CamoQuery: Language-Guided Reasoning Camouflaged Object Segmentation

Tianxin Han, Qing Dong, Xingwei Wang†, Jie Jia, Gang Wu, Bowen Yang, Fu Zhang
School of Computer Science and Engineering, Northeastern University
tianxinhan797@gmail.com, wangxw@mail.neu.edu.cn

Abstract

Although camouflaged object segmentation has advanced rapidly in recent years, existing methods are still confined to visual mask prediction under fixed task assumptions. They cannot interactively respond to user requests, nor can they proactively understand and reason about the user's intent. Our work tackles this issue by proposing a novel task, Language-Guided Reasoning Camouflaged Object Segmentation (LRCOS). Given a camouflaged image and an implicit query text instruction that requires reasoning, LRCOS aims to output intent-consistent segmentation mask. To establish a benchmark for this task, we build CamoQuery, comprising 12,437 image–mask samples and 25971 implicit query text instructions. To better reflect real-world camouflaged scenarios, we additionally collect MCD, a multi-instance camouflage dataset where multiple camouflaged targets coexist within the same scene, increasing the need for reasoning. Building on CamoQuery, we further propose COSA, a vision–language segmentation assistant that segments the intended camouflaged object from implicit queries and produces a reasoning explanation. Experiments on CamoQuery demonstrate that COSA has strong reasoning segmentation capability in camouflaged scenes and exhibits zero-shot capability.

1 Introduction

Real-world visual perception serves as a cornerstone for advancing general artificial intelligence (AGI). As a representative task for visual perception, Camouflaged Object Segmentation (COS) aims to delineate targets that are concealed within their surrounding scenes (Ji et al., 2023; Wu et al., 2025; Lei et al., 2025), and is crucial for medical image analysis (Fan et al., 2020), surface defect inspection (Ma et al., 2021), and pest monitoring (Dong et al., 2024). Existing COS methods treat this task as a vision-only mask prediction problem, relying on appearance cues to decide *what* to segment. While effective under predefined tasks, they fail to understand the contextual factors that make the target camouflaged (*why* it is hard to perceive), and thus cannot adapt to interactive scenarios that require understanding user intent.

Recent advances in Large Vision–Language Models (LVLMs) (Lai et al., 2024; Rasheed et al., 2024) offer a promising direction by enabling language-guided segmentation with strong understanding of complex visual relationships. Their generative flexibility and multimodal input processing allow users to provide nuanced, human-like guidance through text instructions. Nevertheless, most LVLM-based segmentation methods are developed for natural images with clear and salient objects, and thus generalize poorly to highly camouflaged scenes. As shown in Fig. 1, limited in-domain data can lead to spurious grounding and inconsistent predictions, and degrade instruction following in camouflage scenarios. More importantly, in camouflaged scenes, users often do not know the exact category of the target, and thus queries are typically implicit: the query text instruction is not necessarily an explicit reference (eg. "an insect"), but a more complicated description involving complex reasoning or world knowledge (eg. "the camouflaged target that might be poisonous/dangerous and should be avoided"). This setting requires models to reason over complex, implicit text queries jointly with the image and produce segmentation masks. Unfortunately, current research lacks benchmark datasets that support these capabilities, Mainstream COS datasets provide only images and pixel-level masks. *A comprehensive benchmark is therefore needed to evaluate the effectiveness of language-guided reasoning camouflaged object segmentation.*

To this end, we propose a new benchmark task, **Language-Guided Reasoning Camouflaged Object Segmentation (LRCOS)**, where the goal is

† Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22924–22941

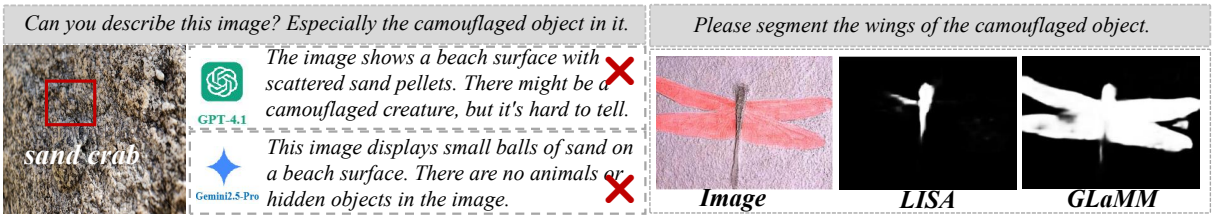

Figure 1: Performance of existing LVLMs in understanding camouflage scenes and grounding segmentation.

Figure 2: We propose a new task, Language-Guided Reasoning Camouflaged Object Segmentation (LRCOS). Our method, COSA, supports simple reasoning, as shown by the first example in the first row, and can also deal with cases involving complex reasoning and world knowledge, as shown by the last example in the third row. It further demonstrates zero-shot capability, as shown by the second example in the second row.

to output segmentation masks given camouflaged images and implicit query text instructions that require reasoning. To support LRCOS, we build the large-scale vision–language COS dataset, **CamoQuery**, consisting of 12,437 camouflaged images with pixel-level masks and 25,971 textual instructions. CamoQuery is built on existing COS resources (Le et al., 2019; Fan et al., 2022; Lv et al., 2021) and is further augmented with our newly collected dataset, **MCD**, which contributes 2,000 images with masks and textual annotations. To complement most COS datasets that are largely single-instance per image, MCD contains more scenes where multiple camouflaged instances co-exist, improving the diversity and realism of camouflaged scenarios while introducing stronger reasoning requirements for identifying the intended target. We annotate two instruction styles for these datasets: *simple* instructions that often require lightweight reasoning, and *complex* instructions that require richer contextual reasoning.

Building upon the above data and task setup, we introduce COSA, a language instructed camouflaged object segmentation assistant for LRCOS that elicits segmentation masks and reasoning explanations. Specifically, we propose a context synthesizer to inject instruction-relevant camouflage cues into language representations, yielding compact embeddings that better ground implicit queries. We further introduce a dedicated [SEG] token under an embedding-as-mask design, and decode its hidden representation into the final mask with a mask decoder, while adapting the LLM via LoRA for efficient reasoning-grounded segmentation in camouflaged scenes. As illustrated in Fig. 2, COSA supports reasoning and exhibits zero-shot generalization. Additionally, given the subjectivity of reasoning explanations, we develop LLM-based metrics to assess explanation quality in terms of Explanation Credibility (EC) and Instruction Alignment (IA). In summary, our contributions are:

- We propose LRCOS, a new task setting that fills a key gap in existing COS resources that lack language interaction and reasoning.
- We build CamoQuery, a multimodal benchmark to support LRCOS, it further includes our newly collected multi-instance camouflaged dataset (MCD) to better reflect camouflaged scenes.
- We propose COSA and conduct extensive experiments on CamoQuery, demonstrating effective reasoning ability and strong zero-shot capability.

2 Related work

COS and Benchmarks. COS aims to identify and segment objects that are visually concealed

Benchmark	Image	Mask	Text	Text Type	Anno.	Input Form	Training?	Zero-shot?	Reasoning?
Vision-only COS Benchmarks									
COD10K (Fan et al., 2022)	5,066	5,066	0	–	M	Image	✓	✗	✗
CHAMELEON (Skurowski et al., 2018)	76	76	0	–	M	Image	✗	✗	✗
NC4K (Lv et al., 2021)	4,121	4,121	0	–	M	Image	✗	✗	✗
CAMO (Le et al., 2019)	1,250	1,250	0	–	M	Image	✓	✗	✗
Attribute-centric Benchmark									
COD-TAX (Zhang et al., 2025a)	4,040	0	4,040	Attribute	M&T	Image + Attribute Text	✓	✗	✗
Category-name Guided COS Benchmark									
OVCamo (Pang et al., 2024b)	11,483	11,483	11,483	Category	M&T	Image + Category Name	✓	✓	✗
Language-guided reasoning COS Benchmark									
CamoQuery (Ours)	12,437	12,437	25,971	Instruction	M&T	Image + Instruction Text	✓	✓	✓

Table 1: Comparison between CamoQuery and related camouflage benchmarks in terms of data scale, annotation type (M/T), input form, training availability, zero-shot capability, and reasoning requirement.

in their surroundings. Existing methods explore multi-scale fusion (Fan et al., 2022), multi-stage refinement (Liu et al., 2023d), weakly supervised learning (Han et al., 2025), uncertainty modeling (Liu et al., 2022), as well as auxiliary cues such as boundaries (Yue et al., 2025), texture (Wang et al., 2024), frequency (Zhang et al., 2025b), and depth (Yu et al., 2024b). However, these approaches and mainstream COS benchmarks (Le et al., 2019; Fan et al., 2022; Lv et al., 2021; Skurowski et al., 2018) remain vision-only, providing only pixel-level masks and thus failing to evaluate intent reasoning and instruction-following for selecting the user-specified target. In contrast, language-guided segmentation datasets in natural scenes (Chng et al., 2024; Kim et al., 2024) rely on salient objects and explicit referring expressions, which transfer poorly to camouflage. Although recent attempts introduce text into COS (Zhang et al., 2025a; Pang et al., 2024b), the language is typically limited to attributes or category prompts that largely pre-define what to segment, reducing the task to prompt-conditioned mask prediction. Therefore, existing benchmarks are still insufficient for evaluating reasoning segmentation under camouflage, motivating CamoQuery. As summarized in table 1, we compare CamoQuery with representative camouflage-related benchmarks in terms of task formulation, modality, and annotation scale.

