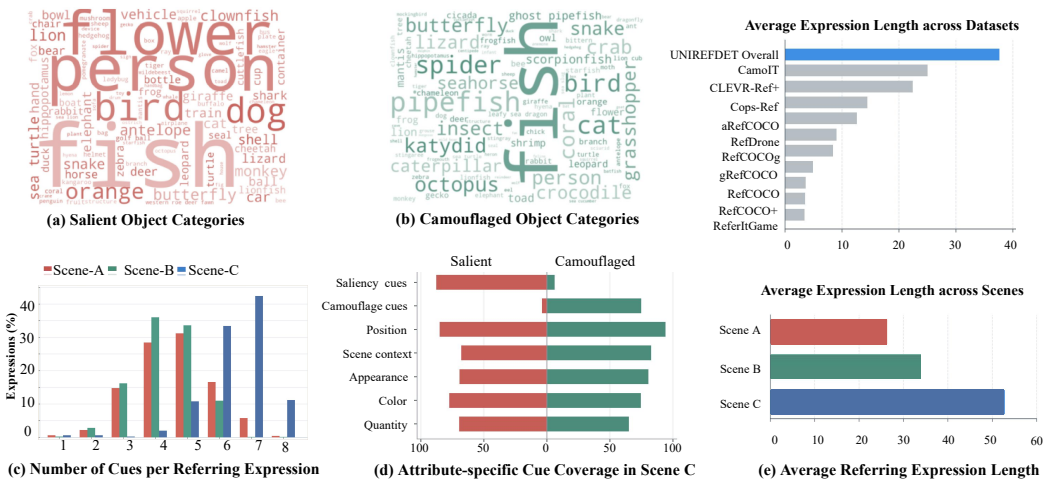

Figure 3: Statistics of UNIREFDET. (a–b) Category distributions of salient and camouflaged objects. (c) Distribution of the number of cues per referring expression across Scenes A–C. (d) Coverage of attribute-specific cues for salient and camouflaged objects in Scene C. (e) Average referring-expression length across benchmark datasets and scene configurations.

Fine-grained referring expressions. Rather than relying on category names or short object prompts, we construct fine-grained referring expressions. Each expression is structured based on eight key dimensions: quantity, category, color, appearance, position, scene context, saliency cues, and camouflage cues. Together, these dimensions capture both target-specific attributes and object-environment relationships, providing complementary information for target localization and how the target visually interacts with its surroundings. Such descriptions are particularly important for salient and camouflaged objects, since the same semantic category may exhibit different visual properties in different environments, as illustrated in Fig. 2.

Dataset annotation and verification. We use a VLM-assisted annotation pipeline to generate initial referring expressions with GPT-5.5 (OpenAI, 2026). We overlay the target-mask contour on the RGB image as spatial guidance. Let T_S and T_C denote the referring expressions for salient and camouflaged targets, respectively. Scene C uses the paired expression $T_{SC} = (T_S, T_C)$, whereas the no-target Scene D uses $T_D =$ "background". To ensure annotation quality, we employ a two-stage verification procedure. An independent vision-language model first examines the consistency among the image, target masks, and generated expression with respect to the eight key dimensions. Subsequently, three annotators perform two rounds of manual review and refinement, correcting factual errors and ambiguous references while removing unsupported or redundant descriptions.

3.2 Dataset Statistics and Analysis

Fig. 3 summarizes the semantic and linguistic statistics of UNIREFDET. The benchmark contains 3,000 salient and 3,000 camouflaged target groups. As shown in Fig. 3(a,b), 105 normalized target categories are shared across the two attributes, with 78.36% of camouflaged categories also appearing as salient targets, highlighting their context-dependent nature. Figure 3(c) shows that 87.4% of Scene C expressions contain six to eight key dimensions. The attribute-specific analysis in Fig. 3(d) further shows that saliency cues appear in 87.8% of T_S, whereas camouflage cues appear in 74.8% of T_C. Finally, the average expression lengths for Scenes A–C are 26.34, 33.67, and 52.85 words, respectively, as shown in Fig. 3(e). These statistics demonstrate the semantic diversity, linguistic richness, and complementary paired supervision of UNIREFDET.

4 COREFNET

Overview. As illustrated in Fig. 4, COREFNET consists of two main components: Language-Aware Semantic Adaptation (LASA) and the Attribute Discrimination Module (ADM). LASA

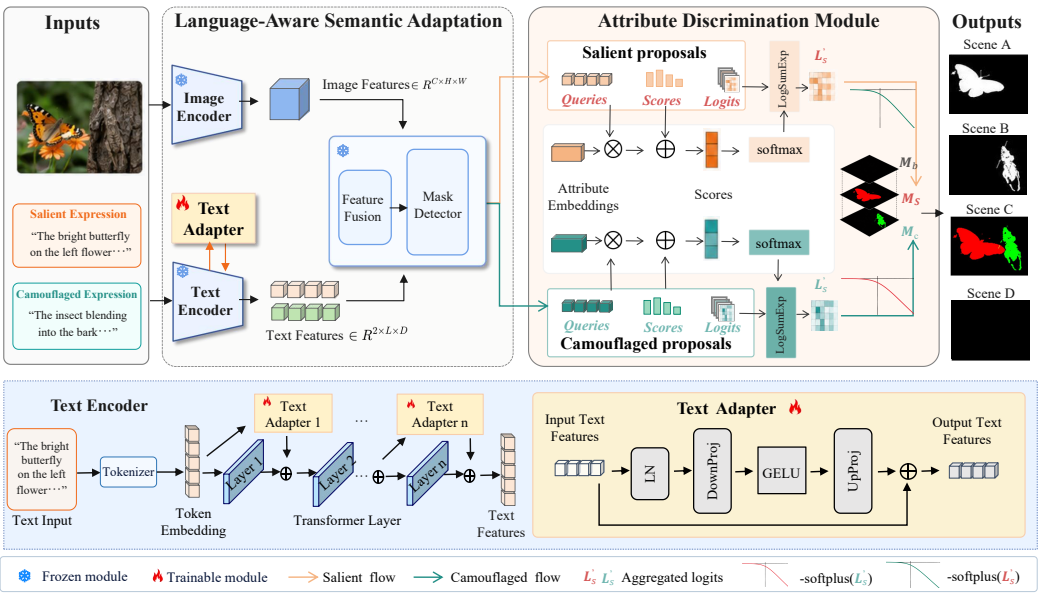

Figure 4: Overview of CoRefNet. Language-Aware Semantic Adaptation (LASA) adapts SAM 3 to fine-grained referring expressions, while the Attribute Discrimination Module (ADM) calibrates proposals and suppresses cross-attribute conflicts in coexistence scenes.

adapts SAM 3 from concept-level segmentation toward fine-grained referring understanding through lightweight Text Adapters. For each input expression, the mask detector produces multiple mask proposals. For $T \in \{T_S, T_C, T_D\}$, attribute-wise proposal calibration aggregates the proposals into a unified prediction. For the coexistence input T_{SC}, the salient and camouflaged proposals are further refined by cross-attribute suppression.

4.1 Language-Aware Semantic Adaptation

Fine-grained referring expressions contain target attributes and object–environment relations that are substantially richer than concept-level prompts. To adapt the pretrained SAM 3 text encoder to this setting, LASA inserts a lightweight Text Adapter into each transformer layer, as illustrated in Fig. 4. Given the features after self-attention, the adapter applies Layer Normalization, down-projection, GELU activation, and up-projection, and adds the resulting residual to the input features. The pretrained transformer layers retain their language-modeling capacity, while the adapters calibrate the representation toward fine-grained referring understanding. The adapted text features are fused with the image features and fed to the mask detector, which produces N proposals for each attribute $a \in \{S, C\}$:

$$\mathcal{P}_a = \{(Q_{a,j}, S_{a,j}, L_{a,j})\}_{j=1}^{N}, \tag{1}$$

where $Q_{a,j}$, $S_{a,j}$, and $L_{a,j}$ denote the query embedding, proposal score, and mask logit corresponding to proposal j, respectively.

