Leveraging Multimodal Large Language Models for Referring Camouflaged Object Detection

1st Xuewei Liu*
Institute of Information Engineering,
Chinese Academy of Sciences
School of Cyber Security,
University of Chinese Academy of Sciences
Beijing, China
liuxuewei@iie.ac.cn

2nd Ziyu Wei*
Beihang University,
Institute for Artificial Intelligence
Beijing, China
weiziyu@buaa.edu.cn

3rd Jizhong Han
Institute of Information Engineering,
Chinese Academy of Sciences
School of Cyber Security,
University of Chinese Academy of Sciences
Beijing, China
hanjizhong@iie.ac.cn
*Corresponding author

***Abstract*—Referring camouflaged object detection aims to identify and segment specified objects hidden within their surroundings, given a text or image as reference. Previous methods have shown limitations in effectively integrating referential and image information, leading to suboptimal performance. Multimodal large language models (MLLM) have demonstrated excellent performance in various tasks, where the rich visual-text information is highly beneficial for referring camouflaged object detection. In this paper, we propose a referring camouflaged object detection framework MLLM-RCOD based on MLLM. We design a Referring Image Packer (RIP) to enhance the extraction of image referential information, employ triple-align pretraining to improve modality alignment, and design a Camouflaged Chain of Thought to guide MLLM in better understanding the characteristics of the Ref-COD task, thereby achieving superior performance. Extensive experiments on the Ref-COD benchmark show that our method achieves new state-of-the-art performance.**

***Index Terms*—Referring Camouflaged Object Detection, Large Multi-Modal Model**

I. Introduction

Camouflaged Object Detection (COD) [1] focuses on detect and segment objects that are visually concealed within their surroundings. This task is inherently challenging due to the high similarity in color and texture between the target objects and their environment, as well as the frequent occlusions caused by surrounding elements. Even humans often struggle to accurately distinguish these objects without explicit cues. To mitigate the difficulty of recognition, Referring Camouflaged Object Detection (Ref-COD) [2] has been introduced, incorporating additional text or images that belong to the same semantic category as the target objects for guidance (as illustrated in Fig. 1). The incorporated information is typically referred to as reference information. However, Ref-COD remains limited by morphological and appearance discrepancies between the reference images and the target objects, as well as modality gaps between textual references and camouflaged images. Effectively leveraging reference information to enhance the detection of camouflaged objects remains a critical research challenge.

Recently, Multimodal Large Language Models (MLLM) have demonstrated remarkable capabilities across various vision perception tasks, including object detection [3] [4], segmentation [5] [6], spatiotemporal grounding [7]. Exploring the potential of MLLM and applying it to Ref-COD holds the promise of significantly enhancing its performance. However, directly applying MLLM to Ref-COD results in suboptimal performance, which primarily stems from two limitations. First, MLLMs lack the domain-specific knowledge required for COD tasks. Unlike the internet datasets used for MLLM training, COD typically deals with natural outdoor environments where target objects are either heavily obscured or visually identical to their surroundings, presenting a distinct visual challenge that current MLLMs are not specifically designed to address. First, there is a lack of domain-specific knowledge. COD images are typically from natural outdoor environments, and the targets to be identified are often obscured or have colors very similar to their surroundings, which is quite different from most of the pertaining dataset of MLLMs. Second, the modality of prompts is limited. Most existing MLLMs only accept text as input prompts (or conditions), such as Qwen-VL [8]'s language-based spatial localization and PSALM [9]'s text-referent segmentation. However, the RefCOD task not only uses text as a reference but also employs reference images to guide the segmentation of camouflaged objects. Therefore, the ability of existing MLLMs to adapt to the input data of Ref-COD is still limited.

To fully leverage the rich multimodal knowledge in MLLM, we designed a unified text-/image-guided Ref-COD framework based on MLLM, called MLLM-RCOD. This framework aims to explore the potential of large models in visual tasks and enhance the performance of Ref-COD under dual-modality references of text and images. To address the issue of single-modality prompts, we designed a Referring Image Packer, which compresses the information of reference images into the text space, thereby unifying text-guided Ref-COD and image-guided Ref-COD into a single segmentation framework.

*Equal Contribution.

Additionally, to enhance the alignment capability of packer features with the text space and to retain the original image information in packer features, we conducted Triple-Align Pretraining. This is based on Triple-Align Loss, which aligns the features among camouflage images, referring images, and referring texts.

To address the issue of lacking domain knowledge, we designed a Camouflage Chain of Thought (CoT) and input it into the MLLM along with the hidden object detection instructions. This approach explicitly guides the MLLM in understanding the scene where the camouflaged object is located, the attributes of the camouflaged object, and thus better segments the camouflaged targets from a mixed background from three aspects: scene awareness, object understanding, and mask prediction.