LVLM Benchmarks. The rapid advancement of LVLMs underscores the critical need for effective benchmarking to reveal model limitations and steer future development (Liu et al., 2024; Yu et al., 2024a). While existing benchmarks mainly evaluate perception and reasoning in natural scenes (Liu et al., 2023b), they largely overlook camouflage scenarios. In camouflaged environments, targets are deliberately concealed with weak boundaries and highly similar foreground–background appearances, as shown in Fig. 1, directly transferring models and evaluation protocols from natural scenes is often infeasible. In contrast, domains such as medicine (Shen et al., 2025; Shao and Hou, 2025), which involve tasks with structural parallels to COS, have benefited from more systematic benchmarking efforts. Consequently, introducing the first reasoning-oriented LVLM benchmark specifically tailored to the camouflage domain is essential to fill this gap and accelerate progress in the field.

3 CamoQuery

Task Settings. To more comprehensively evaluate LRCOS, we establish the CamoQuery benchmark, which contains camouflaged images, pixel-level masks, and implicit query text instructions. CamoQuery is designed to assess both the quality of mask prediction and the quality of reasoning explanations. Accordingly, we define two evaluation tasks that focus on mask generation and explanation generation, respectively: IQS and IRE. IQS (Implicit Query Segmentation) requires a model to generate a high-quality binary mask $m \in \{0,1\}^{H\times W}$ given a camouflaged image I and a text query instruction x. IRE (Intent-Reasoning Explanation) asks the model to produce a concise explanation clarifying *why* the predicted region matches the inferred intent and what visual evidence supports this conclusion.

Image sources. To ensure fair and comparable evaluation, we build CamoQuery upon diverse and high-quality camouflaged resources released in prior COS research, including COD10K (Fan et al., 2022), NC4K (Lv et al., 2021), and CAMO (Le et al., 2019). However, these benchmarks are largely dominated by a "one-image-one-instance" setting, which is insufficient to reflect real scenes where multiple camouflaged targets co-exist and richer inter-object relations arise. In such cases,

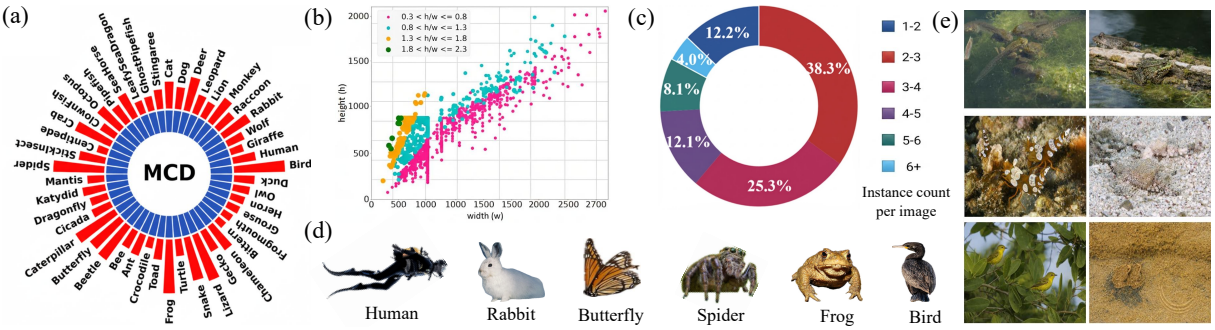

Figure 3: Statistics and camouflaged category examples from MCD dataset. (a) Sunburst visualization of the taxonomy and category frequency distribution. (b) Image resolution distribution. (c) Instance count per image distribution. (d) Examples of sub-classes. (e) Image examples from MCD.

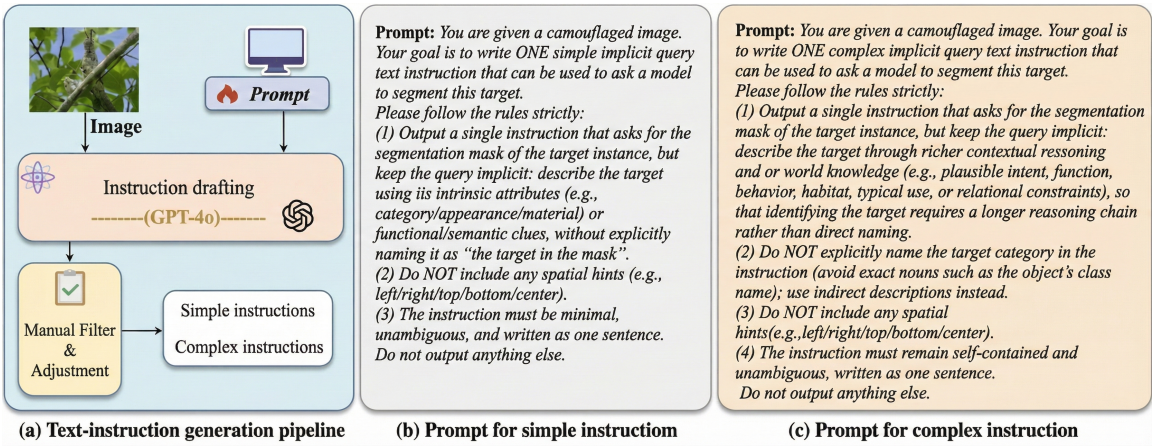

Figure 4: Prompt-based instruction drafting pipeline for CamoQuery. Given a camouflaged image, we use prompt templates to query GPT-4o for drafting implicit query text instructions, followed by human filtering and refinement to produce the final *simple* and *complex* instruction annotations.

implicit instructions often require stronger reasoning over contextual cues and object relations, rather than direct appearance matching. To this end, we additionally curate **MCD (ours)** as a multi-instance complement and include it as one of the image sources of CamoQuery.

MCD: A multi-instance camouflaged dataset. New datasets often significantly advance the frontier of visual understanding and reshape existing paradigms. Motivated by this, we further construct MCD to enrich the diversity of CamoQuery under more complex and realistic camouflage scenarios. The images in MCD are mainly collected from publicly available natural photography resources and online image repositories, with a preference for sources that allow academic research use. MCD contains 2,000 images, each paired with a pixel-level mask and an implicit query text instruction, covering 46 camouflaged categories and diverse backgrounds such as forests, snowfields, grasslands, sky, and seawater. Different from the common "one-image-one-instance" setting in most COS datasets, MCD deliberately selects images where multiple camouflaged targets co-exist in the same scene. This design provides richer inter-object relations and contextual constraints, making implicit instruction following substantially more demanding and increasing the need for reasoning in multi-instance camouflage scenes. The statistics in Fig. 3(c) further show that MCD contains a high proportion of multi-instance co-occurrence samples, which enhances the dataset's real-world relevance and raises the task difficulty, better matching the requirements of LRCOS. More image examples are provided in the Appendix.

Instruction annotation pipeline. Inspired by the practice of prior reasoning dataset construction (Yao et al., 2025), we design two prompts for *simple* and *complex* instructions, as shown in Fig. 4, and use GPT-4o to generate the corresponding *simple/complex* draft instructions. We then perform human filtering and refinement on all drafts to ensure clear semantics and consistency with the target mask. Accordingly, we annotate two styles of im-

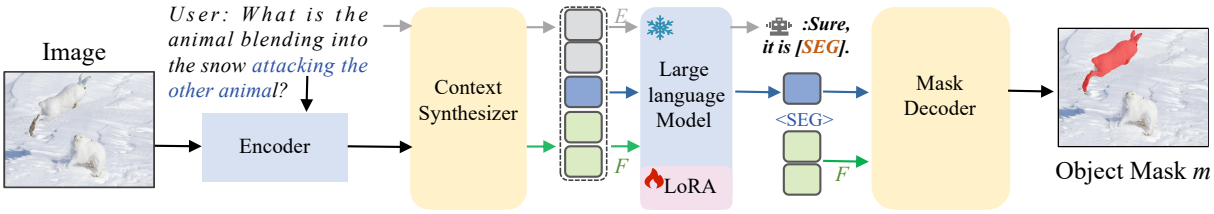

Figure 5: Overall framework of COSA for language-guided reasoning camouflaged object segmentation.

Dataset	Img	Mask	Text	Categories	Images		Text ann.	
					train	test	train	test
CAMO	✓	✓	✗	43	1000	250	–	–
COD10K	✓	✓	✗	68	3040	2026	–	–
NC4K	✓	✓	✗	37	0	4121	–	–
ResCAMO	✓	✓	✓	43	1000	250	2118	512
ResCOD10K	✓	✓	✓	68	3040	2026	6120	4264
ResNC4K	✓	✓	✓	37	0	4121	0	8504
MCD	✓	✓	✓	46	1200	800	2532	1921
all (CamoQuery)	✓	✓	✓	–	5240	7197	10770	15201

Table 2: Comparison between original COS datasets and our CamoQuery.

plicit query text instructions for each image: (1) *simple* instructions that require less reasoning and typically focus on the target's function or implicit attributes. These include functional or behavioral descriptions (e.g., "the predator waiting to ambush prey") or basic local cues. (2) *complex* instructions that require richer contextual reasoning or world knowledge, often involving longer reasoning chains over safety, practicality, or cross-cue integration. Specifically, in camouflaged scenes the instruction is no longer a short, explicit category name (e.g., "a lizard"), but may instead describe the intended target through indirect cues that require reasoning or world knowledge, such as function or behavior (e.g., "the predator waiting to ambush prey") or risk or safety (e.g., "the potentially dangerous camouflaged target to avoid"). Additionally, for images containing multiple camouflaged candidates, we annotate multiple *complex* instructions to cover different reasoning paths. Ultimately, the CamoQuery benchmark contains 12,437 image–mask samples with a total of 25,971 text instructions. More annotation details and quality-control guidelines are provided in the Appendix.

