MR2C: MULTIMODAL REFERRING AND REASONING FOR CAMOUFLAGED OBJECT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Camouflaged Object Segmentation (COS) has achieved significant progress in recent years, yet existing methods mainly focus on vision-only mask prediction and remain limited in understanding flexible user intent in interactive scenarios. To extend the boundaries of COS and facilitate future research in this field, we propose **M**ultimodal **R**eferring and **R**easoning **C**amouflaged **O**bject **S**egmentation (**MRR-COS**), a new task that aims to segment the queried camouflaged object according to multimodal user references. To support this task, we construct **MR2C**, a multimodal benchmark containing **9932** images, **64** categories, **9932** segmentation masks, **19864** textual references, **1920** reference images, and **640** reference sounds. MR2C stands out with three key features: (1) four types of multimodal references, including referring text, reasoning text, reference images, and reference sounds; (2) an emphasis on understanding user intent and target-related sound content rather than merely detecting camouflaged regions; and (3) the incorporation of complex reasoning and world knowledge into its textual references. Furthermore, we introduce **MCOSA**, a multimodal camouflaged object segmentation assistant, to address multimodal understanding, reasoning, and fine-grained segmentation in MR2C. Extensive experiments show that MCOSA outperforms existing methods on MR2C, demonstrating the necessity of the proposed task and the effectiveness of our benchmark and framework. Our code and dataset will be publicly available.

1 INTRODUCTION

Camouflaged Object Segmentation (COS) aims to segment targets that are visually concealed within their surrounding scenes (Xiao et al., 2026; Wu et al., 2025; Lei et al., 2025), and is crucial for medical image analysis (Han et al., 2026), surface defect inspection (Ma et al., 2021), and pest monitoring (Dong et al., 2024). Existing COS methods mainly formulate this task as a vision-only mask prediction problem, where models rely on appearance cues to determine *what* regions should be segmented. Although these methods have achieved promising progress under predefined settings, they still passively segment camouflaged regions from the input image and fail to adapt to interactive scenarios where users may specify the target through diverse instructions.

Recently, Referring Camouflaged Object Detection (Ref-COD) has been introduced to segment specified camouflaged objects using a small set of reference images (Zhang et al., 2025c). This task is important in real-world scenarios where multiple camouflaged objects may coexist, while users often focus on a specific target, such as a particular species or pest. However, Ref-COD remains limited to visual references and cannot cover more flexible user intents beyond reference images. CamoQuery (Han et al., 2026) further explores reasoning-guided camouflaged object segmentation through language-based queries, demonstrating the potential of reasoning cues for identifying camouflaged targets. Nevertheless, existing studies mainly focus on a single form of reference and lack a unified formulation for multimodal referring and reasoning. Meanwhile, Large Vision–Language Models (LVLMs) (Qian et al., 2025; Ying et al., 2025) provide new opportunities for instruction-guided visual understanding, but existing LVLM-based segmentation methods are mostly designed for clear and salient objects in natural images, and thus may suffer from spurious grounding, inaccurate segmentation, and unstable instruction following in highly camouflaged scenes. In practical interactions, users may specify camouflaged objects through textual referring and reasoning, reference images, or object-related sound cues. Therefore, a practical COS system should support mul-

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.

timodal referring and reasoning and generate accurate pixel-level masks according to user intent. However, existing COS datasets mainly provide images and binary masks, while current LVLM-based segmentation methods and evaluation settings are mostly designed for general natural images. *A comprehensive benchmark is therefore needed to evaluate multimodal referring and reasoning ability in camouflaged object segmentation.*

To this end, we propose a task, **Multimodal Referring and Reasoning Camouflaged Object Segmentation** (**MRR-COS**), which aims to segment the queried camouflaged object according to multimodal user references, as shown in Fig. 1. Different from conventional COS, which predicts masks from images alone, our task requires the model to understand user intent from multiple reference forms, identify the referred camouflaged object, generate a concise target-aware description, and produce an accurate segmentation mask. This task provides a more realistic and human-centric formulation for COS.

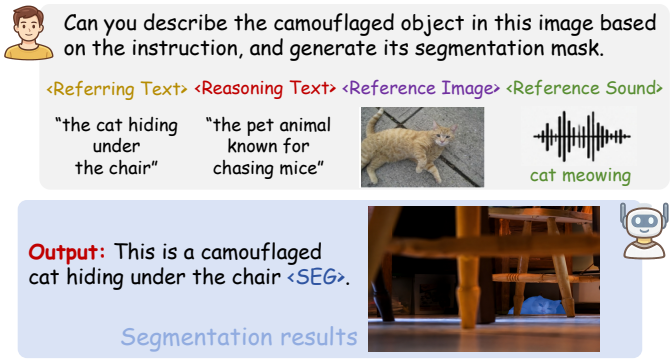


Figure 1: **Multimodal Referring and Reasoning Camouflaged Object Segmentation (MRR-COS).** Our model generates descriptions and pixel-wise masks for referred camouflaged objects. It supports diverse reference modalities, including **Referring Text**, **Reasoning Text**, **Reference Image**, and **Reference Sound** for flexible user interaction.

To support this task, we build **MR2C**, a benchmark for multimodal referring and reasoning camouflaged object segmentation. MR2C provides pixel-level mask annotations and contains multiple reference forms, including referring text, reasoning text, reference images, and reference sound. Specifically, MR2C contains **9932** images, **64** categories, **9932** segmentation masks, **19864** textual references, **1920** reference images, and **640** reference sound annotations. Among them, referring text directly describes the target through appearance, or location; reasoning text refers to the target through implicit clues and world knowledge; reference images provide external visual cues for establishing target correspondence; and reference sounds provide acoustic cues for grounding the corresponding camouflaged target. By introducing these reference forms, MR2C emphasizes not only detecting camouflaged regions, but also understanding user intent, reasoning with world knowledge, and understanding target-related reference sound content, thereby pushing camouflaged object understanding toward interactive, multimodal, and instruction-aware visual perception.

Building upon the above task and benchmark, we further propose **M**ultimodal **C**amouflaged **O**bject **S**egmentation **A**ssistant (**MCOSA**) for MR2C. Specifically, MCOSA encodes multiple reference modalities into a unified space aligned with the visual scene. Then, we introduce a refocusing mechanism to enhance segmentation token representations with contextual information, and a SeekOut Matching (SUM) strategy for multi-scale alignment between language and vision features. Finally, we introduce a dedicated <seg> token following LISA (Lai et al., 2024a), whose hidden representation is decoded into the final segmentation mask. In this way, the proposed model can simultaneously understand multimodal references, describe camouflaged objects, and generate pixel-level masks. As shown in Fig. 2, our method supports both direct referring and complex reasoning, and can generalize to different interaction forms.

Extensive experiments on MR2C and related tasks demonstrate the effectiveness of MCOSA. The results show strong segmentation performance, validating the necessity of the proposed task and the effectiveness of our benchmark and framework. In summary, the main contributions of this work are summarized as follows: **1)** We propose MRR-COS, a new task that fills the gap in existing COS by enabling multimodal interaction, user-intent understanding, and reasoning for more flexible user interactions. **2)** We build MR2C, a multimodal benchmark to support MRR-COS. It provides pixel-level masks and diverse reference forms, including referring text, reasoning text, reference images, and reference sound, covering both single-object and multi-object camouflaged scenes. **3)** We provide MCOSA as a baseline and conduct extensive experiments on MR2C and related tasks, demonstrating effective multimodal reasoning, target-aware description generation, and strong segmentation performance across different reference forms.

