2-2024

Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites

Lei WANG
Singapore Management University, lei.wang.2019@phdcs.smu.edu.sg

Jiabang HE

Shenshen LI

Ning LIU

Ee-peng LIM
Singapore Management University, eplim@smu.edu.sg

Mitigating Fine-Grained Hallucination by Fine-Tuning Large Vision-Language Models with Caption Rewrites

Lei Wang[♠] **Jiabang He**[♣] **Shenshen Li**[♣] **Ning Liu**[♦] **Ee-Peng Lim**[♠]
[♠]Singapore Management University, [♦]Beijing Forestry University
[♠]University of Electronic Science and Technology of China

Abstract

Large language models (LLMs) have shown remarkable performance in natural language processing (NLP) tasks. To comprehend and execute diverse human instructions over image data, instruction-tuned large vision-language models (LVLMs) have been introduced. However, LVLMs may suffer from different types of object hallucinations. Nevertheless, LVLMs are evaluated for coarse-grained object hallucinations only (i.e., generated objects non-existent in the input image). The fine-grained object attributes and behaviors non-existent in the image may still be generated but not measured by the current evaluation methods. In this paper, we thus focus on reducing fine-grained hallucinations of LVLMs. We propose *ReCaption*, a framework that consists of two components: rewriting captions using ChatGPT and fine-tuning the instruction-tuned LVLMs on the rewritten captions. We also propose a fine-grained probing-based evaluation method named *Fine-Grained Object Hallucination Evaluation* (*FGHE*). Our experiment results demonstrate that ReCaption effectively reduces fine-grained object hallucination for different LVLM options and improves their text generation quality. The code can be found at https://github.com/Anonymousanoy/FOHE.

1 Introduction

Large language models (LLMs), such as GPT-3 (Brown et al., 2020) and ChatGPT (OpenAI, 2022), have demonstrated impressive performance in a wide range of natural language processing (NLP) tasks (Qin et al., 2023). To extend LLMs to comprehend and execute both text-only and multi-modal (i.e., vision + text) instructions, new multi-modal large language models have been introduced, exemplified by GPT-4 (OpenAI, 2023). Despite the impressive capabilities of GPT-4 in understanding and processing multi-modal information, the underlying mechanisms responsible for these exceptional abilities remain unclear due to its black-box nature.

To shed light on this mystery, recent research endeavors have focused on extending text-only LLMs to comprehend visual inputs by incorporating vision-language models (LVMs) into text-only LLMs. The resultant model is called the large vision-language model (LVLM). One LVLM research direction involves using vision modality models to provide textual description for visual information, followed by employing closed-source LLMs, such as ChatGPT, to address multi-modal tasks such as visual QA and image caption generation. The LVLM examples using this approach include Visual ChatGPT (Wu et al., 2023), MM-REACT (Yang et al., 2023), and HuggingGPT (Shen et al., 2023). Nevertheless, this approach requires good alignment across modalities to understand specific multi-modal instructions.

Another alternative approach focuses on instruction-tuned large vision-language models (LVLMs), e.g., LLaVA (Liu et al., 2023c), MiniGPT-4 (Zhu et al., 2023), mPLUG-Owl (Ye et al., 2023), and InstructBLIP (Dai et al., 2023), which extend language-only open-source LLMs (e.g., FlanT5 (Chung et al., 2022) and LLaMA (Touvron et al., 2023)) to encompass visual reasoning abilities and instruction execution abilities by training LVLMs on text-image pairs and multi-modal instructions. Instruction-tuned LVLMs demonstrate outstanding capabilities in solving diverse multi-modal tasks (Agrawal et al., 2019; Schwenk et al., 2022; Lu et al., 2022).

Despite the success of instruction-tuned LVLMs, these models are prone to hallucination, which compromises both model performance and user experience in real-world use cases (Ji et al., 2023). Similar to text-only LLMs (Ji et al., 2023; Bang et al., 2023), LVLMs may generate text descriptions that include non-existent or inaccurate objects in the target image also known as object hallucination (Biten et al., 2022; Rohrbach et al., 2018; Ji et al., 2023).

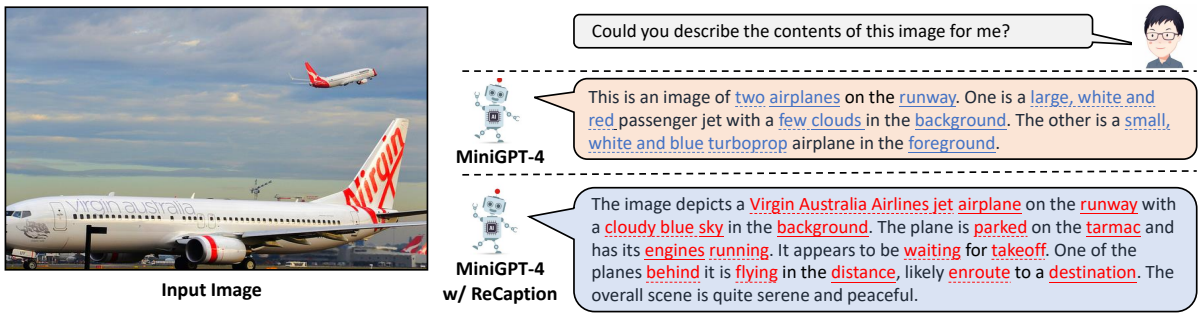


Figure 1: An illustrative example is presented to compare the output of MiniGPT-4 and MinitGPT-4 with ReCaption. The generated caption by MiniGPT-4 contains words in blue that are inconsistent with the given image. In contrast, MinitGPT-4 with ReCaption demonstrates a superior ability to generate words in red that align more closely with the image at a fine-grained level. The words marked with an underline represent objects. The words marked with an dotted underline denote attributes and behaviors.