The contributions of our paper are summarized as follows:

1) We designed a unified text-/image-guided Ref-COD framework based on MLLM, called MLLM-RCOD. Additionally, we developed a Referring Image Packer to better compress the information of reference images into the text space and conducted Triple-Align Pretraining to enhance the alignment capability between the image modality and the text modality.

2) To further activate the MLLM's ability to recognize camouflaged targets, we designed a Camouflage Chain of Thought (Cam-CoT). This explicitly guides the MLLM in understanding the scene where the camouflaged object is located and the attributes of the camouflaged object, thereby improving the segmentation of camouflaged targets from a mixed background.

3) Our algorithm achieves state-of-the-art (SOTA) accuracy in hidden object detection for both image-guided and text-guided scenarios. On the R2C7K dataset, the weighted measure improves by 6% compared to the current SOTA models

II. Related Works

A. Camouflaged Object Detection

Camouflaged Object Detection (COD) [1] aims to detect and segment objects that naturally blend into their environment, making them difficult to distinguish due to their resemblance to the background, variability in size, and indistinct appearance. To this end, many strategies have been proposed to solve this problem, such as Multi-scale-context based strategy [10],Multi-source information fusion strategy [11] [12], Multi-task learning strategy [13] [14],Joint-SOD based strategy [15]. With the rapid development of diffusion models, they have also been applied to COD. For instance, DiffCOD [16] treats the camouflaged object segmentation task as a denoising diffusion process from noise masks to target masks. CamDiff [17] utilizes a latent diffusion model to synthesize salient targets in camouflaged scenes, while leveraging the zero-shot image classification capability of the Contrastive Language-Image Pretraining (CLIP) [18] model to prevent synthesis failures and ensure that the synthesized targets align with the input prompts.

Zhang et al. [2] proposed using textual descriptions of the detected object's category or clear images of camouflaged objects as prompt information to recognize camouflaged targets in the presence of such prompts, and introduced R2CNet as a baseline. Liu et al. [19] proposed a new Reference Prompt Model Adaptation (RPMA) pipeline, which enhances the capability of Ref-COD models by utilizing rich fine-grained semantic knowledge in general segmentation networks. This work aims to explore the visual understanding capabilities of MLLM and the integrated understanding of textual and visual information to improve the performance of RefCOD.

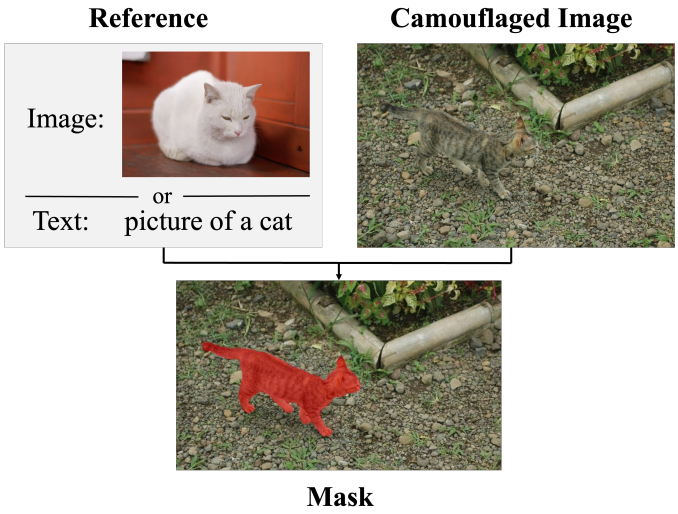

Fig. 1: Illustration of Ref-COD.

B. MLLM

With the maturation of the LLM field, the idea of applying LLM as a foundational model to other domains has been continuously researched and explored. The advent of GPT-4V [20] demonstrates the immense potential of LLM as a foundational model in vision-text bimodal understanding. The emergence of open-source MLLMs such as Flamingo [21], BLIP [22], and LLaVA [23] has ushered in a new era of research in applying LLM to multimodal understanding. Currently, models like LLaVA-OV [24] and Qwen2VL [25] have already shown superior performance in image understanding and multimodal reasoning. Kosmos [26] and vTimeLLM [27] have further introduced spatial object detection and video temporal localization tasks into MLLM, while LLaVA-ST [7] has achieved excellent results in video spatiotemporal joint localization. PSALM [9], based on the LLM framework, has implemented a unified segmentation model that integrates generic segmentation, referent segmentation, and panoptic segmentation.

