🖍️ CoLLaVO: Crayon Large Language and Vision mOdel

Byung-Kwan Lee
KAIST
leebk@kaist.ac.kr

Beomchan Park
KAIST
bpark0810@kaist.ac.kr

Chae Won Kim
KAIST
chaewonkim@kaist.ac.kr

Yong Man Ro[*]
KAIST
ymro@kaist.ac.kr

Abstract

The remarkable success of Large Language Models (LLMs) and instruction tuning drives the evolution of Vision Language Models (VLMs) towards a versatile general-purpose model. Yet, it remains unexplored whether current VLMs genuinely possess quality object-level image understanding capabilities determined from 'what objects are in the image?' or 'which object corresponds to a specified bounding box?'. Our findings reveal that the image understanding capabilities of current VLMs are strongly correlated with their zero-shot performance on vision language (VL) tasks. This suggests that prioritizing basic image understanding is crucial for VLMs to excel at VL tasks. To enhance object-level image understanding, we propose **Cray**on **L**arge **L**anguage **a**nd **V**ision m**O**del (🖍️ **CoLLaVO**), which incorporates instruction tuning with *Crayon Prompt* as a new visual prompt tuning scheme based on panoptic color maps. Furthermore, we present a learning strategy of *Dual QLoRA* to preserve object-level image understanding without forgetting it during visual instruction tuning, thereby achieving a significant leap in numerous VL benchmarks in a zero-shot setting. Code is available in https://github.com/ByungKwanLee/CoLLaVO.

1 Introduction

Spurred by the enduring ambition for artificial general intelligence (AGI) and the success of language models such as BERT (Devlin et al., 2018), GPT-3 (Brown et al., 2020), and LLaMA (Touvron et al., 2023a), there has been a surge in demand for a general-purpose model in a task-unified format via natural language instruction, leading to the emergence of instruction tuning (Wei et al., 2022; Chung et al., 2022). Building on the success of Large Language Models (LLMs) and instruction tuning, InstructBLIP (Dai et al., 2023),
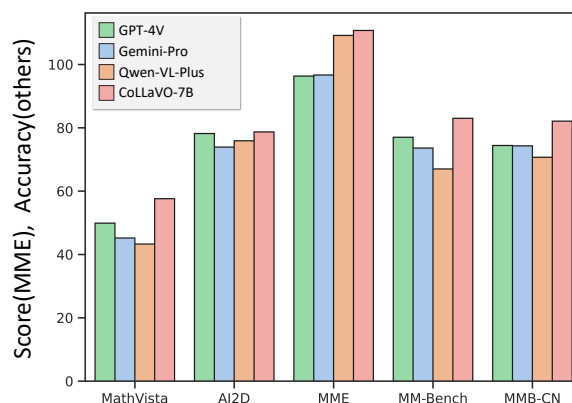

Figure 1: Zero-shot performance of 🖍️ CoLLaVO-7B on challenging VL datasets compared with closed-source VLMs (OpenAI, 2023a,b; Team et al., 2023; Bai et al., 2023). Note: The scores of MME are rescaled by $1/20$ to match the scales with the accuracies of others.

LLaVA1.5 (Liu et al., 2023c,b), and Qwen-VL (Bai et al., 2023) have either directly designed or utilized visual instruction tuning datasets for a wide range of vision language (VL) tasks using natural language instructions. Consequently, they have become paradigm-shifting in Vision Language Models (VLMs), showcasing remarkable zero-shot performance in VL tasks.

However, it is yet uncharted whether the current leading VLMs truly possess a comprehensive understanding of fine-grained object information, and how this understanding influences their zero-shot performance in VL tasks related to each object. Hence, we delve into the analysis of object-level image understanding and zero-shot performance in VL tasks across different objects. To illustrate the behavior of object-level image understanding, we employ four strong baselines: BLIP2 (Li et al., 2023c), InstructBLIP (Dai et al., 2023), LLaVA1.5 (Liu et al., 2023b), and Qwen-VL (Bai et al., 2023). We pose two types of simple questions to gauge their object-level understanding such as: (1) *'Is there any {object name} in this image?' (Class2Binary: C2B)*, and (2) *'Which object is in the specified bounding box [x_{min}, y_{min}, x_{max}, y_{max}]?' (Box2Class: B2C)*. We then evaluate the
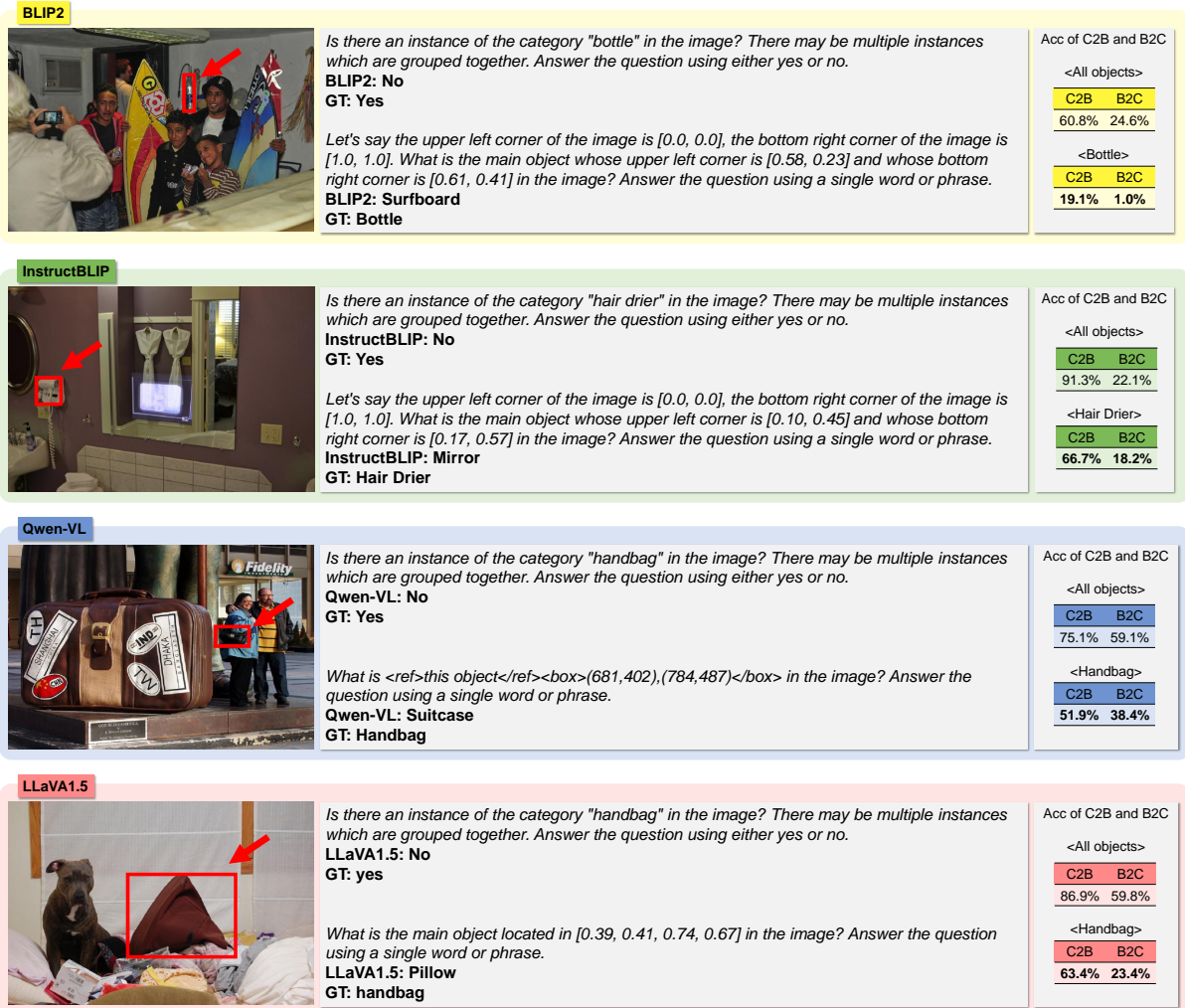
---

[*]Corresponding author.

