Plausible May Not Be Faithful: Probing Object Hallucination in Vision-Language Pre-training

Wenliang Dai[†] **Zihan Liu**[*◇] **Ziwei Ji**[†] **Dan Su**[†] **Pascale Fung**[†]
[†]The Hong Kong University of Science and Technology
[◇]NVIDIA
wenliang.dai@connect.ust.hk, pascale@ece.ust.hk

Abstract

Large-scale vision-language pre-trained (VLP) models are prone to hallucinate non-existent visual objects when generating text based on visual information. In this paper, we systematically study the object hallucination problem from three aspects. First, we examine recent state-of-the-art VLP models, showing that they still hallucinate frequently, and models achieving better scores on standard metrics (e.g., CIDEr) could be more unfaithful. Second, we investigate how different types of image encoding in VLP influence hallucination, including region-based, grid-based, and patch-based. Surprisingly, we find that patch-based features perform the best and smaller patch resolution yields a non-trivial reduction in object hallucination. Third, we decouple various VLP objectives and demonstrate that token-level image-text alignment and controlled generation are crucial to reducing hallucination. Based on that, we propose a simple yet effective VLP loss named ObjMLM to further mitigate object hallucination. Results show that it reduces object hallucination by up to 17.4% when tested on two benchmarks (COCO Caption for in-domain and NoCaps for out-of-domain evaluation).

1 Introduction

Thanks to the advancement of large pre-trained Language Models (LMs) and Vision-Language Pre-training (VLP) methods, models are able to achieve surprisingly good performance in vision-conditioned text generation, e.g., image captioning. However, large LMs are found to generate unfaithful or nonsensical texts given the source input (Ji et al., 2022), which is called hallucination. The hallucination problem is also inherited by VLP models (Alayrac et al., 2022), as they are still LMs that can understand visual information. VLP models often generate fluent and seem appropriate sentences if we only see the text, but wrong when taking the visual input into consideration. One major type of hallucination in VLP is known as object hallucination (Rohrbach et al., 2018), where models generate texts containing non-existent or inaccurate objects from the input images. Object hallucination in VLP models essentially limits their performance and raises safety concerns for industrial applications. For example, in biomedical image captioning (Pavlopoulos et al., 2019), object hallucination reduces the accuracy of diagnosis and may lead to severe consequences for the patient. Despite the limitations and risks caused by object hallucination, this problem has not been studied in contemporary VLP works yet.

To narrow down the aforementioned research gap, in this paper, we systematically investigate four fundamental research questions about object hallucination: 1) how much do modern VLP models hallucinate? 2) how do different forms of image encoding affect object hallucination? 3) what are the effects of common VLP objectives on object hallucination? and 4) how to mitigate object hallucination in VLP models?

For our first question, we examine recent state-of-the-art VLP models on the image captioning task. To evaluate object hallucination, we adopt and expand the CHAIR (Caption Hallucination Assessment with Image Relevance) metric proposed by Rohrbach et al. (2018). Results show that these models still hallucinate frequently with ∼10% of the generated sentences containing at least one hallucinated object. This problem becomes much severer when generating sentences given out-of-domain images. Furthermore, we discover that the widely used optimization method SCST (Rennie et al., 2017) could lead to more hallucination, even if it improves standard metrics like CIDEr (Vedantam et al., 2015).

For our second question, to investigate how different types of image encoding in VLP influence hallucination, we ablate three commonly

[*] Work done during PhD at HKUST.

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2136–2148
May 2-6, 2023 ©2023 Association for Computational Linguistics

used ones, including region-based, grid-based, and patch-based (Kim et al., 2021). Surprisingly, we find that patch-based features perform the best and smaller patch resolution yields a non-trivial reduction in object hallucination.

