

MM-CamObj: A Comprehensive Multimodal Dataset for Camouflaged Object Scenarios

Jiacheng Ruan^{1*}, Wenzhen Yuan^{1*}, Zehao Lin², Ning Liao²,
Zhiyu Li^{2†}, Feiyu Xiong², Ting Liu¹, Yuzhuo Fu^{1†}

¹ Shanghai Jiao Tong University, Shanghai, China

² Institute for Advanced Algorithms Research, Shanghai, China
jackchenruan@sjtu.edu.cn

Abstract

Large visual-language models (LVLMs) have achieved great success in multiple applications. However, they still encounter challenges in complex scenes, especially those involving camouflaged objects. This is primarily due to the lack of samples related to camouflaged scenes in the training dataset. To mitigate this issue, we construct the MM-CamObj dataset for the first time, comprising two subsets: CamObj-Align and CamObj-Instruct. Specifically, CamObj-Align contains 11,363 image-text pairs, and it is designed for VL alignment and injecting rich knowledge of camouflaged scenes into LVLMs. CamObj-Instruct is collected for fine-tuning the LVLMs with improved instruction-following capabilities, and it includes 11,363 images and 68,849 conversations with diverse instructions. Based on the MM-CamObj dataset, we propose the CamObj-Llava, an LVLM specifically designed for addressing tasks in camouflaged scenes. To facilitate our model’s effective acquisition of knowledge about camouflaged objects and scenes, we introduce a curriculum learning strategy with six distinct modes. Additionally, we construct the CamObj-Bench to evaluate the existing LVLMs’ capabilities of understanding, recognition, localization and count in camouflage scenes. This benchmark includes 600 images and 7 tasks, with a total of 9,449 questions. Extensive experiments are conducted on the CamObj-Bench with CamObj-Llava, 8 existing open-source and 3 closed-source LVLMs. Surprisingly, the results indicate that our model achieves a 25.84% improvement in 4 out of 7 tasks compared to GPT-4o.

Code — <https://github.com/JCruan519/MM-CamObj>

Introduction

Benefiting from the revolutionary breakthroughs of large language models (LLMs) (Minaee et al. 2024; Touvron et al. 2023b; Zhu et al. 2024; Yang et al. 2024b; Yu et al. 2024) and the continuous advancements in visual foundational models (Li et al. 2024a), large vision-language models (LVLMs) have achieved remarkable progress (Li et al. 2024a). Via training processes such as vision-language

*These authors contributed equally.

†Zhiyu Li and Yuzhuo Fu are co-corresponding authors.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

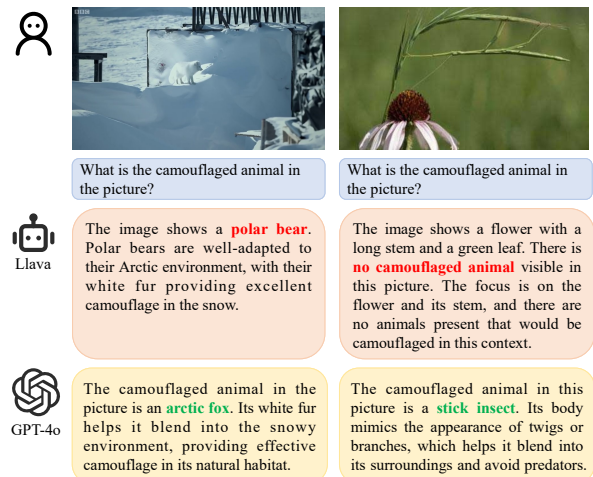


Figure 1: Empirical studies in camouflaged scenarios. We compare Llava-v1.5-7b and GPT-4o in understanding the two classical images with camouflage object existed. There is a certain gap between existing open-source models and advanced closed-source models in camouflage scenes.

alignment (Zhu et al. 2023) and instruction fine-tuning (Lin et al. 2024), LVLMs have demonstrated outstanding performance in various tasks, including image description, visual question answering, object localization, and visual reasoning. These achievements not only showcase the flexibility and adaptability of LVLMs in multimodal scenarios but also drive their widespread applications in fields such as medical imaging (Li et al. 2023), remote sensing (Li et al. 2024b), and autonomous driving (Cui et al. 2024).

Despite the significant progress made by LVLMs in various multimodal applications, they still suffer from significant challenges in complex scenes, particularly those involving camouflaged objects. The primary cause of the difficulty and complexity of tasks in such scenarios lies in the high similarity between the camouflaged objects and their surroundings in terms of texture, shape, and color. As illustrated in Figure 1, our empirical experiments provide a preliminary demonstration of the performance of LVLMs in these sce-

narios. On the left side of Figure 1, Llava misidentifies an Arctic fox as a polar bear, whereas on the right side, Llava fails to recognize a camouflaged stick insect. In contrast, GPT-4o accurately identifies both camouflaged objects. This indicates that existing open-source LVLMs still have shortcomings in understanding scenes with camouflaged objects and lag behind closed-source models. Similar observations can be found in (Jiang et al. 2024b; Ying et al. 2024).