III. Methods

A. MLLM-RCOD

The design of MLLM-RCOD follows the idea of LLaVA [23] and consists of three parts: a vision encoder, an LLM,

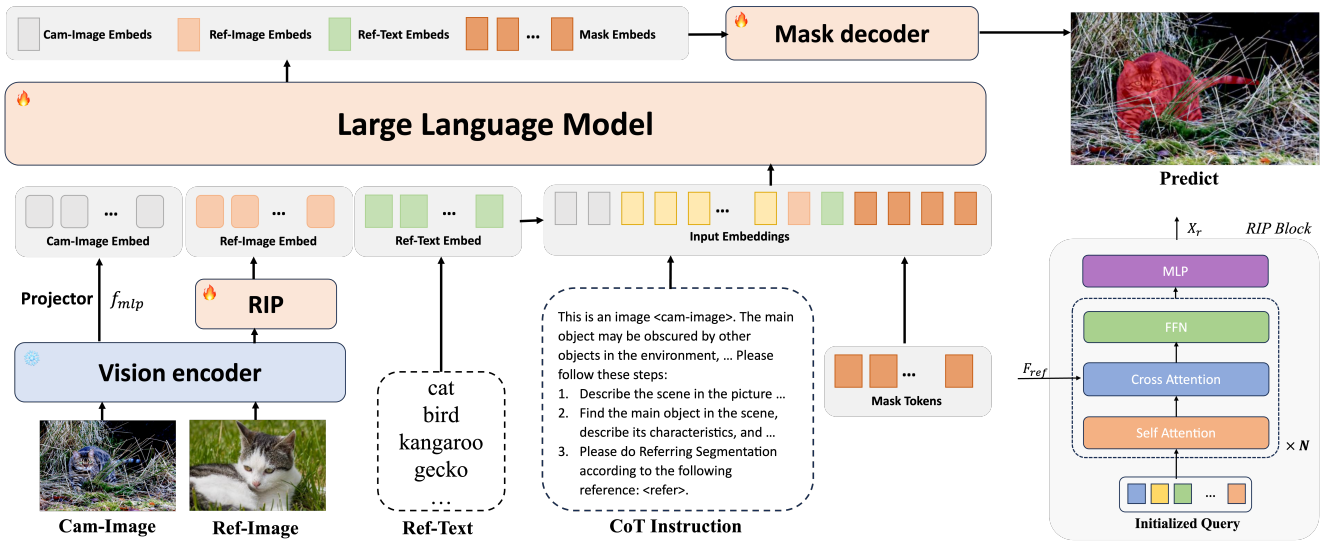

Fig. 2: Overall architecture of our proposed MLLM-RCOD and RIP block. Image information is aligned to the text space through a vision encoder and a specific projector, and then input into the LLM along with text embeddings. Utilizing the rich prior knowledge of the LLM, mask information is extracted, and finally, a mask decoder is used to generate the mask of the camouflaged object.

and a mask decoder. Fig. 2 illustrates the overall artitecture of MLLM-RCOD. The text input is tokenized and represented as embeddings, while the image input is processed using a vision encoder to extract image features, which are then aligned to the text space through a projector layer. During the training process of Ref-COD, there are four inputs: the camouflaged image I_c, the large model text instructions T_i, the reference text T_r, and the reference image I_r. We use a unified vision encoder f_v to extract features from both I_c and I_r.

$$\begin{aligned} F_c &= f_v(I_c) \\ F_r &= f_v(I_r) \end{aligned} \tag{1}$$

The extracted features F_c from I_c are aligned to the LLM's text space using LLaVA [23]'s MLP projector. For I_r, we designed a Referring Image Packer (RIP) as the projector to better extract its referential information, which will be introduced in Section 3.2.

$$\begin{aligned} X_c &= MLP(F_c) \\ X_r &= RIP(F_r) \end{aligned} \tag{2}$$

For T_i and T_r, their corresponding embeddings Z_i and Z_r are obtained using the original tokenizer and embedding layers of the LLM.

For the generation of segmentation masks, we followed the architecture used in PSALM [9]. Traditional LLMs generate the next token autoregressively and are designed specifically for text prediction, making them unable to directly generate segmentation masks. To address this issue, we included N mask tokens in the input. After LLM inference, we extracted the hidden states $\hat{H}_m$ corresponding to these mask tokens and decoded them into segmentation masks using a mask decoder.

$$\hat{H}_{others}, \hat{H}_m = LLM(X_c, Z_i, X_r||Z_r, H_m) \tag{3}$$

$$\{m_i, p_i\}_{i=1}^{N} = MaskDecoder(\hat{H}_m, F_c, X_r||Z_r) \tag{4}$$

Here, m_i is the i-th predicted segmentation mask, and p_i is the corresponding probability.