4.2 Attribute Discrimination Module

In coexistence scenes, LASA produces separate salient and camouflaged proposal sets. For each attribute, the proposal set contains multiple candidates with different confidence scores and complementary predictions. Selecting only one proposal may therefore discard useful information from other candidates. Across attributes, salient and camouflaged logits may both respond strongly to the same pixels, resulting in conflicting attribute assignments. ADM addresses these two issues through Attribute-wise Proposal Calibration and Cross-Attribute Suppression.

Attribute-wise Proposal Calibration. To incorporate attribute information into proposal selection, we introduce a learnable attribute embedding p_a for each $a \in \{S, C\}$ and refine the proposal

score as

$$\tilde{S}_{a,j} = S_{a,j} + \lambda_p \frac{Q_{a,j}^\top p_a}{\|Q_{a,j}\|_2 \|p_a\|_2}, \qquad w_{a,j} = \mathrm{softmax}_{j \in \mathcal{K}_a} \left(\frac{\tilde{S}_{a,j}}{\tau_p} \right), \tag{2}$$

where $\mathcal{K}_a$ contains the top-K proposals ranked by $\tilde{S}_{a,j}$, λ_p scales the attribute similarity, and τ_p is the softmax temperature. The selected proposal logits are then aggregated using a weighted Log-SumExp:

$$L'_a(x) = \log \sum_{j \in \mathcal{K}_a} w_{a,j} \exp(L_{a,j}(x)), \qquad a \in \{S, C\}, \tag{3}$$

yielding the aggregated salient and camouflaged logits L'_S and L'_C.

Cross-Attribute Suppression. At the same pixel, the salient and camouflaged logits may both produce high responses, leading to conflicting attribute predictions. To reduce such conflicts, we increase the separation between the aggregated logits through mutual suppression:

$$M_S(x) = \frac{L'_S(x)}{\tau_S} - \rho_S \,\mathrm{softplus}(L'_C(x)), \tag{4}$$

$$M_C(x) = \frac{L'_C(x)}{\tau_C} - \rho_C \,\mathrm{softplus}(L'_S(x)), \tag{5}$$

where τ_S and τ_C are the corresponding temperature parameters, while ρ_S and ρ_C control the suppression strengths. With the learnable background logit $M_B(x) = b_B$, the final prediction is

$$\hat{Y}(x) = \arg \max_{a \in \{B,S,C\}} M_a(x). \tag{6}$$

4.3 Training Objective

We optimize CoRefNet using pixel-wise cross-entropy loss (Long et al., 2015), Dice loss (Milletari et al., 2016), and a cross-attribute ranking loss:

$$\mathcal{L} = \lambda_{\mathrm{ce}} \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{dice}} \mathcal{L}_{\mathrm{Dice}} + \lambda_{\mathrm{rank}} \mathcal{L}_{\mathrm{Rank}}. \tag{7}$$

Here, $\mathcal{L}_{\mathrm{CE}}$ supervises pixel-wise classification among background, salient, and camouflaged classes, while $\mathcal{L}_{\mathrm{Dice}}$ improves foreground segmentation quality. To further separate salient and camouflaged responses, we compute the ranking loss only for Scene C samples, where both attributes are present:

$$\mathcal{L}_{\mathrm{rank}} = \frac{1}{2} \sum_{a \in \{S,C\}} \frac{1}{|\Omega_a|} \sum_{x \in \Omega_a} [L'_{\bar{a}}(x) - L'_a(x) + m]_+, \tag{8}$$

where $\bar{S} = C$ and $\bar{C} = S$, Ω_a denotes the set of foreground pixels associated with attribute a, and $[\cdot]_+ = \max(\cdot, 0)$. The ranking loss is applied only to Scene C samples and enforces the correct attribute logit to exceed the competing one by a margin m.

5 Experiments

5.1 Experimental Setup

Dataset and Metrics. We conduct experiments on the proposed UniRefDet benchmark, with all compared methods trained on its training split under the same setting. On UniRefDet, we report mean intersection-over-union (mIoU) and overall intersection-over-union (oIoU) (Yang et al., 2022), both in percentage. Scene A:Sal and Scene B:Cam denote the salient and camouflaged targets in Scenes A and B, while Scene C:Sal and Scene C:Cam denote the corresponding targets in the coexistence Scene C; Overall evaluates the unified prediction over background, salient, and camouflaged regions. For cross-dataset evaluation, models are tested on RefCOCO+ (Yu et al., 2016), and TE-COD10K, TE-NC4K from RefTextCOD (Yan et al., 2025), using S_α (Fan et al., 2017), F_β^w (Margolin et al., 2014), E_ϕ^m (Fan et al., 2018), and $\mathcal{M}$ (Perazzi et al., 2012).

Table 1: Quantitative comparisons with representative methods on UNIREFDET. ↑ indicates that higher values are better. The best and second-best results are highlighted in **bold** and underlined, respectively. All metrics are reported as percentages (%).

Method	Scene A:Sal		Scene B:Cam		Scene C:Sal		Scene C:Cam		Overall	
	mIoU ↑	oIoU ↑	mIoU ↑	oIoU ↑	mIoU ↑	oIoU ↑	mIoU ↑	oIoU ↑	mIoU ↑	oIoU ↑
SOD/COD and Unified Segmentation										
PGNet (Xie et al., 2022)	50.23	56.77	26.01	30.25	45.17	49.53	10.17	10.67	42.95	47.02
ZoomNeXt (Pang et al., 2024)	83.10	84.33	68.79	76.04	72.64	77.71	53.24	62.40	75.51	80.07
CamoFormer (Yin et al., 2024)	82.22	85.05	64.69	66.38	71.94	71.40	50.19	50.94	72.99	73.93
VSCode (Luo et al., 2024)	83.07	83.14	62.45	67.24	76.11	77.16	54.26	58.56	73.91	76.00
MDSAM (Gao et al., 2024)	80.05	79.06	55.61	63.65	57.57	59.36	19.19	36.67	62.36	67.67
Samba (He et al., 2025)	82.84	85.32	70.04	76.03	76.17	77.07	59.90	62.20	76.91	79.33
USCNet (Zhou et al., 2025)	75.99	80.29	59.81	70.08	54.34	70.51	26.07	36.92	63.21	71.54
Referring Image Segmentation										
Prompt-RIS (Shang et al., 2024)	82.94	84.90	66.02	69.01	62.76	67.04	51.65	58.58	72.33	75.63
LISA (Lai et al., 2024)	90.26	87.21	74.98	77.16	75.11	75.94	60.48	53.23	80.17	78.71
SAMWISE (Cuttano et al., 2025)	56.54	63.27	31.56	40.07	41.24	51.57	12.95	26.96	43.47	51.76
DETRIS (Huang et al., 2025)	81.91	84.93	69.97	78.18	67.58	73.63	52.15	64.89	74.22	80.28
TALENT (Jin et al., 2026)	85.37	87.10	72.46	78.06	70.70	76.42	54.68	64.07	76.64	81.13
InstructSAM (Yuan et al., 2026)	85.89	86.55	73.89	77.48	63.18	65.66	51.71	50.95	71.93	73.77
COREFNET	**90.56**	**89.81**	**78.25**	**78.28**	**78.14**	**78.79**	**65.97**	**66.03**	**82.53**	**82.54**

Table 2: Zero-shot cross-dataset generalization on three target benchmarks without target-domain adaptation. ↑/↓ indicates that higher/lower is better. The best and second-best results are highlighted in **bold** and underlined, respectively. Rankings are determined from the full-precision scores before rounding.