To evaluate object hallucination in instruction-tuned LVLMs, several evaluation methods (e.g., CHAIR (Rohrbach et al., 2018), POPE (Li et al., 2023) have been proposed. However, these methods only focus on coarse-grained object hallucination but not the hallucinated object attributes and behaviors. We call the latter *fine-grained object hallucination*. Fine-grained object hallucination refers to the phenomenon wherein LVLMs generate captions that include not only non-existent or erroneous objects but also inaccurate object attributes and behaviors.

Consider the example in Figure 1. This example illustrates that MiniGPT-4, an instruction-tuned LVLM, generates fine-grained hallucinations. The input image depicts two airplanes: a smaller one in flight and a larger one parked on the runway, accompanied by clouds. Multi-object hallucination occurs when the generated text mistakenly introduces erroneous or irrelevant relations between objects. Object attribute hallucination refers to generating incorrect attributes for a particular object. Vanilla MiniGPT-4 generates a description such as "small, white, and blue turboprop airplane", even though the color of the turboprop is unknown. Object behavior hallucination pertains to describing incorrect actions for objects. In this example scenario, the smaller airplane is flying, but MiniGPT-4 incorrectly states it is on the runway.

In this paper, we make key contributions to address fine-grained object hallucination. First, we introduce *ReCaption*, a framework that enables instruction-tuned LVLMs to reduce fine-grained object hallucination by fine-tuning them on additional rewritten captions derived from curated high-quality image captions.

ReCaption consists of two components: 1) rewriting captions using ChatGPT and 2) additional training of instruction-tuned LVLM on the rewritten captions. To develop the first component, we employ a two-stage prompting strategy to guide ChatGPT to generate high-quality image-text pairs. In the first stage, we utilize a prompt to tell ChatGPT to extract verbs, nouns, and adjectives from the original input caption. In the second stage, the extracted verbs, nouns and adjectives are merged into a list which is used in another ChatGPT prompt to generate a rewritten image caption that covers the list of words. This caption rewriting process is repeated multiple times, creating a diverse collection of captions that still retain the core content of the original caption. At the end of the first stage, we obtain a set of high-quality image-caption pairs. The second component of ReCaption performs fine-tuning of the instruction-tuned LVLM using the above set of image-caption pairs to strengthen the model's fine-grained alignment between visual and text modalities.

To better evaluate the proposed method, we introduce a new evaluation method called *Fine-Grained Object Hallucination Evaluation (FGHE)*. This method assesses how well any LVLM performs in minimizing fine-grained object hallucination by incorporating another evaluation method POPE with the measurement of hallucinated object attributes and behaviors. Similar to POPE, FGHE converts object hallucination evaluation into a set of binary classification tasks, prompting instruction-tuned LVLMs with simple Yes-or-No questions about the probed objects (e.g., "*Is there a car in the image?*") and about the attributes/behaviors of objects (e.g., "*Is the man's clothing blue in the picture?*"). We

evaluate the fine-grained hallucination reduction of ReCaption using POHE and FGHE. Our evaluation results demonstrate that any LVLM adopting the ReCaption framework can effectively reduce fine-grained object hallucination.

2 Related Work

2.1 Large Vision-Language Models

As text-only LLMs (Brown et al., 2020; Touvron et al., 2023; OpenAI, 2022) show very good results across NLP tasks (Qin et al., 2023), there are many works extending LLMs to comprehend visual inputs, and to the development of large vision-language models (LVLMs). Two primary paradigms have been pursued in this line of research. The first paradigm involves representing visual information through textual descriptions. Closed-source LLMs, such as ChatGPT (OpenAI, 2022), are used to establish connections between vision modality models, enabling subsequent handling of multimodal tasks. Several notable approaches following this paradigm include Visual ChatGPT (Wu et al., 2023), MM-REACT (Yang et al., 2023), and HuggingGPT (Shen et al., 2023).

The second paradigm centers around training LVLMs using vision-language instructions. MultiInstruct (Xu et al., 2022) engages in vision-language instruction tuning, containing various multi-modal tasks involving visual comprehension and reasoning. LLaVA (Liu et al., 2023c) employs self-instruct (Wang et al., 2022) to generate instructions and optimize the alignment network's and LLM's model parameters. MiniGPT-4 (Zhu et al., 2023) integrates a visual encoder derived from BLIP-2 (Li et al., 2022) and trains the model using image captions generated by ChatGPT, ensuring that these captions are longer than the training data of BLIP-2. mPLUG-Owl (Ye et al., 2023) equips LLMs with multimodal abilities through modularized learning, enabling LLMs to support multiple modalities. Lastly, InstructBLIP (Dai et al., 2023) enhances the cross-modal alignment network, empowering the LLM to generate meaningful semantic descriptions for a given image. By exploring these paradigms, researchers enhance the ability of LLMs to process visual information, thus expanding their applicability and effectiveness in various vision-language tasks.

2.2 Hallucnation in Large Vison-Language Mdoels

A recent survey (Ji et al., 2023) has thoroughly analyzed studies examining hallucinations in various tasks, including text summarization (Huang et al., 2021; Maynez et al., 2020; Cao and Wang, 2021; Tang et al., 2021; Chen et al., 2021), dialogue generation (Shuster et al., 2021; Wu et al., 2021; Dziri et al., 2021), and vision-language generation (Rohrbach et al., 2018; Biten et al., 2022; Xiao and Wang, 2021; Dai et al., 2022; Liu et al., 2023b; Gunjal et al., 2023; Wang et al., 2023b; Liu et al., 2023a; Yin et al., 2023; Wang et al., 2023a; Lee et al., 2023). Specifically, within the domain of vision-language generation, object hallucination can be further classified as intrinsic and external hallucinations. Intrinsic hallucinations in vision-language generation refer to generated captions that contain incorrect or non-existent objects in the given image. On the other hand, external hallucinations in vision-language generation refer to generated captions that contain irrelevant objects. To comprehensively investigate the phenomenon of object hallucination in LVLMs, POPE (Li et al., 2023) endeavors to conduct an extensive empirical examination of object hallucinations across various LVLMs. Xu et al. (2023a) add POPE to the proposed comprehensive evaluation benchmark for evaluating VLVMs. While previous works mainly focus on hallucinations about the presence or absence of objects, LVLMs may also generate more fine-grained erroneous or incomplete descriptions for target images, specifically regarding incorrect attributes associated with objects. Therefore, this paper aims to mitigate and evaluate finer-grained hallucinations within LVLMs.