1121

Figure 2: Asking four baselines (BLIP2, InstructBLIP, Qwen-VL, and LLaVA1.5) two types of questions, Class2Binary (C2B) and Box2Class (B2C), and measuring their accuracies on each object category.

accuracy of their responses for 80 object categories (See Section 4 for more details) while assessing their zero-shot performance on VL tasks across the same set of categories.

Following this assessment, Figure 2 illustrates that four strong baselines typically exhibit poor performance on object-level image understanding for several object categories with C2B and B2C accuracies lower than average. This phenomenon arises from various factors, such as biases in co-occurring objects or object size. In Figure 3, we observe a strong correlation between the level of object-level image understanding exhibited by VLMs and their subsequent zero-shot performance. This trend appears consistent across all four baseline VLMs. Consequently, enhancing the object-level image understanding capabilities of VLMs is expected to significantly improve their zero-shot performance in VL tasks.

To improve object-level image understanding, we introduce a new visual prompt called *Crayon Prompt* to assist VLMs in focusing more efficiently on objects. The Crayon Prompt starts from a panoptic segmentation model (Cheng et al., 2022) that generates a panoptic color map for any given image. This map contains semantic information for objects and their numbering. Leveraging this information, we replace both aspects with learnable queries representing semantic and numbering embeddings, correctly termed as the *Crayon Prompt*.

This simple yet effective idea is inspired by the practice of drawing red circles on images (Shtedritski et al., 2023), aiming to direct attention to a specific area. They note that red circles potentially invoke the object-level image understanding of VLMs. However, they may distort image contents, posing a risk to VL tasks, and cannot consider foreground and background objects simultaneously. Instead, the Crayon Prompt encompasses all foreground and background objects simultaneously, thanks to a panoptic color map. Unlike drawing a visual prompt directly on an image, we integrate the Crayon Prompt into image embedding features at every attention module layer in the backbone Multi-
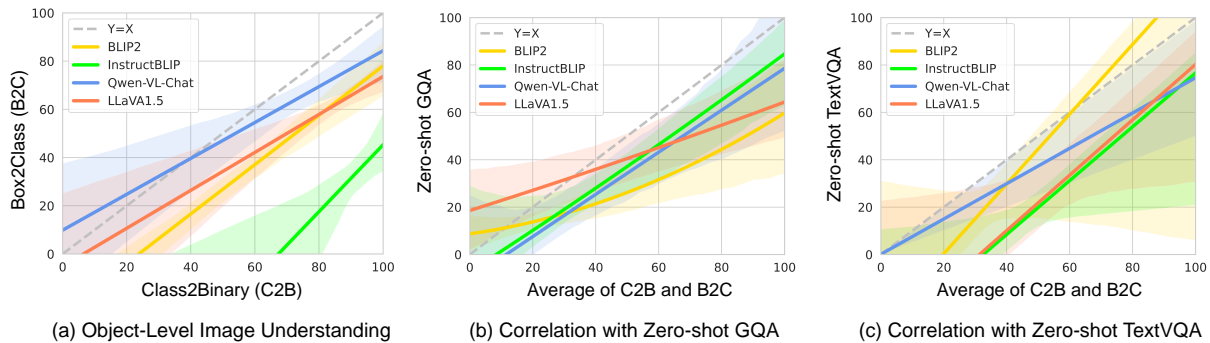
Figure 3: Plotting the regressed relationships between (a) C2B and B2C for each object category, (b) the average of C2B & B2C and zero-shot GQA (Hudson and Manning, 2019) performance for each object category, (c) the average of C2B & B2C and zero-shot TextVQA (Singh et al., 2019) performance for each object category to visualize their correlations. The light-colored areas indicate the vertical span with the probability of confidence interval 0.95.

modal Language Model (MLM) of 📚 CoLLaVO, thereby keeping the raw visual context of the image intact. The Crayon Prompt imparts semantic information about objects and their numbering, akin to how positional embedding (Vaswani et al., 2017) assigns sequential information to token embedding features.

By employing the Crayon Prompt, we create simple crayon instructions to enhance object-level image understanding. Additionally, we utilize the visual instruction tuning datasets (Liu et al., 2023c,b; Chen et al., 2023d) for zero-shot VL tasks. However, conducting visual instruction tuning only may be not sure for the grasp of object-level image understanding. Hence, we propose a learning strategy called *Dual QLoRA* involving two QLoRA (Dettmers et al., 2023) modules. One module is trained for crayon instructions while the other module for visual instruction tuning datasets is frozen, and vice versa. This approach enables efficient fusion of crayon instructions and visual instruction tuning datasets while preserving the capabilities of both object-level image understanding and complex question answering. Pursuing parameter-efficient training, we employ quantized LoRA (QLoRA) instead of LoRA (Hu et al., 2021).

Following the aforementioned methods, we propose a new large language and vision model called **Cray**o**n L**arge **L**anguage **a**nd **V**ision m**O**del (📚 **CoLLaVO**), where the Crayon Prompt and a VLM collaborate to enhance object-level image understanding, which subsequently affects zero-shot VL performance. Our contribution can be summarized as follows:

- To the best of our knowledge, we first reveal the intriguing property of current VLMs, wherein object-level image understanding is strongly correlated with zero-shot VL tasks.

- We propose the *Crayon Prompt* and *Dual QLoRA*, which enhance object-level image understanding and effectively maintain it alongside complex VL performance, respectively.

- By applying all these ingredients, we present an efficient model, 📚 CoLLaVO-7B, which significantly achieves state-of-the-art zero-shot VL performance compared to closed-source VLMs and open-source VLMs.

2 Research Backgrounds

Visual Prompting. Researchers have prioritized enhancing natural language prompts in constructing instruction tuning datasets for LLMs (Wei et al., 2022; Chung et al., 2022; Touvron et al., 2023b). On the other hand, dealing with VLMs offers new opportunities to manipulate both visual and textual aspects of prompts. Earlier studies on visual prompting have focused on techniques such as learnable token embedding concatenated with visual embedding (Jia et al., 2022; Sandler et al., 2022), or learned perturbation patterns directly applied to an input image (Bahng et al., 2022; Chen et al., 2023a; Oh et al., 2023). While these methods aim to find the optimal visual prompt, the learned visual prompts lack human interpretability, hindering the understanding of their effectiveness.