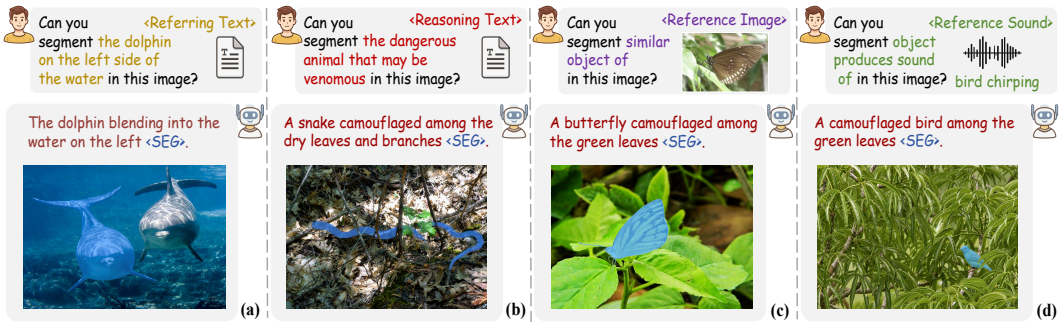


Figure 2: Qualitative examples of **MCOSA** on the proposed **MRR-COS** task under diverse reference modalities, including **(a) referring text**, **(b) reasoning text**, **(c) reference image**, and **(d) reference sound**. MCOSA generates natural language responses and accurate pixel-level segmentation masks for the queried camouflaged objects, demonstrating its strong ability to effectively handle flexible and diverse user references in challenging camouflaged scenes.

Table 1: Statistical comparison between the newly proposed **MR2C** benchmark and other related camouflaged object benchmarks. Avail., Expl., and Ref. Img. are abbreviations for availability of reasoning, explanations, and reference images, respectively.

Dataset	Venue	Content		Referring Format			Reasoning		Statistics					
		Image	Mask	Text	Ref. Img.	Sound	Avail.	Expl.	Image	Category	Mask	Text	Ref. Img.	Sound
COD10K (Fan et al., 2022)	[CVPR'20]	✓	✓	✗	✗	✗	✗	✗	5,066	68	5,066	N/A	N/A	N/A
CHAMELEON (Skurowski et al., 2018)	[N/A]	✓	✓	✗	✗	✗	✗	✗	76	N/A	76	N/A	N/A	N/A
NC4K (Lv et al., 2021)	[CVPR'21]	✓	✓	✗	✗	✗	✗	✗	4,121	N/A	4,121	N/A	N/A	N/A
CAMO (Le et al., 2019)	[CVIU'19]	✓	✓	✗	✗	✗	✗	✗	1,250	8	1,250	N/A	N/A	N/A
COD-TAX (Zhang et al., 2025a)	[ECCV'24]	✓	✗	✓	✗	✗	✗	✗	4,040	N/A	N/A	4,040	N/A	N/A
OVCamo (Pang et al., 2024)	[ECCV'24]	✓	✓	✓	✗	✗	✗	✗	11,483	75	11,483	11,483	N/A	N/A
R2C7K (Zhang et al., 2025c)	[TPAMI'25]	✓	✓	✗	✓	✗	✗	✗	5,015	64	5,015	N/A	1,600	N/A
MMCSBench (Zhang et al., 2025b)	[NeurIPS'25]	✓	✗	✓	✗	✗	✓	✗	22,537	N/A	3,045	76,843	N/A	N/A
CamoQuery (Han et al., 2026)	[ACL'26]	✓	✓	✓	✗	✗	✓	✓	12,437	N/A	12,437	25,971	N/A	N/A
MR2C (ours)	[N/A]	✓	✓	✓	✓	✓	✓	✓	**9932**	**64**	**9932**	**19864**	**1920**	**640**

2 Related Work

COS and Benchmarks. COS aims to identify and segment objects visually concealed in their surroundings. However, existing COS methods and main benchmarks (Le et al., 2019; Fan et al., 2022; Lv et al., 2021) remain vision-only, providing only pixel-level masks and thus failing to evaluate multimodal reference understanding and instruction follow-up to select user-specified camouflaged targets. Although recent attempts introduce references into COS (Zhang et al., 2025a; Pang et al., 2024), they are typically limited to attributes or category prompts, largely pre-defining what to segment. CamoQuery (Han et al., 2026) further introduces reasoning-based segmentation with implicit textual queries, but remains limited to language references. Recent camouflage LVLM benchmarks such as MMCSBench (Zhang et al., 2025b) evaluate high-level perception and language-based reasoning, but lack pixel-level mask annotations and diverse reference forms. Therefore, existing benchmarks are still insufficient to evaluate multimodal reference segmentation in camouflage, motivating MR2C. As summarized in Table 1, we compare MR2C with representative COS benchmarks in terms of task formulation, modality, and annotation scale.

Multimodal Reference Benchmarks. Recent LVLM benchmarks mainly evaluate perception and reasoning in natural scenes (Liu et al., 2024; Nguyen et al., 2026; Ren et al., 2024). Referring segmentation focuses on explicit text references, and reasoning segmentation methods such as LISA (Lai et al., 2024a) introduce reasoning text for mask prediction, but both are primarily designed for natural images. Audio-visual segmentation benchmarks (Zhou et al., 2022) study sound-guided localization, yet lack referring and reasoning references. OISA (Ying et al., 2025) investigates generalized referring segmentation in natural scenes, but does not specifically address camouflage perception and reasoning. In contrast, domains such as medicine (Shen et al., 2025; Shao & Hou, 2025), which also require fine-grained perception under ambiguous visual conditions, have benefited from systematic benchmarking efforts. Therefore, a multimodal reference benchmark for camouflaged object segmentation is needed, motivating MR2C.

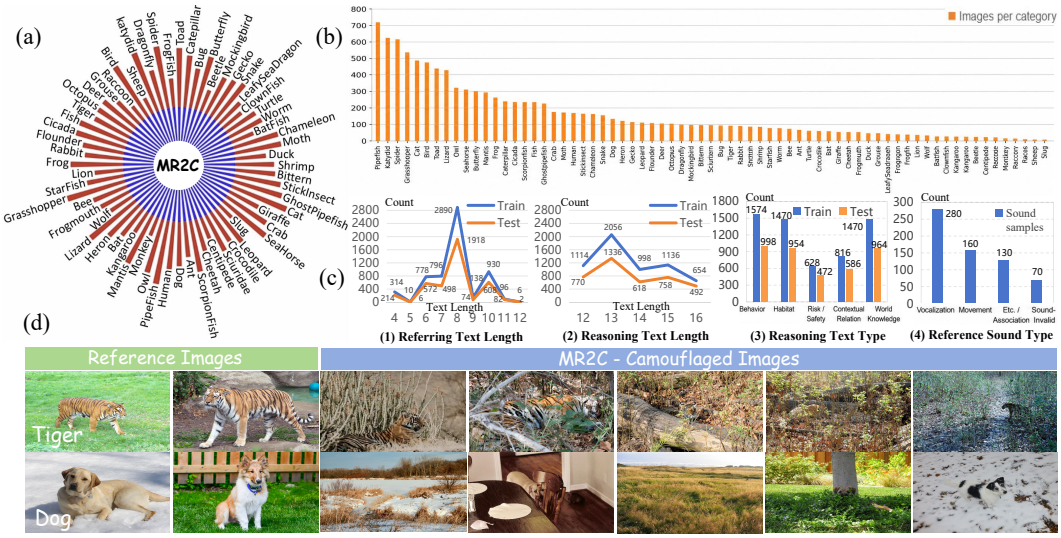


Figure 3: **Overview of MR2C**. **(a)** Object categories in MR2C Benchmark. **(b)** Long-tailed distribution of images per category. **(c)** Statistics of multimodal references, including **(1)** referring text length, **(2)** reasoning text length, **(3)** reasoning type, and **(4)** reference sound type. **(d)** Reference images (left) and corresponding camouflaged images in MR2C (right).

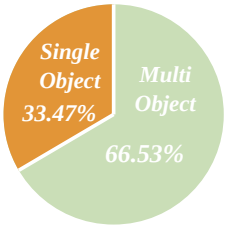
3 BENCHMARK: MR2C

Data Collection and Mask Annotations. To construct the MR2C dataset, we first determine the target camouflaged object categories by referring to several widely used COD datasets, including COD10K, CAMO, and NC4K. Since COD10K is the largest and most comprehensively annotated camouflage dataset, we use it as the primary data source and further incorporate additional samples from related COD datasets to improve overall category coverage.

Table 2: Dataset statistics.

Statistic	Number
Images (train/test)	5,958/3,974
Masks (train/test)	5,958/3,974
Reference Types	4
Referring text	9,932
Reasoning text	9,932
Reference image	1,920
Reference Sound	640
Single Images (train/test)	1,958/1,366
Multi Images (train/test)	4,000/2,608

Figure 4: Image composition stats.