3 ReCaption Framework

In this section, we describe the design of our ReCaption framework, highlighting its key components and strategies. ReCaption is LVLM-agnostic and can be used to reduce fine-grained object hallucination of any LVLMs. There are two components in ReCaption, caption rewriting and training of instruction-tuned LVLM. The latter is conducted using the rewritten captions.

3.1 Caption Rewriting

The caption rewriting component aims to create good quality captions of training images. To generate rewritten image-text pairs, we first randomly se-

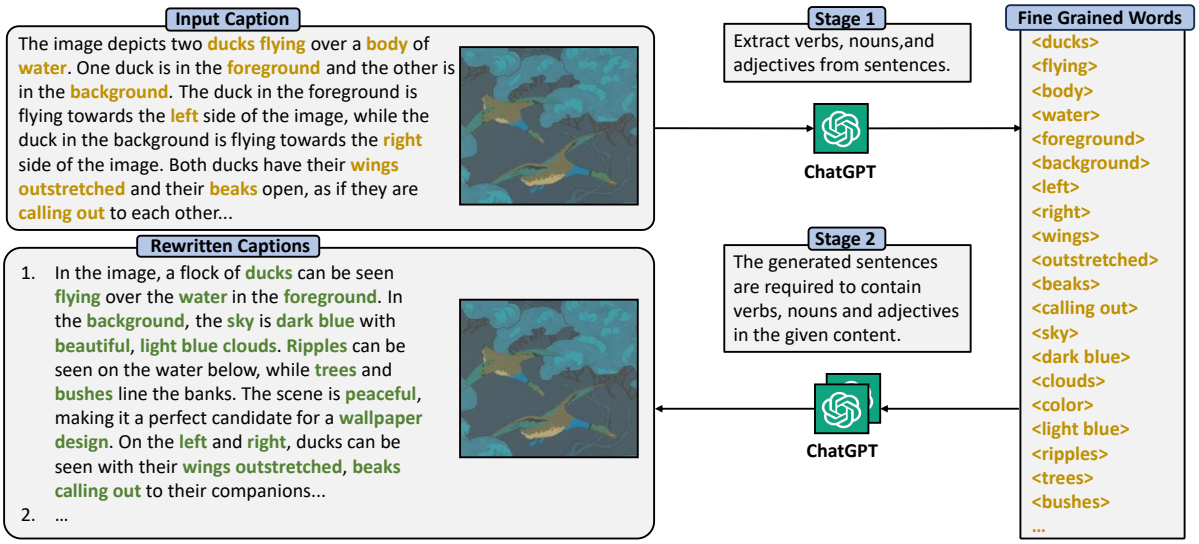


Figure 2: Illustration of rewriting image captions using ChatGPT. Stage 1: Keyword Extraction Prompt (i.e., "Derive verbs, nouns, and adjectives from sentences"), directs ChatGPT to generate verbs, nouns, and adjectives (highlighted in brown) from the original caption. Stage 2: Caption Generation Prompt (i.e., "The resulting sentences must encompass verbs, nouns, and adjectives aligned with the provided content") guides ChatGPT to generate a rewritten caption. By repeating this prompt, multiple rewritten captions will be generated.

lect image-text pairs from the cc_sbu_align dataset, which is curated by MiniGPT-4 (Zhu et al., 2023). As shown in Figure 2, we use a two-stage prompting strategy on the caption of each selected image to generate multiple different captions that convey the essential elements of the original caption but with varied details. We use ChatGPT (OpenAI, 2022) for both stages because it performs well in both stages. We believe other similar LLMs can also be used. This two-stage prompting aims to preserve the original semantics associated with the corresponding image, which is important for enriching the fine-grained alignment between image and text. Note that we provide one demonstration example for each stage to assist LLMs in improving their keyword extraction and caption generation abilities. In the following, we elaborate on the proposed two-stage prompting strategy.

Stage 1: Keyword Extraction. This stage aims to preserve essential information from the caption corresponding to a specific image. In generating descriptions for an input image, LVLMs may produce inaccurate or irrelevant objects, object attributes, and object behaviors. Since verbs, nouns, and adjectives in captions often denote objects, object attributes, and object behaviors, we devise a prompt to facilitate the extraction of these pertinent words from the original caption. Through this prompting, the extracted keywords preserve the caption's essential information. Below is the prompt template for the first-stage prompting:

> `Extract verbs, nouns, and adjectives from [`X`],`

where X denotes the original caption of the input image. Figure 2 depicts the first-stage prompting directly extracts nouns (e.g., "ducks"), adjectives (e.g., "right"), and verbs (e.g., "flying") from the caption. The extracted keywords, denoted as $\{x_1^k, x_2^k, ..., x_n^k\}$, capture essential information in the caption, where k means keywords.