Dataset splits. As summarized in Table 2, CamoQuery contains 12,437 camouflaged images, each paired with a pixel-level segmentation mask, and 25,971 implicit query text instructions, and is organized into four subsets: ResCAMO, ResCOD10K, and ResNC4K derived from CAMO, COD10K, and NC4K, respectively, together with our newly collected MCD. ResCAMO and ResCOD10K follow the standard train/test splits adopted in prior COS work, while ResNC4K is treated as a test-only set consistent with existing benchmarks; MCD provides an additional 1,200/800 split to enrich training and evaluation with newly curated camouflage samples. In total, the benchmark includes 5,240 training images with 10,770 text instructions and 7,197 test images with 15,201 text instructions. CamoQuery provides a systematic and extensible resource for multimodal reasoning camouflaged object detection research.

4 COSA

Model Overview. LRCOS aims to output a binary segmentation mask $\mathbf{m} \in \{0,1\}^{H\times W}$ given a camouflaged image and an implicit query text instruction. This task shares a similar formulation with referring expression segmentation, but it involves complex user query instructions that require reasoning. As shown in Fig. 5, COSA consists of four main components: (1) an encoder, (2) a context synthesizer, (3) a large language model (LLM) adapted with LoRA, and (4) a mask decoder.

Encoder. For an input image I, we use a visual encoder $\mathcal{I}$ to extract visual representations. Specifically, the image is encoded into an embedding $\mathbf{F} \in \mathbb{R}^{N\times C}$, where $N = H_e W_e$ denotes the number of spatial locations and C denotes the embedding dimension. Meanwhile, the user instruction is processed by a text encoder conditioned on the visual embedding $\mathbf{F}$ to generate the text embedding $\mathbf{x}_{txt} \in \mathbb{R}^{M\times C}$, where M denotes the number of queries. This design enables the text embeddings to incorporate instruction-relevant visual cues while maintaining cross-modal consistency. The visual embedding $\mathbf{F}$ captures global image information and is projected into the language space of LLMs through a vision-to-language projection layer $\mathrm{Proj}_{I\to LLM}$.

Context Synthesizer. This module aims to aggregate text-related visual features and inject them to generate text embedding representing the current image, for better reasoning and explanation.

With the text embeddings $\mathbf{x}_{txt}$ and visual features $\mathbf{F}$, context-based aggregation is formulated as:

$$\mathbf{E} = \mathrm{FFN}(\mathrm{CrossAttn}(\mathbf{x}_{txt}, \mathbf{F}, \mathbf{F})), \quad (1)$$

$$\mathbf{E}^c = \mathrm{Concat}\{\mathbf{E}^i\}_{i\in\Omega_K}, \quad (2)$$

where 'CrossAttn' refers to cross-attention operation, while $\mathbf{F}$ is the value and key. By the cross-attention, we further synthesize the visual cues into the refined text embeddings E. However, we argue that not all queries are needed for the refined text embeddings. Unlike QFormer which adopts 32 queries as input LLM tokens, we propose to condense these embeddings into more compact ones. After obtaining $\mathbf{E}$, we compute an attention-response score for each query token as $s_i = \max_j \mathbf{A}_{i,j}$, where $\mathbf{A} \in \mathbb{R}^{M\times N}$ denotes the cross-attention weight matrix between M query tokens and N visual tokens. We then select Ω_K as the indices of the K largest scores and concatenate the corresponding embeddings to form $\mathbf{E}^c$. Therefore, the final input embeddings E^c preserve the most relevant visuals, which helps the LLM better understand the image content under the given query and generate more better explanations.

Segmentation Token. Inspired by LISA (Lai et al., 2024), we adopt the *embedding-as-mask* strategy. Under this design, the LLM outputs a specialized segmentation token [SEG] to encapsulate semantic and spatial information, thereby aligning visual and textual features in a shared embedding space for accurate segmentation and explanation. The [SEG] token serves as a bridge between high-level multimodal reasoning and low-level spatial prediction: it aggregates the discriminative evidence required by the query instruction and compresses such evidence into a compact target representation suitable for subsequent decoding. However, since training an LLM from scratch is prohibitively expensive, we adopt LoRA to fine-tune the LLM. By introducing learnable low-rank adaptations, LoRA better adapts the LLM to camouflaged images and enables efficient fine-tuning.

Mask Decoder. Before decoding, we apply an additional projection layer Φ to adjust the dimension of the [SEG] representation. Let $\mathbf{h}_{[\mathrm{SEG}]} \in \mathbb{R}^d$ denote the last-layer hidden state at the [SEG] position in the LLM output sequence. We obtain $\mathbf{Q} = \Phi(\mathbf{h}_{[\mathrm{SEG}]})$. Then, $\mathbf{Q}$ together with the visual embedding $\mathbf{F}$ are fed into the mask decoder $\mathcal{D}$ to predict the final mask:

$$\hat{\mathbf{m}} = \mathcal{D}(\mathbf{F}, \mathbf{Q}), \qquad \hat{\mathbf{m}} \in [0,1]^{H\times W}, \quad (3)$$

Training Objectives. To jointly optimize reasoning and precise camouflaged object segmentation, we employ three losses to supervise both the language side and the pixel-level mask prediction: a textual loss $\mathcal{L}_{\text{text}}$, a binary cross-entropy loss $\mathcal{L}_{\text{bce}}$, and a Dice loss $\mathcal{L}_{\text{dice}}$. The final training objective is a weighted sum:

$$\mathcal{L} = \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \lambda_{\text{bce}}\mathcal{L}_{\text{bce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}. \quad (4)$$

where λ_{text}, λ_{bce}, and λ_{dice} are set to 1.0, 2.0, and 0.5, respectively, following LISA (Lai et al., 2024).

5 Experiments

Dataset. As described in Section 3, we use the CamoQuery benchmark with its train/test splits. Moreover, following LISA (Lai et al., 2024), we additionally leverage training data from semantic segmentation (ADE20K (Zhou et al., 2017), COCO-Stuff (Caesar et al., 2018)), referring segmentation (RefCOCO/RefCOCO+/RefCOCOg, RefCLEF (Kazemzadeh et al., 2014a)), and visual question answering (VQA; LLaVA-Instruct-150k (Liu et al., 2023c)) for training. Besides, to enhance segmentation for fine-grained object parts, we also leverage part-level segmentation datasets, including PACO-LVIS (Ramanathan et al., 2023), PartImageNet (He et al., 2022), and PASCAL-Part (Chen et al., 2014). We reformat segmentation-style datasets into QA pairs with binary-mask supervision, and convert referring descriptions into vision–text aligned prompts. More implementation details are provided in the Appendix.

Models. We use LLaVA-7B-v1-1 or LLaVA-13B-v1-1 (Liu et al., 2023b) as the base multimodal LLM. We train with AdamW (learning rate 1×10^{-4}, batch size 8) on four NVIDIA RTX 4090 GPUs.

Metrics. We follow prior works (Yao et al., 2025; Lai et al., 2024; Li et al., 2025) and evaluate mask quality for IQS using IoU-based metrics, including gIoU and cIoU. Considering the subjectivity of reasoning-centric outputs, we further adopt an LLM-as-a-judge protocol to evaluate IRE using two criteria: Explanation Credibility (EC) and Instruction Alignment (IA). EF measures whether the explanation is visually supported by the image and is consistent with the predicted mask, while IA assesses whether the explanation aligns with the query intent and clearly connects the predicted mask to the requested target. Detailed criteria and prompts are provided in Appendix.

(a) IQS segmentation quality (gIoU / cIoU)