Method	RefCOCO+ val (3,805)				TE-COD10K (2,026)				TE-NC4K (4,121)			
	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi^m \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi^m \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi^m \uparrow$	$\mathcal{M} \downarrow$
PGNet (Xie et al., 2022)	0.469	0.222	0.491	0.307	0.601	0.305	0.620	0.136	0.599	0.360	0.631	0.165
ZoomNeXt (Pang et al., 2024)	0.498	0.263	0.522	0.269	0.851	0.713	0.869	0.040	0.875	0.790	0.894	0.044
CamoFormer (Yin et al., 2024)	0.501	0.260	0.535	0.258	0.822	0.666	0.855	0.036	0.785	0.659	0.797	0.065
VSCode (Luo et al., 2024)	0.486	0.260	0.516	0.285	0.810	0.677	0.878	0.037	0.818	0.723	0.863	0.056
MDSAM (Gao et al., 2024)	0.521	0.227	0.537	0.228	0.786	0.552	0.774	0.058	0.807	0.627	0.770	0.080
Samba (He et al., 2025)	0.490	0.262	0.513	0.282	0.839	0.722	0.889	0.043	0.872	0.802	0.911	0.044
USCNet (Zhou et al., 2025)	0.517	0.257	0.526	0.267	0.805	0.565	0.765	0.056	0.852	0.707	0.826	0.060
Prompt-RIS (Shang et al., 2024)	0.509	0.265	0.580	0.238	0.813	0.659	0.857	0.043	0.818	0.704	0.848	0.062
LISA (Lai et al., 2024)	0.651	0.472	0.707	0.130	0.855	0.791	0.923	0.025	0.863	0.827	0.922	0.037
SAMWISE (Cuttano et al., 2025)	0.484	0.217	0.533	0.275	0.625	0.316	0.645	0.113	0.678	0.456	0.701	0.128
DETRIS (Huang et al., 2025)	0.513	0.257	0.567	0.233	0.852	0.718	0.894	0.027	0.879	0.806	0.915	0.036
TALENT (Jin et al., 2026)	0.502	0.268	0.531	0.266	0.818	0.669	0.840	0.036	0.863	0.779	0.886	0.041
InstructSAM (Yuan et al., 2026)	0.524	0.263	0.562	0.216	0.812	0.652	0.788	0.040	0.832	0.719	0.821	0.055
COREFNET	**0.652**	**0.479**	**0.732**	**0.121**	**0.907**	**0.880**	**0.955**	**0.018**	**0.907**	**0.895**	**0.948**	**0.026**

Implementation details. We build COREFNET on SAM 3 with Text Adapters of bottleneck width 64. Training consists of a 16-epoch LASA stage, a 2-epoch ADM stage, followed by a 4-epoch joint stage and a 1-epoch final refinement stage. We use AdamW with bfloat16 mixed precision and an effective batch size of 4 on four GPUs. During training, the shorter image side is randomly resized from 480 to 1008 pixels and padded to 1008×1008; evaluation uses a fixed 1008×1008 input. We set $K = 3$ in all main experiments. Full training details are provided in Appendix B.2.

5.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Comparison with representative methods. We compare COREFNET with representative SOD/COD and unified segmentation methods (Xie et al., 2022; Pang et al., 2024; Yin et al., 2024; Luo et al., 2024; Gao et al., 2024; He et al., 2025; Zhou et al., 2025) and referring image segmentation methods (Shang et al., 2024; Lai et al., 2024; Cuttano et al., 2025; Huang et al., 2025; Jin et al., 2026; Yuan et al., 2026). Table 1 shows that COREFNET achieves the best Overall performance among the compared methods, exceeding the respective second-best values by 2.36 and 1.41 points. The gain is particularly clear in coexistence scenes: on Scene C:Cam, the respective second-best mIoU/oIoU values of 60.48%/64.89% increase to 65.97%/66.03%. Figure 5 qualita-

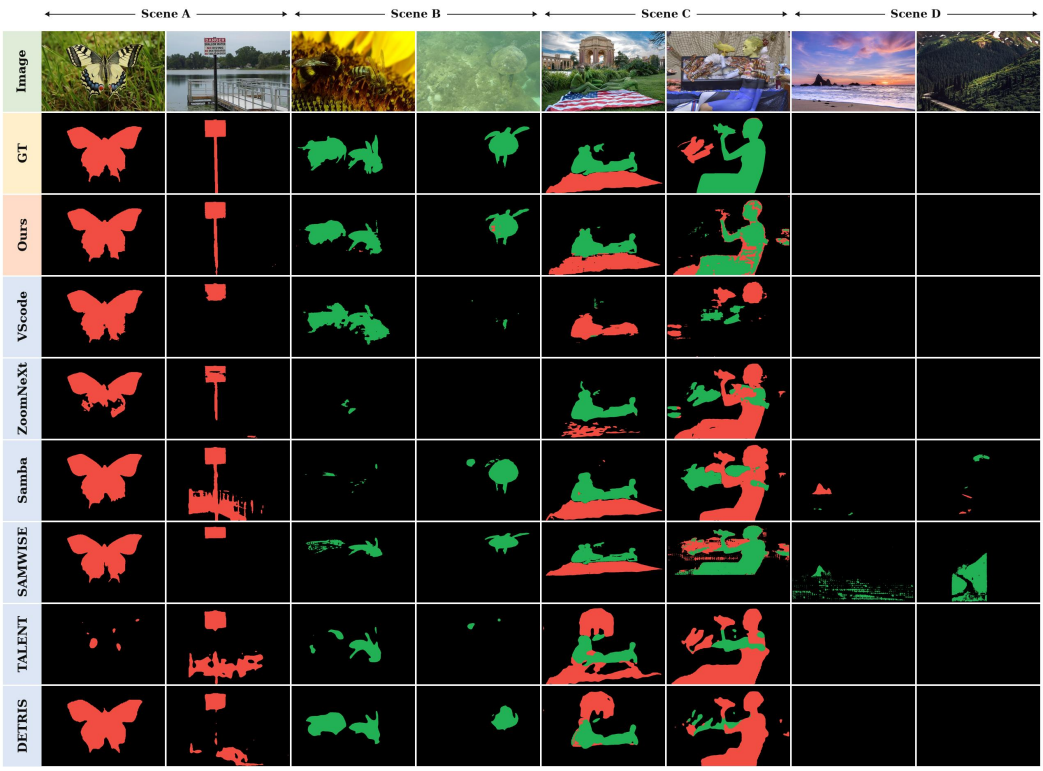

Figure 5: Qualitative comparison on UNIREFDET across all four scene configurations. Red and green denote salient and camouflaged regions, respectively.

Table 3: Ablation study of Language-Aware Semantic Adaptation (LASA). We evaluate the contribution of the Text Adapter and report mIoU/oIoU (%) for each subset as well as the overall performance.

Text Adapter	Segmentation Performance				
	Scene A:Sal	Scene B:Cam	Scene C:Sal	Scene C:Cam	Overall
–	34.70/29.24	27.75/35.28	64.79/49.10	37.20/18.32	51.61/45.34
✓	**90.56/89.81**	**78.25/78.28**	**78.14/78.79**	**65.97/66.03**	**82.53/82.54**

Table 4: Ablation study of the Attribute Discrimination Module (ADM) (mIoU/oIoU,%).

Component		Segmentation Performance		
Prop. Cal.	Cross Supp.	Scene C:Sal	Scene C:Cam	Overall
–	–	76.80/76.54	66.11/57.05	80.69/78.33
✓	–	76.76/77.58	66.07/57.82	80.95/78.42
–	✓	77.69/77.78	65.35/62.70	80.28/78.76
✓	✓	**78.14/78.79**	**65.97/66.03**	**82.53/82.54**

tively shows cleaner separation of salient and camouflaged regions. In the zero-shot evaluation of Table 2, COREFNET ranks first on all four metrics across RefCOCO+ val, TE-COD10K, and TE-NC4K; additional benchmarks and RefCOCO+ test splits are provided in Appendix C.2.

Qualitative Comparison. Figure 5 presents representative results across all four scene configurations. In Scenes A and B, COREFNET produces more complete masks, particularly for weakly contrasted camouflaged regions. In Scene C, competing methods often miss camouflaged targets or confuse the two attributes, whereas COREFNET more reliably separates salient and camouflaged regions. In the background-only Scene D, COREFNET suppresses spurious foreground responses, while several competing methods produce false salient or camouflaged predictions. These results further demonstrate the effectiveness of COREFNET in both cross-attribute discrimination and background suppression.

5.3 ABLATION STUDIES

Language-Aware Semantic Adaptation. Table 3 shows that removing the Text Adapter reduces Overall mIoU/oIoU from 82.53%/82.54% to 51.61%/45.34%, with the largest degradation on cam-

Table 5: Performance across different target sizes. Results are reported as mIoU/oIoU (%).

Configuration	Small $< 1\%$ ($n = 167$)	Medium 1–10% ($n = 783$)	Large $\geq 10\%$ ($n = 650$)
CoRefNet	**57.88/67.46**	**79.58/85.75**	**82.44/87.33**
w/o Text Adapter	29.41/35.64	44.54/49.52	44.90/49.03

Table 6: Effect of relation perturbations in Scene C. Bold indicates the largest performance degradation.