Thirdly, we analyze the effects of commonly adopted vision-language pre-training objectives on object hallucination. Specifically, we decouple and combine the image-text contrastive (ITC) loss, the image-text matching (ITM) loss with and without hard negatives, and the image-conditioned language modeling (ICLM) loss. Counter-intuitively, although ITC and ITM help to bring apart dissimilar images and texts, results show that they do not contribute much to alleviating object hallucination. The generative ICLM loss is the main influential factor of object hallucination and different pre-training datasets lead to distinctive model behaviors. More detailed analysis is described in Section 5.3.

Finally, we propose a simple yet effective new vision-language pre-training loss, namely object-masked language modeling (ObjMLM), to further mitigate object hallucination by enhancing the alignment and restriction between text tokens and visual objects during generation. Code and evaluation setups are released: `https://github.com/wenliangdai/VLP-Object-Hallucination`.

Overall, our contributions are three-fold:

- This is the first paper that systematically studies state-of-the-art VLP models on the object hallucination problem, proving that it is still far from resolved and previous methods that improve standard metrics may reflect in worse hallucination.
- We thoroughly investigate the influence of different VLP losses and image encoding methods on object hallucination. Our findings could be valuable for future work to build more responsible VLP systems.
- We present a new pre-training objective ObjMLM to mitigate object hallucination. Experimental results show that it reduces object hallucination by 17.4% without introducing extra training data.

2 Related Work

2.1 Hallucination in Deep Learning

Generally, the term *hallucination* denotes the appearance of undesirable output that is unfaithful to the conditional input (Maynez et al., 2020), even though it may appear to be fluent or reasonable. In the multimodal field, the hallucination phenomenon refers to the prediction of non-existent or incorrect objects (e.g., in object detection or image captioning) and is called *object hallucination* (Rohrbach et al., 2018; Biten et al., 2022). Despite the success of large pre-trained models, they still suffer the hallucination problem, which degrades the performance and largely hinders practical applications (Ji et al., 2022).

Many works have been proposed to mitigate hallucination in recent years. Nie et al. (2019) applied data refinement with self-training to improve the equivalence between the input and the paired text in the data-to-text generation task. Zhang et al. (2021b) and Zhang et al. (2020) proposes scene graph learning methods to ground the process of visual captioning to reduce hallucination. Ma et al. (2020) reconstruct generated sentences from localized image regions. Xiao and Wang (2021) proposed the uncertainty-aware beam search as an add-on technique to the original beam search, in both image captioning and data-to-text generation. To reduce hallucination in dialog systems, Shuster et al. (2021) introduced knowledge augmentation and Dziri et al. (2021) presented a post-processing method to refine generated outputs. Su et al. (2022) augment models with answer-related information predicted by a machine reading comprehension module to reduce hallucination in the generative question answering task.

2.2 Vision-Language Pre-training

The research on vision-language pre-training (VLP) has progressed vastly in recent years. Due to the demand for large-scale data, most VLP methods use self-supervised pre-training objectives to utilize image-text pairs crawled from the web. In the beginning, BERT (Devlin et al., 2019)-style VLP models (Lu et al., 2019; Tan and Bansal, 2019; Li et al., 2020b; Chen et al., 2020; Yu et al., 2021a; Shen et al., 2022) are trained to perform multimodal understanding tasks, using objectives like image-text matching and masked language modeling. Later, encoder-decoder architectures are introduced to additionally handle multimodal generation tasks with a causal language modeling loss (Li et al., 2021b; Yu et al., 2021b; Lin et al., 2021; Cho et al., 2021; Ding et al., 2021; Li et al., 2022; Wang et al., 2022a). Another line of research uses a dual-stream architecture (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2022; Yao et al., 2022) with separate image and text encoders aligned together through an image-text contrastive loss. They improve the performance of various multimodal downstream tasks by a large step.

Alayrac et al. (2022) show that fatal object hallucination can happen naturally or be provoked by the adversarial prompting in modern VLP models. However, in previous works, how different VLP strategies influence the faithfulness of generated text given images has not been studied. Moreover, the effects of using different types of image encoding are also unclear, including region-based (Li et al., 2020c; Zhang et al., 2021a; Hu et al., 2022), grid-based (Wang et al., 2022b), and patch-based (Kim et al., 2021; Li et al., 2021a).