Method	ResCAMO						ResCOD10K						ResNC4K						MCD					
	Simple		Complex		Overall		Simple		Complex		Overall		Simple		Complex		Overall		Simple		Complex		Overall	
	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
LISA-7B	44.2	40.1	47.0	43.1	45.6	41.6	50.4	46.2	53.0	49.1	51.7	47.6	48.6	44.8	51.2	47.4	49.9	46.1	47.2	43.6	49.8	46.1	48.5	44.8
GSVA-7B	45.6	41.2	48.4	44.3	47.0	42.8	52.1	47.8	54.7	50.6	53.4	49.2	50.0	46.3	52.7	48.9	51.3	47.6	48.6	45.0	51.3	47.7	49.9	46.3
GLaMM-7B	46.1	41.7	48.9	44.8	47.5	43.3	52.8	48.5	55.4	51.3	54.1	49.9	50.6	46.9	53.3	49.6	51.9	48.2	49.2	45.6	51.9	48.3	50.6	46.9
PixelLM-7B	47.8	43.5	50.7	46.6	49.2	45.0	53.8	49.5	56.5	52.4	55.1	50.9	51.4	47.7	54.1	50.4	52.8	49.1	50.0	46.4	52.7	49.1	51.4	47.8
POPEN-7B	49.0	44.8	52.0	48.1	50.5	46.5	55.0	50.9	57.8	53.7	56.4	52.3	52.0	48.3	54.8	51.1	53.4	49.7	51.4	47.8	54.2	50.6	52.8	49.2
Ours-7B	**52.1**	**48.3**	**55.2**	**51.6**	**53.7**	**49.9**	**58.4**	**54.6**	**61.2**	**57.5**	**59.8**	**56.0**	**53.4**	**49.7**	**56.1**	**52.4**	**54.8**	**51.0**	**52.6**	**48.9**	**55.3**	**51.6**	**53.9**	**50.2**
LISA-13B	46.8	42.7	49.6	45.7	48.2	44.2	53.6	49.4	56.4	52.3	55.0	50.8	51.3	47.6	54.1	50.4	52.7	48.9	50.1	46.5	52.9	49.3	51.5	47.9
GSVA-13B	48.3	44.2	51.1	47.2	49.7	45.7	55.2	51.0	58.0	54.0	56.6	52.5	52.8	49.1	55.6	51.9	54.2	50.5	51.6	48.0	54.4	50.8	53.0	49.4
PixelLM-13B	50.1	45.9	52.9	49.0	51.5	47.5	56.2	52.0	59.0	55.0	57.6	53.5	53.7	50.0	56.5	52.8	55.1	51.4	52.1	48.5	54.9	51.3	53.5	49.9
POPEN-13B	51.4	47.3	54.2	50.3	52.8	48.8	57.6	53.5	60.4	56.5	59.0	55.0	55.1	51.4	57.9	54.2	56.5	52.8	53.5	49.9	56.3	52.7	54.9	51.3
Ours-13B	**55.7**	**51.9**	**58.8**	**55.1**	**57.3**	**53.5**	**61.5**	**57.8**	**64.3**	**60.6**	**62.9**	**59.2**	**56.8**	**53.1**	**59.6**	**55.9**	**58.2**	**54.5**	**56.3**	**52.7**	**59.1**	**55.4**	**57.7**	**54.0**

(b) IRE explanation quality (EF / IA)

Method	ResCAMO						ResCOD10K						ResNC4K						MCD					
	Simple		Complex		Overall		Simple		Complex		Overall		Simple		Complex		Overall		Simple		Complex		Overall	
	EC	IA	EC	IA	EC	IA	EC	IA	EC	IA	EC	IA	EC	IA	EC	IA	EC	IA	EC	IA	EC	IA	EC	IA
LISA-7B	3.05	2.90	3.52	3.36	3.36	3.20	3.28	3.12	3.78	3.62	3.60	3.44	3.16	3.00	3.65	3.49	3.48	3.32	3.10	2.95	3.58	3.42	3.42	3.26
GSVA-7B	3.12	2.97	3.60	3.44	3.44	3.28	3.35	3.19	3.86	3.70	3.68	3.52	3.23	3.07	3.73	3.57	3.56	3.40	3.17	3.01	3.66	3.50	3.50	3.34
GLaMM-7B	3.16	3.01	3.65	3.49	3.49	3.33	3.39	3.23	3.90	3.74	3.72	3.56	3.27	3.11	3.77	3.61	3.60	3.44	3.21	3.05	3.70	3.54	3.54	3.38
PixelLM-7B	3.22	3.06	3.74	3.58	3.58	3.42	3.46	3.30	3.98	3.82	3.80	3.64	3.34	3.18	3.86	3.70	3.68	3.52	3.28	3.12	3.78	3.62	3.62	3.46
POPEN-7B	3.30	3.14	3.82	3.66	3.66	3.50	3.54	3.38	4.06	3.90	3.88	3.72	3.42	3.26	3.94	3.78	3.76	3.60	3.36	3.20	3.86	3.70	3.70	3.54
Ours-7B	**3.48**	**3.32**	**4.08**	**3.92**	**3.90**	**3.74**	**3.72**	**3.56**	**4.34**	**4.18**	**4.14**	**3.98**	**3.60**	**3.44**	**4.20**	**4.04**	**4.02**	**3.86**	**3.54**	**3.38**	**4.12**	**3.96**	**3.96**	**3.80**
LISA-13B	3.18	3.02	3.72	3.56	3.56	3.40	3.42	3.26	3.96	3.80	3.78	3.62	3.30	3.14	3.84	3.68	3.68	3.52	3.24	3.08	3.78	3.62	3.62	3.46
GSVA-13B	3.26	3.10	3.80	3.64	3.64	3.48	3.50	3.34	4.04	3.88	3.86	3.70	3.38	3.22	3.92	3.76	3.76	3.60	3.32	3.16	3.86	3.70	3.70	3.54
PixelLM-13B	3.34	3.18	3.90	3.74	3.74	3.58	3.58	3.42	4.12	3.96	3.94	3.78	3.46	3.30	4.00	3.84	3.84	3.68	3.40	3.24	3.94	3.78	3.78	3.62
POPEN-13B	3.42	3.26	3.98	3.82	3.82	3.66	3.66	3.50	4.20	4.04	4.02	3.86	3.54	3.38	4.08	3.92	3.92	3.76	3.48	3.32	4.02	3.86	3.86	3.70
Ours-13B	**3.64**	**3.48**	**4.26**	**4.10**	**4.10**	**3.94**	**3.88**	**3.72**	**4.52**	**4.36**	**4.34**	**4.18**	**3.76**	**3.60**	**4.38**	**4.22**	**4.22**	**4.06**	**3.70**	**3.54**	**4.30**	**4.14**	**4.14**	**3.98**

Table 3: CamoQuery evaluation for LRCOS. *Simple* contains shorter implicit instructions with lower reasoning complexity, whereas *Complex* involves richer contextual reasoning or world knowledge for target grounding. *Overall* reports results on the union of *Simple* and *Complex*. We report both IQS and IRE scores, higher is better.

5.1 Experimental Results

Quantitative results. Table 3 reports quantitative results on LRCOS under implicit query instructions. Specifically, we evaluate (i) IQS, which measures the quality of intent-consistent segmentation masks, and (ii) IRE, which measures the quality of reasoning explanations. Accordingly, we benchmark both mask quality (gIoU/cIoU) and explanation quality (EC/IA), providing a more task-faithful assessment for LRCOS. We compare our method with representative LVLM-based segmentation baselines, including LISA (Lai et al., 2024), GSVA (Xia et al., 2024), GLaMM (Rasheed et al., 2024), PixelLM (Ren et al., 2024), and POPEN (Zhu et al., 2025), under both 7B and 13B model sizes (Liu et al., 2023b). Across all subsets and both instruction styles (Simple/Complex), our approach consistently delivers the strongest overall performance on both IQS and IRE. Notably, COSA improves mask accuracy and explanation quality simultaneously, indicating better intent following in camouflaged scenes. The context synthesizer injects instruction-relevant visual cues to better condition the LLM for explanation generation, while the [SEG] token provides a compact target representation for mask decoding, yielding the high-quality predicted masks.

Qualitative results. Fig. 6 presents visual comparisons between COSA and representative LVLM-based segmentation baselines under diverse implicit instruction settings in camouflaged scenes. As can be observed, existing methods often struggle to correctly interpret indirect and reasoning-intensive queries, especially when the target is visually entangled with the background or surrounded by distracting objects with similar appearance patterns. These limitations usually manifest as incomplete object masks, inaccurate boundary delineation, fragmented predictions, or attention drift toward semantically irrelevant distractors. In contrast, COSA is able to produce more accurate, complete, and instruction-consistent segmentations across a wide range of challenging cases. Even when the queried target is only indirectly specified through functional, relational, or semantic cues, our method can still better associate the implicit instruction with the intended camouflaged region and suppress irrelevant responses. This qualitative evidence indicates that COSA has a stronger capability to bridge implicit intent reasoning and pixel-level localization in complex camouflaged environments.

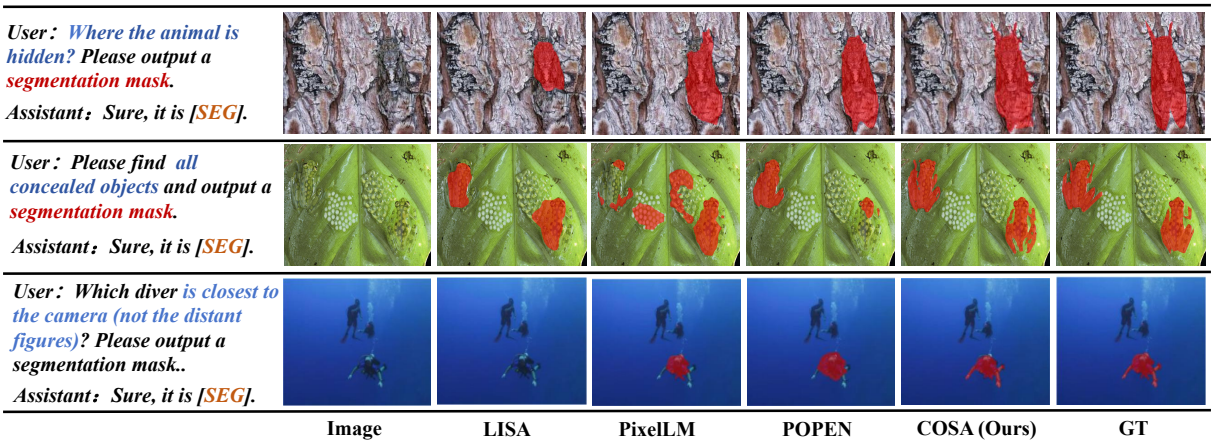

Figure 6: Visual comparison among COSA (ours) and existing related methods.