To address this, current VLMs use human-interpretable visual prompts such as marks (Shtedritski et al., 2023; Yang et al., 2023b; Cai et al., 2023) or semantic masks (Yang et al., 2023a). Shtedritski et al. (2023) draw red circles on images and then demonstrate that CLIP (Radford et al., 2021), by itself, can recognize the simple visual prompts on images, showing improved zero-shot performance for tasks such as referring expressions comprehension and key point localization. By using SEEM (Zou et al., 2023) or SAM (Kirillov
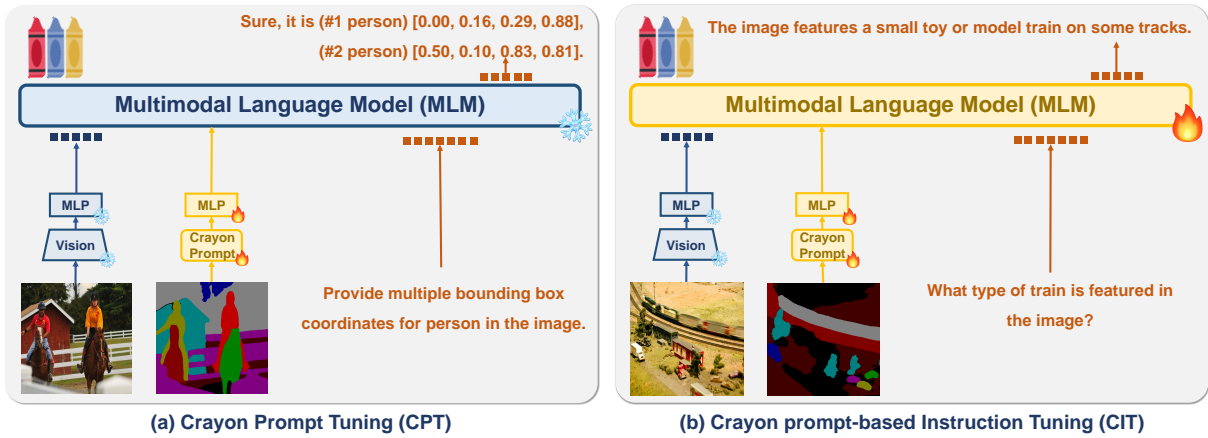
(a) Crayon Prompt Tuning (CPT)

(b) Crayon prompt-based Instruction Tuning (CIT)

Figure 4: Overview of two-step training for 🖍️ CoLLaVO. Note that 'Vision' represents vision encoder, and that the fire symbols represent the modules to learn.

et al., 2023), Yang et al. (2023a) employs special marks including alphanumerics and masks to help VLMs understand fine-grained spatial information. Yang et al. (2023b) uses semantic masks created by an object detection model and SAM, along with visual prompts like contour masks, colorful masks, grayscale reverse masks, and blur reverse masks, to enhance local attention in CLIP.

In brief, previous studies have focused on guiding VLMs towards specific areas using marks and semantic masks. Similar to Yang et al. (2023a), we propose *Crayon Prompt* encompassing all foreground and background objects at once. However, compared with a direct visual prompt on the image (Liu et al., 2023e; Shtedritski et al., 2023; Yang et al., 2023b; Cai et al., 2023; Yang et al., 2023a), the Crayon Prompt is injected into image embedding features at every Transformer (Vaswani et al., 2017) layer in a backbone MLM to keep the image intact and not disrupt its raw visual context. The Crayon Prompt provides semantic information about objects in the image and their numbering, similar to how positional embedding (Vaswani et al., 2017) provides sequential information about the relative orders of token embedding features.

LLMs, VLMs, and Instruction Tuning. Flan (Wei et al., 2022) pioneered the development of instruction tuning by consolidating 62 language datasets, covering a diverse range of tasks. It demonstrates significant improvements in zero-shot performance. In efforts to expand the scope of tasks and the capacity of language models, Chung et al. (2022) introduced Flan-PaLM and Flan-T5, leveraging PaLM (Chowdhery et al., 2023) and T5 (Raffel et al., 2020). Continuing along the trajectory of instruction-tuned LLMs, LLaVA (Liu et al., 2023c) utilizes a language-only

GPT-4 to produce visual dialogues, intricate deductions, and detailed image descriptions for the LLaVA-Instruct-665K dataset. Simultaneously, various VLMs (Dai et al., 2023; Ye et al., 2023a; Li et al., 2023a; Zhu et al., 2023; Chen et al., 2023c; Bai et al., 2023) have developed unique instruction tuning datasets to enhance grounding capability and mitigate hallucinations.

Amidst the current surge of VLMs, we approach them from a fresh angle, notwithstanding the strides made in instruction tuning. Consequently, our focus shifts towards probing whether VLMs effectively grasp object-level image understanding. Should they fall short, we then question whether this inadequacy correlates with their VL performance. In essence, Figure 2-3 emphasize the importance of foundational image understanding and its potential impact on VL performance, in other words, a facet often overlooked in previous studies. Thus, we advocate for a fusion of object-level image understanding and visual instruction tuning.

3 🖍️ CoLLaVO

Model Architecture and Prompt Protocol. The structure of 🖍️ CoLLaVO, as illustrated in Figure 4, comprises a vision encoder, Crayon Prompt, a backbone MLM, and MLP connectors between the vision and language components. CLIP (Radford et al., 2021) is considered as the vision encoder, benefiting from its adeptness in image understanding. The MLM utilized in 🖍️ CoLLaVO is from InternLM-7B (Team, 2023), which is a multilingual foundation model instruction tuned by 1.6T multilingual datasets with RLHF (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022). Moreover, two fully-connected MLPs with GELU activa-
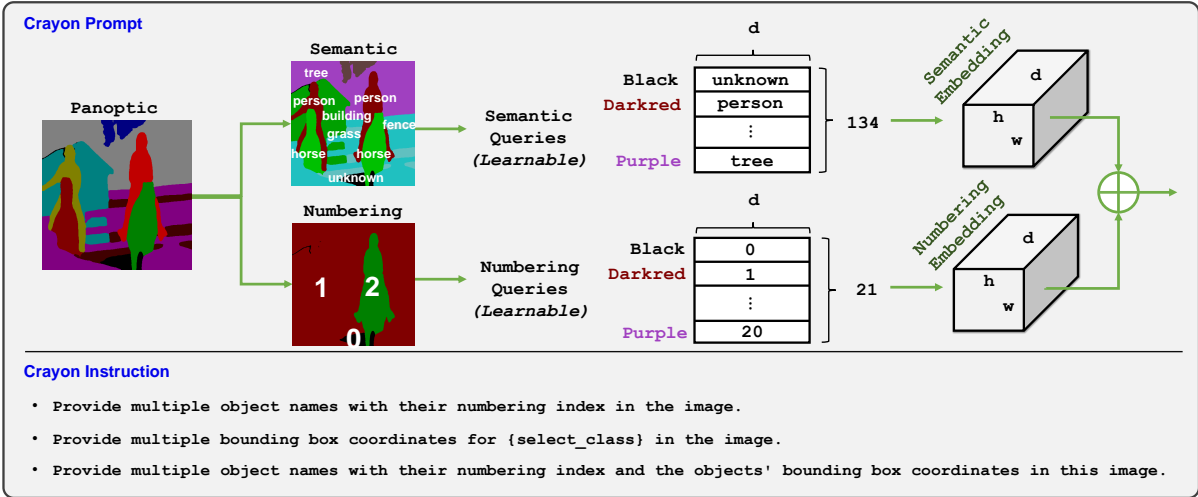
Figure 5: Describing how the Crayon Prompt is generated from a panoptic color map with learnable semantic queries and numbering queries. In addition, crayon instruction examples are given, which are used to conduct CPT and CIT. Note that, '{}' denotes the place where we adaptively input information.

tion function (Hendrycks and Gimpel, 2016) serve as the bridge connector. Regarding 🖍️CoLLaVO input, adherence to a prompt protocol is maintained, where '<image>' signifies a special token for image embedding features, '<stop>' denotes a stop token for text generation, 'User: {}' represents a question template, and 'Assistant: {}' indicates an answer template (See below Figure 5 for an example).