This gap primarily stems from the lack of training samples related to camouflage scenes. To address this issue, we meticulously collect 11,393 images from four classical datasets for camouflage scene understanding (Pang et al. 2023; Fan et al. 2020; Cheng et al. 2022; Yang 2023; Zheng et al. 2018), and leverage the advanced GPT-4o to construct a semantic-rich and feature-diverse dataset called MM-CamObj. This dataset comprises two subsets: CamObj-Align and CamObj-Instruct, which are utilized in the alignment and instruction fine-tuning training phases, respectively. Specifically, CamObj-Align contains 11,393 high-quality image-text pairs, with texts providing detailed descriptions of the camouflage scenes in the images. This subset equips LVLMs with extensive knowledge of camouflage objects and scenarios. CamObj-Instruct consists of 11,393 images with 68,849 diverse conversations, where each image is assigned with at least three high-quality instructions and corresponding responses, and it can further enhance the instruction-following capabilities of LVLMs in camouflage scenarios.

To achieve superior performance in camouflage scene tasks, we propose a novel LVLM, dubbed CamObj-Llava, by further training Llava-v1.5 (Liu et al. 2024a) on our MM-CamObj. However, due to the difficulty of obtaining camouflage images, the training samples in MM-CamObj are relatively few compared to the large-scale datasets used in general scenarios, such as ALLaVA-Instruct-4V (Chen et al. 2024) with 715K samples and Bunny-695K (He et al. 2024). This condition of limited samples poses significant challenges to the model’s training. To address this issue, we introduce a curriculum learning strategy (Soviany et al. 2022). It enhances the model’s capabilities by adopting a learning sequence from easy to difficult, allowing the model to first learn simple samples and then gradually encounter complex ones. In this study, we rank the samples in MM-CamObj according to six different evaluation modes and input them into the LVLM for training in order of increasing difficulty. This strategy enables CamObj-Llava to acquire a more comprehensive understanding of camouflage scenes, improving its performance in complex camouflage scenarios.

To comprehensively evaluate the performance of LVLMs in camouflage scenarios, we introduce the CamObj-Bench, a benchmark specifically designed to assess LVLMs’ performance in multimodal camouflage tasks. CamObj-Bench comprises 600 meticulously curated images, covering 144 species across diverse biological categories, including aquatic animals, mammals, birds, insects, plants, etc., greatly enhancing the richness and diversity of the benchmark. We have established seven core tasks—Easy VQA, Hard VQA, Bbox Location, Image Caption, Count Choice, Mask Match, and Mask TF—to thoroughly evaluate the ca-

pabilities of LVLMs in camouflage scenarios. Each task is rigorously designed to ensure diversity and challenge, posing difficulties even for powerful models like GPT-4o. CamObj-Bench serves not only as a comprehensive tool for evaluating LVLM performance in multimodal camouflage tasks but also provides valuable insights for future advancements in this field.

The main contributions are summarized as follows:

- **MM-CamObj.** We have introduced the MM-CamObj dataset to enhance the existing LVLM’s capabilities in camouflaged object scenarios. This dataset includes two subsets: CamObj-Align and CamObj-Instruct, utilized for the alignment and instruction-tuning stages of LVLM training, respectively.
- **CamObj-Llava.** We have developed CamObj-Llava, an LVLM specifically designed for tasks in camouflaged scenes. By leveraging diverse samples from MM-CamObj dataset and six different curriculum learning methods, CamObj-Llava exhibits outstanding performance across various downstream tasks.
- **CamObj-Bench.** We have established CamObj-Bench, a benchmark aimed at comprehensively and systematically evaluating the performance of LVLMs in camouflaged object scenarios. This benchmark consists of 7 tasks, with a total of 9,449 carefully designed questions.

Related Works

Camouflaged Scene Understanding (CSU) Datasets

In nature, animals use various ingenious camouflage techniques to hide themselves and avoid being detected by predators. Since these camouflage methods are often highly complex, accurately understanding camouflaged scenes and identifying camouflaged objects is a challenging task. Many diverse datasets (Bideau and Learned-Miller 2016; Zheng et al. 2018; Fan et al. 2020; Cheng et al. 2022; Pang et al. 2023; Yang 2023) has propelled the development of the CSU field. For example, COD10K (Fan et al. 2020) is a large-scale dataset for camouflaged object detection that covers 78 categories, aiming at identifying objects that visually blend into the background. PlantCamo (Yang 2023) is an image dataset focused on camouflaged plants, containing images of various camouflaged plants. Each image is accompanied by its corresponding species label and includes mask annotations to display the plant’s contour. CPD1k (Zheng et al. 2018) is a dataset specifically constructed for the task of camouflaged personnel detection. The dataset consists of images extracted from multiple video clips captured in different natural scenes, covering a variety of camouflage patterns and complex lighting and occlusion conditions. MoCA-Mask (Cheng et al. 2022) is a large-scale dataset for video camouflaged object detection. This dataset provides pixel-level manually annotated masks for every five frames of the video to show the contour of camouflaged objects.

Large Vision and Language Models

Large vision language models (LVLMs) are built upon large language models (LLMs) (Touvron et al. 2023a,b; Meta

Source	Num. MM	Num. B
COD-10K (Fan et al. 2020)	4,765	257
MoCA-Mask (Cheng et al. 2022)	4,449	242
PlantCamo (Yang 2023)	1,181	69
CPD-1K (Zheng et al. 2018)	968	32
SUM	11,393	600

Table 1: Image sources for the MM-CamObj dataset (Num. MM) and CamObj-Bench (Num. B).