Condition	Δ mIoU	Δ oIoU	Δ Attr. Acc.
Relation Removed	−2.77	−4.00	−1.16
Same-Attribute Shuffled	−2.28	−4.33	−1.50
Opposite-Attribute Relation	**−5.62**	**−8.59**	**−3.53**

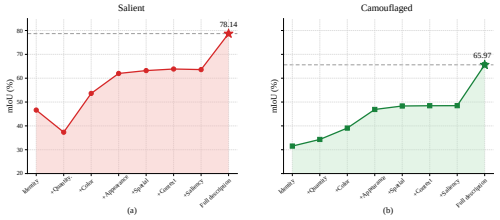

Figure 6: Cumulative linguistic-cue composition in Scene C.

Table 7: Analysis of linguistic cues in Scene C (mIoU, %). Bold highlights the best results in (a) and (c), and the largest degradation in (b).

Language setting	Salient	Camouflaged	Overall
(a) Baselines			
Task prompt	59.32	20.81	40.06
Category only	46.63	31.53	39.08
Full description	**78.14**	**65.97**	**72.06**
(b) Remove one cue from full description			
w/o Quantity	77.06	65.96	71.51
w/o Color	77.99	64.85	71.42
w/o Appearance	77.14	64.86	71.00
w/o Position	**74.48**	**57.10**	**65.79**
w/o Scene Context	78.43	65.63	72.03
w/o Saliency Cues	75.65	66.01	70.83
w/o Camouflage Cues	77.91	65.12	71.52
(c) Add one cue to Category			
+ Quantity	37.35	34.31	35.83
+ Color	**51.76**	41.53	46.65
+ Appearance	51.00	**44.47**	**47.73**
+ Position	43.17	33.10	38.13
+ Scene Context	22.35	17.55	19.95
+ Saliency Cues	47.57	37.84	42.71
+ Camouflage Cues	40.58	32.95	36.77

ouflaged targets in Scene C, where oIoU drops from 66.03% to 18.32%. Table 5 further shows consistent gains across small, medium, and large targets. These results confirm that adapting the pretrained text representation is essential for the fine-grained expressions in UniRefDet.

Attribute Discrimination Module. Table 4 evaluates proposal calibration and cross-attribute suppression. Removing both components reduces Overall oIoU from 82.54% to 78.33%. Proposal calibration improves the use of complementary mask proposals, while cross-attribute suppression is especially important for camouflaged targets in Scene C, where oIoU decreases from 66.03% to 57.82% when it is removed. Together, the two components provide the strongest performance.

5.4 Analysis of Linguistic Cues

Tables 6 and 7, together with Fig. 6, demonstrate the importance of fine-grained language in Scene C. Table 7(a) shows that task prompts and category-only expressions perform substantially worse than full descriptions, while Fig. 6 shows consistent gains as richer cues are incorporated. Table 6 further shows the largest degradation when the object–environment relation is replaced by the opposite attribute. This trend suggests that the benefit comes from combining complementary linguistic cues rather than relying on a single cue type. These results indicate that distinguishing salient and camouflaged objects requires not only object identity but also fine-grained contextual cues.

6 Conclusion

We introduce UniRefDet, a unified referring segmentation benchmark for salient and camouflaged objects in unconstrained scenes, and CoRefNet, a SAM 3-based framework tailored to fine-grained referring expressions. LASA adapts the text representation to detailed target and environmental cues, while ADM calibrates multiple proposals and suppresses conflicting cross-attribute responses. CoRefNet achieves the best performance among the compared methods across all scene configurations on UniRefDet, with clear gains in coexistence scenes, and shows strong zero-shot transfer to cross-dataset COD and referring-segmentation benchmarks. The results further highlight the importance of explicitly modeling object–environment relations when saliency and camouflage are treated as context-dependent visual attributes.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting, translation, language editing, literature organization, methodological and experimental feedback, code-level debugging suggestions, and interpretation of experimental results. In addition, a vision–language model was used as part of the mask-guided annotation pipeline to generate initial referring-expression drafts, as described in Section 3.1; these drafts were subsequently checked by an independent verification model and manually reviewed. All reported numerical results were produced by the authors' experimental pipeline, and all AI-assisted text, analyses, annotations, code suggestions, and scientific claims were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

UNIREFDET is a language augmentation of an existing visual benchmark rather than a collection of new personal or sensitive data. The work does not conduct human-subject experiments or infer sensitive personal attributes. Human involvement is limited to annotation quality control, and the released annotations are task-oriented descriptions of visible objects and their local visual context. Because inherited source data and model-assisted annotations may contain representation or description biases, all generated descriptions are subject to manual review. Any public release of UNIREFDET will follow the licenses and redistribution terms of the underlying source datasets.

REPRODUCIBILITY STATEMENT

Section 3 specifies the task definition, benchmark construction, split sizes, and annotation protocol; Section 4 gives the model formulation and training objectives; and Section 5 together with Appendix B.2 reports the evaluation protocol, principal optimization settings, stage-wise learning rates and loss weights, data augmentation, and the random seed used for controlled ablations. All final experiments use the fixed 4,400/1,600 train/test split. We will release the code, configuration files, split files, and evaluation scripts with the benchmark to facilitate independent reproduction.

REFERENCES

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026.

Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, and Giuseppe Averta. SAMWISE: Infusing wisdom in SAM2 for text-driven video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3395–3405, 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 Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 698–704, 2018. doi: 10.24963/ijcai.2018/97.

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.

Shixuan Gao, Pingping Zhang, Tianyu Yan, and Huchuan Lu. Multi-scale and detail-enhanced segment anything model for salient object detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9894–9903, 2024. doi: 10.1145/3664647.3680650.

Jiahao He, Keren Fu, Xiaohong Liu, and Qijun Zhao. Samba: A unified mamba-based framework for general salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25314–25324, 2025.

Jiaqi Huang, Zunnan Xu, Ting Liu, Yong Liu, Haonan Han, Kehong Yuan, and Xiu Li. Densely connected parameter-efficient tuning for referring image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3653–3661, 2025. doi: 10.1609/aaai.v39i4.32380.

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.

Shuo Jin, Siyue Yu, Bingfeng Zhang, Chao Yao, Meiqin Liu, and Jimin Xiao. TALENT: Target-aware efficient tuning for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 7472–7482, 2026.

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.

Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5455–5463, 2015. doi: 10.1109/CVPR.2015.7299184.

Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23592–23601, 2023a.

Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R. Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18653–18663, 2023b.

Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3917–3926, 2019.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015. doi: 10.1109/CVPR.2015.7298965.

Ziyang Luo, Nian Liu, Wangbo Zhao, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Khan, and Junwei Han. VSCode: General visual salient and camouflaged object detection with 2d prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17169–17180, 2024.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11–20, 2016.

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. doi: 10.1109/CVPR.2014.39.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571. IEEE, 2016. doi: 10.1109/3DV.2016.79.

OpenAI. GPT-5.5 System Card. https://openai.com/index/gpt-5-5-system-card/, April 2026.

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. doi: 10.1109/TPAMI.2024.3417329.

Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 733–740, 2012. doi: 10.1109/CVPR.2012.6247743.

Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. BASNet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7479–7489, 2019.

Chao Shang, Zichen Song, Heqian Qiu, Lanxiao Wang, Fanman Meng, and Hongliang Li. Prompt-driven referring image segmentation with instance contrasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4124–4134, 2024.

Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 136–145, 2017.

Chenxi Xie, Changqun Xia, Mingcan Ma, Zhirui Zhao, Xiaowu Chen, and Jia Li. Pyramid grafting network for one-stage high resolution saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11717–11726, 2022.

Weiqi Yan, Lvhai Chen, Shengchuan Zhang, Yan Zhang, and Liujuan Cao. Scout: Semi-supervised camouflaged object detection by utilizing text and adaptive data selection. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 2170–2178, 2025. doi: 10.24963/ijcai.2025/242.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18155–18165, 2022.

Sheng Ye, Xin Chen, Yan Zhang, Xianming Lin, and Liujuan Cao. ESCNet: Edge-semantic collaborative network for camouflaged object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20053–20063, 2025.