Specifically, we remove several uncommon categories, such as pagurian and crocodile-fish, and finally collect 9,932 camouflaged images covering 64 categories. The dataset split statistics are summarized in Table 2. For images containing a single camouflaged object, we directly adopt the original mask annotations from the source datasets. For images containing multiple camouflaged objects, we retain only the mask annotations corresponding to the referring target and remove the remaining annotated regions. According to the number of camouflaged objects, MR2C can be divided into two groups, i.e., single object images and multiple object images, as illustrated in Fig. 4.

Multimodal Reference Design. MR2C is designed to support multiple forms of target reference in camouflaged images, as shown in Fig. 3. Specifically, we consider four types of references: referring text, reasoning text, reference images, and reference sounds. Referring text directly describes the target through explicit visual or semantic cues, such as category, appearance, position, or local attributes. Reasoning text specifies the target in a more implicit way, requiring the model to infer the intended object from contextual information, common sense, object function, behavior, or potential risk. Reference images provide external visual cues about the referred target, requiring the model to establish visual correspondence between the reference and the camouflaged object in the query image. Reference sounds provide acoustic cues associated with the referred target, requiring the model to establish cross-modal correspondence between the reference sound and the camouflaged object in the query image. These reference forms jointly enable MR2C to evaluate multimodal referring, complex reasoning, and fine-grained pixel-level grounding in camouflaged images.

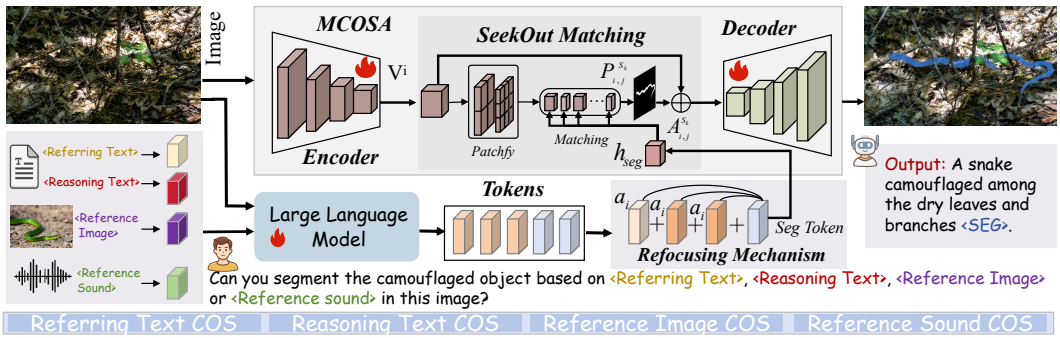


Figure 5: **Overall framework of MCOSA**. MCOSA supports four COS tasks with multimodal references: referring text, reasoning text, reference image, and reference sound. Given an image and a multimodal reference, the large language model produces a <seg> token, whose representation is enhanced by the refocusing mechanism. The enhanced segmentation embedding is further matched with multi-scale visual features through SeekOut Matching and decoded into the final mask.

1) Referring Text Annotation. Referring text provides direct and explicit descriptions of the target camouflaged object. For each image, annotators manually write a short textual description according to the target object and its mask annotation. The referring text is required to clearly indicate the target object using visible attributes, category names, spatial location, or local appearance cues. For example, for an image containing a frog camouflaged among green leaves, the referring text can be annotated as "the camouflaged frog on the leaf". Similarly, for a butterfly blending into dry branches, the referring text can be annotated as "the hidden butterfly near the branch". These referring texts are manually written to specify the camouflaged object to be segmented.

2) Reasoning Text Annotation. Reasoning text provides implicit descriptions of the camouflaged object and requires the model to infer the target from semantic or commonsense cues. Rather than describing the target, annotators describe the object through its function, behavior, habitat, potential risk, interaction with the environment, or world knowledge. For example, for a snake hidden among dry leaves, the reasoning text can be annotated as "the animal that should be avoided because it may inject venom when threatened". For an insect resembling a dry twig, the reasoning text can be annotated as "the living creature that imitates a branch to avoid being discovered by predators".

3) Reference Image Annotation. We construct the reference image annotations of MR2C over the selected 64 categories. For data collection, category names are used only as retrieval indices to collect 30 real-world reference images for each category, where each image clearly presents an object corresponding to the queried camouflaged target. Importantly, the category labels are used only for organizing the reference pool and are not provided to the model. Given a reference image, the model is required to infer its visual semantics, establish correspondence with the target in the query image, and directly predict the corresponding pixel-level mask. Following the image collection protocol of Ref-COD (Zhang et al., 2025c), the reference images are collected from publicly available sources such as Flickr and Unsplash, and samples with low visual quality are removed.

4) Reference Sound Annotation. We further introduce reference sound annotations over 64 categories. Category names are used only as retrieval indices to collect 10 sound clips for each category from audio sources. Each clip contains acoustic characteristics associated with the queried camouflaged target, such as animal calls or other characteristic sounds. Similar to reference images, category labels are not used as model inputs. Instead, the model must interpret the acoustic semantics of the reference sound, associate them with the target in the query image, and directly generate its segmentation mask. All sound clips are manually screened to remove samples with unclear semantics, poor audio quality, or potential copyright issues. The statistics are summarized in Table 2.

4 METHODS: MCOSA

Overview. The overall framework of **MCOSA** comprises a vision encoder, multi-modal feature projection layers, a large language model (LLM), and a mask decoder, as illustrated in Fig. 5. These components form a unified multimodal LLM, supporting various reference formats and generating region-aware grounded responses in camouflaged scenes. Additionally, a special <seg> token is

introduced, bridging language reasoning and pixel-level segmentation through a hybrid grounding mechanism consisting of a refocusing module and a SeekOut Matching (SUM) strategy. Detailed implementation is provided in the supplementary material.

Model Architecture. We adopt the pretrained ViT-L/14 from CLIP (Radford et al., 2021) as the vision encoder, and LLaMA-7B (Touvron et al., 2023) as our LLM. For audio inputs, we use the pretrained audio encoder from ImageBind (Girdhar et al., 2023) to extract audio features. To connect multimodal information beyond text to the LLM, such as images and audio, we adopt vision-language and audio-language projection layers to project image and audio features into the language space. The input image is converted into a fixed number of 16×16 patch embeddings, while the audio is represented as 3 patch embeddings. Both image and audio embeddings are projected to the same dimension as word embeddings. The LLM takes interleaved multimodal embeddings in the same way as language tokens and generates outputs in an auto-regressive manner. For segmentation, we introduce a special <seg> token following LISA (Lai et al., 2024a), whose hidden state from the last LLM layer is used as a query embedding for pixel-level prediction.

Unified Multimodal Reference Representation. To support multimodal references beyond visual inputs, we unify different types of references into fixed-size token representations using special markers, enabling seamless integration within the LLM. For textual references, we consider two forms: *referring text* and *reasoning text*, both directly tokenized and embedded into the LLM input sequence. For visual references, image features are extracted using a CLIP vision encoder and encoded as fixed-length embeddings, which are inserted as contextual visual tokens. For audio references, audio features are extracted using ImageBind and projected into the language embedding space via an audio-language projection layer. All reference modalities are unified into token-level representations and processed by the LLM for multimodal reasoning and segmentation.

Segmentation Token and Refocusing Mechanism. To enable pixel-level segmentation, MCOSA introduces a special segmentation token <seg>, following LISA (Lai et al., 2024a). Given an input image and instruction, the LLM generates a sequence of hidden states, where the hidden state corresponding to the <seg> token is extracted as h_{seg} from the last transformer layer. However, the representation of a single token may contain insufficient contextual information for fine-grained visual localization. To address this issue, we propose a refocusing mechanism that enriches h_{seg} with contextual information from preceding tokens:

$$\hat{h}_{seg} = h_{seg} + \lambda_f \sum_{i<i_{seg}} \bar{a}_i \cdot h_i, \tag{1}$$

where $\bar{a}_i$ denotes the attention weight between token i and the <seg> token, averaged over the attention heads of the last LLM layer and normalized over preceding tokens, and λ_f controls the contribution of contextual refinement. The enhanced representation $\hat{h}_{seg}$ is then projected through a lightweight MLP γ to obtain the segmentation embedding $q_{seg} = \gamma(\hat{h}_{seg})$, which is used as the query input to the segmentation branch for pixel-level mask prediction.

