Investigating and Mitigating Object Hallucinations in Pretrained Vision-Language (CLIP) Models

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) have achieved impressive performance, yet research has pointed out a serious issue with object hallucinations within these models. However, there is no clear conclusion as to which part of the model these hallucinations originate from. In this paper, we present an in-depth investigation into the object hallucination problem specifically within the CLIP model, which serves as the backbone for many state-of-the-art vision-language systems. We unveil that even in isolation, the CLIP model is prone to object hallucinations, suggesting that the hallucination problem is not solely due to the interaction between vision and language modalities. To address this, we propose a counterfactual data augmentation method by creating negative samples with a variety of hallucination issues. We demonstrate that our method can effectively mitigate object hallucinations for CLIP model, and we show the the enhanced model can be employed as a visual encoder, effectively alleviating the object hallucination issue in LVLMs. [1]

1 Introduction

Current Large Vision-Language Models (LVLMs) demonstrate significant potential in tasks requiring joint visual and linguistic perception, such as image captioning (Agrawal et al., 2019b), visual question answering (Antol et al., 2015), visual grounding (Yu et al., 2016), and autonomous agents (Durante et al., 2024; Xi et al., 2023). Despite the success of LVLMs, previous studies have revealed that they commonly suffer from hallucinations in practice, including object hallucinations (Li et al., 2023b; Leng et al., 2023; Zhou et al., 2023), spatial hallucinations (Kamath et al., 2023), attribute hallucinations (Zhang et al., 2024), etc. It is widely believed that hallucinations degrade model performance and reliability, and severely impair the user experience in real-world applications (Ji et al., 2023).

In this work, we focus on investigating the causes of the highly-concerned *object hallucinations*, i.e., LVLMs generate nonexistent objects in the image (Biten et al., 2022). A typical LVLM utilizes a Large Language Model (LLM) as its cognitive foundational model and employs a pre-trained image encoder as its visual perception module (mainly the CLIP encoder). Kamath et al. (2023) investigated the spatial hallucination (e.g., confusing "left of" and "right of") in LVLMs, and they found that various CLIP encoders struggle to recognize simple spatial relationships (achieving only a 55.0% accuracy on benchmarks, whereas humans are 98.8%). Inspired by their findings, we hypothesize that the CLIP visual encoder might also be one of the causes of object hallucinations.

Hence, we first curate the **O**bject **H**allucination **D**etection (**OHD**-Caps) benchmark from subsets of the COCO (Lin et al., 2014), Flickr30K (Young et al., 2014), and Nocaps (as an out-of-domain benchmark because it comprises unseen objects) (Agrawal et al., 2019a) image caption datasets respectively, to more strictly measure the extent of object hallucinations present in CLIP encoders. We randomly select 16k/1k/1.5k (train/dev/test) samples, with each sample containing one image, one positive descriptive text, and 27 negative descriptive texts. The negative samples are perturbations of the positive sample, achieved by *adding* descriptions of nonexistent objects or *reducing* descriptions of existing objects. Theoretically, a CLIP model without object hallucinations should accurately assign the highest CLIP score to the positive sample. However, taking the most commonly used "CLIP ViT-L/14" in LVLMs as an example, it only scores the highest for positive samples in 19.0% of cases. Since we have observed that the CLIP encoder already has a serious issue with object hallucination, how can we mitigate it?

[1] Our benchmark and code are publicly available on https://anonymous.4open.science/r/clip_hallucination-71EC.

In the contrastive pretraining of CLIP, negative samples come from text descriptions of other images within the batch, which makes the distinction between them quite straightforward. However, mitigating object hallucinations requires the CLIP encoder to be able to differentiate between subtle errors at the object level. We further fine-tune the CLIP model using the training set from **OHD**-Caps. By incorporating a fine-grained object-level contrastive loss, we greatly reduce object hallucinations in the CLIP. Then employing the fine-tuned CLIP as the visual encoder, the object hallucinations in our retrained LVLM, LLaVA-1.5, are also diminished.

In this paper, we study the object hallucinations of CLIP models. Our main contributions are,

- we propose a benchmark, **OHD**-Caps, for evaluating object hallucinations in CLIP models.
- we quantitatively evaluate a wide range of encoders from the CLIP family and find that they all exhibit severe object hallucination issues.
- we propose a fine-grained object-level contrastive loss to further fine-tune the CLIP model, significantly alleviating its object hallucination issues (e.g., from 28.7 to 83.2 for "CLIP ViT-B/32") and concurrently reducing the hallucination problems of the LLaVA-1.5 (from 80.3 to 82.4 on Nocaps), which uses it as a visual encoder.