Crayon Prompt Tuning (CPT). To ensure a comprehensive object-level grasp on the entire image, 🖍️CoLLaVO should recognize all distinct objects within it, including both foreground (*e.g.,* person, bus, bottle, hairdryer, and handbag) and background (*e.g.,* sky, road, river, sea, and snow) objects. To achieve this, we employ a panoptic segmentation model (Cheng et al., 2022), which generates a panoptic color map as illustrated in Figure 4(a)-(b). This map enables the discrimination of 133 different object categories (See Appendix A) of foreground and background objects from MS-COCO 2017 (Lin et al., 2014), serving as a visual cue for 🖍️CoLLaVO to focus on all objects within the image.

Notably, the panoptic map contains two crucial pieces of information for each object: semantic information and numbering information. For instance, if an image depicts two people riding horses, as illustrated in Figure 4(a), the panoptic map assigns each object a category label and a numbering index, as shown in Figure 5. The two people receive different numbering indices '1' and '2' but share the same object category 'person'. Other objects, being singular, are all labeled with the numbering index '1'. It is worth noting that the unknown category is assigned the numbering in-

---

Crayon Instruction Example for CPT

User: Provide multiple object names with their numbering index
and the objects' bounding box coordinates in this image.
Assistant: Sure, it is (#1 horse) [0.06, 0.38, 0.27, 0.91],
(#2 horse) [0.56, 0.41, 0.77 , 0.99], (#1 person) [0.00, 0.16, 0.29, 0.88],
(#1 sky) [0.39, 0.00, 0.69, 0.18], (#1 fence) [0.00, 0.42, 0.99, 0.99],
(#1 building) [0.00, 0.14, 0.47, 0.83], (#1 tree) [0.00, 0.00, 0.99, 0.47],
(#1 grass) [0.21, 0.49, 0.99, 0.95], (#2 person) [0.50, 0.10, 0.83, 0.81]<stop>

---

dex '0'. To streamline the next process, we prepare 133+1(**unk**) learnable semantic queries, including the aforementioned 133 categories and an **unk**nown category. In addition, we prepare 20+1('0' for **unk**) learnable numbering queries under the assumption that no more than 20 instances of the same object category appear within one image.

Leveraging 134 semantic queries and 21 numbering queries, we then replace both the semantic and numbering color maps with these queries, akin to generating vector quantized features through a codebook mechanism (Van Den Oord et al., 2017; Esser et al., 2021). This process results in the generation of semantic and numbering embeddings in Figure 5, which are subsequently combined in the backbone MLM. This combined representation is referred to as *Crayon Prompt*. The Crayon Prompt meets the MLP connector, and then its output is added with the image features at every attention module layer in the MLM as shown in Figure 6(a). We then utilize crayon instructions, as shown in the lower half of Figure 5, and perform *Crayon Prompt Tuning* (CPT) to align the Crayon Prompt to the backbone MLM and enhance object-level image understanding. Here, the magenta colored-text is auto-regressively learned, as demonstrated in the crayon instruction example below Figure 5.
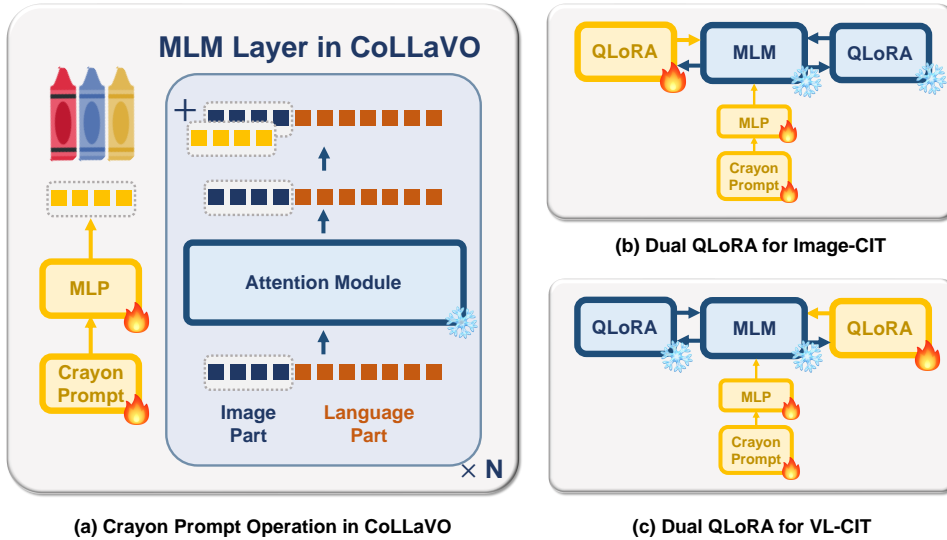
Figure 6: Illuminating (a) how the Crayon Prompt is injected into image embedding features and learning strategies of (b), (c) Dual QLoRA for the object-level image understanding capability (Image-CIT) and VL task capability (VL-CIT) to efficiently coexist without catastrophic forgetting (Luo et al., 2023).

Crayon Prompt-based Instruction Tuning (CIT). CPT focuses solely on learning semantic and numbering queries in the Crayon Prompt and its MLP connector with the MS-COCO 2017 dataset (Lin et al., 2014), aligning them with the backbone MLM to enhance object-level image understanding of 👫 CoLLaVO. On the other hand, *Crayon Prompt-based Instruction Tuning* (CIT) utilizes the visual instruction tuning datasets (Liu et al., 2023c,b; Chen et al., 2023d) as well as crayon instructions to handle complex question answering for VL tasks. It involves training the semantic and numbering queries and the MLP connector again, along with the backbone MLM of 👫 CoLLaVO.

When training the MLM with CIT, we introduce a learning strategy called *Dual QLoRA*, which manages object-level image understanding and complex VL performance, respectively, to effectively maintain both aspects. Figure 6 provides an overview of Dual QLoRA, where *Image-CIT* denotes using crayon instructions to bootstrap object-level image understanding and training only the first QLoRA module, while *VL-CIT* indicates using complex question-answer pairs from visual instruction tuning datasets to achieve zero-shot VL performance and training only the second QLoRA module. During CIT, we present an image in the form of Crayon Prompt to 👫 CoLLaVO, and randomly determine whether to proceed with Image-CIT or VL-CIT. The overarching objective of Dual QLoRA is to efficiently preserve both capabilities of object-level image understanding and complex VL performance. Note that the key distinction between CPT and Image-CIT lies in whether the backbone MLM of

👫 CoLLaVO is trained or not. Further details will be addressed in the following section.