2024; Chiang et al. 2023; Bai et al. 2023a; Yang et al. 2024a) and visual encoders (Radford et al. 2021; Sun et al. 2023; Oquab et al. 2023; Zhai et al. 2023). Through visual alignment and visual instruction tuning, they effectively fuse image and text information, enhancing performance in multi-modal tasks. For instance, Llava (Liu et al. 2024b,a) introduces a linear projection layer between the vision encoder and the LLM to better align visual-language features. Subsequent works (Bai et al. 2023b; Chu et al. 2024; Wang et al. 2023) have further improved the performance of LVLMs through various architectures. However, the performance of these models in complex scenarios, particularly those involving camouflage objects, remains to be thoroughly explored. Besides, recent studies (Wu et al. 2024; Tang et al. 2023; Jiang et al. 2024b; Ying et al. 2024) have evaluated the performance of LVLMs on the camouflaged object detection task, indicating that current LVLMs exhibit limitations in understanding and recognizing such complex scenes.

Datasets and Benchmark

In our study, the images in the MM-CamObj dataset are sourced from publicly available camouflage scene understanding datasets. These datasets not only provide accurate category annotations but also include segmentation masks for camouflaged objects. Specifically, as shown in Table 1, we carefully select 11,963 camouflaged target images from (Pang et al. 2023; Fan et al. 2020; Cheng et al. 2022; Yang 2023; Zheng et al. 2018). Of these, 11,363 images are used to construct CamObj-Align and CamObj-Instruct, while the remaining 600 images are utilized to construct CamObj-Bench.

Metadata Construction

We define a standardized metadata format for MM-CamObj and CamObj-Bench. Each metadata entry contains the following key information: the name of the source dataset, a unique identifier, the basic category of the camouflaged object, the RGB image, and its corresponding camouflaged object mask. Additionally, the metadata records the number of camouflaged objects, their bounding box information, and their proportion within the entire image. With this information, our constructed training samples can better align with the camouflaged scenes in the given images, allowing our benchmark to incorporate more diverse question types. Through a standardized metadata structure, researchers can more easily compare the capabilities of different models us-

System prompt: "You are an AI vision assistant capable of analyzing complex images. ... You need to describe the image through five different sentences, each offering a unique perspective or detail. ... Your descriptions should enable someone who cannot see the image to accurately and completely imagine its content."

User prompt: "The camouflaged object in the given image is the `{class label}`."

Figure 2: An example prompt snippet for constructing CamObj-Align.

System prompt: "You are an AI visual assistant capable of analyzing complex images. ... Your task is to create a conversation with at least `{randomly choose between 3 and 7}` rounds of Q&A between a "human" and the AI assistant ("gpt") based on a given image. The questions should cover visual content such as object quantity, location, and relative positions, as well as knowledge related to the biological and environmental aspects of the objects. ..."

User prompt: "The camouflaged object in the given image is the `{class label}`."

Figure 3: An example prompt snippet for constructing CamObj-Instruct.

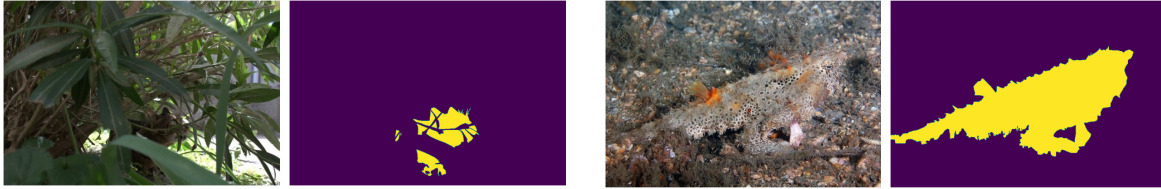
ing a unified evaluation standard, thereby promoting the development of LVLMs in the domain of camouflaged scenes.

CamObj-Align

CamObj-Align dataset consists of 11,363 camouflaged object images along with their corresponding descriptions. To ensure the accuracy of these descriptions, we first use the advanced GPT-4o to generate detailed descriptions of camouflaged objects from five different perspectives, as shown in Figure 2. We then conduct manual reviews and corrections to address errors in the descriptions, such as mistakes in object location and count. Finally, we adopt the instructions from (Liu et al. 2024b) as the detailed image description instructions. An example from CamObj-Align is presented in Figure 4 Left.

CamObj-Instruct

CamObj-Align is utilized during the alignment phase of LVLMs to incorporate knowledge about camouflage objects and scenes into the model. To further enhance the model’s ability to follow instructions, we construct the CamObj-Instruct dataset using 11,363 images and their corresponding camouflage object masks. To ensure diversity and complexity in the instructions, we use GPT-4o to generate dialogues, generating instructions and corresponding responses, as shown in Figure 3. Subsequently, to ensure the accuracy of the dataset, we also conduct manual reviews to correct any hallucinated instructions and responses. Finally, leveraging the information of metadata, we introduce counting and localization instructions and responses in the dialogues to enhance LVM’s ability to perceive camouflage objects. As illustrated in Figure 4 Right, we present an example from CamObj-Instruct dataset.