2 Related Work

2.1 Large Vision-Language Model

Recently, inspired by the success of large language models (LLMs), researchers have begun to dedicate efforts to enhance vision language models (VLMs) by integrating robust LLMs, aiming to broaden the knowledge scope of the model and amplify its linguistic comprehension capabilities.

LVLM architectures typically consist of three components: a visual encoder, a modality connection module, and a LLM. The visual encoder and LLM are typically fixed large pretrained models, the visual encoder is usually a variant of the CLIP model (Radford et al., 2021), used for extract visual features, while the LLM, such as LLaMA (Touvron et al., 2023) and Vicuna (Chiang et al., 2023), is used to integrate image information and text information, and completes the prediction of the target. Research focuses on optimizing modality connection modules, with approaches like Flamingo's (Alayrac et al., 2022) cross-attention module, LLaVA's (Liu et al., 2023b) linear layer, and BLIP2's (Li et al., 2023a) Q-former, diverse yet all boosting VLM performance on various vision-language tasks.

2.2 Hallucination in LVLMs

Despite the fact that LVLMs perform well in solving visual-language tasks, they are also plagued by hallucinations. The problem of hallucinations in LVLMs mainly refers to the mismatch between visual input and textual output. For example, in the image captioning task, hallucination refers to the generation of captions that describe objects that do not exist in the image. Although the hallucination problem of LLMs has been widely studied in the NLP field (Ji et al., 2023), there has not been enough research on mitigating the hallucination issue in LVLMs (Liu et al., 2024). Recent efforts to mitigate hallucination in LVLMs have focused on enhancing each compoment of the model. For example, (Liu et al., 2023a; Hu et al., 2023) constuct instruction-tuning datasets with contrastive question-answer pairs for LVLMs; (Sun et al., 2023; Yu et al., 2023) employ Reinforcement Learning from Human Feedback (RLHF) (Stiennon et al., 2020) to enchance the connection module between the modalities; (Leng et al., 2023) propose a visual contrastive decoding strategy for LLM decoing. Despite the wide application of the CLIP model in VLMs and its in-depth study in pairwise comparison context (Yüksekgönül et al., 2023; Hsieh et al., 2023), there has been little discussion on its evaluation regarding hallucinations. Our research addresses this gap in the literature.

3 The OHD-Caps Benchmark

Recent studies have found that LVLMs are prone to object hallucinations (Li et al., 2023b; Zhou et al., 2023). In response, researchers have developed several datasets to assess the extent of these hallucinations in such models (Li et al., 2023b; Wang et al., 2023). However, there is a relative lack of assessment work regarding the hallucinatory effects of the CLIP model, which is widely used as a visual encoder within LVLMs. In this section, we introduce the **O**bject **H**allucination **D**etection benchmark (OHD-Caps) we create to evaluate the object hallucination problem in CLIP models and the pipeline for evaluations. Figure 1 shows the pipeline of our benchmark creation process.

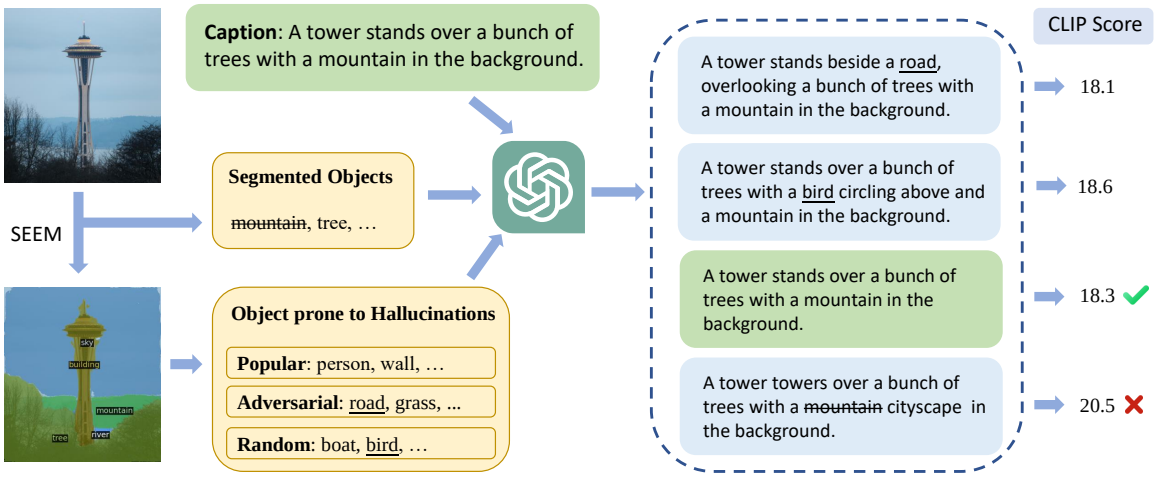

Figure 1: The pipeline of our benchmark creation process. For an image, we first use SEEM (Zou et al., 2023) to identify objects within the image and obtain illusory objects that do not exist in the picture through different sampling strategies. Then we ask GPT to insert or delete objects in the original sentences to create negative samples. We provide both positive and negative samples to the CLIP model to observe if the model predicts the positive samples as having the highest score. This image is from the Nocaps dataset, and the model is CLIP ViT-B/32.