4 Experiments

Implementation Details of 👫 CoLLaVO. To ensure successful reproducibility, we outline the following five crucial technical details of 👫 CoLLaVO: (a) QLoRA, (b) Crayon Prompt, (c) instruction detail of Image-CIT and VL-CIT, (d) training hyper-parameters, and (e) text-generation. **(a)**: we employ Quantized Low-Rank Adaptation (QLoRA) (Hu et al., 2021; Dettmers et al., 2023) since 👫 CoLLaVO pursues efficient training with minimal parameter tuning. Double quantization and normalized float 4-bit (nf4) are used with LoRA of $r = 64$ and $\alpha = 64$. **(b)**: In contrast to CPT with only crayon instructions and images from MS-COCO 2017, CIT is conducted with visual instruction tuning datasets (Liu et al., 2023c,b; Chen et al., 2023d) as well. Hence, many images contain unrecognizable objects, such as text, code, posters, or mathematical symbols. Consequently, a panoptic color map with the unknown category and '0' numbering will be generated, and the semantic query of the **unk** category and numbering query of '0' will operate to create the Crayon Prompt in these cases. **(c)**: Once the color map is given with discernible objects, text descriptions, including object names, their numbering indices, and their bounding box coordinates, are added to the question template. Conversely, if an image contains no objects, the question template includes the phrase "None of detailed object information for image." **(d)**: Regarding training, we train 👫 CoLLaVO with a batch size
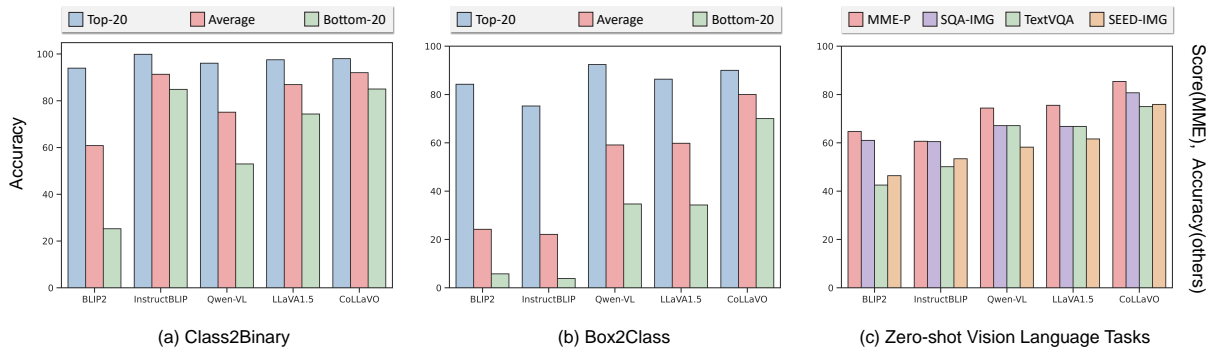
Figure 7: In (a) and (b), there are three metrics for the mean accuracy over Top-20 object categories, Bottom-20, and average of all categories to visualize object-level image understanding of VLMs. In (c), zero-shot performances of VLMs on MME-P (1/20 scaled down of score), SQA-IMG, TextVQA, and SEED-IMG (accuracy) are shown.

of 32 in one epoch using the AdamW (Loshchilov and Hutter, 2019) optimizer, scheduled by cosine annealing (Loshchilov and Hutter, 2016) from a learning rate of 1e-4 to 1e-6 for CPT and from 1e-5 to 1e-6 for CIT, respectively. In addition, $h = 35$, $w = 35$, and $d = 4096$ are used in Figure 5. **(e)**: To find the best performance, 🖍 **CoLLaVO** uses greedy or beam search ($n = 3$) for text generation without any other hyper-parameters.

Object-level Image Understanding. Before delving into validating 🖍 CoLLaVO in VL tasks, it is crucial to ensure its proficiency in object-level image understanding. We assessed the accuracy of 80 object categories classified as 'thing' (See Appendix A) in the MS-COCO 2017 across two directions: Class2Binary (C2B) and Box2Class(B2C), using four strong baselines: BLIP2, InstructBLIP, Qwen-VL, and LLaVA1.5. As illustrated in Figure 7(a)-(b), 🖍 CoLLaVO nearly outperforms the baselines in three cases: Top-20, Bottom-20, and Average for both C2B and B2C. Furthermore, it has the smallest performance gap between the Top-20 accuracy and the Bottom-20 accuracy for both C2B and B2C. Such observation indicates that 🖍 CoLLaVO has a solid object-level image understanding across numerous object classes. Beyond its ability, Appendix B shows zero-shot object grounding performance of 🖍 CoLLaVO for strong generalization to grounding-level understanding.

Zero-shot VL Evaluation. Following improved object-level image understanding, 🖍 CoLLaVO is evaluated to measure zero-shot performance of VL tasks on renowned datasets (See Appendix C). As shown in Figure 1, 7(c), and Table 1, 🖍 CoLLaVO surpasses several closed-source VLMs like GPT-4V, Gemini-Pro, Qwen-VL-Pro, as well as numerous open-source VLMs (See Appendix D for all VLMs used in evaluation). Particularly, noteworthy is its superiority over other models in the follow-

ing benchmarks: MME, MM-Bench, MM-Bench-Chinese, and Q-Bench, which primarily evaluate visual perception and cognition abilities, where 🖍 CoLLaVO demonstrates its significant margins.

The effectiveness of Crayon Prompt and CIT. We ablate the following factors in 🖍 CoLLaVO: semantic embedding in Crayon Prompt, numbering embedding in Crayon Prompt, Dual QLoRA, Image-CIT, and VL-CIT. As illustrated in Table 2, it is evident that the semantic and numbering embedding in the Crayon Prompt significantly boost the zero-shot performance of 🖍 CoLLaVO on MME dataset. It is noteworthy that the semantic embedding alone can improve the zero-shot performance by a large margin, especially in MME-P with 'E&P' scores, implying that injecting object-level semantics helps the model perceive the existence of objects better for solid object-level image understanding. Moreover, the numbering embedding considerably boosts the 'Count' score, demonstrating its effectiveness in differentiating objects of the same category by further refining the performance.

Table 3 demonstrates that Dual QLoRA, Image-CIT, and VL-CIT contribute to improving zero-shot performance, respectively. VL-CIT alone exhibits better performance of 1599.2 in MME-P and 414.1 in MME-C over other open-source VLMs, with the assistance of the Crayon Prompt. Additionally, Image-CIT also enhances performance, albeit to a limited extend without QLoRA, by integrating crayon instructions into CIT as well as CPT. Finally, Dual QLoRA produces the most significant improvement, demonstrating its efficacy in fully leveraging both aspects of Image-CIT and VL-CIT.

5 Discussion and Conclusion

We have shown the effectiveness of 🖍 **CoLLaVO** alongside *Crayon Prompt* and *Dual QLoRA* serving

1127

VLM	GQA	SQA-IMG	TextVQA	POPE	MME-P	MME-C	MM-Bench	MMB-CN	MM-Vet	Q-Bench
BLIP2-13B	42.4	61.0	42.5	85.3	1293.8	290.0	-	-	22.4	-
InstructBLIP-7B	49.2	60.5	50.1	-	-	-	36.0	23.7	26.2	56.7
InstructBLIP-13B	49.5	63.1	50.7	78.9	1212.8	-	-	-	25.6	-
Shikra-13B	-	-	-	-	-	-	58.8	-	-	54.7
IDEFICS-9B	38.4	-	25.9	-	-	-	48.2	25.2	-	-
IDEFICS-80B	45.2	-	30.9	-	-	-	54.5	38.1	-	-
Qwen-VL-7B	59.3	67.1	63.8	-	-	-	38.2	7.4	-	59.4
Qwen-VL-Chat-7B	57.5	68.2	61.5	-	1487.5	360.7	60.6	56.7	-	-
MiniGPT-4-7B	43.5	-	-	-	581.7	-	23.0	-	22.1	-
Otter-7B	-	-	-	-	1292.3	-	48.3	-	24.6	47.2
LLaVA-7B	-	38.5	-	-	807.0	247.9	34.1	14.1	26.7	-
MiniGPT-v2-7B	60.3	-	-	-	-	-	-	-	-	-
MiniGPT-v2-Chat-7B	60.1	-	-	-	-	-	-	-	-	-
LLaVA1.5-7B	62.0	66.8	58.2	85.9	1510.7	293.8	64.3	58.3	30.5	58.7
LLaVA1.5-13B	**63.3**	71.6	61.3	85.9	1531.3	295.4	67.7	63.6	35.4	62.1
mPLUG-Owl-7B	-	-	-	-	967.3	-	46.6	-	-	58.9
mPLUG-Owl2-7B	56.1	68.7	58.2	-	1450.2	-	64.5	-	36.2	62.9
ShareGPT4V-7B	-	68.4	-	-	1567.4	376.4	68.8	62.2	37.6	63.4
CogVLM-17B	56.1	68.7	58.2	-	-	-	65.8	55.9	**54.5**	-
LLaVA-XTuner-20B	-	-	-	-	-	-	75.1	73.7	37.2	-
Intern-XC-7B	-	-	-	-	1528.4	391.1	74.4	72.4	35.2	64.4
CoLLaVO-7B	61.4	**80.7**	**64.2**	**87.2**	**1689.7**	**525.0**	**83.0**	**82.1**	40.3	**67.6**