Stage 2: Caption Generation. The second stage is to rewrite the original caption, conditioned on the extracted words obtained in stage 1. The prompt template for second-stage prompting can be denoted as follows:

> `The generated sentences are required to contain verbs, nouns and adjectives in the given content: [`x_1^k`, `x_2^k`, ..., `x_n^k`].`

Through prompting ChatGPT, we can obtain a new version of the caption, denoted as X'. To keep the characteristic of randomness in the LLM, we use a temperature ratio of 1.0. Further, we prompt ChatGPT R times to generate diverse captions while preserving essential information from the original caption. As shown in Figure 2, each rewritten caption has different details while contextually rele-

vant to the original caption. This caption generation prompt basically elicits the imagination and rewriting abilities of ChatGPT.

3.2 Additional Tuning

By producing R diverse rewritten captions for each original image caption, we can now enhance the implicit fine-grained alignment between the input images and captions. In our experiments, we set $R = 5$. These rewritten image-caption pairs are model-agnostic to language model architecture, thereby enabling their seamless adaptation to various LVLM models, including MiniGPT-4 (Zhu et al., 2023), LLaVA (Liu et al., 2023c), mPLUG-Owl (Ye et al., 2023), and MultiModal-GPT (Gong et al., 2023), with minimal changes. Our later experimental analysis reveals that a small number of rewritten image-text pairs can significantly reduce fine-grained object hallucination and improve caption generation quality. Model tuning using such a small set of pairs entails negligible additional computational cost compared to training of the original LVLM.

The training loss over the images with their rewritten captions can be formulated as follows:

$$\mathcal{L} = \sum_{i=1}^{M} \sum_{j=1}^{R} \mathcal{L}_{\text{CE}}(X'_{i,j}, \hat{X}_i), \qquad (1)$$

where X'_i represents a rewritten caption for image-caption pair i, $\hat{X}_i$ is generated caption of image-caption pair i, R is the total number of rewritten captions for the same image, M denotes the total number of training image-caption examples, and $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss.

4 Hallucination Evaluation of LVLMs

To evaluate the effectiveness of ReCaption framework, we use an existing evaluation dataset and method introduced by the POPE work (Li et al., 2023). As POPE only evaluates coarse-grained hallucination, we introduce another dataset with specific evaluation method. In this section, we will fist introduce POPE dataset and method. We then introduce our proposed dataset and its evaluation method also known as *Fine-Grained Object Hallucination Evaluation (FGHE)*. We leave out the Caption Hallucination Assessment with Image Relevance (CHAIR) metric Rohrbach et al. (2018) which suffers from prompt sensitivity and inaccuracies as reported in (Li et al., 2023).

POPE dataset and evaluation method. In Li et al. (2023), the Polling-based Object Probing Evaluation (POPE) evaluation method was proposed to improve the evaluation of object hallucination in LVLMs. The basic idea behind POPE is to evaluate hallucination by asking LVLMs simple Yes-or-No questions concerning the probed objects. For instance, a sample question could be: "Is there a car in the image?". By including existent or non-existent object into a question of an input image, the evaluation method obtains a question with "yes" or "no" answer. In our experiments, we use popular non-existent objects in the no-questions (i.e., questions with "no" answer). Finally, the POPE dataset involves 3000 questions for the captions of 500 images. By treating a LVLM's answers to the questions as a binary classification task, we obtain the Accuracy, Precision, Recall and F1 scores of the LVLM on this dataset. The higher the scores, the less hallucination. POPE has been adopted by LVLM-eHub, an evaluation benchmark for LVLMs (Xu et al., 2023b).

FGHE dataset and evaluation method. FGHE follows the binary classification approach of POPE to evaluate LVLMs. However, unlike POPE, FGHE requires a different set of binary questions to measure fine-grained hallucination. The FGHE dataset consists of 50 images and 200 binary questions divided into three categories: (a) *multiple-object* question which verifies the relationships between multiple objects in the image; (b) *attribute* question which verifies an attribute of an object in the image; and (c) *behavior* question which verifies a behavior or an object in the image. Figure 3 presents an illustrative comparison between probing questions in FGHE and POPE. All the above questions are manually defined by human annotators on a subset of 50 images from the validation set of MSCOCO dataset. As shown in , the FGHE dataset consists of 100 yes-questions and 100 no-questions. Among the yes-questions, 47, 45 and 8 are multi-object, attribute and behavior questions. Among the no-questions, 51, 42 and 7 are multi-object, attribute and behavior questions. Table 3 only displays few behavior questions as some of them are counted towards multiple objects. Similar to POPE, we finally employ Accuracy, Precision, Recall, and F1 score of all questions as the evaluation metrics.

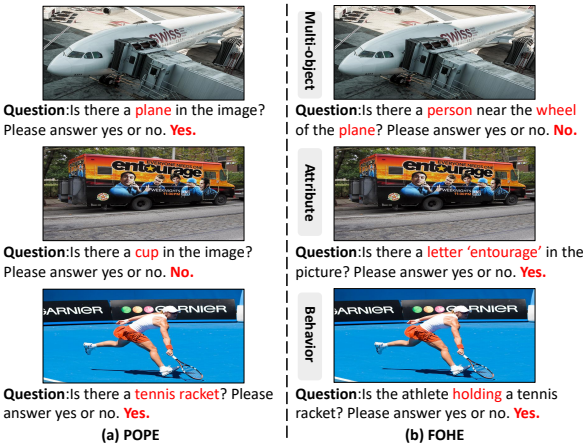


Figure 3: Examples of evaluation data of POPE (a) and FGHE (b).

Type	Questions	Multi-Object	Attribute	Behavior
Yes	100	47	45	8
No	100	51	42	7
Total	200	98	87	15

Table 1: Statistics of FGHE Evaluation Dataset.

5 Experiment

5.1 Experimental Setup

Model Settings. Since the proposed ReCaption is model-agnostic to instruction-tuned LVLMs, we can enrich any LVLMs with ReCaption. In this paper, we choose four open-sourced representative instruction-tuned LVLMs for evaluation: mPLUG-Owl (Ye et al., 2023), LLaVA (Liu et al., 2023c), Multimodal-GPT (Gong et al., 2023), and MiniGPT-4 (Zhu et al., 2023).