SeekOut Matching Strategy (SUM). To further enhance pixel-level alignment in camouflaged scenes, we propose a SeekOut Matching strategy (SUM). Given the segmentation query q_{seg} obtained from the refocusing mechanism, SUM performs multi-scale alignment between the segmentation query and visual features. Specifically, given a visual feature map V_m, SUM extracts local patch representations $P_{i,j}^{s_k}$ at different scales:

$$P_{i,j}^{s_k} = \Phi\left(U_{s_k}(V_m)_{i,j}\right) \in \mathbb{R}^{N_k \times d}, \qquad N_k = s_k^2,\ s_k \in \{s_1, \ldots, s_n\}, \tag{2}$$

where U_{s_k} extracts an $s_k \times s_k$ local patch centered at (i, j), and Φ projects the patch features into the same embedding space as q_{seg}. The matching response and multi-scale feature aggregation are:

$$\alpha_{i,j}^{s_k} = \mathrm{softmax}_{N_k}\left(\frac{P_{i,j}^{s_k} q_{seg}}{\sqrt{d}}\right), \qquad V_m^{\phi} = \psi\left(G\left(V_m \oplus \sum_{k=1}^{n} Z^{s_k}\right)\right). \tag{3}$$

Here, the softmax is performed over the $N_k = s_k^2$ spatial positions within each local patch. $Z^{s_k} \in \mathbb{R}^{d \times H_m \times W_m}$ denotes the scale-specific matching map obtained by aggregating the local patch features with $\alpha_{i,j}^{s_k}$, $\oplus$ denotes channel-wise concatenation, G is a convolutional fusion layer, and ψ denotes the ReLU activation. This multi-scale integration enables the model to localize camouflaged regions across different levels of granularity. More details are provided in the Appendix B.

Table 3: **Comparison on MR2C under different reference modalities.** We report mIoU (%) under three evaluation settings, including all images, single-object images, and multi-object images. Overall denotes the joint use of all four reference types. Unsupported settings are marked with "–".

Method	Referring Text			Reasoning Text			Reference Image			Reference Sound			Overall			METEOR
	All	Single	Multi	All	Single	Multi	All	Single	Multi	All	Single	Multi	All	Single	Multi	
Specialist Segmentation Models																
MLKG	57.8	58.5	49.9	–	–	–	53.2	54.9	43.5	–	–	–	–	–	–	–
R2CNet	–	–	–	–	–	–	60.3	60.8	53.1	–	–	–	–	–	–	–
MMCSBench	64.1	65.5	55.9	61.8	63.2	52.4	–	–	–	–	–	–	–	–	–	43.2
Generalist MLLMs																
LISA-7B	58.4	59.7	49.8	55.1	56.6	46.3	51.8	52.9	43.7	44.5	45.8	36.2	53.0	54.2	44.5	37.6
PixelLM	60.2	61.5	51.6	57.4	58.8	48.5	53.6	54.8	45.1	46.7	48.1	38.4	55.1	56.3	46.5	39.8
GLaMM	62.1	63.4	53.3	59.2	60.6	50.4	55.9	57.2	47.6	48.3	49.7	40.1	57.1	58.4	48.6	41.2
AnyRef	64.5	65.8	55.7	61.3	62.9	52.1	60.4	61.7	51.6	53.8	55.2	45.3	60.9	62.2	52.0	43.7
OISA	63.1	64.4	54.2	60.5	61.9	51.4	58.7	60.1	49.8	56.2	57.5	47.9	60.3	61.7	51.5	42.9
MCOSA (ours)	**68.9**	**70.2**	**59.1**	**65.8**	**67.1**	**56.3**	**65.4**	**66.7**	**56.2**	**59.8**	**61.2**	**50.7**	**69.5**	**70.3**	**59.4**	**46.4**

Decoder and Training Objectives. The decoder takes the SUM-enhanced hierarchical features and predicts the final segmentation mask. MCOSA is trained end-to-end with text generation and mask prediction losses. The text loss follows the next-token prediction objective (Liu et al., 2023b), while the mask loss combines binary cross-entropy and dice losses (Milletari et al., 2016):

$$\mathcal{L} = \lambda_{text}\mathcal{L}_{text} + \lambda_{bce}\mathcal{L}_{bce} + \lambda_{dice}\mathcal{L}_{dice}. \tag{4}$$

Following the training setup of LiSA (Lai et al., 2024a), we set $\lambda_{text} = 1.0$, $\lambda_{bce} = 2.0$, and $\lambda_{dice} = 0.5$.

5 Experiments

Implementation Details. We use CLIP ViT-L/14 (Radford et al., 2021) as the vision encoder, ImageBind-H (Girdhar et al., 2023) as the audio encoder, and LLaMA-7B (Touvron et al., 2023) as the LLM. The vision-language projection layer is initialized from LLaVA (Liu et al., 2023b), while the audio-language projection layer and new special tokens are randomly initialized. SAM-H (Kirillov et al., 2023) is used for mask prediction. Images are resized to 224×224 for the MLLM and 1024×1024 for segmentation. Audio inputs are 5-second clips encoded into 3 fixed-size embeddings. During training, we freeze the vision and audio encoders and fine-tune the LLM with LoRA. We train MCOSA on 8 NVIDIA V100 GPUs with batch size 6 per GPU, AdamW (Loshchilov & Hutter, 2019) optimizer, learning rate 5×10^{-5}, weight decay 0.01, and 10K training steps.

Datasets. Following prior works (Lai et al., 2024a; He et al., 2024), we train MCOSA on a mixture of segmentation and multimodal reference datasets, including semantic/instance segmentation datasets (COCO-Stuff (Caesar et al., 2018), ADE20K (Zhou et al., 2017), PACO-LVIS (Ramanathan et al., 2023)), referring expression segmentation datasets (RefClef, RefCOCO, RefCOCO+ (Yu et al., 2016), RefCOCOg (Mao et al., 2016), PhraseCut (Wu et al., 2020)), image-level referring segmentation data built from COCO, PascalVOC (Everingham et al., 2010), and PhraseCut (Wu et al., 2020), audio-visual segmentation data from AVSBench (Zhou et al., 2022), and our MR2C dataset with text, image, and sound references. We evaluate on MR2C using the split in the Benchmark section.

Evaluation Metrics. Following previous work (Lai et al., 2024a; He et al., 2024), we adopt mIoU for segmentation evaluation, where IoU measures region similarity between the predicted mask and ground truth. We use METEOR (Banerjee & Lavie, 2005) to evaluate text explanation generation, with human evaluation in the supplementary material.

5.1 MR2C Benchmark Results

Quantitative Results. In Table 3, we report the MR2C benchmark results under four reference modalities and an Overall setting that jointly uses all four reference types. We evaluate all methods under All, Single, and Multi settings according to the number of camouflaged objects. For camouflage-related specialist methods, we evaluate MLKG (Cheng et al., 2023), R2CNet (Zhang et al., 2025c), and MMCSBench (Zhang et al., 2025b) under their supported settings. R2CNet only supports reference-image-based Ref-COD, MLKG supports text and image reference cues but lacks reasoning text and sound references, and MMCSBench only supports language-based camouflage

Table 4: Ablation study on the refocusing weight λ_f and SUM.