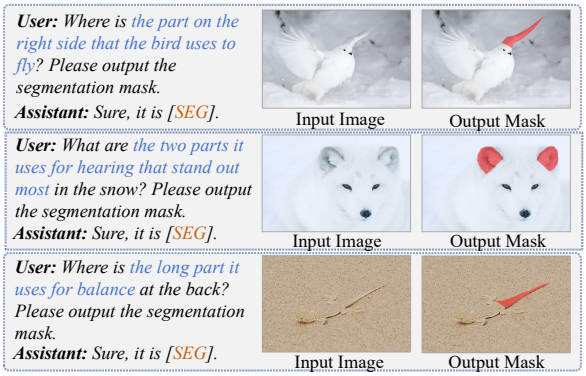

Figure 7: Visualization of zero-shot segmentation.

Moreover, we further provide zero-shot qualitative results on part-level segmentation in Fig. 7. Although COSA is not explicitly trained for fine-grained part segmentation, it still shows promising generalization ability in identifying object parts specified by language queries without additional in-domain training. Compared with conventional baselines, COSA produces part-level masks that are more structurally coherent and better aligned with the queried semantic content. These results suggest that the reasoning and representation abilities learned by COSA are not limited to whole-object segmentation, but can also transfer to finer-grained understanding scenarios, further demonstrating its robustness and generalization capacity.

5.2 Ablation Analysis

We conduct ablations on the MCD subset of CamoQuery to quantify the contribution of key COSA components (Table 4). Starting from the baseline without the context synthesizer or the [SEG] token (ID 1), adding the context synthesizer improves both gIoU and cIoU (ID 2). Introducing the [SEG] token further boosts performance (ID 3). Table 5 ablates the query condensation strategy in the context synthesizer. Using all M queries yields lower performance, and randomly selecting K queries brings only marginal gains. In contrast, our attention-response based selection achieves the best results. More ablation studies are reported in the appendix, including the design choice of the SAM backbone and an ablation on the impact of additional training data sources.

ID	Context Sythesizer	ToKen	gIoU	cIoU
1	✗	✗	45.9	42.4
2	✓	✗	49.8	46.3
3	✓	✓	**53.9**	**50.2**

Table 4: Ablation study of key components on the MCD subset of CamoQuery. (Overall setting).

ID	Condensation strategy	gIoU	cIoU
1	None (use all M queries)	51.6	49.1
2	Random-K queries	52.1	49.5
3	Ours (attention-response selection)	**53.9**	**50.2**

Table 5: Ablation on attention-response based query condensation on the MCD (Overall setting).

5.3 Referring Expression Segmentation

The referring expression segmentation (RES) task (Kazemzadeh et al., 2014b) takes as input an image and a natural-language expression that refers to a specific object in the image (e.g., "Please segment the apple in the image."), and outputs the segmentation mask of the referred object. As a classic setting in semantic segmentation, RES provides an intuitive measure of a model's language-conditioned visual localization ability. RefCOCO (Kazemzadeh et al., 2014c), RefCOCO+ (Kazemzadeh et al., 2014c), and Re-

Model	RefCOCO			RefCOCO+			RefCOCOg	
	Val	TestA	TestB	Val	TestA	TestB	Val	Test
GRIS	70.5	73.2	66.1	65.3	68.1	53.7	59.9	60.4
LAVT	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
GRES	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
LISA	74.1	76.5	72.3	65.1	70.8	58.1	67.9	70.6
PixelLM	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5
GSVA	76.4	77.4	72.8	64.5	67.7	58.6	71.1	72.0
Ours	**78.2**	**80.1**	**73.4**	**72.1**	**74.2**	**62.4**	**72.6**	72.0

Table 6: Results on the RES benchmark (cIoU↑).

fCOCOg (Mao et al., 2016) provide widely used evaluation benchmarks for this task.

To assess whether COSA generalizes beyond camouflaged scenes, we further evaluate it on these classical RES benchmarks and compare it with representative segmentation-oriented baselines. Following prior work (Lai et al., 2024), we report cIoU on the validation and test splits of RefCOCO, RefCOCO+, and RefCOCOg. As shown in Table 6, we include comparisons with representative RES methods, including CRIS (Wang et al., 2022), LAVT (Yang et al., 2022), and GRES (Liu et al., 2023a). COSA achieves strong overall performance and consistently outperforms recent LVLM-based segmenters on most splits. These results suggest that the reasoning and localization ability learned from camouflaged scenes can transfer effectively to conventional referring segmentation, without sacrificing general RES capability.

5.4 Failure Case Analysis

This section presents representative qualitative examples that correspond to the limitations (Fig. 8). First, COSA follows a one-image-per-instruction interface and may fail when a single query is intended to operate over an image set. As shown in Fig. 8(a), it segments one image correctly but cannot consistently localize the corresponding targets in the remaining images due to the set-level scope mismatch. Second, COSA is developed for static images and may be unreliable on videos or multi-frame inputs: in Fig. 8(b), predictions drift over time and become temporally inconsistent. Third, when the queried object is absent, COSA may still produce a non-empty mask; Fig. 8(c) shows a spurious prediction on a salient region, suggesting limited rejection ability and imperfect intent calibration. Overall, COSA and our benchmark provide a strong and practical baseline for LRCOS; we hope these findings and resources will facilitate future research and inspire further breakthroughs, such as supporting cross-image instruction execution, enforcing temporal consistency for dynamic inputs, and improving no-target rejection via curated negative supervision.

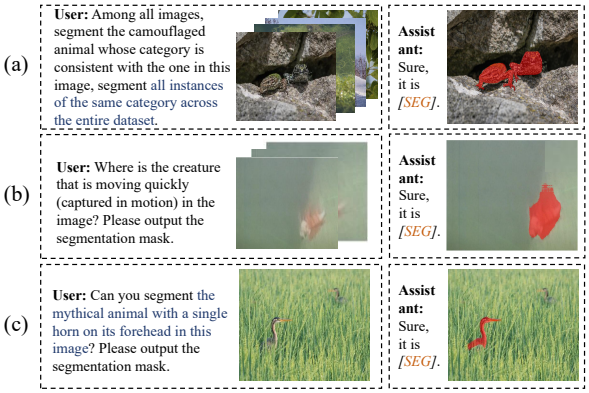

Figure 8: Failure cases of COSA. (a) Cross-image instruction mismatch. (b) Dynamic/motion-related query where temporal cues are implied but the input is a single static frame. (c) Absent-target query where the model produces a spurious non-empty mask.

6 Conclusion

In this work, we introduce a new segmentation task, LRCOS, which extends camouflaged object segmentation from purely visual perception to language-guided reasoning and controllable target understanding. To support this task, we construct a new benchmark, CamoQuery, and further augment it with MCD to better capture challenging multi-instance camouflaged scenarios. Building upon this benchmark, we present COSA, a segmentation framework designed to associate implicit language cues with subtle visual evidence and produce instruction-consistent masks in complex camouflaged scenes. Extensive experiments demonstrate the effectiveness of COSA on both the proposed benchmark and several downstream generalization settings. We hope this work can promote future research on reasoning-oriented camouflaged segmentation, and contribute to real-world applications such as medical image analysis, surface defect inspection, and ecological monitoring.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62432003 and No. U25A20431. The first and second authors made comparable contributions to this work. AI tools were used for language polishing to improve the readability of this manuscript.

Limitations

Despite the strong performance of our approach, several failure cases suggest room for improvement. First, our current setting assumes one image per instruction. Therefore, COSA may struggle when a single instruction is intended to operate over multiple images (e.g., "segment the poisonous plants in all images"), since it is designed to process one input image and output one mask per query. This work provides a strong baseline, and future extensions could explore instruction-level batching or task-level aggregation to support such cross-image queries. Second, COSA is primarily developed for static images and may be less reliable in dynamic scenarios, such as videos or multi-frame inputs, where maintaining temporal consistency becomes critical. Third, when an instruction refers to an object that is absent from the image (e.g., 'segment the unicorn"), the model may still produce a spurious mask. Incorporating a no-target rejection token and curating more diverse instructions (including negative cases) could mitigate this issue and improve robustness and generalization. We provide qualitative examples in the appendix. Overall, our work establishes a strong benchmark for LRCOS, and we hope it will encourage the community to further explore and overcome these limitations.

References

Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. 2009. Frequency-tuned salient region detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1597–1604. IEEE.

Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. 2015. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2018. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1209–1218.

Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. 2014. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1971–1978.

Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. 2024. Mask grounding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26573–26583.

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

Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. 2017. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. 2018. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, pages 698–704. International Joint Conferences on Artificial Intelligence Organization.

Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. 2022. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6024–6042.

Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. 2020. Pranet: Parallel reverse attention network for polyp segmentation.

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

Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. 2023a. Camouflaged object detection with feature decomposition and edge reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22046–22055.

Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jieneng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan L. Yuille. 2022. Partimagenet: A large, high-quality dataset of parts. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, volume 13668 of *Lecture Notes in Computer Science*, pages 128–145. Springer.

Ruozhen He, Qihua Dong, Jiaying Lin, and Rynson W. H. Lau. 2023b. Weakly-supervised camouflaged object detection with scribble annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 781–789.

Xihang Hu, Xiaoli Zhang, Fasheng Wang, Jing Sun, and Fuming Sun. 2024. Efficient camouflaged object detection network based on global localization perception and local guidance refinement. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):5452–5465.

Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, HuaiXin Chen, Jie Qin, and Huan Xiong. 2023. Feature shrinkage pyramid for camouflaged object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5557–5566.