Question: Illustrate the image through a descriptive explanation.
Answer: The image primarily depicts a dense thicket of green foliage with elongated, pointed leaves, and a network of intertwining branches and stems. Surrounding the feline, the foliage is thick and lush, with several layers of leaves creating a natural cover that conceals the rest of the cat's body effectively. On the left side of the image, the foliage continues with varied leaf sizes and shapes, creating a dense backdrop that adds to the complexity of the scene. In the bottom portion of the image, sunlight breaks through, casting light onto some leaves and highlighting the vibrant green of the foliage, contrasting with the shadowy areas where the cat is hidden.

Question: Please identify the number of camouflaged objects in the image and provide the corresponding bounding box coordinates.
Answer: There is 1 camouflaged target in the image, and the bounding box coordinates are [0.0, 0.245779, 0.87875, 0.834897].
Question:
Answer:
Question: What is the texture of the seafloor where the BatFish is located?
Answer: The seafloor in the image has a rough, granular texture, covered with small rocks, pebbles, and seaweed, providing ample cover for the BatFish.

Figure 4: *Left:* A sample example snippet in CamObj-Align. *Right:* A sample example snippet in CamObj-Instruct. The corresponding mask images are solely for better visualization of the camouflaged objects. These masks are not included in the CamObj-Align and CamObj-Instruct datasets and will not be used during the training process.

CamObj-Bench

Statistical Information To comprehensively evaluate the capabilities of LVLMs in camouflage scene tasks, we propose CamObj-Bench. This benchmark comprises 600 meticulously selected images along with their corresponding camouflage object masks, covering 144 different biological species. As shown in Table 1, these images are sourced from four major public datasets: COD-10K (Fan et al. 2020), MoCA-Mask (Cheng et al. 2022), PlantCamo (Yang 2023), and CPD-1K (Zheng et al. 2018).

CamObj-Bench exhibits significant diversity in the number of camouflaged objects, presenting varying levels of challenges to LVLMs. Specifically, 83.7% of the images contain one camouflaged object, 8.8% contain two objects, 2.3% contain three objects, 1.8% contain four objects, and 3.3% contain more than four camouflaged objects. Additionally, there is notable variation in the size of camouflaged objects within the images. Among these, 69.3% of the objects occupy 0% to 10% of the image area, 18.0% occupy 10% to 20%, 7.5% occupy 20% to 30%, and 5.2% occupy more than 30% of the image area. These variations in relative size further complicate the recognition task.

Tasks Construction We have designed seven tasks in CamObj-Bench based on metadata information. As shown in Figure 5, these tasks include Easy-VQA, Hard-VQA, Bbox Location, Image Caption, Count Choice, Mask Match, and Mask TF. The number of images and questions for each task are shown in Table 2. The first five tasks utilize single-image input, while the last two involve multi-image input.

Easy VQA: Given a camouflaged scene image, the LVLm is required to select the name of the object from four options. The distractor options are generated by GPT-4o and are distinctly different from the correct answer. For example, if the correct answer is ‘Spider’, the distractor options might be ‘Bird’, ‘Frog’, and ‘Fish’. Accuracy is used as the evaluation metric for this task.

Hard VQA: Similar to Easy VQA, but with more challenging distractor options that are closer to the correct an-

Task	#Images	#Questions
Easy-VQA	600	3,000
Hard-VQA	599	2,995
Bbox Location	482	482
Image Caption	600	600
Count Choice	572	572
Mask Match	600	600
Mask TF	600	1,200

Table 2: Image and Question Counts for CamObj-Bench.

swer. For instance, if the correct answer is ‘Spider’, the distractor options might be ‘Water Spider’, ‘Dragonfly’, and ‘Mosquito’, making it more difficult for LVLMs to correctly identify the object. This task also uses accuracy as the evaluation metric.

Bbox Location: Given an image of a camouflaged object, the LVLm needs to detect the location of the object and output it in the format of a bounding box $[x1, y1, x2, y2]$, where $x1, y1$ represent the coordinates of the top-left corner, and $x2, y2$ represent the coordinates of the bottom-right corner. Mean Intersection over Union (mIoU) is used as the evaluation metric for this task.

Image Caption: In this task, the LVLm is required to provide a detailed description of the image based on the camouflaged scene. The ground truth answers are generated by GPT-4o and Gemini-1.5-pro (Reid et al. 2024). Subsequently, we manually inspect and combine the two responses, correcting the hallucinations. We employ BGE-v1.5-en (Xiao et al. 2023) to calculate the CLIPScore between the image and the description, using this as the evaluation metric.

Count Choice: This task also employs a single-choice format, where the LVLm is given an image and must identify the number of camouflaged objects within the image and select the correct answer from four options. Accuracy is used as the evaluation metric for this task.

Mask Match: Given an image of a camouflaged object



Figure 5: Some typical examples from the CamObj-Bench. Our benchmark consists of seven tasks to evaluate LVLMs’ capabilities in recognition, classification, localization, counting, and scene understanding in the camouflaged object scenarios.

and two mask images, where one mask is correct and the other is randomly selected and incorrect, the LVLm is required to select the correct mask image.