Table 1: Evaluating zero-shot performances of 🖍️ CoLLaVO on ten vision language datasets compared with the current powerful VLMs such as InstructBLIP, Qwen-VL, LLaVA1.5, and so forth.
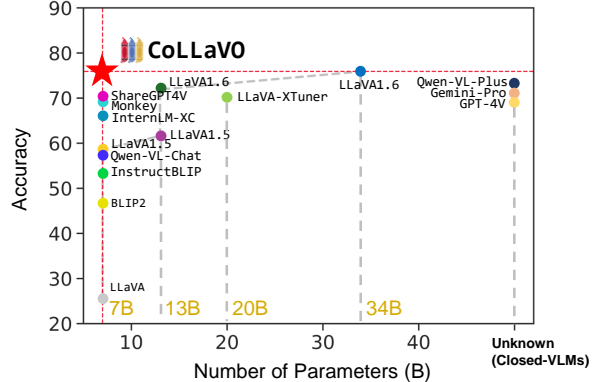
Crayon Prompt		MME			
Sem-Query	Num-Query	MME-P	MME-C	E&P	Count
✗	✗	1553.4	375.0	288.3	141
✓	✗	1636.7	482.1	310.6	147
✓	✓	**1689.7**	**525.0**	**341.6**	**160**

Table 2: Controlling semantic and numbering queries in crayon prompt. Note: 'E&P' denotes the score of the existence and position, and 'Count' denotes the score to understand the numbering.
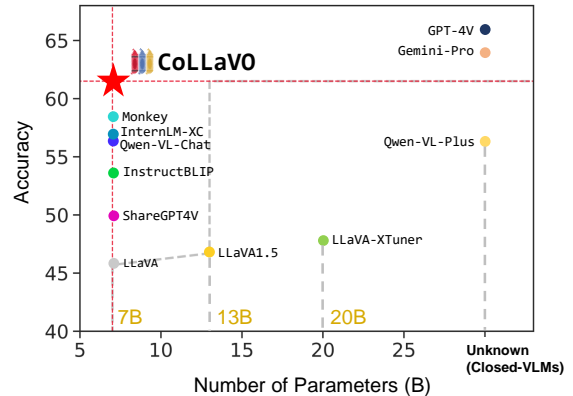
CIT			MME			
Dual Q-LoRA	Image-CIT	VL-CIT	MME-P	MME-C	E&P	Count
✗	✗	✓	1599.2	414.1	298.6	145
✗	✓	✓	1620.5	456.2	308.3	146
✓	✓	✓	**1689.7**	**525.0**	**341.6**	**160**

Table 3: Controlling Dual QLoRA, Image-CIT, and VL-CIT in conducting CIT.

(a) SEED-IMG

(b) HallusionBench

Figure 8: Demonstrating the efficiency and effectiveness of 🖍️ CoLLaVO compared with those of other VLMs. Note that accuracy is measured on SEED-IMG and HallusionBench dataset.

as a key in enhancing the object-level image understanding. Notably, Figure 8(a) illustrates the impressive ability of 🖍️ **CoLLaVO** achieving cutting-edge zero-shot performance with a relatively small size, thanks to its grasp of object-level understanding validated in SEED-IMG (Li et al., 2023b) with 9 types of questions on spatial understandings of images. Even from the perspective of hallucination, Figure 8(b) and Appendix E demonstrate that 🖍️ **CoLLaVO** reduces hallucination due to improved object-level image understanding, satisfactorily compared to both closed-source and open-source VLMs on POPE (Li et al., 2023d) and HallusionBench (Liu et al., 2023a). This suggests that

while many researchers have dramatically scaled up their models and curated their own visual instruction tuning datasets, tackling object-level image understanding proves to be an effective strategy.

Acknowledgments

6 Limitations

Crayon Prompts, relying on a panoptic color map, which is an external source beyond VLMs, may be constrained by the performance of the segmentation model and its encompassing number of object classes. Despite this, we have achieved commendable scores across all zero-shot tasks. It is expected for 📚 **CoLLaVO** to further improve once it incorporates a plethora of visual prompts obtained from diverse sources like robust object classification or image captioning models (Lee et al., 2020, 2022, 2023; Kim et al., 2023c), object-centric causally human-interpretable information (Kim et al., 2021, 2023b), open object detection (Zhang et al., 2023a), visual grounding (Liu et al., 2023d; Ren et al., 2024), interactive or unsupervised segmentation (Kirillov et al., 2023; Kim et al., 2023a), optical characteristic recognition model (Bautista and Atienza, 2022), and other fascinating approaches (Lee et al., 2024a,b; Park et al., 2024b,a; Kim et al., 2024). Beyond its limitation, we believe our promising direction for crayon prompt-like visual cues surely further improve on image understanding for human-like AGI.

7 Ethics Statement

We affirm that all research presented in this paper adheres to the principles of ethical conduct and integrity. The experiments conducted and the results reported are based on rigorous scientific methods and strive to contribute positively to the field of vision language models. All datasets used in this study: MS-COCO 2017 (Lin et al., 2014) and visual instruction datasets (Liu et al., 2023c,b; Chen et al., 2023d) were obtained and analyzed in compliance with relevant regulations and guidelines for research ethics and data privacy. In addition, any potential limitations have been transparently discussed, so we are committed to upholding the highest standards of integrity, accountability, and respect for communities affected by our research.

References

Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. 2022. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Darwin Bautista and Rowel Atienza. 2022. Scene text recognition with permuted autoregressive sequence models. In *European Conference on Computer Vision*, pages 178–196. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2023. Making large multimodal models understand arbitrary visual prompts. *arXiv preprint arXiv:2312.00784*.

Aochuan Chen, Yuguang Yao, Pin-Yu Chen, Yihua Zhang, and Sijia Liu. 2023a. Understanding and improving visual prompting: A label-mapping perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19133–19143.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023b. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023c. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023d. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

XTuner Contributors. 2023. Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1821–1831.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer.

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

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

Junho Kim, Byung-Kwan Lee, and Yong Man Ro. 2021. Distilling robust and non-robust features in adversarial examples by information bottleneck. *Advances in Neural Information Processing Systems*, 34:17148–17159.

Junho Kim, Byung-Kwan Lee, and Yong Man Ro. 2023a. Causal unsupervised semantic segmentation. *arXiv preprint arXiv:2310.07379*.

Junho Kim, Byung-Kwan Lee, and Yong Man Ro. 2023b. Demystifying causal features on adversarial examples and causal inoculation for robust network by adversarial instrumental variable regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12302–12312.