Implementation Details. A total of 500 image-caption pairs are randomly selected from a high-quality well-aligned image-text pair dataset named cc_sbu_align, which is curated by MiniGPT-4 (Zhu et al., 2023), for our study. To increase the diversity of the data, we generated 5 rewritten versions for each original caption. We employed the AdamW optimizer with a beta value of (0.9, 0.98) for optimization purposes. The learning rate and weight decay were set to 0.0001 and 0.1, respectively. During the training process, we initiated a warm-up phase consisting of 2,000 steps, after which we applied the cosine schedule to decay the learning rate. As for the input image, it was randomly resized to dimensions 224×224. We additionally fine-tune LVLMs 20 epochs with the batch size 256, and the learning rate is set to 0.00002.

5.2 Main Results

LVLMs with ReCaption have reduced hallucination. Overall, a remarkable improvement in performance can be observed across various LVLMs and evaluation metrics when the LVLM is used with ReCaption, as compared to the original LVLM. For instance, Mini-GPT4 with ReCaption demonstrates a significant improvement of F1 over Mini-GPT4 without ReCaption (7% improvement on POPE and 10% on FGHE). Among the LVLMs, mPLUG-Owl with ReCaption enjoys the least improvement in F1 score (with 1.32% improvement on POPE but more substantial 3.71% improvement on FGHE).

The results above demonstrate that our proposed ReCaption framework enhances the generation quality of LVLMs. It effectively reduce both coarse-grained and fine-grained hallucinations. The improvement on fine-grained hallucinations can be attributed to a strong alignment between images and text descriptions at the fine-grained level. Additionally, the proposed ReCaption approach is model-agnostic, allowing for effortless integration as a plug-and-play component during the training of LVLMs.

LVLMs with ReCaption reduce over-confidence. As shown in Table 2, based on the Recall of POPE and FGHE, it is evident that mPLUG-Owl, LLaVA, and MultiModel-GPT show a strong inclination to respond with the affirmative answer "Yes". For instance, both LLaVA and MultiModel-GPT provide "Yes" responses to most questions. The mPLUG-Owl also achieve a high recall rate of 94.34% on POPE and 84.42% on FGHE. It suggests that certain LVLMs exhibit a high degree of over-confidence and struggle to accurately identify objects, object attributes, and object behaviors in the given images. With ReCaption, LLaVA and MultiModal-GPT reduces their over-confidence.

FGHE serves as a different hallucination evaluation compared to POPE. Table 2 reveals different performances of LVLMs evaluated on POPE and FGHE, suggesting that LVLMs possess varying degrees of coarse-grained and fine-grained object hallucinations. For example, the mPLUG-Owl model attains an F1 score of 69.89% on POPE but only 68.42% on FGHE, indicating that mPLUG-Owl may excel in identifying objects rather than attributes or behaviors within a given image. Among the LVLMs, MultiModal-GPT performs the best for both coarse-grained and fine-grained object hallucinations.

Model	POPE				FGHE			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
mPLUG-Owl	59.37	55.51	94.34	69.89	55.56	57.52	84.42	68.42
w/ ReCaption	**61.79**	**57.11**	**94.61**	**71.22**	**62.22**	**62.26**	**85.71**	**72.13**
	(+2.42)	(+1.60)	(+0.27)	(+1.33)	(+6.66)	(+4.74)	(+1.29)	(+3.71)
LLaVA	50.00	50.00	**100.00**	66.67	56.83	56.52	**100.00**	72.22
w/ ReCaption	**61.87**	**57.33**	92.80	**70.88**	**67.63**	**65.42**	89.74	**75.68**
	(+11.87)	(+7.33)	(−7.80)	(+4.21)	(+10.80)	(+8.90)	(−10.26)	(+3.46)
MultiModal-GPT	50.00	50.00	**100.00**	66.67	57.56	56.93	**100.00**	72.56
w/ ReCaption	**62.40**	**57.60**	93.93	**71.41**	**75.83**	**72.31**	83.73	**77.60**
	(+12.40)	(+7.60)	(−6.07)	(+4.74)	(+18.27)	(+15.38)	(−16.27)	(+5.04)
MiniGPT-4	54.23	58.24	31.97	41.29	53.01	59.62	63.27	61.39
w/ ReCaption	**57.69**	**62.54**	**39.67**	**48.55**	**60.24**	**62.12**	**83.67**	**71.30**
	(+3.46)	(+4.30)	(+7.70)	(+7.26)	(+7.23)	(+2.50)	(+20.40)	(+9.91)

Table 2: Results of LVLMs w/o ReCaption and LVLMs w/ ReCaption on POPE and FGHE datasets. The best results are denoted in bold.

Method	Multi-Object	Attribute	Behavior
mPLUG-Owl	72.36	68.14	61.59
w/ ReCaption	**74.23**	**71.55**	**67.90**
	(+1.87)	(+3.41)	(+6.31)
LLaVA	74.26	71.49	66.45
w/ ReCaption	**76.70**	**75.18**	**71.82**
	(+2.44)	(+3.69)	(+5.37)
MultiModal-GPT	74.82	70.56	68.13
w/ ReCaption	**76.84**	**78.92**	**74.22**
	(+2.02)	(+8.36)	(+6.09)
MiniGPT-4	63.30	60.16	56.72
w/ ReCaption	**71.02**	**72.16**	**67.93**
	(+7.72)	(+12.00)	(+11.21)

Table 3: Evaluation over different hallucination categories in terms of F1 score of FGHE.