3.1 Dataset Construction

CLIP is a versatile neural network that excels at image understanding and can predict text for images in a zero-shot manner. To evaluate the CLIP model's ability to handle object hallucinations in paired comparison scenarios, given an image with a correct caption, we create incorrect captions containing hallucinatory content. The purpose is to observe whether the model can accurately select the correct text without hallucinations.

Inserting Hallucinatory Objects Previous work (Li et al., 2023b; Zhou et al., 2023) show that LVLMs are more prone to generate hallucinatory responses for objects that frequently appear in the dataset. Inspired by this, we create negative samples by inserting objects prone to hallucination into the correct captions. To collect object annotations, we first use SEEM (Zou et al., 2023) to automatically segment objects in the images. Three kinds of hallucinatory objects are collected: *random objects* which are sampled randomly, *popular objects* which are the top frequent objects in the whole dataset, and *adversarial objects* which are the top frequent objects with the segmented objects. Each category contains three objects. To create examples with varying levels of hallucinations, we attempt to insert one to three objects for each category, resulting in each type of hallucination containing a total of 7 ($\sum_{r=1}^{3} C_3^r$) samples.

Given a caption text and several hallucinatory objects, we insert the objects into the appropriate locations in the caption, which can be effectively achieved by the help of GPT4. In an automatical way, the caption and objects are fed to the GPT4, with the prompt as follows: *Given a sentence {caption}, generate a new sentence and includes each object from the list {objects}. Make the changes to the original sentence as minimal as possible. Ensure that the new sentence is coherent, natural, semantically smooth and free of grammatical errors.*

Removing existing Objects Except from inserting hallucinatory objects, we also remove objects from the captions to create negative samples. We randomly select 1 or 2 segmented objects in the image which results in 6 negative samples ($\sum_{r=1}^{2} C_3^r$), and ask GPT4 to remove them from the caption with the prompt: *Given a sentence {caption}, generate a new sentence and remove each object from list {objects} to make the semantics of the sentence different. Ensure that the new sentence is coherent, natural, semantically smooth and free of grammatical errors.* To account for scenarios where the identified objects are not present in the title text, we ask GPT to alter elements like objects, colors, and properties in the original caption: *Given a sentence {caption}, choose to modify the objects, colors, attributes, etc., within the sentence to make the semantics of the sentence different. Make the changes to the original sentence as minimal as possible. Ensure that the new sentence is coherent,*

natural, semantically smooth and free of grammatical errors.

we construct a dataset of 500 samples for each of the COCO (Lin et al., 2014), Flickr30K (Young et al., 2014), and the out of domain subset of NoCaps Validation datasets (Agrawal et al., 2019a), with 27 negative samples for each image. Specifically, the out of domain subset of NoCaps comprises objects not seen in the COCO dataset, commonly used to measure a model's ability to generalize to unseen classes. The average length of the captions in the datasets is shown in Table 7.

3.2 Evaluation and Analysis

We study several models to evaluate their performance on our benchmark. Each image is paired with a correct caption and 27 negative samples, and models are required to calculate the similarity between the image and the caption candidates and select the correct caption.

Models We evaluate a variety of models on our benchmark, including CLIP (Radford et al., 2021) ViT-B/32 and ViT-L/14; RoBERTaCLIP (Ilharco et al., 2021) which is a CLIP ViT-B/32 model initialized with RoBERTa-pretrained (Liu et al., 2019) weights; NegCLIP (Yüksekgönül et al., 2023), an improved model based on CLIP ViT-B/32, which enhances the understanding of relationships between objects, attributes, and the sequence of words by swapping phrases; CECLIP (Zhang et al., 2023) which further develop enhanced negative samples and employ contrastive loss to enhance compositional reasoning; FLAVA (Singh et al., 2022) which is a single unified foundation model which can work across vision, language as well as vision-and-language multi-modal tasks; CoCa (Yu et al., 2022) is a pretrained model with contrastive and generative learning objectives; XVLM (Zeng et al., 2021) which aligns the visual concept and textual input in a multi-grained manner with 14M and 16M pretrained images; BLIP (Li et al., 2022) which effectively utilizes the noisy web data by bootstrapping the captions with 14M and 129M pretrained images; BLIP2 (Li et al., 2023a) which bridges the gap between the visual and textual modalities with a Q-former.[2]

We also evaluate the performance of the models after fine-tuning on downstream tasks: CoCa finetuned on COCO captioning, and XVLM 14M and BLIP models respectively finetuned on Flickr30K retrieval and COCO retrieval.