B. Referring Image Packer

For feature extraction and alignment of the referring image, a basic approach is to use a vision encoder and projector shared with the camouflaged image. However, from a usage perspective, the information we focus on differs between the referring image and the camouflaged image. For the camouflaged image, we are interested in global fine-grained information, including scene details and contextual object information. In contrast, for the referring image, we focus solely on referential information, such as object categories and attribute features, which generally constitute only a part of the global fine-grained information. Therefore, for the design of the projector, we designed a dedicated **Referring Image Packer (RIP)** module to extract focused referential information. The RIP uses a q-former module to extract a specific number, N_q, of features, which are then aligned to the LLM space through an MLP layer. This approach allows us to concentrate on extracting referential information, reducing interference from other information, and simultaneously compressing the number of features, thereby improving operational efficiency.

The design of the q-former follows the approach used in BLIP-2 [28]. It utilizes an initialized set of queries $Q_0 = \{q_i\}_{i=1}^{N_q}$ and employs self-attention along with cross-attention with F_r to extract referential information.

$$Q_r = QFormer(Q_0, F_r) \tag{5}$$

$Q_r = \{q_i\}_{i=1}^{N_q}$ represents the extracted referential information, which is then mapped to the LLM space through an MLP layer.

$$X_r = MLP(Q_r) \tag{6}$$

In summary,

$$\begin{aligned} X_r &= RIP(Q_r) \\ &= MLP(QFormer(Q_0, F_r)) \end{aligned} \tag{7}$$

To achieve better alignment between the image and text spaces, we conducted **Triple-Align Pretraining** using a contrastive learning approach before formally applying the RIP module. Inspired by BLIP-2 [28], we designed a Triple-Align loss to measure the degree of alignment among the representation spaces of the three sources: camouflaged images, referring images, and referring texts. As shown in Fig. 3

RefI-RefT Alignment. For aligning referring images with referring texts, we designed three losses to measure the similarity of their features: Image-Text Contrastive Loss (ITC), Image-Text Matching Loss (ITM), and Image-Text Generation Loss (ITG).

$$\begin{aligned} L_{rr}(\{f_I^i\}_{i=1}^{N_i}, \{f_T^j\}_{j=1}^{N_i}) = &\sum_{i,j} c_{ij} \frac{f_I^i \cdot f_T^j}{||f_I^i||\ ||f_T^j||} \\ &+ \sum_{i,j=i} -c_{ij} \log P(f_T^j | f_I^i) \\ &+ \sum_{ij} \mathcal{L}_{CE}(f_{cls}(f_I^i, f_T^j), c_{ij}) \end{aligned} \tag{8}$$

Here, f_I^i represents the features of the referring image extracted by the RIP module, and f_T^j represents the features of the referring text extracted after passing through transformer blocks that share parameters with the RIP module. The first term in the equation is the ITC loss, which measures the similarity between features using cosine similarity. The second term is the ITG loss, adopting a text generation loss form. The third term is the ITM loss, where f_I and f_T pass through a classification head with an MLP layer to obtain category labels, which are then compared with ground truth using a cross-entropy loss. When f_I^i and f_T^j belong to the same class, c_{ij} is 1; otherwise, it is 0.

ComI-RefI Alignment. For aligning camouflaged images with referring images, we designed two losses to measure the similarity between their features: Image-Image Contrastive Loss (IIC) and Image-Image Matching Loss (IIM).

$$\begin{aligned} L_{ci}(\{f_I^i\}_{i=1}^{N_i}, \{f_I^j\}_{j=1}^{N_i}) = &\sum_{i,j} c_{ij} \frac{f_I^i \cdot f_I^j}{||f_I^i||\ ||f_I^j||} \\ &+ \sum_{ij} \mathcal{L}_{CE}(f_{cls}(f_I^i, f_I^j), c_{ij}) \end{aligned} \tag{9}$$

Here, f_I^i and f_I^j represent the features extracted by the RIP module for the camouflaged image and the referring image, respectively.

ComI-RefT Alignment. For aligning camouflaged images with referring texts, we designed three losses to measure the similarity between their features: Image-Image Contrastive Loss (IIC) and Image-Image Matching Loss (IIM).

$$\begin{aligned} L_{ct}(\{f_I^i\}_{i=1}^{N_i}, \{f_T^j\}_{j=1}^{N_i}) = &\sum_{i,j} c_{ij} \frac{f_I^i \cdot f_T^j}{||f_I^i||\ ||f_T^j||} \\ &+ \sum_{i,j=i} -\log P(f_T^j | f_I^i) \\ &+ \sum_{ij} \mathcal{L}_{CE}(f_{cls}(f_I^i, f_T^j), c_{ij}) \end{aligned} \tag{10}$$

Here, f_I^i represents the features of the camouflaged image extracted by the RIP module, and f_T^j represents the features of the referring text extracted after passing through transformer blocks that share parameters with the RIP module.