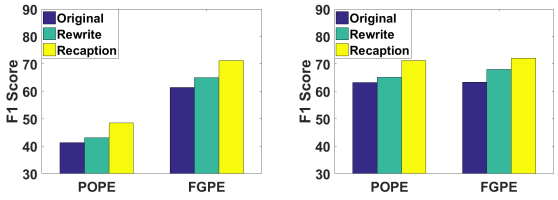


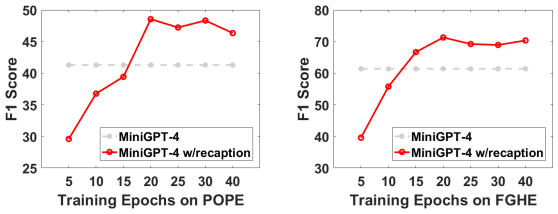
Figure 4: Performance comparison of MiniGPT-4 training with different caption rewriting strategies. Original means MiniGPT-4 without rewriting. Rewrite denotes MiniGPT-4 with a simple rewriting prompting (e.g, "Rewrite the following image description"). ReCaption means MiniGPT-4 with our proposed strategy.

5.3 Further Analysis

Break-down Study of Fine-Grained Object Hallucination. FGHE includes three types of binary questions, namely multiple objects, object attributes, and object behaviors. Table 2 depicts the F1 Scores of FGHE, combining the three types of questions. We further examine the hallucination organized by the three types of questions using the F1 scores.

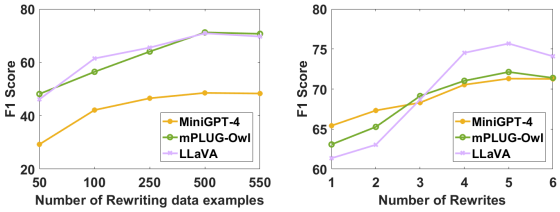
Table 3 presents the results, demonstrating that adding ReCaption to any LVLM reduces fine-grained hallucinations measured by the F1 scores computed over three question categories. This finding suggests that ReCaption is effective for reducing fine-grained hallucination generation by enriching alignment between images and texts through additional training with rewritten captions. Furthermore, ReCaption is also more effective in reducing hallucination in object attributes and behaviors than in multiple objects. One possible reason is that the rewritten captions may contain more verb and adjective keywords, thereby reducing hallucination, especially in object attributes and behaviors.

Varying Rewriting Strategies. Figure 4 depicts a comparison between our proposed two-stage caption rewriting strategy and a simple rewriting method (called Rewrite) employing ChatGPT (Fan et al., 2023). We select MiniGPT-4 as the model of choice since the data used for rewriting is based on image-text pairs collected by MiniGPT-4. The latter uses the prompt "Rewrite the following image description" to generate a new caption without any constraint. The performance of this simple rewriting technique exhibits some improvements compared to the original LVLMs without rewrites. However, this improvement is relatively minor. In contrast, our proposed two-stage rewriting prompting approach yields substantial improvements, highlighting multiple diverse rewritten captions with the same keywords to guide the LVLM



(a) (b)

Figure 5: F1 Score Curves for MiniGPT-4 with and without ReCaption over training epochs. The evaluation datasets used are POPE (a) and FGHE (b).



(a) (b)

Figure 6: F1 Score Curves of three LVLMs with our ReCaption as the number of images used for rewriting and the number of rewrites per image increase. We evaluate hallucination degree of models using FGHE.

in establishing a more robust alignment between images and captions.

Number of Training Epochs. We examine the impact of the number of additional training epochs on hallucination reduction of LVLMs with ReCaption. Figure 5 presents the F1 Score curves in relation to the training epochs for MiniGPT-4 with ReCaption on POPE and FGHE. The figure illustrates that the F1 Score of MiniGPT-4 with ReCaption increases as the number of training epochs increases. This suggests that conducting further training using this limited number of rewritten data pairs can enhance the fine-grained alignment between images and text, consequently reducing the generation of hallucinations by LVLMs.

Number of Images Used for Rewriting. Figure 6a illustrates the impact of the number of images utilized for rewriting in additional training of LVLMs, including MiniGPT-4, mPLUG-Owl, and LLaVA. All are evaluated in terms of F1 Score of FGHE. In the figure, a value of 50 indicates LVLMs trained with ReCaption incorporating 50 distinct images and their respective rewritten captions. The remaining hyperparameters, such as the number of rewritten captions per image, remain unchanged and fixed. The results strongly indicate that our proposed ReCaption consistently enhances the performance of all LVLMs as more images are employed for rewriting.

Number of Rewrites per Image. In Figure 6b, we observe the effect of the number of rewritten captions per image on LVLMs, such as MiniGPT-4, mPLUG-Owl, and LLaVA. We use the F1 Score of FGHE to evaluate them. Increasing the number of rewritten captions per image is expected to enhance the alignment between input images and output captions. This is because more output captions containing the same keywords are utilized for training the mapping from image to text. The findings strongly support that our proposed ReCaption consistently improves the hallucination reduction ability of all LVLMs as more rewritten captions per image are used during training.

Case Study. In the appendix, we show more examples of image caption generation using four vanilla LVLMs and these LVLMs with our ReCaption. Overall, adding ReCaption to LVLMs yields superior quality captions with fewer hallucinated objects, object attributes, and object behaviors. Conversely, based on the provided examples, Mini-GPT4 performs poorly in generating accurate descriptions of the given image, while MultiModal-GPT and LLaVA are prone to generating irrelevant content. Despite vanilla mPlug-Owl displaying a relatively higher-quality caption generation compared to the other three LVLMs, ReCaption still allows it to generate more accurate image captions.