3 Evaluation Setup

In this section, we first introduce the CHAIR evaluation metric and our proposed improvements to it in Section 3.1. Then, in Section 3.2, we describe two datasets used for evaluation and explain how to calculate CHAIR scores under such settings.

3.1 Evaluation Metric

To automatically measure object hallucination, we adopt the CHAIR (Caption Hallucination Assessment with Image Relevance) metric proposed by Rohrbach et al. (2018). CHAIR calculates what proportion of generated object words are not in the image (i.e., hallucinated) according to the ground truth. CHAIR has two variants: CHAIR$_i$ (instance-level) and CHAIR$_s$ (sentence-level), which are formulated as follows:

$$\text{CHAIR}_i = \frac{\#\,\{\text{hallucinated objects}\}}{\#\,\{\text{all objects in prediction}\}},$$

$$\text{CHAIR}_s = \frac{\#\,\{\text{hallucinated sentences}\}}{\#\,\{\text{all sentences}\}}.$$

As formulated, CHAIR$_i$ represents the proportion of hallucinated objects over all golden objects in all data samples. It can be seen as the probability of a generated object to be a hallucination. On the other hand, CHAIR$_s$ measures the proportion of generated sentences that contain at least one hallucinated object. Therefore, to calculate CHAIR$_i$ and CHAIR$_s$, we need a pre-defined list of golden object categories to recognize objects in the text. We illustrate dataset-specific calculation details in Section 3.2.

3.2 Evaluation Datasets

To evaluate models' performance on object hallucination with CHAIR, we adopt two widely used benchmarks: Microsoft COCO Caption (Lin et al., 2014) and NoCaps (Agrawal et al., 2019). For all models, the COCO Caption training set is used for the finetuning of the image captioning task, and COCO Caption test set and NoCaps valid set are used for in-domain and out-of-domain evaluation, respectively. In the following, we introduce statistics of each dataset and how to calculate CHAIR on them.

3.2.1 COCO Caption

The COCO Caption (Lin et al., 2014) is a large-scale and widely used dataset for the training and evaluation of the image captioning task. We use the Karpathy split (Karpathy and Fei-Fei, 2017), in which 82K, 5K, and 5K images are in the train, validation, and test sets, respectively. Each image is annotated with at least five ground truth captions.

To calculate CHAIR scores on this dataset, we follow the setting proposed in Rohrbach et al. (2018). In practice, we first tokenize each sentence and then singularize each word. Then, we use a list of synonyms from Lu et al. (2018) to map fine-grained objects to the pre-defined 80 coarse-grained MSCOCO object categories (e.g., mapping "puppy", "chihuahua", "poodle" to the "dog" category). The purpose of doing this mapping is to ensure that we do not detect hallucinated objects by mistake. For example, when the ground-truth caption only has the "puppy" object, the CHAIR metrics will undesirably consider the "dog" object generated by models as a hallucinated object if we do not perform the mapping.

3.2.2 NoCaps

The NoCaps (Agrawal et al., 2019) dataset aims to evaluate models trained on the training set of COCO Caption to examine how well they generalize to a much larger variety of visual concepts, i.e., unseen object categories. There are 4,500 images in the validation set and 10,600 images in the test set. Images are taken from the Open Images V4 (Kuznetsova et al., 2020) dataset, which contains 600 object classes. Due to the unavailability of ground truth captions of the test set, we use the valid set of NoCaps.