Seongyeop Kim, Hyung-Il Kim, and Yong Man Ro. 2024. Improving open set recognition via visual prompts distilled from common-sense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2786–2794.

Yeonju Kim, Junho Kim, Byung-Kwan Lee, Sebin Shin, and Yong Man Ro. 2023c. Mitigating dataset bias in image captioning through clip confounder-free captioning network. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1720–1724. IEEE.

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

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. 2023. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*.

Byung-Kwan Lee, Chae Won Kim, Beomchan Park, and Yong Man Ro. 2024a. Meteor: Mamba-based traversal of rationale for large language and vision models. *arXiv preprint arXiv:2405.15574*.

Byung-Kwan Lee, Junho Kim, and Yong Man Ro. 2022. Masking adversarial damage: Finding adversarial saliency for robust and sparse network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15126–15136.

Byung-Kwan Lee, Junho Kim, and Yong Man Ro. 2023. Mitigating adversarial vulnerability through causal parameter estimation by adversarial double machine learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4499–4509.

Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro. 2024b. Moai: Mixture of all intelligence for large language and vision models. *arXiv preprint arXiv:2403.07508*.

Byung-Kwan Lee, Youngjoon Yu, and Yong Man Ro. 2020. Towards adversarial robustness of bayesian neural network through hierarchical variational inference.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023b. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023d. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco:

Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023d. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. 2023e. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19434–19445.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023f. Mm-bench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

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

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Changdae Oh, Hyeji Hwang, Hee-young Lee, YongTaek Lim, Geunyoung Jung, Jiyoung Jung, Hosik Choi, and Kyungwoo Song. 2023. Blackvip: Black-box visual prompting for robust transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24224–24235.

OpenAI. 2023a. Gpt-4v(ision) system card. `https://openai.com/research/gpt-4v-system-card`, Last accessed on 2024-02-13.

OpenAI. 2023b. Gpt-4v(ision) technical work and authors. `https://openai.com/contributions/gpt-4v`, Last accessed on 2024-02-13.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Sungjune Park, Hyunjun Kim, and Yong Man Ro. 2024a. Integrating language-derived appearance elements with visual cues in pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Sungjune Park, Hyunjun Kim, and Yong Man Ro. 2024b. Robust pedestrian detection via constructing versatile pedestrian knowledge bank. *Pattern Recognition*, 153:110539.

Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. 2023. Jack of all tasks, master of many: Designing general-purpose coarse-to-fine vision-language model. *arXiv preprint arXiv:2312.12423*.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

Mark Sandler, Andrey Zhmoginov, Max Vladymyrov, and Andrew Jackson. 2022. Fine-tuning image transformers using learnable memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12155–12164.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does clip know about a red circle? visual prompt engineering for vlms. *arXiv preprint arXiv:2304.06712*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM-techreport.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. 2023. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023a. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.

Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. 2023b. Fine-grained visual prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023a. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023b. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2024. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. 2023a. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1020–1031.

Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. 2023b. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. 2023. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*.

A   COCO Classes for Panoptic Color Map

COCO Panoptic Classes			
person	bicycle	car	motorcycle
airplane	bus	train	truck
boat	traffic light	fire hydrant	stop sign
parking meter	bench	bird	cat
dog	horse	sheep	cow
elephant	bear	zebra	giraffe
backpack	umbrella	handbag	tie
suitcase	frisbee	skis	snowboard
sports ball	kite	baseball bat	baseball glove
skateboard	surfboard	tennis racket	bottle
wine glass	cup	fork	knife
spoon	bowl	banana	apple
sandwich	orange	broccoli	carrot
hot dog	pizza	donut	cake
chair	couch	potted plant	bed
dining table	toilet	tv	laptop
mouse	remote	keyboard	cell phone
microwave	oven	toaster	sink
refrigerator	book	clock	vase
scissors	teddy bear	hair drier	toothbrush
banner*	blanket*	bridge*	cardboard*
counter*	curtain*	door*	floor-wood*
flower*	fruit*	gravel*	house*
light*	mirror*	net*	pillow*
platform*	playingfield*	railroad*	river*
road*	roof*	sand*	sea*
shelf*	snow*	stairs*	tent*
towel*	wall-brick*	wall-stone*	wall-tile*
wall-wood*	water*	window-blind*	window*
tree*	fence*	ceiling*	sky*
cabinet*	table*	floor*	pavement*
mountain*	grass*	dirt*	paper*
food*	building*	rock*	wall*
rug*			

∗: Object class that is not classified as 'thing' (countable) but 'stuff' (uncountable)

B Grounding-level Image Understanding

Model	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
Shikra-7B	87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19
Ferret-7B	87.49	91.35	82.45	80.78	87.38	73.14	83.93	84.76
VistaLLM-7B	88.10	91.50	83.00	82.90	89.80	74.80	83.60	84.40
Qwen-VL-7B	89.36	92.26	85.34	83.12	88.25	77.21	85.58	85.48
CogVLM-Grounding-17B	92.76	94.75	88.99	88.68	92.91	83.39	89.75	90.79
CoLLaVO-7B	87.34	91.08	82.39	80.87	86.36	73.20	82.44	82.33

Table 4: Comparing object grounding performances of Shikra (Chen et al., 2023c), Ferret (You et al., 2024), VistalLLM (Pramanick et al., 2023), Qwen-VL (Bai et al., 2023), CogVLM-Grounding (Wang et al., 2023), and CoLLaVO on several object grounding benchmarks: RefCOCO, RefCOCO+, and RefCOCOg (Kazemzadeh et al., 2014). Even though CoLLaVO did not use object grounding dataset (RefCOCO) in training phase, CoLLaVO shows comparable zero-shot object grounding performances, compared with (no zero-shot) other models specifically targeting object grounding task trained with RefCOCO grounding dataset.

C Zero-shot Vision Language Datasets used in Evaluation

- **GQA** (Hudson and Manning, 2019) is a visual question answering dataset comprising real-world images annotated with scene graphs. It tackles the issue of semantic compositionality by utilizing semantic representations of scenes and questions. It encompasses 22 million questions covering a wide array of images, each associated with structured representations of image objects, attributes, and relations.

- **SQA-IMG** (Iyyer et al., 2017), a subset of the ScienceQA (SQA) dataset that includes image context, comprises 10,332 multiple-choice questions sourced from elementary and high school science education materials, covering diverse sub-fields. A majority of the questions in the SQA dataset are accompanied by supplementary lectures (83.9%) and detailed explanations (90.5%), enriching understanding with broader knowledge and specific reasoning for correct answers.

- **TextVQA** (Singh et al., 2019) is a large-scale complex benchmark to analyze and understand text embedded within images in order to respond to associated questions. This involves integrating textual information present within images and reasoning over it to provide answers. The dataset comprises 28,408 images sourced from OpenImages, accompanied by 45,336 questions and 453,360 corresponding ground truth answers.

- **POPE** (Li et al., 2023d) serves as a polling-based binary classification query dataset, tailored to assess object hallucination challenges within VLMs. It comprises three distinct subsets, *i.e.,* random, popular, and adversarial, each crafted using varied sampling techniques, resulting in a total of 8,910 entries.

- **MME** (Fu et al., 2023) is introduced as a novel comprehensive benchmark aimed at assessing the performance of VLMs by measuring both perception and cognition abilities across 14 sub-tasks. To mitigate potential data leakage issues associated with public datasets, all annotations for instruction-answer pairs are manually designed.