6 Conclusion

In this paper, we aim to address the issue of fine-grained object hallucinations in instruction-tuned large vision-language models (LVLMs). We introduced a framework called ReCaption, which comprises two components: caption rewriting using ChatGPT and fine-tuning of LVLMs based on the rewritten captions. To evaluate the effectiveness of our approach, we proposed a fine-grained probing-based evaluation dataset and method called Fine-Grained Object Hallucination Evaluation (FGHE). Our experimental results demonstrate that ReCaption can effectively mitigate fine-grained object hallucinations across various LVLMs, thereby enhancing text generation quality. Future work could focus on refining and expanding the evaluation metrics to support more comprehensive evaluation of LVLM performance in hallucination reduction.

One should also explore more effective rewriting techniques to enrich alignment between unnatural images, such as invoices and cartoon pictures, and longer output text.

7 Limitations

Although ReCaption effectively mitigates fine-grained hallucination in VLVMs, it remains subject to certain limitations. Firstly, the rewriting solely relies on keywords, such as verbs, nouns, and adjectives, disregarding structural information and the interrelations among the keywords. This limitation hinders the preservation of crucial details from the original image caption. Secondly, to train any LVLM, a small corpus of alignment-enriched rewrites is required, despite the minimal training costs involved.

8 Ethics Statement

To evaluate the proposed ReCaption, this paper presents a small sef to evaluation dataset FOHE, and we discuss some related ethical considerations here. (1) Intellectual property. FOHE is collected from the validation set of MSCOCO, which is licensed under Creative Commons Attribution 4.0. This license lets you build upon our evaluation. (2) Treatments. We annotate the small evaluation data by our research group with agreed salaries and workloads. (3) Controlling Potential Risks. The texts in FOHE do not involve private information and social issues.

References

Harsh Agrawal, Peter Anderson, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8947–8956. IEEE.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. 2022. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 2473–2482. IEEE.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NeurIPS*.

Shuyang Cao and Lu Wang. 2021. Cliff: Contrastive learning for improving faithfulness and factuality in abstractive summarization. *EMNLP*.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2022. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. *ArXiv*, abs/2210.07688.

Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *EMNLP*.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving clip training with language rewrites. *arXiv preprint arXiv:2305.20088*.

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*.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2023. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2023. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

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

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*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

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

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

OpenAI. 2022. Introducing chatgpt. https://openai.com/blog/chatgpt.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *EMNLP*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13668 of *Lecture Notes in Computer Science*, pages 146–162. Springer.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving AI tasks with chatgpt and its friends in huggingface. *CoRR*, abs/2303.17580.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *EMNLP*.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2021. Confit: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. *arXiv preprint arXiv:2112.08713*.

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

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023a. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. 2023b. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671.

Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, et al. 2021. A controllable model of grounded response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14085–14093.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023a. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023b. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2022. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *arXiv preprint arXiv:2212.10773*.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. MM-REACT: prompting chatgpt for multimodal reasoning and action. *CoRR*, abs/2303.11381.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yi Zhou, Junyan Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qiang Qi, Ji Chao Zhang, and Feiyan Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

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

A Appendix: More Examples

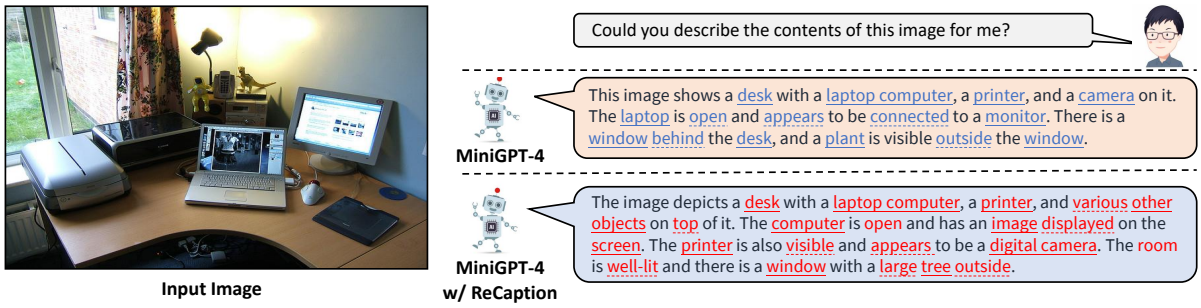


Figure 7: An illustrative example is presented to compare the output of MiniGPT-4 and MinitGPT-4 with ReCaption.

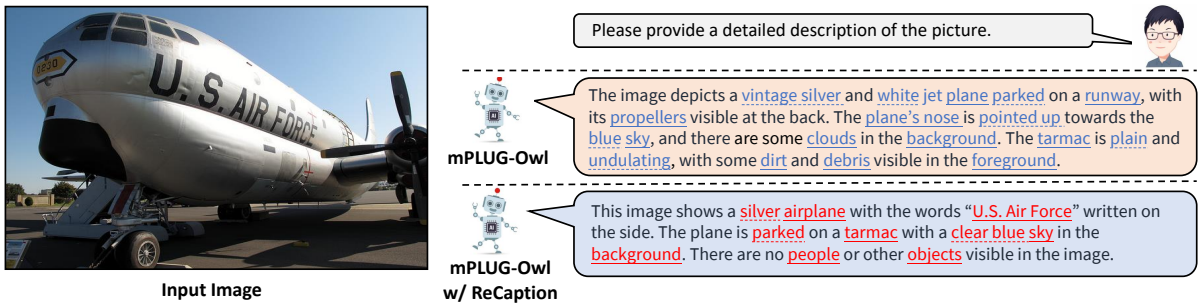


Figure 8: An illustrative example is presented to compare the output of mPLUG-Owl and mPLUG-Owl with ReCaption.