Setting	Referring Text	Reasoning Text	Reference Image	Reference Sound	Overall	METEOR
λ_f=0	66.1	63.0	62.7	56.8	66.7	45.1
λ_f=0.05	67.5	64.3	64.0	58.3	68.0	45.2
λ_f=0.1	**68.9**	**65.8**	**65.4**	**59.8**	**69.5**	**46.4**
λ_f=0.5	67.9	64.9	64.6	58.9	68.6	45.7
λ_f=1	66.8	63.7	63.5	57.6	67.4	45.8
Without SUM	64.8	61.6	61.3	55.4	66.3	44.3
Single-scale SUM	67.2	64.1	63.8	58.1	67.8	45.1
Multi-scale SUM	**68.9**	**65.8**	**65.4**	**59.8**	**69.5**	**46.4**

reasoning without reference image or reference sound. For generalist MLLMs, LISA (Lai et al., 2024a), PixelLM (Ren et al., 2024), GLaMM (Rasheed et al., 2024), AnyRef (He et al., 2024), and OISA (Ying et al., 2025) are evaluated for comparison. LISA, PixelLM, and GLaMM mainly support text-guided segmentation but lack reference image and reference sound perception; we therefore equip them with the same visual-reference and audio-reference projection layers as MCOSA while preserving their original segmentation paradigms. AnyRef and OISA are adapted to MR2C with the same four reference formats for fair comparison. MCOSA achieves the best performance across individual reference modalities and the Overall setting, and also obtains better METEOR scores, indicating stronger multimodal understanding, segmentation, and explanation generation.

Qualitative Results and Failure Cases. In Fig. 6, we present qualitative results under four reference modalities. Existing methods often produce incomplete masks or miss camouflaged targets, while MCOSA achieves more accurate segmentation by leveraging textual, visual, and acoustic cues. MCOSA successfully identifies the bats from referring text, reasoning text, and reference image inputs. For reference sound, although MCOSA correctly associates the bat calls with the corresponding camouflaged targets, it may still miss one highly camouflaged instance. This indicates that sound based camouflaged object segmentation remains challenging, and we hope MR2C can serve as a useful benchmark to promote future research on more effective sound-guided segmentation methods.

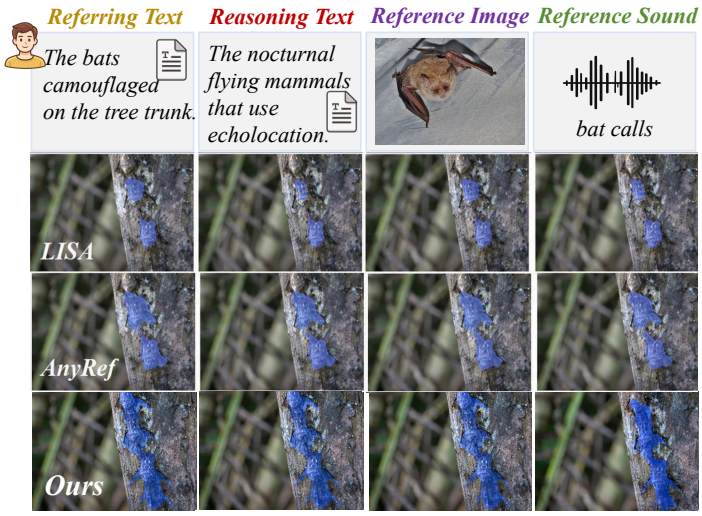


Figure 6: Qualitative results of MCOSA.

5.2 Ablation Study

We conduct ablation studies on MR2C under the Overall setting to examine key components.

Core Components. We first evaluate the core components of MCOSA in Table 5, including the <seg> token, refocusing mechanism, and SeekOut Matching (SUM). The <seg> token connects LLM reasoning with mask prediction, refocusing enhances its contextual representation, and SUM further improves multi-scale visual matching. Combining all components achieves the best performance.

Table 5: Ablation study of **core components** under the Overall setting. Seg.: <seg> token; Refocus.: refocusing mechanism. w/o Seg.: last output token as the segmentation query.

ID	Core Components			Overall Metrics			
	Seg.	Refocus.	SUM	All	Single	Multi	METEOR
I	✗	✗	✗	58.0	58.9	47.9	39.7
II	✓	✗	✗	63.7	64.5	54.1	42.5
III	✓	✓	✗	66.3	67.1	56.5	44.3
IV	✓	✓	✓	**69.5**	**70.3**	**59.4**	**46.4**

Refocusing Mechanism and SeekOut Matching. We study the refocusing weight λ_f and different SUM designs in Table 4. Setting $\lambda_f = 0$ removes the refocusing mechanism, while a moderate value

Table 6: Ablation study on training datasets under the Overall setting.

ID	MR2C	Sem./Inst. Seg.	Ref. Exp. Seg.	Image Ref. Seg.	Audio-Visual Seg.	mIoU	METEOR
1	✓	✗	✗	✗	✗	63.0	44.6
2	✓	✓	✗	✗	✗	64.7	45.1
3	✓	✓	✓	✗	✗	66.6	45.6
4	✓	✓	✓	✓	✗	67.9	46.2
5	✓	✓	✓	✓	✓	**69.5**	**46.4**

Table 7: Comparison of MCOSA with related methods on audio-visual, referring, and reasoning segmentation benchmarks. Left: audio-visual segmentation results on AVSBench. Right: cIoU results on refCOCO, refCOCO+ (Yu et al., 2016), refCOCOg (Mao et al., 2016), and ReasonSeg (Lai et al., 2024b). Bold and underline indicate the best and second-best results, respectively.

Method	Single-source		Multi-source	
	mIoU	F-score	mIoU	F-score
AVS (Zhou et al., 2022)	78.7	87.9	54.0	64.5
BG (Hao et al., 2024)	81.7	90.4	55.1	66.8
AVSegFormer (Gao et al., 2024)	82.1	89.9	<u>58.4</u>	<u>69.3</u>
AnyRef (He et al., 2024)	82.8	<u>90.8</u>	55.6	66.3
AV2T-SAM Lee et al. (2025)	<u>83.1</u>	90.1	55.8	63.4
MCOSA (ours)	**83.4**	**91.6**	**59.6**	**70.6**

(a) Audio-Visual Segmentation

Method	refCOCO			refCOCO+			refCOCOg		ReasonSeg	
	val	testA	testB	val	testA	testB	val_U	test_U	val	test
LAVT (Yang et al., 2022)	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1	–	–
GRES (Liu et al., 2023a)	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0	–	–
LISA-7B (Lai et al., 2024a)	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.4	54.0	48.4
AnyRef (He et al., 2024)	74.1	75.5	70.8	64.1	68.7	57.5	68.1	69.9	–	–
READ (Qian et al., 2025)	**78.1**	**80.2**	<u>73.2</u>	<u>68.4</u>	**73.7**	**60.4**	**70.1**	<u>71.4</u>	<u>67.6</u>	**58.6**
MCOSA (ours)	<u>76.4</u>	<u>78.4</u>	**73.4**	**68.7**	<u>71.9</u>	<u>58.8</u>	<u>69.4</u>	**71.8**	**68.8**	<u>57.6</u>

(b) Referring and Reasoning Image Segmentation

enhances the <seg> token with useful contextual information. For SUM, single-scale matching improves visual correspondence, and the proposed multi-scale SUM achieves the best performance by exploiting visual features at different granularities. Therefore, we set $\lambda_f = 0.1$ and use multi-scale SUM in all experiments.

Training Datasets. We ablate the training data in Table 6 by progressively adding semantic/instance segmentation, referring expression segmentation, image-level referring segmentation, and audio-visual segmentation datasets to MR2C. These additional datasets provide complementary supervision for generic object understanding, language-guided localization, cross-image referring, and audio-visual correspondence. The full mixture achieves the best performance.

5.3 Results on Related Tasks

Audio-Visual Segmentation. This task requires localizing and segmenting sounding objects according to audio-visual correspondence. As shown in Table 7 (a), MCOSA performs favorably against existing methods, demonstrating its ability to associate acoustic semantics with visual targets. In particular, the consistent gains across single-source and multi-source settings indicate robust cross-modal alignment under different numbers of sounding objects. These results further support the effectiveness of reference sounds in multimodal camouflaged object segmentation.

Referring and Reasoning Image Segmentation. This task requires segmenting image regions specified by explicit referring expressions or implicit reasoning instructions. As shown in Table 7 (b), MCOSA achieves competitive results, showing strong text-guided segmentation ability. The consistent performance across multiple referring and reasoning benchmarks further demonstrates its capability to understand diverse linguistic instructions and accurately localize corresponding targets.