In summary, the Triple-Align loss can be expressed as:

$$L = L_{ci} + L_{ct} + L_{it} \tag{11}$$

By minimizing the Triple-Align loss, the RIP module undergoes pretraining, thereby achieving better alignment between the image and text spaces.

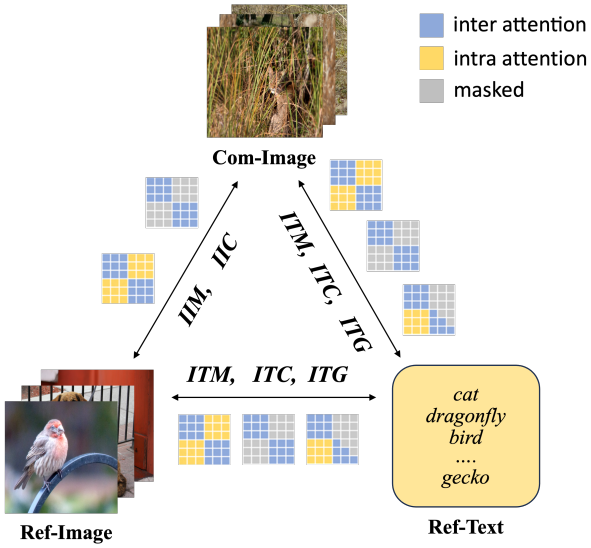

Fig. 3: **Triple Align Loss**. ITM, IIM use inter-source attention to compute the matching score between two sources. ITC, IIC use intra-source attention to avoid information leakage between sources. ITG employs unidirectional attention for text generation.

C. Camouflage CoT

In the ref-COD task, the color of the object to be segmented often closely resembles the surrounding environment or is obscured by other objects in the environment, making segmentation particularly challenging. Therefore, in designing the instructions, we placed particular emphasis on this aspect, highlighting the characteristics of the task and the points that

Models	Overall				Single-obj				Multi-obj			
	S-measure ↑	α E-measure↑	w F-measure ↑	MAE↓	S-measure ↑	α E-measure↑	w F-measure ↑	MAE↓	S-measure ↑	α E-measure↑	w F-measure ↑	MAE↓
PreyNet-RefS [?]	0.817	0.900	0.704	0.032	0.822	0.900	0.709	0.032	0.763	0.898	0.645	0.041
PreyNet-RefT [?]	0.816	0.901	0.705	0.033	0.821	0.900	0.710	0.032	0.759	0.902	0.648	0.041
DGNet-RefS [?]	0.821	0.891	0.696	0.032	0.827	0.890	0.703	0.031	0.748	0.879	0.607	0.045
DGNet-RefT [?]	0.824	0.891	0.701	0.032	0.830	0.892	0.709	0.031	0.745	0.873	0.596	0.046
SINetV2-RefS [?]	0.823	0.888	0.700	0.033	0.828	0.889	0.705	0.032	0.771	0.874	0.634	0.043
SINetV2-RefT [?]	0.822	0.887	0.696	0.033	0.827	0.888	0.702	0.032	0.766	0.866	0.629	0.043
ZoomNet-RefS [?]	0.834	0.886	0.720	0.029	0.839	0.887	0.726	0.029	0.781	0.876	0.652	0.038
ZoomNet-RefT [?]	0.835	0.897	0.725	0.029	0.839	0.897	0.731	0.028	0.783	0.889	0.661	0.038
BSANet-RefS [?]	0.830	0.912	0.727	0.030	0.827	0.913	0.733	0.030	0.774	0.895	0.655	0.039
BSANet-RefT [?]	0.830	0.914	0.730	0.030	0.834	0.915	0.734	0.029	0.784	0.898	0.674	0.036
BGNet-RefS [?]	0.840	0.909	0.738	0.029	0.844	0.910	0.742	0.029	0.792	0.887	0.679	0.036
BGNet-RefT [?]	0.840	0.912	0.739	0.029	0.844	0.914	0.745	0.028	0.791	0.888	0.677	0.038
RPMA-RefS	0.862	0.930	0.784	0.023	0.867	0.934	0.791	0.023	0.806	0.894	0.718	0.033
RPMA-RefT	0.861	0.928	0.783	0.024	0.867	0.931	0.789	0.023	0.802	0.890	0.717	0.034
MLLM-RCOD-RefS(ours)	**0.890**	**0.956**	**0.849**	**0.017**	**0.894**	**0.958**	**0.855**	**0.017**	**0.834**	**0.916**	**0.768**	**0.028**
MLLM-RCOD-RefT(ours)	**0.890**	**0.956**	**0.850**	**0.017**	**0.894**	**0.956**	**0.855**	**0.017**	**0.836**	**0.912**	**0.770**	**0.028**