Ge-Peng Ji, Deng-Ping Fan, Yu-Cheng Chou, Dengxin Dai, Alexander Liniger, and Luc Van Gool. 2023. Deep gradient learning for efficient camouflaged object detection. 20(1):92–108.

Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. 2022. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4713–4722.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014a. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014b. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014c. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar. Association for Computational Linguistics.

Seoyeon Kim, Minguk Kang, Dongwon Kim, Jaesik Park, and Suha Kwak. 2024. Extending CLIP's image-text alignment to referring image segmentation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4611–4628, Mexico City, Mexico. Association for Computational Linguistics.

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

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

Cheng Lei, Jie Fan, Xinran Li, Tian-Zhu Xiang, Ao Li, Ce Zhu, and Le Zhang. 2025. Towards real zero-shot camouflaged object segmentation without camouflaged annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11990–12004.

Kaiyu Li, Zepeng Xin, Li Pang, Chao Pang, Yupeng Deng, Jing Yao, Guisong Xia, Deyu Meng, Zhi Wang, and Xiangyong Cao. 2025. Segearth-r1: Geospatial pixel reasoning via large language model. ArXiv preprint: 2504.09644.

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

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

Haotian Liu and 1 others. 2023c. Llava-instruct-150k (llava visual instruct 150k). Hugging Face Datasets. Accessed: 2026-01-03.

Jiawei Liu, Jing Zhang, and Nick Barnes. 2022. Modeling aleatoric uncertainty for camouflaged object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1445–1454.

Yan Liu, Kaihua Zhang, Yaqian Zhao, Hu Chen, and Qingshan Liu. 2023d. Bi-rrnet: Bi-level recurrent refinement network for camouflaged object detection. *Pattern Recognition*, 139:109514.

Yu Liu, Haihang Li, Juan Cheng, and Xun Chen. 2023e. Mscaf-net: A general framework for camouflaged object detection via learning multi-scale context-aware features. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4934–4947.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision (ECCV)*.

Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. 2021. Simultaneously localize, segment and rank the camouflaged objects.

Yixuan Lyu, Hong Zhang, Yan Li, Hanyang Liu, Yifan Yang, and Ding Yuan. 2023. Uedg: Uncertainty-edge dual guided camouflage object detection. *IEEE Transactions on Multimedia*, 26:4050–4060.

Tianjiao Ma, Jing Bai, Tiantian Li, Shuai Chen, Xiaodong Ma, Jie Yin, and Xuesong Jiang. 2021. Light-driven dynamic surface wrinkles for adaptive visible camouflage. *Proceedings of the National Academy of Sciences*, 118(48):e2114345118.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 11–20.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. 2022. 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 (CVPR)*, pages 2160–2170.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. 2024a. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Please add volume/number/pages/DOI if required by your bibliography style.

Youwei Pang, Xiaoqi Zhao, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. 2024b. Open-vocabulary camouflaged object segmentation.

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, Amir Mousavi, Yiwen Song, Abhimanyu Dubey, and Dhruv Mahajan. 2023. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7141–7151.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. 2024. Glamm: Pixel grounding large multimodal model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13009–13018.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. 2024. Pixellm: Pixel reasoning with large multimodal model.

Hao Shao and Qibin Hou. 2025. Medseg-r: Medical image segmentation with clinical reasoning. ArXiv preprint 2506.18669.

Yiqing Shen, Chenjia Li, Bohan Liu, Cheng-Yi Li, Tito Porras, and Mathias Unberath. 2025. Operating room workflow analysis via reasoning segmentation over digital twins. ArXiv preprint: 2503.21054.

Przemysław Skurowski, Hassan Abdulameer, J Błaszczyk, Tomasz Depta, Adam Kornacki, and P Kozieł. 2018. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*, 2(6):7.

Ze Song, Xudong Kang, Xiaohui Wei, Haibo Liu, Renwei Dian, and Shutao Li. 2023. Fsnet: Focus scanning network for camouflaged object detection. *IEEE Transactions on Image Processing*, 32:2267–2278.

Tingran Wang, Zaiyang Yu, Jianwei Fang, Jinlong Xie, Feng Yang, Huang Zhang, Liping Zhang, Minghua Du, Lusi Li, and Xin Ning. 2025. Multidimensional fusion of frequency and spatial domain information for enhanced camouflaged object detection. *Information Fusion*, 117:102871.

Yaming Wang, Jiatong Chen, Xian Fang, Mingfeng Jiang, and Jianhua Ma. 2024. Dual cross perception network with texture and boundary guidance for camouflaged object detection. *Computer Vision and Image Understanding*, 248:104131.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. 2022. CRIS: CLIP-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11686–11695.

Ranwan Wu, Tian-Zhu Xiang, Guo-Sen Xie, Rongrong Gao, Xiangbo Shu, Fang Zhao, and Ling Shao. 2025. Uncertainty-aware transformer for referring camouflaged object detection. *IEEE Transactions on Image Processing*, 34:5341–5354.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. 2024. Gsva: Generalized segmentation via multimodal large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3858–3869.

Haozhe Xing, Shuyong Gao, Hao Tang, Tsui Qin Mok, Yanlan Kang, and Wenqiang Zhang. 2023. Tinycod: Tiny and effective model for camouflaged object detection. In *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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

Mingde Yao, King Man Tam, Menglu Wang, Lingen Li, and Rei Kawakami. 2025. Language-guided reasoning segmentation for underwater images. *Information Fusion*, 122:103177.

Bowen Yin, Xuying Zhang, Deng-Ping Fan, Shaohui Jiao, Ming-Ming Cheng, Luc Van Gool, and Qibin Hou. 2024. Camoformer: Masked separable attention for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Please add volume/number/pages/DOI if required by your bibliography style.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024a. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 57730–57754.

Zhenni Yu, Xiaoqin Zhang, Li Zhao, Yi Bin, and Guobao Xiao. 2024b. Exploring deeper! segment anything model with depth perception for camouflaged object detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 4322–4330, New York, NY, USA. Association for Computing Machinery.

Zhenni Yu, Xiaoqin Zhang, Li Zhao, Yi Bin, and Guobao Xiao. 2024c. Exploring deeper! segment anything model with depth perception for camouflaged object detection. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, pages 4322–4330.

Guanghui Yue, Shangjie Wu, Tianwei Zhou, Gang Li, Jie Du, Yu Luo, and Qiuping Jiang. 2025. Progressive region-to-boundary exploration network for camouflaged object detection. *IEEE Transactions on Multimedia*, 27:236–248.

Hong Zhang, Yixuan Lyu, Qian Yu, Hanyang Liu, Huimin Ma, Ding Yuan, and Yifan Yang. 2025a. Unlocking attributes' contribution to successful camouflage: A combined textual and visual analysis strategy. In *Computer Vision – ECCV 2024*, pages 315–331, Cham. Springer Nature Switzerland.

Shizhou Zhang, Dexuan Kong, Yinghui Xing, Yue Lu, Lingyan Ran, Guoqiang Liang, Hexu Wang, and Yanning Zhang. 2025b. Frequency-guided spatial adaptation for camouflaged object detection. *IEEE Transactions on Multimedia*, 27:72–83.

Xiaoqin Zhang, Zhenni Yu, Li Zhao, Deng-Ping Fan, and Guobao Xiao. 2025c. Comprompter: Reconceptualized segment anything model with multiprompt network for camouflaged object detection. *Science China Information Sciences*, 68(1):112104.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 633–641.

Lanyun Zhu, Tianrun Chen, Qianxiong Xu, Xuanyi Liu, Deyi Ji, Haiyang Wu, De Wen Soh, and Jun Liu. 2025. Popen: Preference-based optimization and ensemble for lvlm-based reasoning segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30231–30240.

Appendix Overview. This supplementary material provides additional details for our proposed task and method. We first expand the discussion of segmentation tasks, which covers both referring expression segmentation and evaluation on standard COS benchmarks for fair comparison. Next, we present extended dataset details of MCD (a multi-instance camouflaged dataset), including representative examples. We then describe the implicit query instruction annotation pipeline, followed by the LLM-as-a-Judge metrics used to evaluate IRE. We futher describe the training data formulation used in our framework, including how semantic segmentation, referring segmentation, and VQA datasets are converted into a unified instruction-following format. We also provide additional ablation studies to analyze key design choices in COSA, and conclude with a failure case analysis that highlights current limitations and future directions. The content is organized as follows:

- Evaluation on Standard COS Benchmarks
- MCD: A Multi-instance Camouflaged Dataset
- Details of implicit query instruction annotation
- LLM-as-a-Judge Metrics for IRE
- Auxiliary Training Data Formulation
- Ablation Analysis (Additional)

A Evaluation on Standard COS Benchmarks

Motivation and protocol. Our main task, Language-Guided Reasoning Camouflaged Object Segmentation (LRCOS), differs fundamentally from conventional Camouflaged Object Segmentation (COS). LRCOS requires the model to follow a user-provided implicit query text instruction and segment the intended camouflaged target, which may involve reasoning under ambiguous cues and can further extend to previously unseen categories. In contrast, COS aims to segment the camouflaged object(s) in an image without any language input. Therefore, directly comparing LRCOS performance on CamoQuery with COS performance on standard benchmarks is not meaningful.