To calculate CHAIR scores on NoCaps, we setup a similar setting as used in COCO Caption. Specifically, we map the fine-grained classes defined in

Model	CIDEr Optim (SCST)	# Pretrain Image-Text Pairs	COCO Caption Karpathy Test						NoCaps Validation Out-of-domain			
			B@4↑	C↑	M↑	S↑	CH_i↓	CH_s↓	C↑	S↑	CH_i↓	CH_s↓
OSCAR *Base**	✗	6.5M	34.4	117.6	29.1	21.9	7.1	13.0	-	-	-	-
OSCAR *Base**	✓	6.5M	39.6	134.2	29.8	23.5	7.2	13.5	-	-	-	-
VinVL *Base*	✗	6.5M	38.2	129.3	30.3	23.6	5.3	10.0	83.1	10.8	12.1	21.2
VinVL *Base*	✓	6.5M	40.9	140.4	30.9	25.1	5.7	10.9	87.5	11.7	17.4	32.1
VinVL *Large*	✗	6.5M	38.5	130.8	30.4	23.4	5.5	10.5	-	-	-	-
VinVL *Large*	✓	6.5M	41.0	140.9	31.1	25.2	5.6	10.6	-	-	-	-
BLIP *Base*	✗	129M	39.7	133.3	31.0	23.8	4.9	8.9	112.1	14.2	6.6	10.5
BLIP *Large*	✗	129M	40.4	136.7	31.1	24.3	4.7	8.8	115.3	14.4	6.4	10.5
OFA *Large*	✗	21M†	41.7	140.5	31.2	24.2	4.7	8.9	103.2	13.3	6.4	10.2
OFA *Large*	✓	21M†	43.8	149.5	31.8	25.9	4.2	8.1	113.1	15.2	7.1	12.4

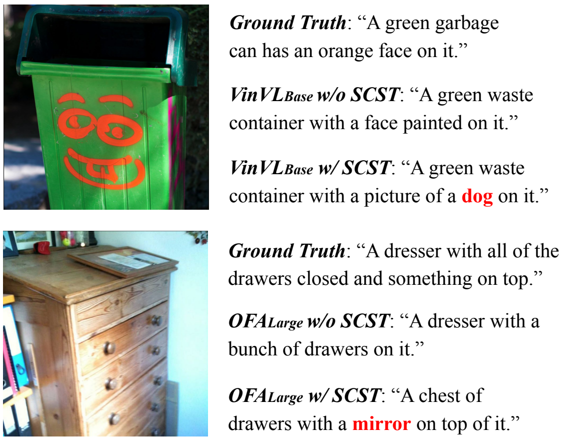
Table 1: Image captioning results of recent state-of-the-art VLP models (Li et al., 2020c; Zhang et al., 2021a; Li et al., 2022; Wang et al., 2022a) on the COCO Caption Karpathy test set and NoCaps validation set. Here, B@4, C, M, S, and CH denote BLEU-4, CIDEr, METEOR, SPICE, and CHAIR, respectively. CIDEr Optim indicates whether the SCST CIDEr optimization is used or not. All results are generated by using their officially provided checkpoints and hyper-parameters, * means the model is finetuned by us as the provided one is broken. † denotes the model also uses unimodal data besides image-text pairs.

NoCaps to coarse-grained categories based on the hierarchical object relationship[1] to improve the effectiveness of CHAIR metrics. We only add two types of object categories to our final object list: 1) super-categories that have sub-categories, and 2) object categories that have neither super-category nor sub-categories. Eventually, we construct a list of 139 coarse-grained object categories from the 600 classes.

4 Object Hallucination in VLP Models

Benefitting from the vast advancement of various VLP methods, the performance of image captioning has been improved a lot by following a pretrain-then-finetune schema. Generally, the performance is measured by metrics like CIDEr (Vedantam et al., 2015), SPICE (Anderson et al., 2016), METEOR (Banerjee and Lavie, 2005), and BLEU (Papineni et al., 2002), which consider the semantic and syntactic similarity or n-gram-based fluency between the model generated and ground truth captions. However, the faithfulness of captions generated by VLP models is neglected.

In this section, we provide a thorough analysis of recent VLP models to investigate how much they hallucinate when generating text conditioned on visual information. The results are shown in Table 1. Models are finetuned on the COCO Caption training set and evaluated on both the COCO Caption test set and the NoCaps valid set.



Ground Truth: "A green garbage can has an orange face on it."