TABLE I: Comparison with previous state-of-the-art Ref-COD methods. "-RefS": Methods with reference images. "-RefT": Methods with reference text. "Single-obj": Scenes of a single camouflaged object. "Multi-obj": Scenes of multiple camouflaged objects. "Overall": All scenes containing camouflaged objects. "↑": The higher the better. "↓": The lower the better.

require attention for the model. P_i = *"This is an image ... The object may be obscured by other objects in the environment ... so that the object is difficult to see."*

D. Implementation details

We used Swin-B [29] as the vision encoder, pi-1.5-1.3B [30] as the LLM, and the architecture of Mask2Former [31] for the mask generator. The values of N and N_q are 100 and 16, respectively. The vision encoder, LLM, and mask decoder were initialized based on the PSALM [9] model, and the RIP module was initialized using parameters pre-trained with Triple-Align. During the training phase, we fixed the vision encoder and fine-tuned the RIP and mask decoder, while the LLM was fine-tuned using the LORA [32] method. Training was conducted for 3 hours on 8 A100 GPUs. We used AdamW as the optimizer, with cosine learning rate decay and a warm-up phase. The learning rate was set to $1e^{-4}$.

Furthermore, if the model can fully understand the scene content within the image, it will be able to better distinguish between the foreground and background, thereby more effectively differentiating the main object from other objects in the environment, including occluders. This can significantly reduce the difficulty of segmentation and improve the accuracy. Based on this, we guided MLLM-RCOD through three stages:

1) **Scene Perception** aims to guide the model in accurately perceiving the scene of the image, including the environment described by the scene, the objects present, and the relationships between objects, as well as between objects and the environment. For this stage, we designed CoT prompts P_{sp} = "*Describe the scene in the picture. Pay attention to ... objects and the relationship between the environment.*.

2) **Object Understanding** aims to guide the model in accurately comprehending the main objects within the scene, including their categories and attributes such as color and shape. We designed CoT prompts for this stage: P_{sp} = "*Find the main object in the scene ... describe its characteristics.*"

3) **Mask Prediction** involves the model performing segmentation of the hidden targets by integrating its scene awareness and object understanding of the image. We designed CoT prompts for this stage, with P_{sp} = "*Based on your observation of the scene and ... do Referring Segmentation according to the following text/image:*"

Through the three stages of CoT guidance, MLLM-RCOD can better understand the scene content and the characteristics of hidden objects, thereby achieving accurate segmentation of hidden objects.

IV. Experiment

A. Dataset and metrics

We used the R2C7K dataset [2], which consists of Camo and Ref subsets. The Camo subset includes images with hidden targets, covering 64 animal categories, totaling 5,015 images. The Ref subset includes referent images, providing 25 salient images for each animal category, totaling 1,600 referent images. Following R2CNet [2], we employed four evaluation metrics: Structure Measure (S_m), Adaptive Measure (αE), Weighted Measure (ωF), and Mean Absolute Error (MAE).

B. Comparison with sota

We compared the accuracy of our algorithm with existing ref-COD algorithms, and the results are shown in Table I. The table reports results for two referent settings—image referent and text referent—and two segmentation settings—single-object segmentation and multi-object segmentation—as well as their average values. It can be observed that our method significantly outperforms previous methods across all four metrics on the entire test set, achieving a 6% improvement in the weighted measure. Our method delivers excellent results for both image and text referents, and in both single-object and multi-object segmentation, indicating that our method is capable of effectively handling complex camouflaged scenes.