[2]We use the image-text matching head for both BLIP and BLIP2.

Model	#Params	OHD-Caps Benchmark			
		COCO	Flickr30K	Nocaps	Avg.
CLIP ViT-B/32	151M	15.2	17.6	10.2	14.3
CLIP ViT-L/14	428M	22.4	22.6	12.0	19.0
RoBERTaCLIP	213M	1.0	1.6	1.0	1.2
NegCLIP	151M	32.8	28.0	25.0	28.6
CECLIP	151M	52.8	40.8	23.4	39.0
FLAVA	350M	28.0	28.4	16.6	24.3
CoCa	2.1B	26.0	24.4	20.0	23.5
XVLM 4M	216M	46.4	35.8	34.0	38.7
XVLM 16M	216M	41.8	19.4	21.8	27.7
BLIP 14M	583M	51.4	**48.0**	**42.0**	47.1
BLIP 129M	583M	40.8	38.0	31.2	36.7
BLIP2	3.4B	**62.6**	42.2	41.2	**48.7**
CoCa-Caption	2.1B	6.8	5.6	6.8	6.4
XVLM-Flickr30K	216M	62.6	**60.4**	41.6	**54.9**
XVLM-COCO	216M	**68.2**	47.6	**47.6**	54.5
BLIP-Flickr30K	583M	53.6	52.0	38.4	48.0
BLIP-COCO	583M	59.2	47.2	41.2	49.2

Table 1: Results of varied models on our benchmark: models in the first section are evaluated in zero-shot, and models in the second section have been finetuned on some downstream task: COCO captioning, image-text retrieval on Flickr30K or COCO.

Results Table 1 shows the results of the models on our benchmark. From the results, we could find that,

- First of all, the vanilla CLIP models (CLIP ViT-B/32, CLIP ViT-L/14, RoBERTaCLIP) perform poorly across all three datasets, indicating their limited ability to recognize illusory objects in images. On the other hand, NegCLIP attempts to enhance the model's understanding of text by parsing and substituting phrases, but it only achieves a marginal improvement compared to the original CLIP model. CECLIP exhibits relatively better performance, which is mainly due to the constructed negative samples enhancing the model's comprehension of the combined semantics of sentences. The NegCLIP and CECLIP models are trained on the COCO training set to distinguish between positive samples and enhanced negative samples. This might contribute to CECLIP's good performance on the COCO dataset, owing in part to the model's memory of the original correct text. However, their performance on the Nocaps dataset indicates that these models lack the ability to effectively differentiate hallucinated objects.

- Secondly, generative vision-language models typically achieve higher performance than vanilla CLIP models due to their more precise alignment

of image and text representations. Furthermore, it is generally observed that the larger the model parameters, the better the performance. In particular, BLIP2, which has the highest number of parameters, performs best across all three datasets. In comparison, the XVLM 4M model has relatively fewer parameters but still demonstrates good performance. This indicates that XVLM's strategy of multi-scale alignment indeed assists the model in more accurately capturing the fine-grained details within images.

- Furthermore, the overall trend among different models is consistent across the three datasets, with their performance typically being the lowest on the Nocaps dataset. Although fewer objects are recognized on the Nocaps dataset than Flickr30K, the performance is the lowest there due to the inclusion of categories that are out-of-domain. The BLIP 14M model demonstrates the best performance on both Flickr and Nocaps, which indicates its strong generalization capabilities.

- Finally, under normal circumstances, models usually experience a improvement in performance after being fine-tuned on downstream tasks, with the CoCa model being an exception. Moreover, these performance enhancements can also be generalized to other datasets.

Analysis The inability of models to recognize hallucinated objects primarily stems from the data used and the learning methods employed. The vanilla CLIP model is trained with a large number of image-caption pairs collected from the internet, using a contrastive loss function for optimization. Those captions are often brief and noisy, and the model is optimized to differentiate between correct and a multitude of incorrect image-text pairs. However, because the incorrect pairs are usually significantly different from the correct ones, the model can easily distinguish them. This means that the model does not need to learn the rich details in the pictures to make accurate predictions. To address this issue, we need to make improvements to the original CLIP model in terms of data utilization and learning methodologies.