VinVL*$_{Base}$ *w/o SCST: "A green waste container with a face painted on it."

VinVL*$_{Base}$ *w/ SCST: "A green waste container with a picture of a **dog** on it."

Ground Truth: "A dresser with all of the drawers closed and something on top."

OFA*$_{Large}$ *w/o SCST: "A dresser with a bunch of drawers on it."

OFA*$_{Large}$ *w/ SCST: "A chest of drawers with a **mirror** on top of it."

Figure 1: Comparison of image captioning examples generated by VinVL$_{Base}$ and OFA$_{Large}$ with and without the SCST CIDEr optimization. Red color denotes the occurrence of object hallucination.

Overall, we observe two noteworthy insights. Firstly, similar to the findings in Rohrbach et al. (2018), for all CHAIR scores, they are not proportional to standard evaluation metrics. Although standard metrics (e.g., the cosine similarity in CIDEr) could potentially penalize the wrong object prediction, they do not directly reflect faithfulness. Captions can still have good scores from standard metrics as long as they contain sufficient accurate objects to fulfill coverage, even if hallucinated objects exist. For example, VinVL$_{Large}$ achieves higher CIDEr and BLEU-4 scores than VinVL$_{Base}$, but its CHAIR scores are also higher.

[1] https://github.com/nocaps-org/image-feature-extractors/blob/master/data/oi_categories.json

Therefore, it is important to have a supplementary metric like CHAIR to reflect faithfulness besides other metrics.

Secondly, the Self-Critical Sequence Training (SCST) (Rennie et al., 2017) for the CIDEr optimization method harms the faithfulness of generated captions. SCST is a reinforcement learning algorithm that has been widely adopted as the second-stage finetuning after the standard cross-entropy optimization for image captioning (Anderson et al., 2018; Zhou et al., 2020; Li et al., 2020c; Zhang et al., 2021a; Hu et al., 2022; Wang et al., 2022a). It calculates the reward based on the CIDEr score by sampling captions during training without the need of another baseline. Although SCST can significantly boost performance on previous standard metrics, it encourages models to generate more hallucinated objects in the captions. For example, applying SCST improves the CIDEr score by 11.1 and BLEU-4 score by 2.7 for VinVL$_{Base}$, yet it also increases 0.9 CHAIR$_s$ score on the COCO Caption dataset.

While Rennie et al. (2017) also observed this phenomenon by testing small scale models, we show that SCST hurts VLP models less. When the model is pre-trained very well, the side effect of SCST is alleviated (e.g., the OFA large model). Moreover, we demonstrate that this problem becomes more serious on out-of-domain images. For the VinVL$_{Base}$ model, there are 10.9% more generated captions containing at least one hallucinated object after using SCST. We speculate that the CIDEr-based optimization encourages models to generate more words or phrases that have higher cosine similarities to the ground truth captions in the vision-language representation space, which can be plausible but not faithful.

We show a case study in Figure 1. After finetuned by SCST, models will take a bigger risk to generate more detailed yet incorrect information (e.g., in the second example in Figure 1, the sentence with hallucination generates the detailed information “mirror”, which cannot be found in the image). This will further amplify the object hallucination problem on out-of-domain images, as models may have lower confidence in unfamiliar visual concepts.

5 Probing Image Encoding Methods and VLP Objectives

In this section, we systematically study two determinants in VLP that are intuitively influential to the severity of the object hallucination problem. Firstly, we study how different types of image encoding affect object hallucination, as they are the key components of models to interpret visual information. Specifically, we ablate three encoding approaches including region-based, grid-based, and patch-based. Secondly, we analyze how different VLP objectives influence object hallucination. We ablate three commonly used ones: image-text contrastive (ITC), image-text matching (ITM), and image-conditioned language modeling (ICLM). Implementation details are described in Appendix A.