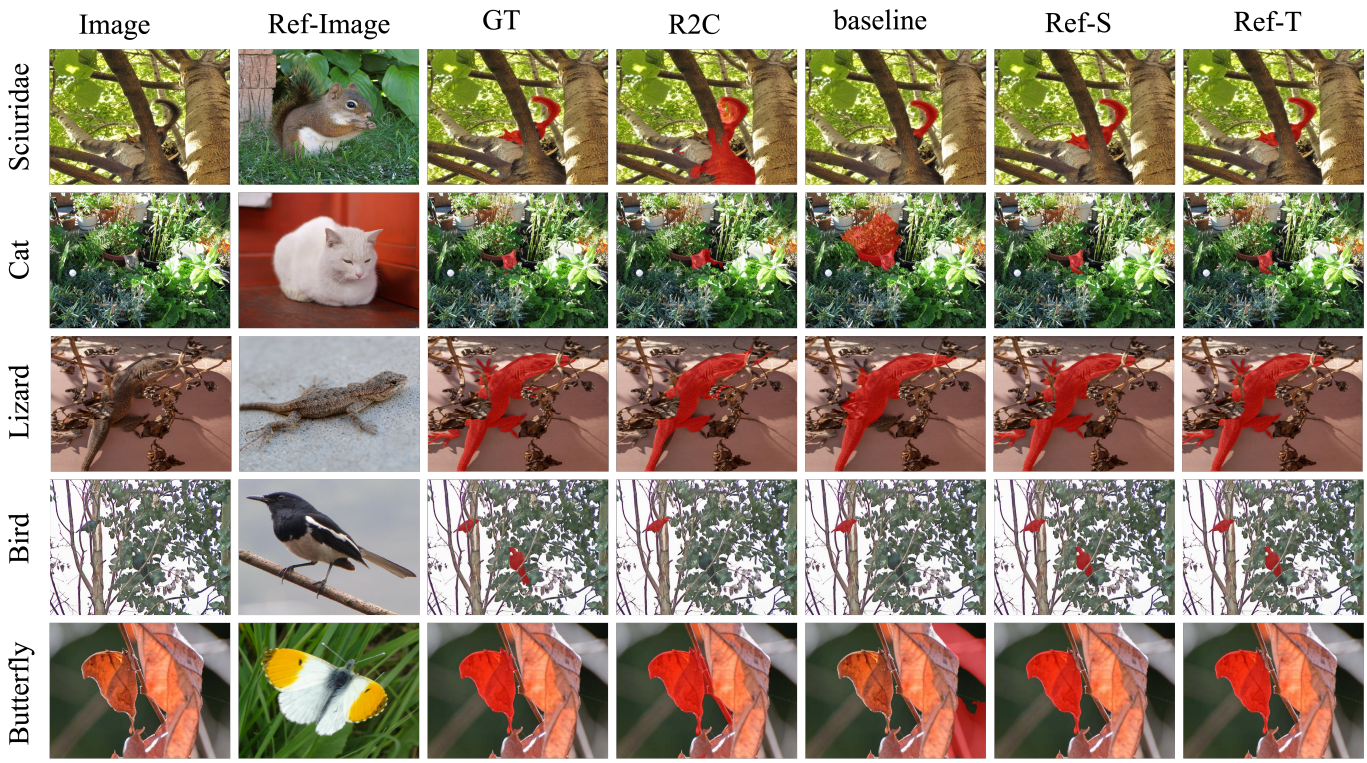

Fig. 4: Qualitative comparison of predictions between our MLLM-RCOD, R2CNet [2], and PSALM [9] as Baseline. '-S': Results with reference images. '-T': Results with reference text. The segmentation masks are shown in red.

C. Ablation study

About RIP. To verify the effectiveness of the RIP module, we removed it and directly used an MLP layer as the projector for the ref-image. We compared this setup with the original model on the image-referring COD, as shown in Table II. It can be seen that the removal of RIP resulted in a performance decrease of 4% of S_m, indicating that the RIP module more effectively extracts the referential information of the reference image.

TABLE II: Ablation about RIP module.

Method	$\mathbf{S}_m \uparrow$	$\alpha\mathbf{E} \uparrow$	$\omega\mathbf{F} \uparrow$	MAE $\downarrow$
MLLM-RCOD	0.890	0.956	0.849	0.017
w/o. RIP	0.853	0.920	0.812	0.024
$N_q = 64$	0.889	0.939	0.833	0.018
$N_q = 8$	0.877	0.945	0.841	0.017

About Triple-Align Pretrain. For the initialization of the RIP module, we compared three settings on image-referring COD: no pretraining, pretraining using only Ref-I and Ref-I alignment, and triple pretraining with Ref-I, Ref-T, and Com-I. As shown in Table III, it can be observed that triple-align pretraining can effectively improve segmentation performance, indicating that triple-align pretrain helps in better extraction of referential information.

About CoT To verify the effectiveness of CoT, we removed the guidance questions related to scene awareness and object understanding from the instructions, and directly instructed the model to segment the camouflaged targets. We compared the performance in text-referring COD, as shown in Table IV. It can be observed that CoT helps the model better understand the image content, thereby achieving better segmentation results.

TABLE III: Ablation about Triple-Align Pretrain.

Method	$\mathbf{S}_m \uparrow$	$\alpha\mathbf{E} \uparrow$	$\omega\mathbf{F} \uparrow$	MAE $\downarrow$
MLLM-RCOD	0.890	0.956	0.849	0.017
w/o. TAP	0.871	0.910	0.822	0.019
RefI-RefT Align	0.886	0.934	0.838	0.017

TABLE IV: Ablation about CoT.