4 Methodology

We first revisit the training process of vanilla CLIP model. Let I be the image and T be the text, the training objective of CLIP is to maximize the similarity between the image and text pairs, and minimize the similarity between the image and text pairs that are not matched. The loss function is defined as:

$$\begin{aligned}\mathcal{L}_{i2t} &= -\log \frac{\exp(I \cdot T^{+}/\tau)}{\sum_{T^{-}} \exp(I \cdot T^{-}/\tau)}, \\ \mathcal{L}_{t2i} &= -\log \frac{\exp(T \cdot I^{+}/\tau)}{\sum_{I^{-}} \exp(T \cdot I^{-}/\tau)}, \\ \mathcal{L}_{0} &= \frac{1}{2}\big(\mathcal{L}_{i2t} + \mathcal{L}_{t2i}\big),\end{aligned} \tag{1}$$

where T^{+} and I^{+} are the correct text and image, and T^{-} and I^{-} are the incorrect text and image, respectively.

With the addition of the negative samples T^{neg} created as in the previous section, we can expand T^{-} as $T^{*} = \{T^{-}, T^{neg}\}$. Then we could modify the loss $\mathcal{L}_{i2t}$ as:

$$\mathcal{L}_{i2t} = -\log \frac{\exp(I \cdot T^{+}/\tau)}{\sum_{T^{*}} \exp(I \cdot T^{*}/\tau)}. \tag{2}$$

To further enhance the model's ability to distinguish between positive and negative samples, we additionally introduce a margin loss. This is to ensure that the distance between an image and its corresponding correct text is smaller than the distance to incorrect text by a specific threshold. This concept can be formulated as:

$$\mathcal{L}_{1} = \max(0, \tau_{1} - I \cdot T^{+} + I \cdot T^{*}), \tag{3}$$

where τ_1 is the margin threshold.

Additionally, we generate enhanced negative samples by introducing perturbations to the original positive samples. Such negative samples are typically more challenging to distinguish than other negative samples within the batch. To encourage the model to recognize the partially correct information contained in the enhanced negative samples, resulting in a higher similarity to the positive samples compared to other negative samples within the batch, we introduce a margin loss between the in-batch negative samples and the enhanced negative samples:

$$\mathcal{L}_{2} = \max(0, \tau_{2} - I \cdot T^{-} + I \cdot T^{neg}), \tag{4}$$

where τ_2 is the margin threshold.

Next, we assign different weights to the aforementioned loss terms, allowing the model to learn adaptively. Consequently, the final loss function can be expressed as follows:

$$\mathcal{L} = \frac{1}{2}\big(\mathcal{L}_{t2i} + \mathcal{L}_{i2t}\big) + \lambda_{1}\mathcal{L}_{1} + \lambda_{2}\mathcal{L}_{2}. \tag{5}$$

5 Experiments

Training Datasets In order to enable the model to possess not only compositional understanding capability but also the ability to recognize illusory objects in images, we combine data featuring compositional understanding with a dataset for hallucination recognition dataset that we create. We start with the COCO dataset's training set, [3] following the methods (Zhang et al., 2023), we generate four types of negative samples for each image. These negative samples are designed to enhance the model's recognition of relationship, attribution, actions, and objects, respectively. To create negatives samples for relationship, we use Spacy (Honnibal and Montani, 2017) to get Parts-of-Speech (POS) tag and swap the positions of two noun words in the sentence. For the enhancement of attribution, actions, and objects, we randomly mask adjectives, verbs, or nouns in the sentences and employ the RoBERTa model to fill in these masked words. For hallucination recognition, we sample 8k images from training set of COCO and 8k images from Flickr30k datasets, then generate negative samples for each image as in Section 3. Additionally, we randomly select ∼2k samples from the COCO dataset's validation set as our dev set for compositional understanding (∼1k) and hallucination recognition (∼1k).

Training Details We utilize the CLIP ViT/32-B and CLIP ViT/14-L-336px implemented by Huggingface (Wolf et al., 2020) as the initial models and conduct fine-tuning for three epochs. The best-performing model is selected based on its performance on the validation set. The training process is carried out on a single A100 GPU, with batch sizes of 64 and 16 set for the base and large models, respectively, and the learning rate is set at 1e-5. The selection of hyper-parameters is determined by their performance on the validation set, where λ_1 and λ_2 are set as 0.3 and 0.2, τ_1 and τ_2 are set as 5.

Evaluation We evaluate our fine-tuned CLIP models on two common Visual Language (VL) combination benchmarks: ARO (Yüksekgönül et al., 2023) and SugarCrepe (Hsieh et al., 2023), and the OHD-Caps benchmark we create. The ARO benchmark contains more than 50,000 test cases, is designed to systematically assess the capabilities of VLMs in comprehending various types of relationships, attributes, and sequential information through tests focused on object properties and relational understanding within the Visual Genome dataset (Krishna et al., 2017). The SugarCrepe benchmark is an enhanced version of CREPE (Ma et al., 2023) that mitigates bias issues which uses large language models to generate hard negatives with human validation. Both ARO and SuparCrepe datasets require classifying positive and negative captions for a given image, with a random success probability of 50%.