6 Conclusion

We present the MRR-COS task and MR2C, a multimodal camouflaged object segmentation benchmark supporting four complementary reference forms. Based on MR2C, we further propose MCOSA, a unified MLLM framework for multimodal camouflaged object segmentation. MCOSA unifies heterogeneous reference modalities into token-level representations, introduces a special <seg> token for explicit mask prediction, and further improves target localization and segmentation through the refocusing mechanism and SeekOut Matching. Extensive experiments on MR2C and multiple related tasks demonstrate the effectiveness, robustness, and generalization ability of MCOSA across diverse reference modalities. Performance may vary under highly ambiguous cross-modal cues. Future work will explore stronger temporal, acoustic, and interactive reasoning capabilities for more comprehensive multimodal camouflaged object understanding.

AI USE STATEMENT

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

REPRODUCIBILITY STATEMENT

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

REFERENCES

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

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

Shupeng Cheng, Ge-Peng Ji, Pengda Qin, Deng-Ping Fan, Bowen Zhou, and Peng Xu. Large model based referring camouflaged object detection. *arXiv preprint arXiv:2311.17122*, 2023.

Qing Dong, Lina Sun, Tianxin Han, Minqi Cai, and Ce Gao. Pestlite: A novel yolo-based deep learning technique for crop pest detection. *Agriculture*, 14(2), 2024. ISSN 2077-0472.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88 (2):303–338, 2010.

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

Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12155–12163, 2024. doi: 10.1609/aaai.v38i11.29104.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Tianxin Han, Qing Dong, Xingwei Wang, Jie Jia, Gang Wu, Bowen Yang, and Fu Zhang. CamoQuery: Language-guided reasoning camouflaged object segmentation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22924–22941, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1050. URL https://aclanthology.org/2026.acl-long.1050/.

Dawei Hao, Yuxin Mao, Bowen He, Xiaodong Han, Yuchao Dai, and Yiran Zhong. Improving audio-visual segmentation with bidirectional generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2067–2075, 2024. doi: 10.1609/aaai.v38i3.27978.

Junwen He, Yifan Wang, Lijun Wang, Huchuan Lu, Jun-Yan He, Jin-Peng Lan, Bin Luo, and Xuansong Xie. Multi-modal instruction tuned llms with fine-grained visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13980–13990, 2024.

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

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

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, 2024b.

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. ISSN 1077-3142. doi: 10.1016/j.cviu.2019.04.006. URL https://doi.org/10.1016/j.cviu.2019.04.006.

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

Chang Liu et al. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23592–23601, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

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. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision (ECCV)*, 2024.

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

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

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

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

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. URL https://doi.org/10.1109/3DV.2016.79.

Tuan Dung Nguyen, Minh Khoi Ho, Qi Chen, Yutong Xie, Cam-Tu Nguyen, Minh Khoi Nguyen, Dang Huy Pham Nguyen, Anton van den Hengel, Johan Verjans, Phi Le Nguyen, and Vu Minh Hieu Phan. Beyond the global scores: Fine-grained token grounding as a robust detector of lvlm hallucinations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40235–40244, June 2026.

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

Rui Qian et al. Reasoning to attend: Try to understand how <seg> token works. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

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. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7141–7151, June 2023.

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

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

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

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

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10216–10225, 2020.

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

Feng Xiao, Jianhua Zhang, Guang Yang, Ruyu Liu, Xu Cheng, and Shengyong Chen. Zoomingcod: Zooming-like behavior for camouflage object detection based on cross-axis perception and adaptive receptive field. *IEEE Transactions on Multimedia*, pp. 1–13, 2026. doi: 10.1109/TMM.2026.3668536.

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 (CVPR)*, pp. 18155–18165, 2022.

Kaining Ying, Henghui Ding, Guangquan Jie, and Yu-Gang Jiang. Towards omnimodal expressions and reasoning in referring audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22575–22585, 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. Springer, 2016.

Hong Zhang, Yixuan Lyu, Qian Yu, Hanyang Liu, Huimin Ma, Ding Yuan, and Yifan Yang. Unlocking attributes' contribution to successful camouflage: A combined textual and visual analysis strategy. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, volume 15113 of *Lecture Notes in Computer Science*, pp. 315–331, Cham, 2025a. Springer Nature Switzerland. ISBN 978-3-031-73001-6. doi: 10.1007/978-3-031-73001-6_18.

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, 2025b.

Xuying Zhang, Bowen Yin, Zheng Lin, Qibin Hou, Deng-Ping Fan, and Ming-Ming Cheng. Referring camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3597–3610, 2025c. doi: 10.1109/TPAMI.2025.3532440.

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

Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation. In *European Conference on Computer Vision*, pp. 386–403. Springer, 2022.

A More Details of MR2C

A.1 Reference Image Collection

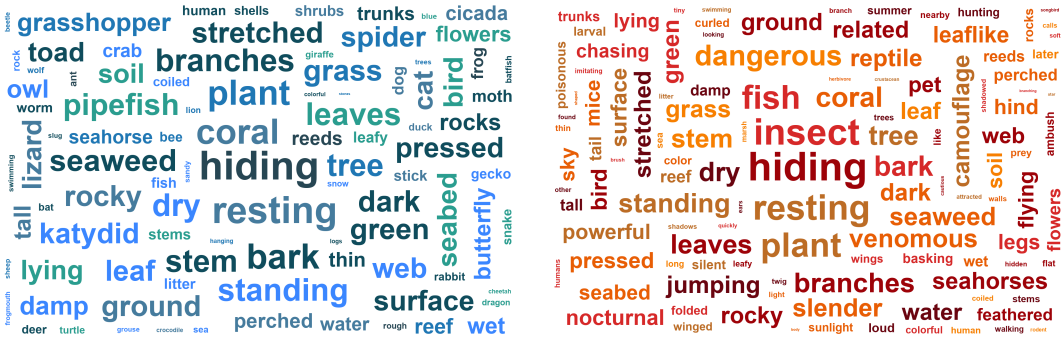
For each category in MR2C, we collect reference images from publicly available online resources. The reference images are selected to clearly present the corresponding object category, so that they can serve as visual references for identifying camouflaged targets in query images. We use category names and manually curated aliases as search queries, and remove images with low resolution, heavy occlusion, ambiguous category labels, or irrelevant content.

To ensure source traceability, we record the metadata of each accepted reference image, including its category, file path, source page, source URL, creator information when available, matched query, and collection note. We only retain resources whose licenses or usage terms allow academic research use, and remove samples with unclear sources or potential copyright restrictions. All accepted reference images are manually checked to ensure that their visual content matches the corresponding MR2C category. Detailed source information is provided together with the released dataset.

A.2 Reference Sound Collection

For each category in MR2C, we collect reference sounds from publicly available audio resources. The reference sounds are expected to reflect acoustic characteristics related to the target category, such as animal calls or other category-related sounds. We use category names and curated aliases as search queries, and manually screen the collected audio clips to remove samples with unclear semantics, severe background noise, poor audio quality, or mismatched content.

For each accepted sound clip, we record detailed source information, including its category, file path, title, source page, source URL, creator, matched query, file size, and collection note. These metadata records make the source of each reference sound traceable and verifiable. We retain only audio resources whose licenses or usage terms permit academic research use, and discard clips with missing source information or potential copyright restrictions. Detailed source information is provided together with the released dataset.



(a) Referring Text **(b) Reasoning Text**

Figure 7: Word cloud visualizations of textual references in MR2C. Referring texts mainly focus on object categories and visual attributes, while reasoning texts contain more implicit semantic cues related to behavior, function, habitat, and world knowledge.

A.3 Textual Reference Annotation

Referring Text. For reasoning-text annotation, annotators are provided with the camouflaged image and the target mask, which indicate the referred object without prescribing a textual category label. Annotators are asked to formulate an implicit description based on target-specific semantic cues, including behavior, function, habitat, potential risk, interaction with the environment, or world knowledge. For example, a frog hidden on leaves can be annotated as "the camouflaged frog on the leaf", and a butterfly near branches can be annotated as "the hidden butterfly near the branch". After annotation, each referring text is manually checked together with the image-mask pair. Expressions are revised if they are vague, mismatched with the target, or unable to distinguish the queried object from the background.