Bowen Yin, Xuying Zhang, Deng-Ping Fan, Shaohui Jiao, Ming-Ming Cheng, Luc Van Gool, and Qibin Hou. Camoformer: Masked separable attention for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10362–10374, 2024. doi: 10.1109/TPAMI.2024.3438565.

Zuyao You, Lingyu Kong, Lingchen Meng, and Zuxuan Wu. FOCUS: Towards universal foreground segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9580–9588, 2025.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pp. 69–85, 2016. doi: 10.1007/978-3-319-46475-6_5.

Yuqian Yuan, Wentong Li, Zhaocheng Li, Yutong Lin, Juncheng Li, Siliang Tang, Jun Xiao, Yueting Zhuang, and Wenqiao Zhang. Instructsam: Segment any instance with any instructions. *arXiv preprint arXiv:2605.26102*, 2026.

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*, volume 38, 2025. doi: 10.52202/085713-1028.

Zhangjun Zhou, Yiping Li, Chunlin Zhong, Jianuo Huang, Jialun Pei, Hua Li, and He Tang. Rethinking detecting salient and camouflaged objects in unconstrained scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22372–22382, 2025.

A BENCHMARK DETAILS

We provide a detailed characterization of UNIREFDET to complement the benchmark construction in Sec. 3. Beyond the number of images in each scene, we analyze category frequency, referring-expression length, and image provenance. These statistics make the benchmark composition explicit and provide context for the scene-wise and overall evaluations, especially the coexistence setting where salient and camouflaged targets must be distinguished within the same image.

Figure 7 summarizes the global composition of UNIREFDET. The four scene configurations are balanced with 1,500 images each, preventing any scene from dominating the benchmark by sample count. Target categories, however, remain naturally long-tailed, preserving the heterogeneous object distribution inherited from the source data rather than imposing an artificial class-balanced prior.

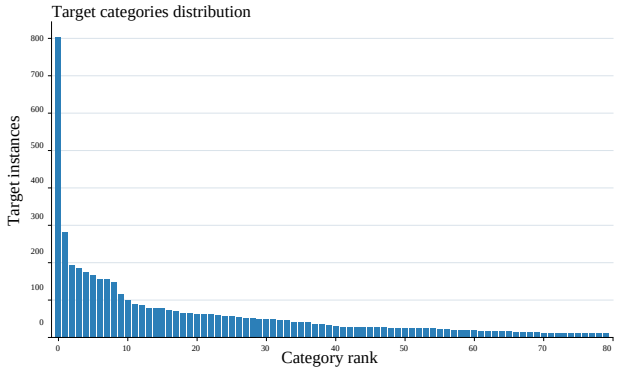

Figure 7: Overall composition of UNIREFDET. The four scene configurations are balanced, while target categories follow a long-tailed distribution.

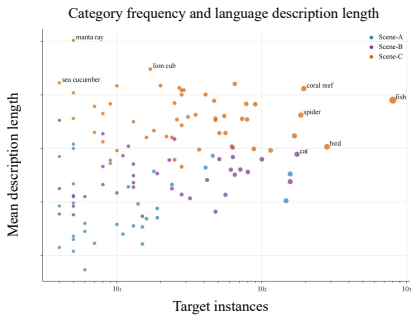

Figure 8: Category frequency and referring-expression length. Category frequency and mean English expression length are shown for Scenes A–C.

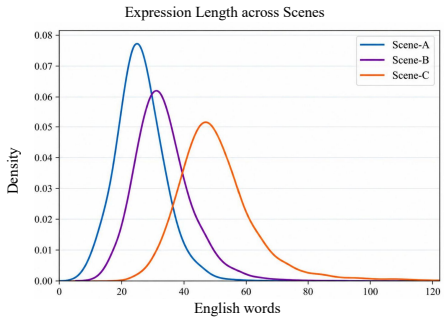

Figure 9: Distribution of expression lengths across scenes. Expression length is measured by English words for Scenes A–C.

The language statistics reveal a complementary source of difficulty. Figure 8 relates category frequency to mean expression length, showing that linguistic complexity is not explained solely by category frequency. Figure 9 further compares expression-length distributions across Scenes A–C. Scene C shifts toward longer descriptions, consistent with the coexistence setting where the expression must identify the target while disambiguating it from an object with the opposite visual attribute.

Thus, UNIREFDET varies in both visual composition and the linguistic information needed to resolve the target.

UNIREFDET is derived from USC12K (Zhou et al., 2025) by randomly sampling 6,000 images from its original training and test splits while preserving split assignments, yielding 4,400 training images and 1,600 held-out test images. The scene configurations inherit USC12K's manually verified object-presence labels and are not determined by source dataset. Figure 10 reports image provenance to make this construction transparent.

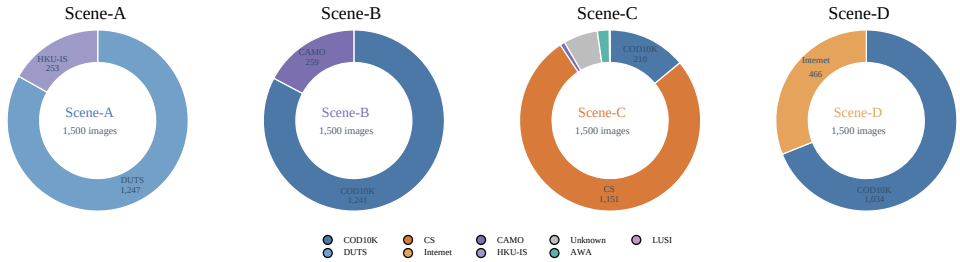

Figure 10: Source composition of UNIREFDET. The figure reports the number of sampled images inherited from each source across the four scene configurations.

Source composition is not uniform across scenes, reflecting USC12K and the availability of suitable images in its constituent sources. We therefore complement internal UNIREFDET evaluation with zero-shot transfer to external camouflage and referring-segmentation benchmarks in Sec. C.2, providing complementary evidence of transfer beyond the UNIREFDET test distribution.

Finally, Figs. 11 and 12 provide representative qualitative examples of UNIREFDET. Figure 11 illustrates the four scene configurations together with their corresponding referring annotations, while Fig. 12 highlights different ways in which camouflaged targets visually interact with their surroundings. These examples complement the aggregate statistics by making the target–environment relations encoded in the benchmark explicit.



Figure 11: Representative examples from UNIREFDET. Scenes A–D correspond to salient-only, camouflaged-only, coexistence, and background configurations, respectively. The examples illustrate how referring annotations describe target appearance, location, and target–environment relations.

B EVALUATION AND IMPLEMENTATION DETAILS

This section specifies the evaluation protocol used throughout the paper and summarizes the implementation details required to reproduce the benchmark and our final model.

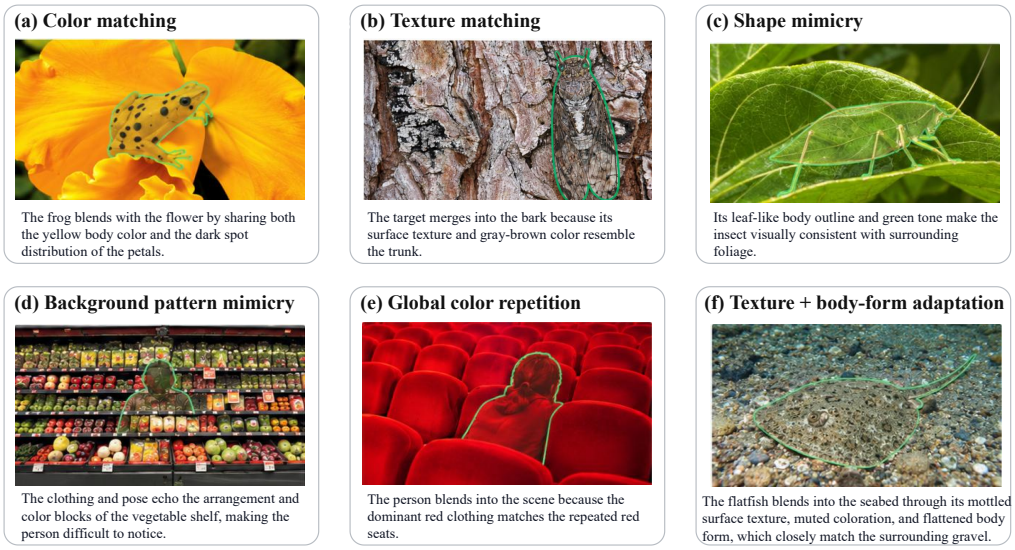

Figure 12: Representative camouflage mechanisms in UNIREFDET. The examples illustrate different object–environment relations, including texture matching, background-pattern mimicry, global color repetition, and texture/body-form adaptation.