Mask TF: Given a camouflaged object image and a mask, the LVLm is required to determine whether the contour in the mask image matches the camouflaged object in the original image, answering either True or False. For each image in CamObj-Bench, we generate one ‘True’ question with the correct mask and one ‘False’ question with an incorrect, randomly selected mask.

Methods

Due to the challenging of acquiring high-quality camouflage images, our CamObj-Align and CamObj-Instruct, contain a limited number of training samples compared to general domain datasets. Therefore, in this section, we introduce a curriculum learning strategy to fully leverage the limited samples, enhancing the learning efficacy and improving the downstream performance of CamObj-Llava.

Determination of Curriculum Learning Metrics

The core idea of curriculum learning is to train models first on simpler samples and gradually increase the complexity samples, thereby improving the model’s learning efficiency and generalization capability (Wang, Chen, and Zhu 2021; Soviany et al. 2022). This strategy simulates the human learning process, which progresses from simple to complex, helping to accelerate model convergence and enhance performance on complex tasks. However, determining the difficulty level of samples is a prerequisite for implementing curriculum learning.

In our CamObj-Align and CamObj-Instruct datasets, each sample consists of an image with the question-answer conversations between a human and an AI assistant. The ‘Question’, posed by a human, can be regarded as an instruction, while the ‘Answer’, provided by the AI assistant, can be considered a response. Consequently, we naturally employ

Mode	Easy VQA	Hard VQA	Bbox Location
-	90.3	68.9	52.8
Len_{ins}	90.9	67.8	52.2
Len_{rsp}	91.3	70.6	52.8
Len_{all}	90.9	69.8	56.7
Sim_{ins}	90.8	<u>70.3</u>	51.8
Sim_{rsp}	91.3	<u>69.1</u>	54.7
Sim_{all}	91.5	69.3	<u>54.9</u>

Table 3: Preliminary studies with various curriculum learning modes. Bold indicates the best, underline indicates the second best.

the sentence lengths of instructions and responses as metrics to assess the difficulty level of the samples (Platanios et al. 2019; Tay et al. 2019). Typically, longer sentence lengths are viewed as indicative of more challenging samples (Wang, Chen, and Zhu 2021). We examine three modes: instruction length (Len_{ins}), response length (Len_{rsp}), and their sum (Len_{all}). Additionally, we introduce image-text similarity as an extra difficulty metric, where higher image-text similarity often implies lower sample difficulty. We consider the similarity between the instruction and image (Sim_{ins}), the response and image (Sim_{rsp}), and the entire dialogue and image (Sim_{all}). We utilize BGE-M3 (Chen et al. 2023) and BGE-v1.5-en (Xiao et al. 2023) to obtain embeddings for the image and text, calculating the cosine similarity between them. Subsequently, we average the similarity scores from these two sets of BGE models to measure the final image-text similarity.

Preliminary Experiment

After determining the difficulty level of the samples, we sort them from easy to difficult and gradually introduce them into the training process according to the principles of curriculum learning. Following the setup of Llava (Liu et al. 2024a), we conduct a two-stage training on the Llava-v1.5-

Model	Size	Easy VQA	Hard VQA	Bbox Location	Image Caption	Count choice
<i>Open-source LVLMs</i>						
Fuyu	8B	35.9	26.1	-	69.1	24.1
Idenfics	9B	79.2	69.1	-	63.5	44.1
InstructBLIP	11B	72.0	55.0	-	68.2	68.4
Kosmos2	1.6B	25.3	25.2	-	70.4	28.8
Llava-v1.5-7B	7B	78.2	59.0	37.4	74.4	70.5
Llava-v1.5-13B	13B	83.4	63.0	40.8	74.7	82.9
MiniCPM-V-2.5	8B	66.1	41.2	33.3	77.1	78.7
Qwen-VL-Chat	7B	77.8	59.7	21.2	74.9	67.3
CamObj-Llava-7B	7B	90.9	69.8	56.7	85.0	81.5
CamObj-Llava-13B	13B	91.8	71.9	59.4	85.7	83.0
<i>Closed-source LVLMs</i>						
GPT-4o	-	88.5	79.9	32.0	81.8	70.8
GPT-4o-mini	-	86.8	72.4	28.1	80.6	58.4
Gemini-1.5-pro	-	95.6	86.4	45.0	79.5	65.3

Table 4: Validation results on single-image tasks in CamObj-Bench. We use accuracy (%) as the evaluation metric for Easy VQA, Hard VQA, and Count choice, mIoU (%) as the evaluation metric for Bbox Location, and CLIPScore (%; extend the range from 0 to 1 into percentage units.) as the evaluation metric for Image Caption.

7b. First, during the alignment stage, we introduce knowledge related to camouflage scenes using the CamObj-Align dataset. Second, in the instruction fine-tuning stage, we enhance the model’s ability to follow instructions with the CamObj-Instruct dataset. Different from Llava, we employ LoRA (Hu et al. 2021) fine-tuning method in both stages.