Reasoning Text. For reasoning text annotation, annotators are also provided with the camouflaged image, the mask, and the category. Different from referring text, annotators are asked to describe the target through implicit semantic clues rather than direct visual appearance alone. These clues include object behavior, function, habitat, potential risk, interaction with the environment, or world knowledge. For example, a hidden snake can be annotated as "the animal that may inject venom when threatened", while an insect resembling a branch can be annotated as "the living creature that imitates a twig to avoid predators". Each reasoning text is manually verified together with the image-mask pair to ensure that the implicit cues are semantically consistent with the referred target and sufficiently informative for target grounding. Texts with incorrect commonsense, overly generic clues, or ambiguous target descriptions are revised during verification.

A.4 Word Cloud Analysis of Textual References

Referring Text. As shown in Fig. 7(a), the word cloud of referring texts mainly contains object names and visual attributes, such as category, color, shape, location, and appearance-related words. This indicates that referring texts describe the target in an explicit manner and provide direct cues for identifying the queried camouflaged object. They mainly evaluate whether the model can follow clear user references and ground the target according to visible attributes.

Reasoning Text. As shown in Fig. 7(b), the word cloud of reasoning texts contains more words related to behavior, function, habitat, risk, and commonsense. This shows that reasoning texts describe the target in a more implicit way and require the model to infer the queried object from semantic clues. They further evaluate whether the model can use world knowledge beyond direct visual descriptions.

B More Details of MCOSA

B.1 Refocusing Mechanism

MCOSA uses a special segmentation token <seg> to connect multimodal reasoning with pixel-level mask prediction. Given an input image and a multimodal reference, we extract the hidden state

$\boldsymbol{h}_{seg} \in \mathbb{R}^{d_l}$ corresponding to the <seg> token from the last LLM layer. Since a single token representation may contain insufficient contextual information for fine-grained grounding, we further aggregate preceding tokens according to their attention to the segmentation token. Specifically, let $\boldsymbol{A}^{(L,r)}$ denote the self-attention matrix of the r-th head in the last LLM layer L. We average the attention scores over all H heads and normalize them over the tokens preceding <seg>:

$$a_i = \frac{1}{H}\sum_{r=1}^{H} A^{(L,r)}_{i_{seg},i}, \qquad \bar{a}_i = \frac{a_i}{\sum_{j<i_{seg}} a_j + \epsilon}, \qquad \hat{\boldsymbol{h}}_{seg} = \boldsymbol{h}_{seg} + \lambda_f \sum_{i<i_{seg}} \bar{a}_i \boldsymbol{h}_i. \tag{5}$$

Here, i_{seg} denotes the position of the segmentation token, $\boldsymbol{h}_i$ is the hidden state of the i-th preceding token, and λ_f controls the contextual contribution. We set $\lambda_f = 0.1$ in all experiments. The enhanced representation is then projected by a lightweight MLP γ and used as the segmentation query, i.e., $\boldsymbol{q}_{seg} = \gamma(\hat{\boldsymbol{h}}_{seg})$. Together with the segmentation visual features $\mathcal{V}_{seg}(\boldsymbol{x}_{img})$, this query is passed to the mask decoder for final prediction. In this way, the segmentation query preserves the target semantics encoded by the LLM while incorporating contextual cues that are strongly related to the queried object.

B.2 SeekOut Matching Strategy

Camouflaged objects often exhibit weak boundaries and strong appearance similarity with their surroundings, making direct visual grounding difficult. To strengthen the correspondence between the segmentation query and local visual structures, we introduce SeekOut Matching (SUM), which performs query-aware matching over local visual neighborhoods at multiple scales. Let $V_m \in \mathbb{R}^{C_m \times H_m \times W_m}$ denote the m-th visual feature map and $\boldsymbol{q}_{seg} \in \mathbb{R}^d$ the segmentation query obtained from the refocusing mechanism. For each scale $s_k \in S$, we extract an $s_k \times s_k$ neighborhood around spatial location (i, j) using stride 1 and padding $\lfloor s_k/2 \rfloor$. After projection and flattening, each local patch contains $N_k = s_k^2$ visual tokens. The local matching and aggregation are formulated as:

$$\begin{aligned} P^{s_k}_{i,j} &= \Phi(U_{s_k}(V_m)_{i,j}) \in \mathbb{R}^{N_k \times d}, \qquad N_k = s_k^2, \\ \alpha^{s_k}_{i,j} &= \mathrm{softmax}_{N_k}\left(\frac{P^{s_k}_{i,j}\boldsymbol{q}_{seg}}{\sqrt{d}}\right), \\ \boldsymbol{z}^{s_k}_{i,j} &= \sum_{p=1}^{N_k} \alpha^{s_k}_{i,j,p} P^{s_k}_{i,j,p}. \end{aligned} \tag{6}$$

The softmax is therefore performed over the N_k spatial positions inside each local patch. Collecting $\boldsymbol{z}^{s_k}_{i,j}$ over all (i, j) yields a scale-specific matching map $Z^{s_k} \in \mathbb{R}^{d \times H_m \times W_m}$. Because stride-1 unfolding with scale-dependent padding preserves the same $H_m \times W_m$ spatial resolution for every scale, the matching maps can be directly aggregated and fused with the original visual feature:

$$Z^{\mathrm{MS}} = \sum_{k=1}^{n} Z^{s_k}, \qquad V^{\phi}_m = \psi\big(G\big(V_m \oplus Z^{\mathrm{MS}}\big)\big). \tag{7}$$

Here, $\oplus$ denotes channel-wise concatenation, G is a lightweight convolutional fusion layer, and ψ denotes ReLU. The resulting V^{ϕ}_m is fed into the mask decoder for final segmentation. In contrast to directly fusing multi-scale features, SUM first performs target-aware local matching at each scale and then aggregates the matched responses, allowing the segmentation query to capture both fine object details and larger contextual structures.

C Human Evaluation of Explanation Quality

Automatic metrics such as METEOR mainly evaluate surface-level lexical overlap and may not fully reflect whether an explanation is semantically correct, target-aware, and consistent with the predicted segmentation. Therefore, we further conduct a human evaluation to assess the explanation quality of different methods.

Specifically, we randomly sample 20 examples from the MR2C test set, covering four reference settings, including referring text, reasoning text, reference image, and reference sound. Each example

Method	METEOR	Human-1	Human-2	Human-3	Human-4	Human-5	Avg.
MMCSBench (Zhang et al., 2025b)	43.2	4.0	4.1	3.9	4.0	4.0	4.0
LISA-7B (Lai et al., 2024a)	37.6	3.2	3.3	3.1	3.3	3.2	3.2
PixelLM (Ren et al., 2024)	39.8	3.4	3.5	3.3	3.5	3.4	3.4
GLaMM (Rasheed et al., 2024)	41.2	3.6	3.7	3.5	3.6	3.6	3.6
AnyRef (He et al., 2024)	43.7	4.1	4.2	4.0	4.1	4.1	4.1
OISA (Ying et al., 2025)	42.9	4.0	4.1	3.9	4.0	4.0	4.0
MCOSA (ours)	**46.4**	**4.3**	**4.3**	**4.2**	**4.3**	**4.3**	**4.3**
Human	–	**4.4**	**4.8**	**4.6**	**4.7**	**4.5**	**4.6**

Table 8: Automatic and human evaluation of explanation quality on MR2C. Human denotes manually written reference explanations and serves as the upper bound. Human scores are averaged over five annotators, with higher values indicating better target relevance, factual correctness, and reasoning consistency.

is evaluated by five human annotators. For each evaluated sample, the annotators are shown the input image, the reference prompt, the predicted mask, and the generated explanation. The method name is hidden during evaluation to avoid potential bias. In addition, we include human-written explanations as an upper-bound reference.

Each explanation is rated on a 1–5 scale according to three criteria: target relevance, factual correctness, and reasoning consistency. The detailed scoring protocol is summarized in Table 9. For each method, we report the average score over all evaluated samples and annotators in Table 8.

As shown in Table 8, human-written explanations achieve the highest average score, serving as the upper bound of explanation quality. Among automatic methods, MCOSA obtains the best human evaluation result, indicating that it generates more target-aware and semantically grounded explanations. MMCSBench, AnyRef, and OISA also achieve competitive scores, showing that these methods can produce generally reasonable explanations for camouflaged object understanding. In contrast, generalist segmentation MLLMs such as LISA-7B, PixelLM, and GLaMM receive relatively lower scores, mainly because their explanations are less stable in fine-grained camouflaged scenes and may not always align well with the queried target. These results complement the automatic METEOR evaluation and further demonstrate that MCOSA improves not only mask prediction but also explanation quality under multimodal reference settings.