B.1 EVALUATION PROTOCOL

Scene-wise metrics. Let g denote one evaluation unit in

$$\mathcal{G} = \{\text{A:Sal}, \text{B:Cam}, \text{C:Sal}, \text{C:Cam}, \text{D:Bg}\}.$$

For sample i in group g, let P_i^g and G_i^g denote the predicted and ground-truth binary masks for the evaluated attribute, respectively. The per-image IoU is

$$\text{IoU}_i^g = \frac{|P_i^g \cap G_i^g|}{|P_i^g \cup G_i^g|}. \tag{9}$$

We report scene-wise mIoU by averaging the per-image IoUs,

$$\text{mIoU}_g = \frac{1}{N_g} \sum_{i=1}^{N_g} \text{IoU}_i^g, \tag{10}$$

whereas scene-wise oIoU first accumulates intersections and unions over the whole evaluation unit,

$$\text{oIoU}_g = \frac{\sum_i |P_i^g \cap G_i^g|}{\sum_i |P_i^g \cup G_i^g|}. \tag{11}$$

Overall mIoU. Overall mIoU is the macro-average over the five scene-attribute evaluation units:

$$\text{Overall-mIoU} = \frac{1}{5} \sum_{g \in \mathcal{G}} \text{mIoU}_g \,. \tag{12}$$

Thus, Scene A:Sal, Scene B:Cam, Scene C:Sal, Scene C:Cam, and Scene D:Bg contribute equally, independently of their pixel counts.

Overall oIoU. Overall oIoU is computed differently from the scene-balanced Overall mIoU. All test images from Scenes A–D are evaluated jointly in the unified three-class label space

$$\mathcal{Y} = \{B, S, C\},$$

where B, S, and C denote background, salient, and camouflaged pixels, respectively.

Let $y_{n,p}$ and $\hat{y}_{n,p}$ denote the ground-truth and predicted labels of pixel p in image n. We accumulate a single pixel-level confusion matrix $\mathbf{M} \in \mathbb{N}_0^{3\times 3}$ over the complete test set:

$$M_{ij} = \sum_n \sum_p \mathbf{1}[y_{n,p} = i \wedge \hat{y}_{n,p} = j], \qquad i, j \in \{B, S, C\}. \tag{13}$$

For class k, its globally accumulated union is

$$U_k = \sum_j M_{kj} + \sum_i M_{ik} - M_{kk}, \qquad k \in \{B, S, C\}. \tag{14}$$

Overall oIoU is then computed by globally aggregating intersections and unions across all three semantic classes:

$$\text{Overall-oIoU} = \frac{\sum_{k\in\{B,S,C\}} M_{kk}}{\sum_{k\in\{B,S,C\}} U_k}. \tag{15}$$

For diagnostic purposes, we additionally compute the class-specific global IoU

$$\text{IoU}_k^{\text{global}} = \frac{M_{kk}}{U_k}, \qquad k \in \{B, S, C\}, \tag{16}$$

but their class-wise arithmetic mean is not used as the reported Overall mIoU or Overall oIoU.

Therefore, Overall mIoU provides a scene-balanced view in which the five scene–attribute conditions contribute equally, whereas Overall oIoU provides a pixel-level global view over the complete unified three-class test set. The foreground-only aggregate (FG Overall) used in Appendix D macro-averages only $\{\text{A:Sal}, \text{B:Cam}, \text{C:Sal}, \text{C:Cam}\}$ and is used only for diagnostic sensitivity analysis.

Complementary metrics. For additional binary-map evaluation, let $S \in [0,1]^{H\times W}$ be a normalized prediction map and $G \in \{0,1\}^{H\times W}$ the corresponding binary ground truth. We report AUC, maximal F-measure $F_{\max}^{\beta}$, weighted F-measure F_{β}^{w}, mean/max E-measure $E_{\phi}^{m}/E_{\phi}^{\max}$, structure measure S_{α}, mean absolute error $\mathcal{M}$, and mIoU@0.5. For threshold τ,

$$\text{Prec}(\tau) = \frac{TP(\tau)}{TP(\tau) + FP(\tau)}, \qquad \text{Rec}(\tau) = \frac{TP(\tau)}{TP(\tau) + FN(\tau)}, \tag{17}$$

and

$$F_{\beta}(\tau) = \frac{(1+\beta^2)\,\text{Prec}(\tau)\,\text{Rec}(\tau)}{\beta^2\,\text{Prec}(\tau) + \text{Rec}(\tau)}, \qquad F_{\max}^{\beta} = \max_{\tau} F_{\beta}(\tau). \tag{18}$$

AUC integrates the true-positive rate against the false-positive rate over thresholds; $\mathcal{M}$ is the pixel-wise mean absolute error; and mIoU@0.5 binarizes S at 0.5 before averaging the per-image IoUs. We follow the original definitions and official protocols for F_{β}^{w} (Margolin et al., 2014), E_{ϕ} (Fan et al., 2018), and S_{α} (Fan et al., 2017). SI-AUC and SI-$F_{\max}^{\beta}$ are their size-invariant counterparts used to reduce the dominance of large objects. For RefCOCO+ and RefTextCOD-derived test sets, we use the same score normalization and metric implementation for all methods and perform no target-domain fine-tuning.

B.2 Implementation Details

Benchmark setup. All compared methods are retrained on the 4,400-image UniRefDet training split and evaluated on the held-out 1,600-image test split. Every method uses the same train/test lists, target masks, and evaluation scripts. Referring-segmentation methods additionally receive the corresponding expressions, whereas SOD/COD and unified baselines use the supervision supported by their original formulations. Test images are never used for optimization or model selection.

Backbones and baseline adaptation. We retain the model-specific backbone and adaptation strategy of each method. PGNet uses a ResNet-18/Swin dual branch, ZoomNeXt PVTv2-B5, CamoFormer PVTv2-B4, VSCode Swin-T, MDSAM SAM ViT-B, Samba VMamba-S, and USCNet SAM ViT-H. Prompt-RIS combines CLIP ViT-B/16 with SAM ViT-B; LISA uses a LLaVA/LLaMA-7B language backbone with CLIP ViT-L/14 and SAM ViT-H; SAMWISE uses SAM2 Base+ and

RoBERTa; DETRIS and TALENT combine CLIP ViT-B/16 with DINOv2 ViT-B/14; Instruct-SAM uses a Qwen3-VL 2B-scale multimodal backbone with SAM3 segmentation components; and CoRefNet is built on SAM 3. Because these methods were originally developed under different task formulations, semantic models are adapted to predict the three benchmark attributes directly, whereas prompt- and language-conditioned methods retain their task- or query-conditioned prediction mechanisms.

Unified three-class inference. All predictions are converted into mutually exclusive background, salient, and camouflaged class maps before evaluation. For methods that produce separate salient and camouflaged predictions for the same Scene-C image, we combine the two predictions at the score level: each foreground score is retained for its corresponding attribute, the background score is conservatively combined across the two queries, and each pixel is assigned the class with the highest score. Scene D contains neither salient nor camouflaged targets and is evaluated as a background-only condition; for CoRefNet, we use the dedicated background expression defined in our formulation. The Overall aggregation rules then follow the definitions in Appendix B.

CoRefNet default configuration. The default configuration used throughout the main experiments builds CoRefNet on SAM 3 with Text Adapters of bottleneck width 64 and residual scale 1.0, together with the ADM described in Sec. 4.2. During training, the shorter image side is randomly resized from 480 to 1008 pixels in steps of 32 and then padded to 1008×1008; evaluation uses a fixed 1008×1008 input. Data augmentation includes scale augmentation and bounding-box perturbation with standard deviation 0.1 and maximum displacement 20. Inputs are normalized with mean and standard deviation both set to 0.5. The default number of aggregated proposals is fixed to $K = 3$.