To validate the effectiveness of curriculum learning, we evaluate its performance on three tasks within the CamObj-Bench: Easy VQA, Hard VQA, and Bbox Location. As shown in Table 3, introducing curriculum learning generally enhances model performance. For instance, in the Len_{all} mode, performance on the Bbox Location task improves by 3.9% compared to the baseline without curriculum learning. However, in the Len_{ins} mode, performance on the Hard VQA and Bbox Location tasks is even lower than without curriculum learning. This outcome aligns with expectations because, similar to (Liu et al. 2024b,a), the human instructions used in the CamObj-Align dataset are randomly selected from a fixed template library, resulting in similar semantic information and sentence length between instructions. After comprehensive consideration, we choose Len_{all} as the training strategy for CamObj-Llava to maximize the utilization of limited training data and achieve better performance in downstream tasks.

Experiments

Baseline

In this section, we conduct a comprehensive evaluation of LVLMs’ performance across the seven tasks of the CamObj-Bench. We test a range of open-source LVLMs, including Fuyu-8B (Bavishi et al. 2023), Idenfics (Laurençon et al. 2024), InstructBLIP-FlanT5-XXL (Dai et al. 2023), Kosmos2 (Peng et al. 2023), Llava-v1.5-7B (Liu et al. 2024a), Llava-v1.5-13B (Liu et al. 2024a), MiniCPM-Llama3-V-2.5 (Yao et al. 2024), and Qwen-VL-Chat (Bai et al. 2023b). In addition, we evaluate several advanced closed-source mod-

els such as GPT-4o¹, GPT-4o-mini², and Gemini-1.5-pro (Reid et al. 2024). The selection of these models aims to provide a comprehensive performance benchmark to analyze the performance of different models across various tasks.

Training Details

In our study, we build the CamObj-Llava-7B/13B by continuously training on the CamObj-Align and CamObj-Instruct datasets, based on the Llava-v1.5-7B/13B (Liu et al. 2024a). The training consists of two stages: alignment and instruction fine-tuning. During the alignment stage, we set the learning rate to $5e-4$, and for the instruction fine-tuning stage, the learning rate is adjusted to $2.5e-4$. Both stages are trained for 1 epoch using the AdamW (Loshchilov and Hutter 2017) optimizer, with a cosine decay strategy (Loshchilov and Hutter 2016) to dynamically adjust the learning rate. Specifically, we apply LoRA (Hu et al. 2021) modules to each linear layer of the large language model, with the rank r and scaling factor α set to 128 and 256. All experiments on conducted on 8xNVIDIA A800 GPUs.

Main Results

Single-image Tasks Table 4 shows the performance of existing LVLMs and our CamObj-Llava on five single-image tasks. Our CamObj-Llava-7B/13B models, trained with CamObj-Align and CamObj-Instruct, outperform all other open-source models in these tasks. Compared to Llava-v1.5-7B/13B, our CamObj-Llava-7B/13B achieves an average improvement of 23.2% and 16.9% across these tasks.

CamObj-Llava excels in camouflaged object recognition. In the Easy and Hard VQA tasks, we evaluate the ability of LVLMs to recognize and classify camouflaged objects. Although our model underperforms compared to the closed-source models in the Hard VQA task, CamObj-Llava-7B

¹<https://platform.openai.com/docs/models/gpt-4o>

²<https://platform.openai.com/docs/models/gpt-4o-mini>

Model	Mode	Mask Match	Mask TF
<i>Open-source LLMs</i>			
Fuyu	merge	51.8	50.4
Idefics	sequence	43.1	47.4
InstructBLIP	merge	43.2	47.4
Kosmos2	merge	50.0	50.0
Llava-v1.5-7B	merge	50.8	50.0
Llava-v1.5-13B	merge	45.7	47.0
MiniCPM-V-2.5	sequence	51.7	48.3
Qwen-VL-Chat	merge	50.0	50.0
CamObj-Llava-7B	merge	50.0	50.0
CamObj-Llava-13B	merge	50.0	49.5
<i>Closed-source LLMs</i>			
GPT-4o	sequence	95.8	88.3
GPT-4o-mini	sequence	84.8	69.0
Gemini-1.5-pro	sequence	87.8	72.0

Table 5: Validation results on multiple-images tasks in CamObj-Bench. We use accuracy (%) as the evaluation metric for Mask Match and Mask TF.

surpasses GPT-4o by 2.4% in the Easy VQA task, with CamObj-Llava-13B further improving upon this by 0.9%.

CamObj-Llava surpasses closed-source LLMs in object localization, scene understanding, and counting. In the Bbox Location task, which primarily evaluates the localization ability of LLMs for camouflaged objects, our CamObj-Llava-7B/13B outperforms all advanced closed-source models. Compared to the top-performing Gemini-1.5-pro, CamObj-Llava-13B achieves a 14.4% improvement in the mIoU metric. The Image Caption task focuses on assessing the model’s understanding of camouflaged scenes. In this task, CamObj-Llava-7B/13B also outperforms leading closed-source models. Specifically, CamObj-Llava-7B, with only 7B parameters, surpasses GPT-4o by 3.2% in CLIP-Score. The Count choice task evaluates the counting abilities of LLMs, and once again, our CamObj-Llava exceeds all closed-source models.