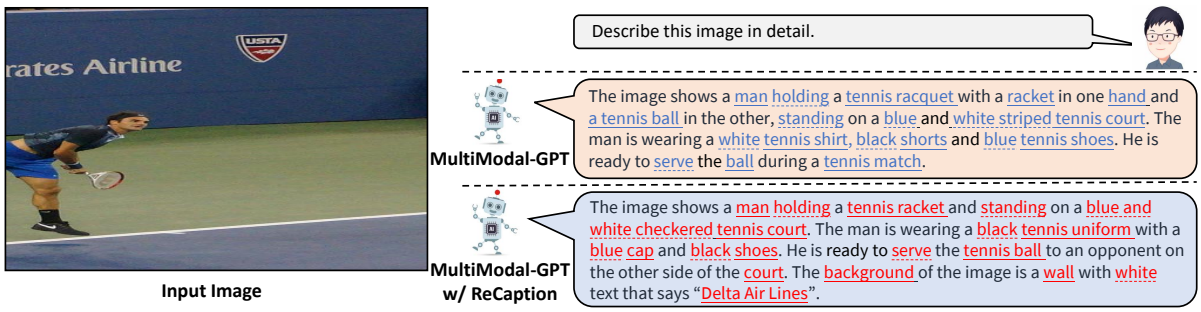


Figure 9: An illustrative example is presented to compare the output of MultiModal-GPT and MultiModal-GPT with ReCaption.

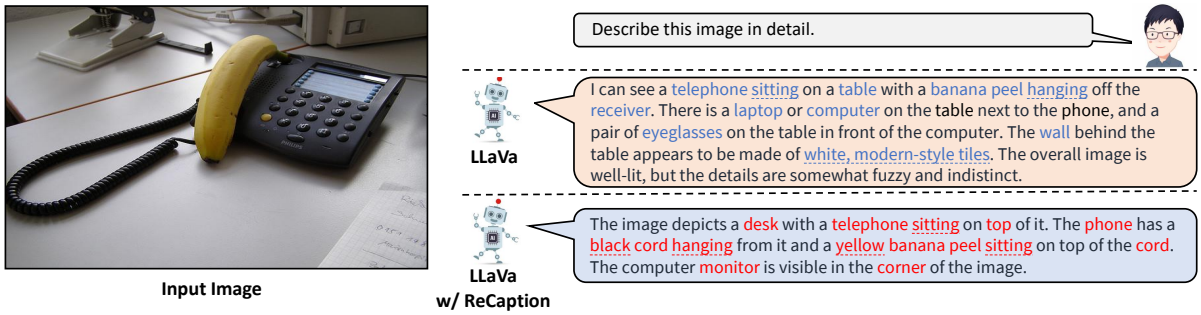
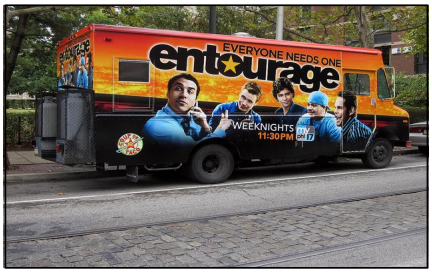


Figure 10: An illustrative example is presented to compare the output of LLaVa and LLaVa with ReCaption.



Input Image

Describe this image in detail.

MiniGPT-4

The image depicts a large red and black truck parked on a city street. The truck has the words "Entourage" written on the side in white letters. There are several people standing on the sidewalk in front of the truck. A few cars are parked on the street behind the truck. The background of the image is a mix of buildings and trees.

MiniGPT-4 w/ ReCaption

The image shows a large red and black advertising truck with the word "Entourage" written on it parked on a dirt road next to a building. The building has windows and doors, and trees are visible in the background. The truck is parked on the side of the road

Figure 11: An illustrative example is presented to compare the output of MiniGPT-4 and MinitGPT-4 with ReCaption.

Input Image

Could you describe the contents of this image for me?

mPLUG-Owl

The image is of a white boat in the middle of a large body of water. The boat appears to be in the process of moving through the water, with its propeller spinning in the water. There is a reflection of the sky and a cloud in the water behind the boat. The image is taken from a low angle, looking up at the boat from below the water's surface. There is a hint of a wake behind the boat, indicating that it is moving through the water at some speed.

mPLUG-Owl w/ ReCaption

The image depicts a small, sleek white car driving smoothly on a smooth, white body of water with waves gently lapping against the side. The car has a roof that can be raised or lowered according to the driver's preference.

Figure 12: An illustrative example is presented to compare the output of mPLUG-Owl and mPLUG-Owl with ReCaption.

Input Image

Please provide a detailed description of the picture.

MultiModal-GPT

This image appears to show a group of people on skis in a snowy forest. The people are wearing ski gear and are holding poles. One person is carrying a backpack. The trees in the background are covered in snow, and there are tracks in the snow leading off to the right of the image. There is a clear blue sky in the background, and the sun is shining from the left side of the image.

MultiModal-GPT w/ ReCaption

The image depicts a group of people on skis walking through a wooded area with trees on either side of a path. The people are dressed in ski gear, including jackets, pants, gloves, and hats, with poles in their hands. The person in the middle is wearing a cap and has a beard, while the others are older men with more hair. The sky is cloudy and the sun is shining through the clouds. The terrain is covered in snow and there are trees on either side of the path.

Figure 13: An illustrative example is presented to compare the output of MultiModal-GPT and MultiModal-GPT with ReCaption.

Input Image

Take a look at this image and describe what you notice.

LLaVa

This image is of a restaurant sign hanging from a pole in front of a building. The sign reads "Restaurant" and is made of metal with black letters on a green background. There is also a smaller sign above it that reads "Pizza" in white letters on a red background. There are trees and a sky in the background.

LLaVa w/ ReCaption

The image shows a sign for a restaurant called "Rafael's Restaurant" located on a street corner. The sign is mounted on a pole and features green and white lettering with black accents. Above the sign, a leafy tree can be seen, and the building it is outside of is a modern structure.The image was taken on a clear day, with a bright blue sky and a few clouds visible in the background.

Figure 14: An illustrative example is presented to compare the output of LLaVa and LLaVa with ReCaption.