To enable a fair comparison under the COS setting, we additionally evaluate COSA on standard COS benchmarks by removing the query-specific requirement and instructing COSA to segment the camouflaged object in each image, i.e., aligning the evaluation goal with COS. Specifically, for each test image, we provide a generic instruction that asks COSA to segment the camouflaged target in the scene, and then compare the predicted mask with the ground-truth COS annotation. This evaluation isolates the core mask prediction capability of COSA and makes the results directly comparable to prior COS methods.

Datasets. We follow the standard COS evaluation protocol on three widely used benchmarks: CAMO (Le et al., 2019), COD10K (Fan et al., 2022), and NC4K (Lv et al., 2021). For each dataset, we use the official test split and the pro-

Method	Pub.	CAMO				COD10K				NC4K			
		S_m	$F_\beta^{\max}$	$E_\phi^{\max}$	M	S_m	$F_\beta^{\max}$	$E_\phi^{\max}$	M	S_m	$F_\beta^{\max}$	$E_\phi^{\max}$	M
ZoomNet (Pang et al., 2022)	CVPR_{22}	0.820	0.805	0.889	0.066	0.837	0.777	0.896	0.029	0.852	0.826	0.903	0.044
SegMAR (Jia et al., 2022)	CVPR_{22}	0.816	0.803	0.884	0.071	0.833	0.774	0.906	0.034	0.841	0.826	0.907	0.046
UEDG (Lyu et al., 2023)	TMM_{23}	0.863	0.856	0.929	0.048	0.858	0.812	0.934	0.025	0.879	0.864	0.935	0.035
TinyCOD (Xing et al., 2023)	ICASSP_{23}	0.822	0.807	0.899	0.066	0.811	0.742	0.903	0.036	0.843	0.817	0.910	0.047
MSCAF-Net (Liu et al., 2023e)	TCSVT_{23}	0.873	0.867	0.937	0.046	0.865	0.823	0.936	0.024	0.887	0.874	0.942	0.032
FSPNet (Huang et al., 2023)	CVPR_{23}	0.856	0.846	0.928	0.050	0.851	0.794	0.930	0.026	0.879	0.859	0.937	0.035
FEDER (He et al., 2023a)	CVPR_{23}	0.802	0.789	0.873	0.071	0.822	0.768	0.905	0.032	0.847	0.833	0.915	0.044
FSNet (Song et al., 2023)	TIP_{23}	0.880	0.878	0.941	**0.041**	0.870	0.833	0.948	0.023	0.891	0.880	0.948	0.031
CRNet (He et al., 2023b)	AAAI_{23}	0.735	0.707	0.830	0.092	0.733	0.636	0.845	0.049	–	–	–	–
PRNet (Hu et al., 2024)	TCSVT_{24}	0.872	0.867	0.930	0.050	0.873	0.839	0.943	0.022	0.891	0.881	0.942	0.031
CamoFormer (Yin et al., 2024)	TPAMI_{24}	0.817	0.801	0.883	0.068	0.838	0.786	0.928	0.029	0.854	0.829	0.908	0.042
$\text{ZoomNeXt}_{\text{pvtv2b2}}$ (Pang et al., 2024a)	TPAMI_{24}	0.874	0.873	0.931	0.047	0.887	0.856	0.945	**0.019**	0.892	0.884	0.941	0.030
DSAM (Yu et al., 2024c)	ACM MM_{24}	0.832	0.834	0.920	0.061	0.845	0.805	0.930	0.033	0.872	0.864	0.942	0.040
MAMIFNet (Wang et al., 2025)	IF_{25}	0.872	0.870	0.935	0.045	0.869	0.826	0.940	0.023	0.890	0.878	0.943	0.031
COMPrompter (Zhang et al., 2025c)	SCIS_{25}	0.853	0.856	0.931	0.054	0.860	0.826	0.946	0.027	0.880	0.876	0.946	0.036
COSA (Ours)	–	**0.888**	**0.884**	**0.938**	0.045	**0.901**	**0.885**	**0.952**	0.020	**0.906**	**0.897**	**0.949**	**0.030**

Table 7: Quantitative comparison on standard COS benchmarks (CAMO, COD10K, and NC4K). ↑ higher is better and ↓ lower is better.

vided pixel-level ground-truth masks.

Metrics. To be consistent with the COS literature, we report standard image-level segmentation metrics, including S_α (Fan et al., 2017), $F_\beta^{\max}$ (Achanta et al., 2009), $E_\phi^{\max}$ (Fan et al., 2018), and MAE (M) (Borji et al., 2015). These metrics are computed between the predicted mask and the ground-truth mask for each test image, following common COS evaluation practice.

Discussion. This COS-style evaluation should be viewed as an auxiliary analysis: it does not measure instruction following or reasoning (which are central to LRCOS), but it verifies that COSA remains competitive when its objective is aligned to the conventional COS formulation. The main conclusions of our paper are drawn from LRCOS-specific evaluation on CamoQuery, while the COS benchmark results provide an additional reference for the general mask segmentation capability of our framework. As shown in Table 7, COSA demonstrates consistently superior performance across the three standard COS benchmarks (CAMO, COD10K, and NC4K), achieving state-of-the-art or near state-of-the-art results on most metrics. It attains notably high scores in both structure-based metrics (S_m) and region-based metrics ($F_\beta^{\max}$, $E_\phi^{\max}$), while maintaining one of the lowest mean absolute errors (M). These results indicate that COSA possesses robust segmentation ability and strong generalization under conventional COS settings, even though its primary objective focuses on reasoning-driven language-guided segmentation.

B MCD: A Multi-instance Camouflaged Dataset

Motivation. The emergence of new tasks and datasets has consistently accelerated progress in computer vision. In the camouflage domain, existing COS benchmarks largely follow a single-instance-per-image setting, which is insufficient to reflect realistic scenes where multiple camouflaged targets co-exist and richer inter-object relations arise. With this in mind, our goals for collecting MCD are: (1) to introduce more realistic and challenging multi-instance camouflage scenes that increase reasoning demand, (2) to complement existing COS data with richer scenario diversity, and (3) to provide high-quality pixel-level masks so that MCD can be directly used in standard COS pipelines. Examples of MCD are shown in Fig. 9.

Image Collection Following common practice in dataset construction, we prioritize data diversity and annotation feasibility. MCD contains 2,000 camouflaged images collected from publicly available natural photography resources and web image repositories. We intentionally select images where multiple camouflaged targets may appear in the same scene, covering diverse environments such as forest, grassland, snowfield, wetland, and underwater backgrounds. This collection strategy increases the prevalence of multi-instance co-occurrence, making the dataset more representative of real scenes and more demanding for reasoning-oriented segmentation.

Professional Mask Annotation To ensure the reliability and usability of MCD for COS-style evaluation, we employ professional annotators to produce

Figure 9: Qualitative examples from the MCD dataset. Each image contains multiple camouflaged target instances.

pixel-level ground-truth masks. Each camouflaged target instance is annotated with a high-quality binary mask, and images containing multiple targets are labeled with instance-level masks for all camouflaged candidates. Annotators carefully delineate subtle object boundaries under low contrast, heavy texture confusion, and partial occlusions, resulting in accurate object contours suitable for training and benchmarking.

C Details of implicit query instruction annotation.

To produce high-quality implicit query text instructions for LRCOS, we adopt a two-stage annotation pipeline consisting of *LLM-based draft generation* and *human filtering and refinement*.

Step-1: Prompt-driven draft generation. Given a camouflaged image and a target instance defined by its pixel-level mask, we query GPT-4o with prompt templates (see Fig. 4) to generate instruction drafts. We design two prompt strategies to cover different reasoning depths: (1) *simple* prompts generate *lightweight-reasoning* instructions that typically rely on identifiable cues such as implicit attributes, functional hints, or weak semantic evidence to refer to the target, while explicitly requesting the model to output the segmentation mask and keeping the query *implicit* (i.e., avoiding direct category naming or explicit phrases like "the target in the mask"); (2) *complex* prompts generate *deeper-reasoning* instructions that require richer contextual reasoning or world knowledge, e.g., using behavior/usage, relational constraints to surrounding elements, or cross-cue integration, so that identifying the target requires a longer reasoning chain rather than direct naming. To ensure instructions are actionable in camouflaged scenes, we impose unified constraints in both prompts: no absolute spatial hints (e.g., left/right/top/bottom/center), no anchor-region references, and the instruction should remain a single, executable sentence.