- **MMBench, MMBench-Chinese** (Liu et al., 2023f) establish a comprehensive evaluation framework spanning multiple modalities. These frameworks encompass around 3000 multiple-choice questions addressing 20 distinct capability dimensions in both English and Chinese languages. An innovative approach is introduced through the integration of ChatGPT into the evaluation process.

- **MM-Vet** (Yu et al., 2023) is a multi-modal assessment benchmark that assesses a broad range of capabilities essential for handling real-world scenarios, such as solving mathematical problems or interpreting visual humor. The dataset consists of 187 images collected from diverse online platforms and presents 205 questions, each requiring the application of one or more capabilities for an answer. These questions vary in type and necessitate open-ended responses of varying lengths.

- **Q-Bench** (Wu et al., 2023) evaluates VLMs across three dimensions relevant to low-level vision: perception, description, and assessment. To assess perception, the framework utilizes 2,990 diverse images, each accompanied by a human-generated question focusing on its low-level attributes. For evaluating VLMs' description regarding low-level information, human-labeled textual descriptions for 499 images are utilized, alongside a comparison pipeline involving GPT. Additionally, the framework assesses VLMs' visual quality assessment abilities, aiming to align with human opinion scores.

- **MathVista** (Lu et al., 2023) assesses VLMs' mathematical reasoning ability within visual contexts, with 6,141 examples sourced from 28 existing multimodal datasets on mathematics. MathVista provides a comprehensive evaluation platform, requiring meticulous visual comprehension and compositional reasoning, posing challenges even to state-of-the-art foundational models.

- **AI2D** (Kembhavi et al., 2016), or AI2 Diagrams, is a dataset comprising over 5,000 grade school science diagrams. It includes comprehensive annotations of constituents and relationships, along with rich syntactic parses and over 15,000 corresponding multiple-choice questions.

- **SEED-IMG** (Li et al., 2023b) comprises a subset of SEED-Bench, focusing on the image modality. The original SEED-Bench includes 19,000 multiple-choice questions with precise human annotations, covering 12 evaluation dimensions, including comprehension of both image and video modalities.

- **HallusionBench** (Liu et al., 2023a) introduces a comprehensive benchmark tailored for evaluating image-context reasoning abilities. It prioritizes nuanced comprehension and interpretation of visual information. The benchmark consists of 346 images accompanied by 1129 expert-crafted questions, enabling a quantitative analysis of model response tendencies, logical consistency, and diverse failure modes.

D Vision Language Models used in Evaluation

- **BLIP2** (Li et al., 2023c) introduces Q-Former that serves as an intermediary between frozen unimodal models, extracting pertinent visual features from a frozen image encoder and providing them to a frozen large language model to generate text.

- **InstructBLIP** (Dai et al., 2023) presents a vision-language instruction tuning framework designed to address the challenges of generalizing to diverse tasks, through a systematic study involving 26 datasets transformed into instruction tuning format across 11 task categories.

- **Shikra** (Chen et al., 2023c) proposes a unified model designed for referential dialogue tasks, which encompass various vision-language tasks such as VQA, image captioning, and location-related tasks like referring expression comprehension and PointQA.

- **IDEFICS** (Laurençon et al., 2023) introduces a curated web-scale dataset comprising 141 million multimodal English web documents, each containing associated images and text, totaling 353M images and 115B tokens. They aim to provide full multimodal documents preserving the natural context of images within web pages.

- **Qwen-VL, Qwen-VL-Chat** (Bai et al., 2023) introduces Qwen-VL series, a collection of highly performant and versatile vision-language models based on Qwen language model. They support multiple languages and handling of multi-image inputs, and fine-grained visual understanding capabilities.

- **MiniGPT-4** (Zhu et al., 2023) presents a vision-language model that combines Vicuna with freezed pre-trained vision components of Q-Former from BLIP2, aiming to replicate the exceptional capabilites demonstrated by GPT-4.

- **MiniGPT-v2** (Chen et al., 2023b) is designed to effectively handle multiple vision-language tasks by employing a task-oriented instructiom training scheme, through three training stage and utilization of higher-resolution images.

- **Otter** (Li et al., 2023a) addresses the gap between DeepMind Flamingo by employing OpenFlamingo and multi-modal in-context instruction tuning (MIMIC-IT) dataset.

- **LLaVA** (Liu et al., 2023c,b) first introduces the concept of visual instruction tuning, extending language only instruction tuning to vision language instruction tuning to develop a general-purpose visual assistant.

- **LLaVA-XTuner** (Contributors, 2023) is a tool to fine-tune LLaVA to achieve general-purpose model.

- **mPLUG-Owl** (Ye et al., 2023a) introduces a modularized training paradigm for large multi-modal language models capable of supporting multiple modalities simultaneously. Inspired by modularization concepts, their method integrates pre-trained language models, visual knowledge modules, and visual abstractor modules to achieve effective alignment between images and text.

- **mPLUG-Owl2** (Ye et al., 2023b) features a modularized network design to handle both modality collaboration and interference. They introduce shared functional modules to promote collaboration and a modality-adaptive module to manage different modalities effectively.

- **ShareGPT4V** (Chen et al., 2023d) argues that current Large multi-modal models face sub-optimal modality alignment due to the lack of high-quality image-text pairs. To address this issue, they collected high-quality captions on a larger scale in two phases. This effort led to the creation of the ShareGPT4V dataset, comprising 100K GPT4-Vision generated captions and 1.2M captions crafted by their caption model.

- **CogVLM** (Wang et al., 2023) handles challenges of the lack of direct equivalence between visual and textual input spaces. They introduce a trainable visual expert to the language model, where it allows for the retention of natural language processing capabilities while enhancing visual understanding abilities.

- **Intern-XC** (Zhang et al., 2023b) is trained to generate long-form content interleaved with contextually relevant images, based on a multilingual vision-language dataset comprising over 11M semantic concepts collected from public websites, thereby enhancing vision-language interactions.

- **MM-GPT** (Gong et al., 2023) fine-tune OpenFlamingo using comprehensive datasets of image and text instructions to conduct multi-turn image-text dialogues more closely aligned with human preferences. A perceiver resampler is used for efficient visual information extraction and gated cross-attention layers for image-text interactions.

E   Detail of POPE dataset for Hallucination

Types	Metrics	LLaVA	MiniGPT-4	MM-GPT	mPLUG-Owl	InstructBLIP	Shikra	CoLLaVO
Adversarial	Accuracy	49.7	65.2	50.0	50.7	72.1	83.1	**86.8**
	Precision	49.6	61.2	50.0	50.3	65.1	85.6	**96.3**
	Recall	99.1	82.9	**100.**	99.3	95.1	79.6	76.5
	F1-Score	66.3	70.4	66.7	66.8	77.3	82.5	**85.2**
Random	Accuracy	50.4	49.7	50.1	54.0	**88.6**	86.9	87.4
	Precision	50.2	78.2	50.1	52.1	84.1	94.4	**98.1**
	Recall	99.1	82.2	**100.**	99.6	95.1	79.3	76.2
	F1-Score	66.6	80.1	66.7	68.4	**89.3**	86.2	85.8
Popular	Accuracy	49.9	69.7	50.0	50.9	82.8	84.0	**87.6**
	Precision	49.9	65.9	50.0	50.5	76.3	87.6	**98.0**
	Recall	99.3	81.9	**100.**	99.4	95.1	79.2	76.8
	F1-Score	66.4	73.0	66.7	66.9	84.7	83.2	**86.1**

Table 5: Measuring four metrics: Accuracy, Precision, Recall, F1-score on three types of question answering to evaluate hallucination of vision language models: Adversarial, Random, and Popular in POPE.