Method	$\mathbf{S}_m \uparrow$	$\alpha\mathbf{E} \uparrow$	$\omega\mathbf{F} \uparrow$	MAE $\downarrow$
MLLM-RCOD	0.890	0.956	0.849	0.017
w/o. CoT	0.879	0.937	0.825	0.018

D. Qualatative results

In Fig. 4, we visualize the segmentation results of our algorithm and qualitatively compare them with our baseline model and the R2CNet method. By effectively utilizing the reference information, our method has achieved significant improvements in the recognition and segmentation accuracy of camouflaged targets. For instance, in the fourth row, while both

the baseline and R2C only segmented one bird, our method successfully segmented two birds. In the fifth row, the baseline failed to recognize the butterfly with a color similar to the leaves, and the R2C method confused the butterfly's body with a branch, whereas our method correctly segmented the butterfly. These results effectively demonstrate the superiority of our method in the Ref-COD task.

REFERENCES

[1] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2777–2787.

[2] X. Zhang, B. Yin, Z. Lin, Q. Hou, D.-P. Fan, and M.-M. Cheng, "Referring camouflaged object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[3] H. Wang, Y. Ye, Y. Wang, Y. Nie, and C. Huang, "Elysium: Exploring object-level perception in videos via mllm," in *European Conference on Computer Vision*. Springer, 2024, pp. 166–185.

[4] C. Ma, Y. Jiang, J. Wu, Z. Yuan, and X. Qi, "Groma: Localized visual tokenization for grounding multimodal large language models," in *European Conference on Computer Vision*. Springer, 2024, pp. 417–435.

[5] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, "Lisa: Reasoning segmentation via large language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9579–9589.

[6] C. Liu, H. Ding, and X. Jiang, "Gres: Generalized referring expression segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 23 592–23 601.

[7] H. Li, J. Chen, Z. Wei, S. Huang, T. Hui, J. Gao, X. Wei, and S. Liu, "Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding," *arXiv preprint arXiv:2501.08282*, 2025.

[8] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization," *Text Reading, and Beyond*, vol. 2, 2023.

[9] Z. Zhang, Y. Ma, E. Zhang, and X. Bai, "Psalm: Pixelwise segmentation with large multi-modal model," in *European Conference on Computer Vision*. Springer, 2024, pp. 74–91.

[10] Z. Huang, H. Dai, T.-Z. Xiang, S. Wang, H.-X. Chen, J. Qin, and H. Xiong, "Feature shrinkage pyramid for camouflaged object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5557–5566.

[11] Z. Wu, D. P. Paudel, D.-P. Fan, J. Wang, S. Wang, C. Demonceaux, R. Timofte, and L. Van Gool, "Source-free depth for object pop-out," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 1032–1042.

[12] L. Wang, J. Yang, Y. Zhang, F. Wang, and F. Zheng, "Depth-aware concealed crop detection in dense agricultural scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 201–17 211.

[13] G.-P. Ji, D.-P. Fan, Y.-C. Chou, D. Dai, A. Liniger, and L. Van Gool, "Deep gradient learning for efficient camouflaged object detection," *Machine Intelligence Research*, vol. 20, no. 1, pp. 92–108, 2023.

[14] C. Xie, C. Xia, T. Yu, and J. Li, "Frequency representation integration for camouflaged object detection," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 1789–1797.

[15] A. Li, J. Zhang, Y. Lv, B. Liu, T. Zhang, and Y. Dai, "Uncertainty-aware joint salient object and camouflaged object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 071–10 081.

[16] Z. Chen, R. Gao, T.-Z. Xiang, and F. Lin, "Diffusion model for camouflaged object detection," in *ECAI 2023*. IOS Press, 2023, pp. 445–452.

[17] X.-J. Luo, S. Wang, Z. Wu, C. Sakaridis, Y. Cheng, D.-P. Fan, and L. Van Gool, "Camdiff: Camouflage image augmentation via diffusion model," *arXiv preprint arXiv:2304.05469*, 2023.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[19] X. Liu, S. Huang, R. Wu, H. Zhao, D. Xu, X. Wei, J. Han, and S. Liu, "Reference prompted model adaptation for referring camouflaged object detection," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[20] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[21] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[22] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[23] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[24] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[25] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[26] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra *et al.*, "Language is not all you need: Aligning perception with language models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 72 096–72 109, 2023.

[27] B. Huang, X. Wang, H. Chen, Z. Song, and W. Zhu, "Vtimellm: Empower llm to grasp video moments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 271–14 280.

[28] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[29] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[30] Y. Li, S. Bubeck, R. Eldan, A. Del Giorno, S. Gunasekar, and Y. T. Lee, "Textbooks are all you need ii: phi-1.5 technical report," *arXiv preprint arXiv:2309.05463*, 2023.

[31] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[32] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.