[3] To prevent information leakage, we exclude 8k samples that are subsequently used to create the hallucination dataset.

5.1 Main Results

We present the results for ARO dataset and our self-constructed dataset in Table 2, and SuparCrepe in Table 3. From the results, we could find:

- Our model shows comparable performance to previously state-of-the-art model (CECLIP) on both datasets for compositional understanding and achieves significant improvements in hallucination recognition. The performance of the CLIP models on the ARO dataset, as well as the hallucination detection dataset, is relatively poor and close to the performance of random guessing, indicating that the model lacks a fine-grained understanding of images. NegCLIP and CECLIP enhance the model's capability of understanding composites by constructing negative samples, and also make progress on the hallucination detection dataset, achieves a moderate improvement on OHD-Caps benchmark, with performance rising from 14.3% to 39.0%. Our model, while being comparable in compositional understanding to CECLIP, further enhances the performance of hallucination detection to 83.2%.

- Our model also demonstrates strong generalization capabilities in hallucination recognition. NegCLIP, CECLIP, and our model are all fine-tuned on the training set of the COCO dataset. Although they show varying degrees of performance improvement in COCO-related hallucination tests (NegCLIP at 32.8%, CECLIP at 52.8%), their performances are worse when facing unknown categories (NegCLIP at 25.0%, CECLIP at 23.4% for Nocaps images), indicating limited generalization capabilities of the models. In contrast, our model performs consistently across three different datasets, at approximately 83%. This result verifies that our model can effectively distinguish hallucinated objects in dif-

Model	ARO			OHD-Caps			
	Relation	Attribute	Avg.	COCO	Flickr30k	Nocaps	Avg.
Radom Chance	50.0	50.0	50.0	3.6	3.6	3.6	3.6
CLIP ViT-B/32	59.3	62.8	61.1	15.2	17.6	10.2	14.3
NegCLIP	80.2	70.5	75.4	32.8	28.0	25.0	28.6
CECLIP	83.0	**76.4**	79.7	52.8	40.8	23.4	39.0
Ours w/o object	83.7	74.7	79.2	39.8	24.2	22.0	28.7
Ours	**83.8**	76.3	**80.1**	**82.6**	**85.0**	**82.0**	**83.2**
CLIP ViT-L/14-336px	62.7	62.0	62.4	26.0	27.0	16.8	23.3
Ours w/o object	**85.2**	**76.3**	**80.8**	50.6	35.2	23.4	36.4
Ours	84.6	**76.3**	80.4	**89.0**	**88.0**	**81.6**	**86.2**

Table 2: Results(%) on the ARO dataset and our OHD-Caps benchmark. The ARO dataset evaluates the model's accurate understanding of relationships and attributes by swapping the positions of two objects. The table is divided into two sections, which respectively show the results obtained from fine-tuning on the CLIP ViT-B/32 and CLIP ViT-L/14-336px configurations. 'w/o object' means without the data we create for object hallucination. In each section, the best results are highlighted in bold.

Model	REPLACE				SWAP			ADD		
	Object	Attribute	Relation	Avg.	Object	Attribute	Avg.	Object	Attribute	Avg.
Human	100.0	99.0	97.0	98.7	99.0	100.0	99.5	99.0	99.0	99.0
Random	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
CLIP ViT-B/32	90.9	80.1	69.2	80.1	61.2	64.0	62.6	77.2	68.8	73.0
NegCLIP	92.7	85.9	76.5	85.0	75.5	75.4	75.5	88.8	82.8	85.8
CECLIP	93.1	88.8	79.0	87.0	72.8	77.0	74.9	**92.4**	**93.4**	**92.9**
Ours w/o object	92.4	88.6	75.7	85.6	**77.1**	**77.8**	**77.5**	82.2	84.4	83.3
Ours	**94.1**	**89.6**	**80.6**	**88.1**	76.3	77.0	76.7	89.9	85.1	87.5
CLIP ViT-L/14-336px	94.5	80.6	66.7	80.6	63.7	62.3	63.0	81.3	74.1	77.7
Ours w/o object	94.7	87.3	79.4	87.1	**78.0**	**77.5**	**77.7**	82.3	**85.8**	**84.1**
Ours	**95.2**	**89.8**	**82.9**	**89.2**	76.7	76.9	76.8	**86.1**	78.0	82.1

Table 3: Results(%) on SugarCrepe dataset. The SuparCrep dataset aims to test the model's ability to comprehend combinations by replacing, swapping, and augmenting concepts within the dataset.

ferent datasets and possesses the capability to generalize across datasets.