Optimization schedule. We use AdamW with bfloat16 automatic mixed precision and gradient clipping at 0.1. Training uses 4-GPU distributed data parallelism with one sample per GPU, giving an effective batch size of 4. All stages use random seed 123.

Training begins with a 16-epoch Stage-A LASA phase on the complete training split, in which only the Text Adapter is optimized. Stage B trains only ADM for 2 epochs using Scene-C coexistence samples. Stage C then jointly optimizes the Text Adapter and ADM for 4 epochs on the complete training split, followed by a 1-epoch Stage-D refinement with the same trainable modules. Stage C is initialized from Stage B, and Stage D from Stage C. The pretrained SAM 3 image encoder and mask decoder remain frozen in all four stages. We use an inverse-square-root learning-rate schedule with timescale 3. Stage A uses scheduler warmup/cooldown parameters of $3/3$, whereas Stages B–D use $1/1$, respectively. The stage-wise learning rates and ADM loss weights are summarized in Table 8.

Table 8: Stage-wise optimization settings of CoRefNet. The CE/Dice/Rank. columns denote the ADM loss weights used in Stages B–D. The pretrained image encoder and mask decoder remain frozen throughout training.

Stage	Training data	Trainable modules	CE/Dice/ Rank.	ADM LR	Text Adapter LR
A	All scenes	Text Adapter	–	–	6.4×10^{-4}
B	Scene C	ADM	10 / 5 / 3	2×10^{-4}	–
C	All scenes	ADM + Text Adapter	6 / 3 / 1.8	1×10^{-4}	2×10^{-6}
D	All scenes	ADM + Text Adapter	8 / 4 / 2.4	2×10^{-5}	8×10^{-7}

C Additional Quantitative Results

C.1 Complementary Metrics on UniRefDet

The main paper uses mIoU and oIoU as primary measures, but a single metric family cannot fully characterize dense prediction quality. We therefore reevaluate representative methods on the complete 1,600-image UniRefDet test split using complementary measures. AUC and $F^{\beta}_{\max}$ capture threshold-dependent discrimination; SI-AUC and SI-$F^{\beta}_{\max}$ reduce object-size bias; F^{w}_{β} emphasizes

Table 9: Additional quantitative comparison on the UNIREFDET test split. ↑/↓ indicates that higher/lower is better. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	AUC↑	SI-AUC↑	$F^{\beta}_{\max}$ ↑	SI-$F^{\beta}_{\max}$ ↑	F^{w}_{β} ↑	E^{m}_{ϕ} ↑	$E^{\max}_{\phi}$ ↑	S_{α} ↑	$\mathcal{M}$ ↓	mIoU@0.5↑
PGNet (Xie et al., 2022)	0.721	0.656	0.427	0.545	0.384	0.653	0.655	0.605	0.107	0.329
ZoomNeXt (Pang et al., 2024)	0.889	0.842	0.797	0.781	0.773	0.891	0.893	0.838	0.036	0.713
CamoFormer (Yin et al., 2024)	0.811	0.771	0.702	0.634	0.646	0.774	0.782	0.744	0.084	0.568
VSCode (Luo et al., 2024)	0.886	0.832	0.779	0.769	0.752	0.891	0.894	0.828	0.036	0.695
MDSAM (Gao et al., 2024)	0.788	0.744	0.623	0.585	0.586	0.729	0.730	0.714	0.094	0.523
Samba (He et al., 2025)	0.909	0.861	0.813	0.802	0.791	0.909	0.911	0.854	0.032	0.737
USCNet (Zhou et al., 2025)	0.837	0.795	0.698	0.685	0.676	0.795	0.797	0.782	0.049	0.622
Prompt-RIS (Shang et al., 2024)	0.877	0.837	0.762	0.755	0.725	0.878	0.888	0.815	0.046	0.669
LISA (Lai et al., 2024)	0.915	0.865	0.828	0.823	0.810	0.920	0.922	0.863	0.031	0.761
SAMWISE (Cuttano et al., 2025)	0.719	0.654	0.511	0.545	0.467	0.839	0.841	0.636	0.089	0.396
DETRIS (Huang et al., 2025)	0.873	0.818	0.782	0.761	0.755	0.884	0.887	0.825	0.033	0.679
TALENT (Jin et al., 2026)	0.890	0.840	0.794	0.774	0.772	0.897	0.900	0.840	0.029	0.708
InstructSAM (Yuan et al., 2026)	0.883	0.844	0.789	0.779	0.769	0.879	0.881	0.833	0.046	0.710
COREFNET	**0.932**	**0.898**	**0.859**	**0.858**	**0.843**	**0.927**	**0.930**	**0.885**	**0.027**	**0.796**

spatially meaningful errors; $E^{m}_{\phi}/E^{\max}_{\phi}$ and S_{α} measure alignment and structural consistency; MAE measures pixel-wise deviation; and mIoU@0.5 evaluates masks at a fixed threshold. Together, these metrics test whether the mIoU/oIoU trends remain stable under different evaluation criteria. Higher is better except for MAE.

Table 9 shows that the main mIoU/oIoU conclusions are preserved across all ten complementary measures. COREFNET obtains the best AUC/SI-AUC (0.932/0.898), $F^{\beta}_{\max}$/SI-$F^{\beta}_{\max}$ (0.859/0.858), F^{w}_{β} (0.843), $E^{m}_{\phi}/E^{\max}_{\phi}$ (0.927/0.930), S_{α} (0.885), and mIoU@0.5 (0.796), together with the lowest MAE (0.027). The consistent gains on standard and size-invariant metrics indicate that the advantage is not driven only by large foreground regions, while the strong E-measure and S_{α} values confirm improved structural agreement. The lowest MAE and best mIoU@0.5 further show that the trend holds for both continuous score maps and thresholded masks.

C.2 CROSS-DATASET GENERALIZATION

The main paper reports cross-dataset results on RefCOCO+ val, TE-COD10K, and TE-NC4K in Table 2. We extend the analysis to TE-CHAMELEON, TE-CAMO, and RefCOCO+ testA/testB. The camouflage benchmarks test transfer across visual domains, while RefCOCO+ evaluates language-conditioned segmentation on a standard referring benchmark outside UNIREFDET. All methods use the same continuous score-map protocol as in the main paper, with no target-domain fine-tuning, model selection, or adaptation on the evaluation split; we refer to this setting as *zero-shot*.

Table 10 summarizes the two additional camouflage benchmarks and therefore measures transfer under changes in image statistics and target appearance. Across TE-CHAMELEON and TE-CAMO, COREFNET obtains the best value on every reported metric. It reaches 0.921/0.914/0.966/0.019 on TE-CHAMELEON, with InstructSAM second, and 0.872/0.858/0.925/0.044 on TE-CAMO, with DETRIS second. The gains therefore hold across structure-aware, alignment-based, weighted F-measure, and pixel-error metrics. Together with TE-COD10K and TE-NC4K in the main paper, the results show consistent transfer across four camouflage test domains.

We next evaluate language-conditioned transfer on RefCOCO+ testA and testB. The main paper already reports RefCOCO+ val, so Table 11 adds the two held-out test splits without duplicating validation results. On RefCOCO+ testA and testB, COREFNET again achieves the best results on all four metrics, with LISA consistently second. The margins are smaller than on the camouflage benchmarks, but the ordering remains stable across testA, testB, and the validation split in Table 2. This supports transfer of the learned language-conditioned segmentation capability beyond a single RefCOCO+ split and complements the camouflage-domain results above.