Score	Description
1	Incorrect explanation; fails to describe the queried target or contains obvious factual errors.
2	Partially related explanation; mentions some relevant cues but includes clear mistakes or misses key information.
3	Mostly correct explanation; identifies the target but provides limited or incomplete reasoning.
4	Correct explanation; describes the target with relevant visual or semantic cues and only minor omissions.
5	High-quality explanation; accurately describes the target and provides clear, consistent, and well-grounded reasoning.

Table 9: Scoring protocol for human evaluation.

D MULTIPLE-REFERENCE COMBINATION.

We further investigate whether multiple reference types provide complementary information for target segmentation. As shown in Table 10, combining heterogeneous references generally improves performance over using a single reference. The improvement becomes more evident as additional complementary modalities are introduced. In particular, jointly using all four reference types achieves the best performance, reaching 69.5, 70.3, and 59.4 mIoU under the All, Single, and Multi settings, respectively. These results demonstrate that MCOSA can effectively integrate textual, visual, and acoustic cues for multimodal target grounding.

Setting	Single Reference				Two References						Three References				Four References
	RT	ReT	RI	RS	RT+ReT	RT+RI	RT+RS	ReT+RI	ReT+RS	RI+RS	RT+ReT+RI	RT+ReT+RS	RT+RI+RS	ReT+RI+RS	All Four
All	68.9	65.8	65.4	59.8	69.2	69.3	69.0	66.4	65.9	65.8	69.4	69.2	69.3	66.6	**69.5**
Single	70.2	67.1	66.7	61.2	70.2	70.2	70.0	67.6	67.2	67.0	70.2	70.1	70.2	67.8	**70.3**
Multi	59.1	56.3	56.2	50.7	59.2	59.3	59.1	56.8	56.4	56.5	59.3	59.2	59.3	57.0	**59.4**

Table 10: Evaluation of different reference combinations on MR2C. RT, ReT, RI, and RS denote referring text, reasoning text, reference image, and reference sound, respectively.

E Analysis of Multimodal Reference Integration

To further examine whether MCOSA genuinely exploits heterogeneous references rather than relying predominantly on textual cues, we conduct controlled inference-time experiments by modifying only the reference inputs while keeping the trained model unchanged. We first remove one modality at a time from the complete four-reference input and additionally evaluate text-only (RT+ReT) and non-textual (RI+RS) references. We then replace the correct reference image or sound with an unrelated one to examine whether MCOSA actually relies on their semantic content. We further corrupt reference sounds with Gaussian noise at 10 dB and 0 dB SNR to evaluate robustness to degraded acoustic cues. Finally, we construct conflicting multimodal inputs by keeping the textual references correct while replacing the reference image, reference sound, or both with unrelated references.

As shown in Table 11, removing individual reference modalities consistently reduces performance, while the complete four-reference setting achieves the best results. Removing RT causes the largest degradation, whereas removing ReT, RI, or RS also leads to measurable drops, indicating that the Overall result does not arise from textual input alone. The text-only RT+ReT setting is also inferior to All Four, further showing that visual and acoustic references provide complementary information. More importantly, replacing the correct reference image or sound with an unrelated reference causes a substantial performance decrease, demonstrating that MCOSA relies on the semantic content of these references rather than simply ignoring non-textual inputs. The decrease is more pronounced on the Multi subset, where reference-conditioned grounding is more challenging. Performance also decreases progressively as reference sounds are corrupted from clean audio to 10 dB and 0 dB, showing that MCOSA can exploit acoustic information under moderate noise while remaining sensitive to severe degradation. Finally, conflicting references produce clear performance drops, and simultaneously corrupting both image and sound references leads to the largest decrease among the conflicting settings. These observations indicate that MCOSA actively integrates textual, visual, and acoustic cues rather than depending exclusively on the strongest textual reference. Together with the reference-combination results in Table 10, these controlled experiments provide additional evidence for effective multimodal reference integration in MCOSA.

Table 11: Robustness analysis of multimodal reference integration on MR2C. Results are reported in mIoU (%).

Setting	All	Single	Multi
Modality Dropout			
All Four	**69.5**	**70.3**	**59.4**
w/o RT	66.6	67.8	57.0
w/o ReT	69.3	70.2	59.3
w/o RI	69.2	70.1	59.2
w/o RS	69.4	70.2	59.3
RT+ReT	69.2	70.2	59.2
RI+RS	65.8	67.0	56.5
Reference Perturbation			
RI (Clean)	65.4	66.7	56.2
RI (Shuffled)	50.6	52.0	41.2
RS (Clean)	59.8	61.2	50.7
RS (Shuffled)	46.8	48.1	37.9
Acoustic Corruption			
RS (10 dB)	58.4	59.8	49.2
RS (0 dB)	55.7	57.1	46.4
Conflicting References			
Conflict-RI	67.1	68.2	56.7
Conflict-RS	68.4	69.4	58.1
Conflict-RI+RS	66.3	67.4	55.4

F Comparison with Standard Multi-scale Baselines

To further determine whether the improvement of SUM mainly comes from additional multi-scale computation, we compare it with two standard multi-scale baselines. **MS-Fusion** resizes visual features from different scales to the same spatial resolution and directly fuses them by concatenation followed by a 1×1 convolution. **MS-Attention** first performs the same multi-scale fusion and then applies standard cross-attention between the fused visual features and the segmentation embedding. In contrast, SUM performs query-aware local matching between the segmentation embedding and visual patches at each scale before multi-scale aggregation. For a fair comparison, all variants use the same backbone, decoder, training schedule, and multi-scale visual features.

Method	MS	Query	All	Single	Multi
w/o SUM	✗	✗	66.3	67.2	56.5
MS-Fusion	✓	✗	66.7	67.3	56.8
MS-Attention	✓	✓	67.1	68.0	57.0
S-SUM	✗	✓	67.8	68.7	57.6
M-SUM	✓	✓	**69.5**	**70.3**	**59.4**

Table 12: Comparison of SUM with standard multi-scale baselines under the Overall setting. MS denotes multi-scale processing.

As shown in Table 12, directly introducing multi-scale visual features through MS-Fusion improves the Overall mIoU from 65.3 to 66.4, indicating that multi-scale information itself is beneficial. Adding standard query-aware attention further improves the result to 67.1, showing the importance of aligning visual features with the segmentation query. S-SUM achieves 67.8 mIoU by performing local query-aware matching, outperforming both standard alternatives. Finally, the proposed M-SUM reaches 69.5 mIoU by performing such matching across multiple visual scales. These results indicate that the gain of M-SUM cannot be attributed solely to additional multi-scale computation; the proposed scale-specific query-aware matching provides additional improvements over standard multi-scale fusion and attention.

G Limitations and Future Work

Although MR2C provides a multimodal benchmark for camouflaged object segmentation with diverse reference forms, there remains room for further extension. First, the current reference images and sounds are collected at the category level. They provide useful visual and acoustic cues for target identification, but may not fully cover all fine-grained intra-class variations. Future versions of MR2C can include more diverse reference samples across different viewpoints, environments, and acoustic conditions. Second, reference sounds are naturally more challenging than text and image references, since some categories have weak or less distinctive acoustic cues. Therefore, sound-based reference segmentation may still be affected by background noise and semantic ambiguity. This provides a useful direction for future research on more robust sound-guided camouflaged object understanding.

Finally, although MR2C provides a multimodal benchmark for referring and reasoning camouflaged object segmentation, it is still centered on camouflaged scenes. This focused design allows MR2C to systematically evaluate fine-grained target grounding under camouflage, where objects often share similar color, texture, or shape with the background. However, the current benchmark does not cover other challenging segmentation scenarios, such as transparent objects, heavily occluded objects, low-light scenes, or general salient objects. Future work can extend the benchmark to a broader set of visually challenging scenes, enabling a more comprehensive evaluation of multimodal referring and reasoning segmentation.