The results across these five single-image tasks demonstrate that our CamObj-Align and CamObj-Instruct datasets provide comprehensive knowledge of camouflaged objects and scenes, significantly enhancing LLMs’ capabilities in recognition, classification, localization, counting, and scene understanding.

Multiple-images tasks In handling multiple-images tasks, we draw inspiration from previous research (Jiang et al. 2024a) and employ two different processing methods, depending on whether the LLM supports multiple-images input. The first strategy, termed ‘merge’, involves horizontally concatenating multiple images into a single image for processing. The second strategy, termed ‘sequence’, involves inputting multiple images in a sequential manner.

Table 5 shows the performance of existing LLMs and our CamObj-Llava on the Mask Match and Mask TF tasks. For open-source LLMs, including our model, the performance on these multiple-images tasks is close to random selection. This phenomenon could also be observed in (Fu

Ratio	Easy VQA	Hard VQA	Bbox Location
-	78.2	59.0	37.4
10%	84.6	61.7	42.7
20%	86.0	63.9	44.2
50%	87.8	65.5	49.4
100%	90.9	69.8	56.7

Table 6: Ablation studies on the Easy/Hard VQA and Bbox Location tasks with various training samples. Ratio refers to the proportion of training samples randomly selected from CamObj-Align and CamObj-Instruct datasets. ‘-’ stands for Llava-v1.5-7B without training on our datasets.

et al. 2024; Jiang et al. 2024a). One possible reason is that during training, only a few or no multiple-images input modes and related instructions are used, resulting in poor performance on multiple-images tasks. In contrast, advanced closed-source models have better adaptability to multiple-images tasks. However, even the advanced GPT-4o-mini achieves only 69.0% accuracy on the Mask TF task, only 19.0% higher than random selection. This further indicates that our CamObj-Bench presents new challenges for existing LLMs.

Ablation Studies

To validate the performance of our CamObj-Llava-7B with limited data, we randomly selected 10%, 20%, and 50% of the samples from the CamObj-Align and CamObj-Instruct datasets for training, while keeping the rest of the experimental settings consistent with those described in Sec. ‘Training Details’. The results are shown in Table 6. Even with only 10% of the data samples, CamObj-Llava-7B surpasses Llava-v1.5-7B in the Easy VQA, Hard VQA, and Bbox Location tasks, with improvements of 6.4%, 2.7%, and 5.3%, respectively. This demonstrates the significant advantage of our training dataset in providing LLMs with rich knowledge about camouflage scenarios.

Conclusion

In this paper, we aim to enhance the capabilities of large vision-language models in complex scenarios, particularly those involving camouflaged objects. To achieve this, we have collected and constructed the MM-CamObj dataset, which includes CamObj-Align for alignment training and CamObj-Instruct for instruction fine-tuning. Additionally, to effectively utilize the limited training samples, we introduce six different modes of curriculum learning and continue training based on Llava-v1.5, resulting in our CamObj-Llava. To comprehensively validate the performance of existing LLMs and our model in camouflaged scenarios, we develop CamObj-Bench, a benchmark consisting of seven challenging tasks. Extensive experiments demonstrate that our CamObj-Llava surpasses the advanced GPT-4o in four of these tasks. In the future, we plan to expand the dataset and introduce multi-image training for our CamObj-Llava.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61977045).

References

- Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.
- Bavishi, R.; Elsen, E.; Hawthorne, C.; Nye, M.; Odena, A.; Somani, A.; and Taşlılar, S. 2023. Fuyu-8B: A multimodal architecture for AI agents.
- Bideau, P.; and Learned-Miller, E. 2016. It's moving! a probabilistic model for causal motion segmentation in moving camera videos. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, 433–449. Springer.
- Chen, G. H.; Chen, S.; Zhang, R.; Chen, J.; Wu, X.; Zhang, Z.; Chen, Z.; Li, J.; Wan, X.; and Wang, B. 2024. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*.
- Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; and Liu, Z. 2023. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. *arXiv:2309.07597*.
- Cheng, X.; Xiong, H.; Fan, D.-P.; Zhong, Y.; Harandi, M.; Drummond, T.; and Ge, Z. 2022. Implicit motion handling for video camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13864–13873.
- Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3): 6.
- Chu, X.; Qiao, L.; Zhang, X.; Xu, S.; Wei, F.; Yang, Y.; Sun, X.; Hu, Y.; Lin, X.; Zhang, B.; et al. 2024. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*.
- Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Zhou, Y.; Liang, K.; Chen, J.; Lu, J.; Yang, Z.; Liao, K.-D.; et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 958–979.
- Dai, W.; Li, J.; LI, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 49250–49267. Curran Associates, Inc.
- Fan, D.-P.; Ji, G.-P.; Sun, G.; Cheng, M.-M.; Shen, J.; and Shao, L. 2020. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2777–2787.
- Fu, X.; Hu, Y.; Li, B.; Feng, Y.; Wang, H.; Lin, X.; Roth, D.; Smith, N. A.; Ma, W.-C.; and Krishna, R. 2024. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*.
- He, M.; Liu, Y.; Wu, B.; Yuan, J.; Wang, Y.; Huang, T.; and Zhao, B. 2024. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*.
- Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.
- Jiang, D.; He, X.; Zeng, H.; Wei, C.; Ku, M.; Liu, Q.; and Chen, W. 2024a. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.
- Jiang, Y.; Yan, X.; Ji, G.-P.; Fu, K.; Sun, M.; Xiong, H.; Fan, D.-P.; and Khan, F. S. 2024b. Effectiveness assessment of recent large vision-language models. *arXiv preprint arXiv:2403.04306*.
- Laurençon, H.; Saulnier, L.; Tronchon, L.; Bekman, S.; Singh, A.; Lozhkov, A.; Wang, T.; Karamcheti, S.; Rush, A.; Kiela, D.; et al. 2024. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36.
- Li, C.; Gan, Z.; Yang, Z.; Yang, J.; Li, L.; Wang, L.; Gao, J.; et al. 2024a. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2): 1–214.
- Li, X.; Wen, C.; Hu, Y.; Yuan, Z.; and Zhu, X. X. 2024b. Vision-language models in remote sensing: Current progress and future trends. *IEEE Geoscience and Remote Sensing Magazine*.
- Li, X.; Zhang, L.; Wu, Z.; Liu, Z.; Zhao, L.; Yuan, Y.; Liu, J.; Li, G.; Zhu, D.; Yan, P.; et al. 2023. Artificial general intelligence for medical imaging. *arXiv preprint arXiv:2306.05480*.
- Lin, J.; Yin, H.; Ping, W.; Molchanov, P.; Shoeybi, M.; and Han, S. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26689–26699.
- Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.
- Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