- With the increase in model parameters, upgrading from CLIP ViT-B/32 to CLIP ViT-L/14-336px, the model generally performs better on datasets involving the compositional understanding as well as the recognition of hallucinations, with a slight enhancement in performance. The only exception is observed in the SuparCrepe dataset, where there is a decline in performance on the subset that involves the insertion of attributes and objects. We observe that even without incorporating our constructed hallucination detection data, there is still a decline in performance during the evaluations. This could be due to an increased number of negative examples resulting in a reduced batch size.

5.2 Evaluation for LVLM

To verify the effectiveness of the enhanced CLIP model compared to the original CLIP in assisting large vision-language models to mitigate the issue of object hallucination, we replace the CLIP ViT-L/14-336px baseline model in LLaVA-1.5 with our fine-tuned version. We train LLaVA (Liu et al., 2023b) from scratch using the hyper-parameters specified in the original paper.

We conduct an evaluation of object hallucination phenomena on the expanded POPE dataset. The POPE dataset is created by selecting samples from the COCO validation set and constructing questions about hallucinated objects of various categories. The format of the questions is 'Is there a X in the image?', where X refers to the name of the object. The questions in the dataset are designed such that the objects are present and absent

Dataset	Criterions	LLaVA	Ours
COCO	Accuracy (↑)	85.2	**86.7**
	Precision (↑)	82.1	**86.5**
	Recall (↑)	**90.7**	87.5
	F1 Score (↑)	86.1	**86.9**
	Yes (→50%)	55.5	**50.8**
Flickr30K	Accuracy (↑)	73.3	**79.9**
	Precision (↑)	67.3	**75.4**
	Recall (↑)	**96.6**	91.2
	F1 Score (↑)	78.9	**82.2**
	Yes (→50%)	73.3	**61.3**
Nocaps	Accuracy (↑)	77.1	**81.3**
	Precision (↑)	71.7	**79.0**
	Recall (↑)	**91.7**	86.7
	F1 Score (↑)	80.3	**82.4**
	Yes (→50%)	64.7	**55.4**

Table 4: Results on expanded POPE datasets. Yes denotes the proportion of answering "Yes" to the given question. The best results in each block are denoted in bold.

in equal measure, therefore the ideal 'yes' response rate should be around 50%. To comprehensively assess the model's performance on various datasets, particularly on out-of-domain datasets, we expand the Flickr30k and Nocaps datasets following the original setup. Each dataset contains 500 images, with 18 questions associated with each image.

The results are shown in Table 4. It reveals that the LLaVA model, trained with the enhanced CLIP, achieves an improvement in the F1 score across three datasets, with the average performance increasing from 81.8 to 83.8. Apart from the Recall metric, our model surpasses the original LLaVA model in all other metrics. Compared to the original, it attains a better balance between accuracy and recall and also approaches a more ideal balance in the proportion of "Yes" responses. Moreover, although both models perform less impressively on the Flickr30k and Nocaps datasets compared to the COCO dataset, our model demonstrates a more significant advantage on these two datasets, thereby evidencing its superior generalization capability.

5.3 Ablation Study

In this subsection, we present ablation studies to examine the impact of our model's different components. We conduct these experiments on CLIP Vit-B/32 model.

Losses As demonstrated in Table 5, inclusion of the $\mathcal{L}_0$ loss alone significantly improve both the ARO and Object metrics over the baseline. Subsequently, iterative incorporation of $\mathcal{L}_1$ and $\mathcal{L}_2$ provide incremental benefits, with the full combination yielding the highest average performance. Compared to $\mathcal{L}_1$ loss, $\mathcal{L}_2$ loss has a more significant effect on improving model performance. This suggests that by increasing the distance between constructed negative samples and other negative samples in the batch, the model can achieve a more refined understanding.

Model	$\mathcal{L}_0$	$\mathcal{L}_1$	$\mathcal{L}_2$	ARO	Object	Avg.
CLIP				61.1	14.3	37.7
Ours	✓			78.0	82.1	80.1
	✓	✓		78.2	82.5	80.4
	✓		✓	80.0	83.1	81.6
	✓	✓	✓	**80.1**	**83.3**	**81.7**

Table 5: Ablation of losses on CLIP ViT-B/32.

λ_1 Values	λ_2 Values		
	0.2	0.3	0.4
0.2	81.3	81.5	81.6
0.3	**81.7**	81.4	81.5
0.4	81.6	81.2	81.3

Table 6: Ablation of λ_1 and λ_2 Values on Vit-B/32. The results are averaged on ARO and Object Datasets.