5.1 Model Architecture

CLIP. CLIP (Radford et al., 2021) is a dual-stream VLP model that consists of an image encoder and a text encoder. It is pre-trained on 400 million image-text pairs data using a cross-modal contrastive loss. Specifically, CLIP explores the image encoder with different sizes of two architectures[2], including the ResNet (He et al., 2016) and the Vision Transformer (ViT) (Dosovitskiy et al., 2021). The resulting image and text encoders are aligned in the same multimodal feature space.

BERT. BERT (Devlin et al., 2019) is a Transformer (Vaswani et al., 2017) model pre-trained on a large corpus by the masked language modeling (MLM) and sentence permutation losses. It is shown to have excellent performance on various downstream tasks after finetuning. Moreover, BERT can also handle generation tasks when the self-attention layers are restricted to the left-to-right direction to generate text auto-regressively. In this paper, we refer to this variant as BertLM.

We design a flexible architecture that can plug in various visual encoders and fit modern VLP objectives without introducing extra influential factors. As shown in Figure 4, the model consists of two parts, a visual encoder to encode images and a text decoder to generate sentences conditioned on the image representations. We use two separate modules rather than a unified single-stream model, as it is convenient to alter the visual encoder while keeping the text decoder the same. Specifically,

[2] https://github.com/openai/CLIP/blob/main/model-card.md

Visual Encoder	#Params	COCO Karpathy Test			NoCaps Val Out-of-domain		
		C↑	CH$_i$↓	CH$_s$↓	C↑	CH$_i$↓	CH$_s$↓
Region features							
BUTD-RN101	45M	110.6	9.1	15.9	40.5	36.7	49.0
ResNeXt-152	60M	115.9	7.1	12.9	45.1	30.5	41.1
Grid features							
RN50×4	83M	107.6	11.2	19.1	41.6	37.5	49.9
RN50×16	160M	111.6	9.0	15.8	47.5	33.1	45.2
RN50×64	401M	115.8	7.5	13.2	56.2	26.3	36.6
Patch features							
ViT-B/32	84M	108.9	10.3	17.9	44.4	34.7	46.8
ViT-B/16	82M	111.8	8.1	14.7	51.9	30.3	42.3
ViT-L/14	290M	120.7	6.4	11.6	59.8	24.2	33.5

Table 2: Results of different types of visual encoders with the same BertLM text decoder on the COCO Karpathy test set and NoCaps validation set (out-of-domain).

VLP Objectives	COCO Karpathy Test			NoCaps Val Out-of-domain		
	C↑	CH$_i$↓	CH$_s$↓	C↑	CH$_i$↓	CH$_s$↓
(a) None	120.7	6.4	11.6	59.8	24.2	33.5
Discriminative Objectives						
CC3M						
(b) ITC	120.5	6.5	11.7	59.9	24.4	33.8
(c) ITC$_{Late}$	121.2	6.2	11.3	60.5	23.8	32.9
(d) ITC$_{Late}$ + ITM	121.0	6.3	11.5	60.2	23.9	33.1
(e) ITC$_{Late}$ + ITM$_{Hard}$	120.9	6.6	11.7	59.9	24.2	33.3
Generative Objectives						
Visual Genome						
(f) LM	120.3	5.5	9.8	62.8	9.0	13.9
(g) LM + ObjectMLM	121.9	5.3	9.2	63.8	8.8	13.1
CC3M						
(h) LM	122.3	6.0	10.9	92.1	8.3	14.5
(i) LM + ObjectMLM	124.5	5.1	9.0	94.0	8.0	13.1
(c) + (i)	**125.1**	**4.9**	**8.8**	**94.5**	**7.9**	**12.5**

Table 3: Comparison of the effects of different VLP objectives and their combination on object hallucination.

for region-based image features, we explore the Faster R-CNN object detector (Ren et al., 2015) with two different backbones: the ResNet-101 used in BUTD (Anderson et al., 2018) and the ResNeXt-152 (Xie et al., 2017) used by Zhang et al. (2021a). They are both pre-trained on COCO (Lin et al., 2014) and Visual Genome (Krishna et al., 2016) datasets for object detection. For the grid-based and patch-based image features, we use the CLIP ResNet variants and CLIP ViT variants, respectively. The reason for using CLIP is that all its variants are pre-trained on the same data and there is a wide range of different model sizes. For all visual encoders, we use the same BertLM as the text decoder.