- Meta, A. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.
- Minaee, S.; Mikolov, T.; Nikzad, N.; Chenaghlu, M.; Socher, R.; Amatriain, X.; and Gao, J. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.
- Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.
- Pang, Y.; Zhao, X.; Zuo, J.; Zhang, L.; and Lu, H. 2023. Open-Vocabulary Camouflaged Object Segmentation. *arXiv preprint arXiv:2311.11241*.
- Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.
- Platanios, E. A.; Stretcu, O.; Neubig, G.; Poczos, B.; and Mitchell, T. M. 2019. Competence-based curriculum learning for neural machine translation. *arXiv preprint arXiv:1903.09848*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lillcrap, T.; Alayrac, J.-b.; Soricut, R.; Lazaridou, A.; Firat, O.; Schrittwieser, J.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Soviany, P.; Ionescu, R. T.; Rota, P.; and Sebe, N. 2022. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6): 1526–1565.
- Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; and Cao, Y. 2023. Evalclip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.
- Tang, L.; Jiang, P.-T.; Shen, Z.; Zhang, H.; Chen, J.; and Li, B. 2023. Generalization and Hallucination of Large Vision-Language Models through a Camouflaged Lens. *arXiv preprint arXiv:2311.11273*.
- Tay, Y.; Wang, S.; Tuan, L. A.; Fu, J.; Phan, M. C.; Yuan, X.; Rao, J.; Hui, S. C.; and Zhang, A. 2019. Simple and effective curriculum pointer-generator networks for reading comprehension over long narratives. *arXiv preprint arXiv:1905.10847*.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.
- Wang, X.; Chen, Y.; and Zhu, W. 2021. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 4555–4576.
- Wu, J.; Zhong, M.; Xing, S.; Lai, Z.; Liu, Z.; Wang, W.; Chen, Z.; Zhu, X.; Lu, L.; Lu, T.; et al. 2024. Vision-LLM v2: An End-to-End Generalist Multimodal Large Language Model for Hundreds of Vision-Language Tasks. *arXiv preprint arXiv:2406.08394*.
- Xiao, S.; Liu, Z.; Zhang, P.; and Muennighof, N. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.
- Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. 2024a. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.
- Yang, H.; Lin, Z.; Wang, W.; Wu, H.; Li, Z.; Tang, B.; Wei, W.; Wang, J.; Tang, Z.; Song, S.; et al. 2024b. Memory³: Language Modeling with Explicit Memory. *arXiv preprint arXiv:2407.01178*.
- Yang, J. 2023. Plantcamo dataset. Accessed: 2024-07-20.
- Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800*.
- Ying, K.; Meng, F.; Wang, J.; Li, Z.; Lin, H.; Yang, Y.; Zhang, H.; Zhang, W.; Lin, Y.; Liu, S.; et al. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*.
- Yu, X.; Wang, Y.; Chen, Y.; Tao, Z.; Xi, D.; Song, S.; and Niu, S. 2024. Fake Artificial Intelligence Generated Contents (FAIGC): A Survey of Theories, Detection Methods, and Opportunities. *arXiv preprint arXiv:2405.00711*.
- Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11975–11986.
- Zheng, Y.; Zhang, X.; Wang, F.; Cao, T.; Sun, M.; and Wang, X. 2018. Detection of people with camouflage pattern via dense deconvolution network. *IEEE Signal Processing Letters*, 26(1): 29–33.
- Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.
- Zhu, T.; Qu, X.; Dong, D.; Ruan, J.; Tong, J.; He, C.; and Cheng, Y. 2024. Llama-moe: Building mixture-of-experts from llama with continual pre-training. *arXiv preprint arXiv:2406.16554*.