Table 10: Zero-shot cross-dataset generalization on TE-CHAMELEON and TE-CAMO. ↑/↓ indicates that higher/lower is better. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	TE-CHAMELEON (76)				TE-CAMO (250)			
	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi^m \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi^m \uparrow$	$\mathcal{M} \downarrow$
PGNet (Xie et al., 2022)	0.585	0.341	0.603	0.194	0.534	0.282	0.559	0.206
ZoomNeXt (Pang et al., 2024)	0.872	0.771	0.885	0.042	0.830	0.732	0.847	0.072
CamoFormer (Yin et al., 2024)	0.849	0.733	0.876	0.044	0.754	0.626	0.766	0.089
VSCode (Luo et al., 2024)	0.801	0.682	0.850	0.053	0.750	0.638	0.786	0.093
MDSAM (Gao et al., 2024)	0.820	0.648	0.817	0.061	0.760	0.580	0.723	0.110
Samba (He et al., 2025)	0.859	0.770	0.907	0.042	0.828	0.747	0.870	0.076
USCNet (Zhou et al., 2025)	0.835	0.643	0.804	0.058	0.825	0.667	0.785	0.091
Prompt-RIS (Shang et al., 2024)	0.821	0.688	0.849	0.054	0.756	0.630	0.779	0.097
LISA (Lai et al., 2024)	0.838	0.783	0.900	0.039	0.807	0.761	0.876	0.065
SAMWISE (Cuttano et al., 2025)	0.658	0.427	0.677	0.138	0.623	0.406	0.637	0.166
DETRIS (Huang et al., 2025)	0.874	0.780	0.916	0.032	0.865	0.793	0.900	0.050
TALENT (Jin et al., 2026)	0.883	0.799	0.915	0.031	0.854	0.777	0.887	0.057
InstructSAM (Yuan et al., 2026)	0.910	0.830	0.918	0.030	0.825	0.723	0.818	0.070
CoRefNet	**0.921**	**0.914**	**0.966**	**0.019**	**0.872**	**0.858**	**0.925**	**0.044**

Table 11: Zero-shot cross-dataset generalization on RefCOCO+ testA and testB. ↑/↓ indicates that higher/lower is better. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	RefCOCO+ testA (1,975)				RefCOCO+ testB (1,798)			
	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi^m \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi^m \uparrow$	$\mathcal{M} \downarrow$
PGNet (Xie et al., 2022)	0.479	0.244	0.485	0.313	0.469	0.217	0.511	0.300
ZoomNeXt (Pang et al., 2024)	0.508	0.282	0.530	0.265	0.488	0.251	0.520	0.280
CamoFormer (Yin et al., 2024)	0.509	0.282	0.532	0.262	0.493	0.244	0.536	0.258
VSCode (Luo et al., 2024)	0.497	0.235	0.514	0.281	0.476	0.252	0.509	0.302
MDSAM (Gao et al., 2024)	0.538	0.262	0.546	0.240	0.508	0.198	0.529	0.216
Samba (He et al., 2025)	0.498	0.279	0.510	0.280	0.483	0.252	0.515	0.290
USCNet (Zhou et al., 2025)	0.527	0.275	0.537	0.260	0.509	0.250	0.510	0.283
Prompt-RIS (Shang et al., 2024)	0.524	0.290	0.580	0.235	0.485	0.238	0.563	0.258
LISA (Lai et al., 2024)	0.677	0.534	0.732	0.099	0.612	0.418	0.661	0.155
SAMWISE (Cuttano et al., 2025)	0.491	0.229	0.541	0.272	0.476	0.210	0.531	0.280
DETRIS (Huang et al., 2025)	0.522	0.278	0.572	0.234	0.504	0.245	0.559	0.242
TALENT (Jin et al., 2026)	0.513	0.287	0.537	0.261	0.492	0.260	0.525	0.283
InstructSAM (Yuan et al., 2026)	0.540	0.295	0.583	0.212	0.516	0.247	0.551	0.219
CoRefNet	**0.686**	**0.539**	**0.755**	**0.098**	**0.615**	**0.425**	**0.693**	**0.149**

D Ablation and Sensitivity Analysis

We conduct diagnostic sensitivity analyses around the fixed default configuration used in all main experiments: Text Adapter bottleneck width 64, $K = 3$ aggregated proposals, and 1008×1008 evaluation resolution. These analyses are intended to characterize robustness to reasonable implementation variations rather than to select the reported main configuration. We vary one factor at a time while keeping all remaining settings fixed.

The sensitivity variants are evaluated under the same held-out test protocol as the main results solely for diagnostic comparison; they do not determine model selection or modify the default configuration reported in the main experiments. These studies use FG Overall, a foreground-only diagnostic macro-average over $\{\mathrm{A{:}Sal}, \mathrm{B{:}Cam}, \mathrm{C{:}Sal}, \mathrm{C{:}Cam}\}$ that excludes Scene D:Bg.

Figure 13 evaluates robustness to changes in adapter capacity and proposal aggregation around the fixed defaults. For the Text Adapter, performance remains stable over bottleneck widths 16–64; the default width 64 obtains 78.48% mIoU and 77.20% oIoU, while increasing the width further to 128 slightly reduces both metrics. This indicates that the method is not sensitive to moderate variations in adapter capacity. For proposal aggregation, the default $K = 3$ obtains 78.28% mIoU and 76.99% oIoU. Increasing K beyond 3 provides no consistent additional gain, suggesting that a small set of high-confidence proposals already captures most of the useful target evidence. Overall,

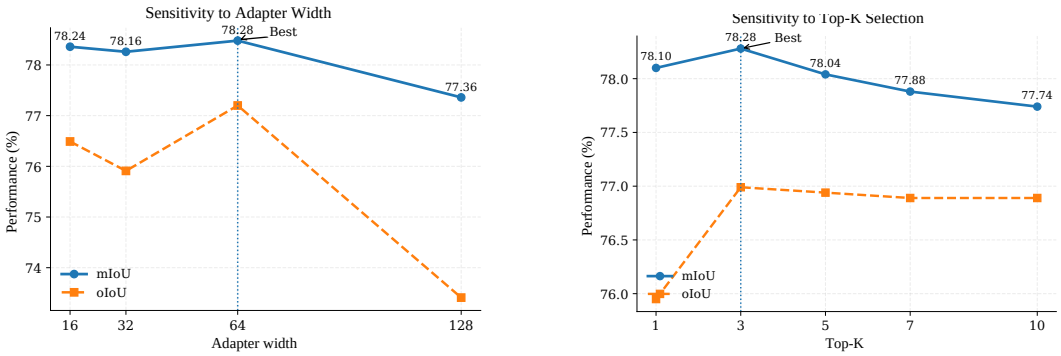

Figure 13: Sensitivity to adapter width (left) and the number of aggregated proposals K (right). FG Overall denotes the macro-average over the four foreground evaluation units; mIoU and oIoU are reported in percentage.

Table 12: Sensitivity to input resolution. Time denotes measured wall-clock time per image group under the same implementation and evaluation setting.

Resolution	FG mIoU↑	FG oIoU↑	Time↓
672	76.72	75.67	0.234 s
840	<u>77.26</u>	<u>76.26</u>	<u>0.227 s</u>
1008	**78.28**	**76.99**	**0.190 s**

the performance trend is stable around the default proposal setting rather than relying on a large number of proposals. Table 12 further examines sensitivity to input resolution.

Table 12 evaluates lower-resolution variants relative to the fixed 1008×1008 default. Reducing the input resolution from 1008 to 840 and 672 pixels gradually decreases FG mIoU from 78.29 to 77.26 and 76.72, and FG oIoU from 77.00 to 76.26 and 75.67, respectively. This trend is consistent with the need to preserve small structures and ambiguous object boundaries. In our implementation, the default 1008-pixel setting also has the lowest measured wall-clock time (0.190 s per image group). We emphasize that this timing reflects hardware utilization and kernel behavior rather than lower theoretical complexity. The resolution analysis therefore serves as a diagnostic robustness check and does not determine the default configuration.

E Additional Qualitative Results

Figure 14 provides additional qualitative comparisons on UniRefDet across Scenes A–D. Red and green denote salient and camouflaged regions, respectively. In the single-attribute Scenes A and B, CoRefNet accurately localizes the corresponding targets while suppressing responses from the opposite attribute. Scene C is more challenging because salient and camouflaged targets coexist in the same image; the examples show that CoRefNet better preserves their attribute-specific regions with less cross-attribute confusion. In the background-only Scene D, where neither attribute is present, CoRefNet produces no spurious foreground response in the shown example. Overall, these visualizations complement the quantitative results and illustrate the behavior of different methods under the four scene configurations.



Figure 14: Additional qualitative comparison on UNIREFDET. Red and green indicate salient and camouflaged regions, respectively. Scenes A–D correspond to salient-only, camouflaged-only, coexistence, and background-only conditions. The examples include representative predictions from SOD/COD, unified, referring-segmentation, and foundation-model-based methods.