Step-2: Human filtering and refinement. To ensure the final annotations meet the needs of training and evaluation, we recruit annotators with visual annotation experience to review and refine all LLM drafts. They follow the quality-control criteria below:

- **Relevance**: The instruction must provide discriminative clues for the intended target instance; overly generic or scene-irrelevant descriptions are removed.
- **Clarity**: The instruction must be concise and unambiguous, especially under multi-instance scenes where multiple candidates may partially match the text.
- **Reasonability**: The instruction should remain verifiable through visual evidence and contextual reasoning without explicitly naming the category; *simple* instructions keep lightweight

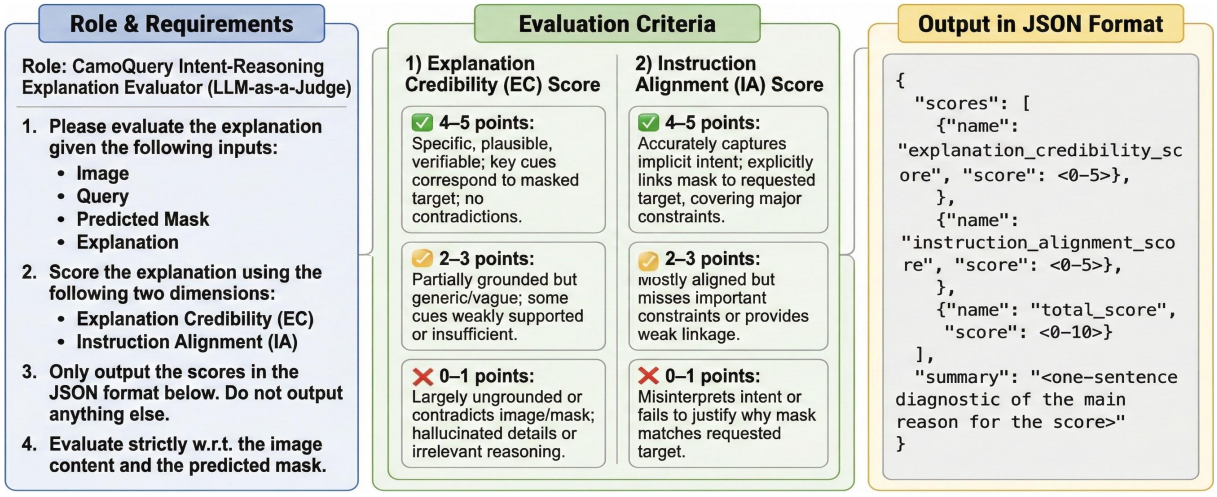

Figure 10: LLM-as-a-Judge prompt template for IRE explanation evaluation.

reasoning, while *complex* instructions reflect a longer reasoning chain, but both avoid unverifiable speculation.

- **Consistency**: The referred target must strictly match the given mask; drafts that do not align with the mask are rewritten.

Organization of final annotations. Each image is annotated with one *simple* and one *complex* implicit instruction as the default setting. For a subset of images, we additionally annotate multiple *complex* instructions to increase linguistic diversity and provide alternative reasoning paths. These extra *complex* instructions are intentionally added for challenging cases, *e.g.*, images with highly ambiguous camouflage cues, multiple plausible candidates, or fine-grained part-level targets, where a single complex query may not sufficiently cover diverse yet valid intents. Accordingly, we prioritize such images for extra annotation to support part-oriented zero-shot evaluation under implicit intents and to better stress-test reasoning robustness. Finally, we retain the *simple*/*complex* pair as the basic unit for each image–mask sample. This yields two reasoning levels over the same image–mask foundation, enabling systematic evaluation of *intent-consistent mask generation* and *intent-reasoning explanations* in camouflaged scenes.

C.1 LLM-as-a-Judge Metrics for IRE

Traditional text-overlap metrics (e.g., BLEU, METEOR) are not suitable for evaluating intent-reasoning explanations, as they often fail to reflect whether an explanation is plausible, instruction-consistent, and properly justified with respect to the predicted mask. Therefore, we adopt an LLM-as-a-Judge protocol to assess the explanation quality for IRE. Concretely, given the camouflaged image, the implicit query instruction (simple or complex), the predicted mask, and the model-generated explanation, the judge scores the explanation along two complementary dimensions:

Explanation Credibility (EC). EC measures whether the explanation is specific, plausible, and verifiable under the visual evidence and the predicted mask. High EC requires that the key camouflage cues described in the explanation are consistent with what can be inspected in the image/mask (e.g., texture/color similarity, boundary ambiguity, low local contrast), while hallucinated details or contradictions are penalized.

Instruction Alignment (IA). IA evaluates whether the explanation correctly captures the implicit intent of the instruction and explicitly links the predicted mask to the requested camouflaged target, covering major constraints implied by the query. Explanations that miss important constraints or provide weak linkage between the query intent and the selected mask receive lower scores.

Scoring and output format. Both EC and IA are rated on a 5-point scale with three intervals (0–1, 2–3, 4–5) using clearly defined criteria (Fig. 10). The total score is computed as:

$$s_{total} = s_{EC} + s_{IA}, \quad s_{total} \in [0, 10]. \tag{5}$$

To reduce randomness, we repeat the judging process multiple times (e.g., 5 runs) and report the averaged scores. The judge is instructed to output only a JSON object containing the EC score, IA

Dataset	M	T	QA	Role in training
CamoQuery (Ours)	✓	✓	✗	In-domain LRCOS
RefCOCO (Kazemzadeh et al., 2014c)	✓	✓	✗	Vision–text alignment with referring expressions
RefCLEF (Kazemzadeh et al., 2014a)	✓	✓	✗	Vision–text alignment with referring expressions
ADE20K (Zhou et al., 2017)	✓	✗	✗	Dense pixel supervision (converted to QA-style prompts)
COCO-Stuff (Caesar et al., 2018)	✓	✗	✗	Dense pixel supervision (converted to QA-style prompts)
LLaVA-Instruct-150k (Liu et al., 2023c)	✗	✓	✓	General instruction-following and visual reasoning
PACO-LVIS (Ramanathan et al., 2023)	✓	✗	✗	Part-level masks for fine-grained segmentation
PartImageNet (He et al., 2022)	✓	✗	✗	Part-level masks for fine-grained segmentation
PASCAL-Part (Chen et al., 2014)	✓	✗	✗	Part-level masks for fine-grained segmentation

Table 8: Auxiliary training data in our method. "M", "T", and "QA" denote pixel-level masks, text instructions/expressions, and question–answer supervision, respectively.

ID	Incremental setting	CamoQuery	SemSeg	RefSeg	VQA	PartSeg	gIoU	cIoU
1	CamoQuery only	✓	✗	✗	✗	✗	49.4	45.8
2	+ SemSeg (ADE20K, COCO-Stuff)	✓	✓	✗	✗	✗	51.1	47.2
3	+ RefSeg (RefCOCO/+/g, RefCLEF)	✓	✓	✓	✗	✗	51.7	48.0
4	+ VQA (LLaVA-Instruct-150k)	✓	✓	✓	✓	✗	52.1	48.5
5	+ PartSeg (PACO-LVIS, PartImageNet, PASCAL-Part)	✓	✓	✓	✓	✓	53.9	50.2

Table 9: Stepwise ablation on training data types. Starting from CamoQuery, we progressively add semantic segmentation (SemSeg), referring segmentation (RefSeg), VQA, and part-level segmentation (PartSeg).

score, the total score, and a one-sentence diagnostic summary, following the template shown in Fig. 10.

D Auxiliary Training Data Formulation (Additional).

As summarized in Table 8, auxiliary training data are drawn from widely-used public datasets, including semantic segmentation, referring segmentation, part-level segmentation, and visual question answering (VQA) datasets. The auxiliary sources and their reformulation into a unified instruction-following format are originally curated by LISA (Lai et al., 2024); this work adopts the same auxiliary data and follows the identical data formulation with binary-mask supervision. Further implementation details can be found in the LISA paper.

Semantic segmentation data. In LISA (Lai et al., 2024), semantic segmentation datasets provide images with multi-class pixel annotations. During training, one (or several) category(ies) are randomly sampled per image to construct a binary target mask for the selected category. Each sample is then converted into a QA pair using templates such as:

> USER: ⟨IMAGE⟩ Can you segment the {class name} in this image?
> ASSISTANT: It is ⟨SEG⟩.

where ⟨IMAGE⟩ denotes the image-token placeholder and ⟨SEG⟩ corresponds to the output mask. The binary mask is used as ground truth to supervise the mask loss. The adopted datasets include ADE20K and COCO-Stuff.

Referring segmentation data. In LISA (Lai et al., 2024), referring segmentation datasets provide an image and an explicit textual description of the target, and are reformulated into QA pairs via:

> USER: ⟨IMAGE⟩ Can you segment {description} in this image?
> ASSISTANT: Sure, it is ⟨SEG⟩.

where {description} is the provided referring expression. The adopted datasets include RefCOCO, RefCOCO+, RefCOCOg, and RefCLEF.

Visual question answering (VQA) data. In LISA, VQA data are included to preserve the original VQA ability of the multimodal LLM. LLaVA-Instruct-150k is used for LLaVA v1, and the language modeling objective is optimized on the answer texts.

Part-level segmentation data. In LISA (Lai et al., 2024), part-level segmentation datasets are additionally used to enhance fine-grained object-part segmentation. Each annotated part is treated as a target region with a binary mask, and is reformulated with the same QA templates by replacing {class name} with {part name}. The adopted

datasets include PACO-LVIS, PartImageNet, and PASCAL-Part.

In this work, these LISA-curated sources are used as auxiliary training data to complement the in-domain supervision, following the same reformulation and binary-mask supervision strategy.

E Ablation Analysis (Additional)

Contribution of Training Data Types. Table 9 reports a stepwise ablation on auxiliary training data types. Starting from CamoQuery-only training (ID 1), progressively adding each data type yields consistent gains. Incorporating semantic segmentation data improves mask quality, indicating that dense pixel-level supervision helps stabilize boundary prediction in camouflaged scenes. Adding referring segmentation data further boosts performance, suggesting better vision–text alignment when grounding target descriptions. Introducing VQA data brings additional improvements, likely due to enhanced general instruction-following and visual reasoning ability. Finally, adding part-level segmentation data provides the strongest improvement in the later stages, highlighting the importance of fine-grained part supervision for improving mask completeness and local details. Overall, combining all auxiliary data types achieves the best result, validating that these data sources are complementary and jointly strengthen COSA's segmentation accuracy under implicit-query camouflage supervision. All results are reported on the MCD.