Weight of Losses Table 6 illustrates the changes in model performance when different loss weights are applied. The experimental results indicate that the sensitivity of model performance to weight changes is relatively low. The model demonstrates the best performance when the values of λ_1 and λ_2 are set to 0.3 and 0.2, respectively.

6 Conclusion

Our study investigate the reasons behind object hallucination in LVLMs. We construct a benchmark specifically for the evaluation of hallucinations and find that the visual perception module commonly used in current LVLMS, i.e., the CLIP model, cannot effectively discriminate hallucinated text. By designing negative samples and optimizing the contrastive loss function, we achieve a significant improvement in model performance on the hallucination detection dataset. Moreover, replacing the original CLIP model with our improved model can effectively alleviate the issue of object hallucination in LLaVA model.

7 Limitations

Although we conduct a series of explorations, our research still has its limitations. Firstly, our focus is solely on the issue of object hallucination within LVLMs, and we do not extend our research to other types of hallucinations. Secondly, the benchmark we propose, comprises over 20 negative samples. Due to budgetary constraints, the size of this dataset is much smaller compared to the datasets used for evaluating compositional understanding, e.g. ARO dataset (Yüksekgönül et al., 2023). Thirdly, we only evaluate the visual encoders of most LVLMs, i.e. the CLIP models, but we do not conduct research on encoders used by some other models, for instance, the variant of ResNet called NFNet-F6 (Brock et al., 2021) used by Flamingo (Alayrac et al., 2022).

References

Harsh Agrawal, Peter Anderson, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019a. 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.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019b. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

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

Andy Brock, Soham De, Samuel L. Smith, and Karen Simonyan. 2021. High-performance large-scale image recognition without normalization. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 1059–1071. PMLR.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. 2024. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. 2023. CIEM: contrastive instruction evaluation method for better instruction tuning. *CoRR*, abs/2309.02301.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma,

Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *CoRR*, abs/2311.16922.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 292–305. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. 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*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Aligning large multi-modal model with robust instruction tuning. *CoRR*, abs/2306.14565.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *CoRR*, abs/2402.00253.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. @ CREPE: can vision-language foundation models reason compositionally? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 10910–10921. IEEE.

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, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15617–15629. IEEE.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning large multimodal models with factually augmented RLHF. *CoRR*, abs/2309.14525.

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

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. 2023. Evaluation and analysis of hallucination in large vision-language models. *CoRR*, abs/2308.15126.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz,

Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *Trans. Mach. Learn. Res.*, 2022.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2023. RLHF-V: towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *CoRR*, abs/2312.00849.

Mert Yüksekgönül, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *CoRR*, abs/2111.08276.

Le Zhang, Rabiul Awal, and Aishwarya Agrawal. 2023. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic fine-grained understanding. *CoRR*, abs/2306.08832.

Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2024. Debiasing multimodal large language models.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *CoRR*, abs/2310.00754.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. 2023. Segment everything everywhere all at once. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

A Statistics on the Datasets

Dataset	Size	#Negative Samples	#Avg Length
		Train	
COCO	8000	27	16.0
Flickr30K	8000	27	18.4
		Dev	
COCO	990	27	15.6
		Test	
COCO	500	27	16.3
Flickr30K	500	27	21.1
Nocaps	500	27	19.1

Table 7: Statistics of the datasets used in our benchmark.

The statistical information of the dataset is presented in the Table 7, which is divided into three parts: training, testing, and validation. The average length displayed in the table refers to the average length of the negative examples in the dataset.

B More Examples

We present more examples in Figure 2. It can be observed that our method can seamlessly integrate objects that are not present in the original image into the text. The names of the added objects are highlighted in red. Removing objects that are present in the picture can be accomplished with minimal adjustments. As for the removal of objects not depicted in the image, such as the "food" mentioned in the third figure, the negative samples typically involve modifications to the objects, attributes, and other content in the positive samples.

Caption: A person on a snowboard weaves down a mountain slope.

Add 'backpack': A person with a backpack on a snowboard weaves down a mountain slope.

Add 'car': A person in a car weaves down a mountain slope.

Delete 'person': A snowboard glides down the mountain slope.

Caption: A barber is trimming the neckline of a man on the side of the street.

Add 'sky': A barber is trimming the neckline of a man under the sky on the side of the street.

Add 'river': A barber is trimming the neckline of a man by the side of the river.

Caption: Two cans of redbull along with several other energy drink supplements and a starbucks coffee cup.

Add 'person': A person holding two cans of Redbull, along with several other energy drink supplements and a Starbucks coffee cup.

Delete 'food': Three bottles of green tea along with several other herbal tea bags and a porcelain tea cup.

Figure 2: Examples from our benchmark OHD-Caps. The three images in the figure are from the COCO, Flickr, and Nocaps datasets, respectively.