5.2 Effects of Different Image Features

Recognizing visual objects correctly is crucial for avoiding object hallucination. In Table 2, we compare the performance of different visual encoders with the same text decoder on COCO (in-domain) and NoCaps (out-of-domain) datasets.

Overall, patch-based visual encoders attain the best performance in terms of avoiding object hallucination. Models with grid features hallucinate more frequently when achieving comparable CIDEr scores to the other models. For example, on COCO, RN50×16 has a similar CIDEr score to ViT-B/16 but higher CHAIR$_s$, which is also observed between RN50×64 and ResNeXt-152. We conjecture that the inductive biases (Cohen and Shashua, 2017) of the Convolutional Neural Network (CNN), such as locality and translation invariance, weaken the connection of different characteristics of a single object and thus lead to more hallucination. Oppositely, regional or patch-level features are obtained by directly dividing images into different parts and further associating them through positional embeddings. In addition, we see that a smaller patch resolution helps to reduce object hallucination without enlarging the model size.

For region-based visual encoders, although they achieve modest results on COCO with relatively small model sizes, their performance of object hallucination on out-of-domain images drops dramatically. One important reason is that the output of such encoders only contains representations of detected visual objects rather than the whole image, which may amplify detection errors as there is much less context. Moreover, as the object detector is pre-trained separately from the whole model and its parameters are fixed during finetuning, this gap could also aggravate object hallucination on unseen images.

5.3 Effects of Different VLP Objectives

Based on the best performing ViT-L/14 baseline, we explore three commonly used vision-language pre-training objectives and their variants that could potentially affect object hallucination.

5.3.1 Pre-training Datasets

We explore two pre-training datasets with image-text pairs: 1) the VG Caption from the Visual Genome (Krishna et al., 2016) dataset, which contains 10K images and each image has multiple corresponding descriptions; and 2) a more large-scale dataset CC3M (Sharma et al., 2018) that contains three millions of image-text pairs.

5.3.2 Image-Text Contrastive (ITC) Loss

The cross-modal contrastive loss is shown to be fairly effective in representation learning (Tian et al., 2020; Sigurdsson et al., 2020) and VLP (Radford et al., 2021; Li et al., 2021a, 2022). It aligns the visual and textual representations into the same multimodal feature space by shortening the distance between an image and a text if they are paired, or enlarging if they are not.

Counter-intuitively, as shown in Table 3 (b), ITC has negligible influence on the faithfulness of generated captions. We speculate that it only enhances the model's understanding of global-level representations rather than token-level alignment between images and texts. To verify, we further test the ITC with a more fine-grained token-level late interaction (ITC$_{Late}$) proposed by Yao et al. (2022). As shown in Table 3 (c), ITC$_{Late}$ is more effective than the vanilla ITC and slightly reduces object hallucination. We think this benefits from the word-patch alignment ability enabled by ITC$_{Late}$.

5.3.3 Image-Text Matching (ITM) Loss

ITM is a widely used objective in VLP (Li et al., 2020a; Chen et al., 2020; Zhou et al., 2021). It is a binary classification task that aims to make the model learn whether an image and a sentence are paired or not. Based on that, ITM with hard negatives (ITM$_{Hard}$) is introduced to increase the difficulty of the task, which is shown to be very effective on representation learning (Kalantidis et al., 2020; Robinson et al., 2021; Li et al., 2021b). We follow the ITM loss proposed by Li et al. (2022), in which an in-batch negative example is sampled either uniformly (normal negative) or from the similarity distribution of image-text pairs computed by ITC (hard negative).

The results are exhibited in Table 3 (d) (e). Both ITM and ITM$_{Hard}$ are not highly correlated with the object hallucination problem. They only slightly reduce hallucination in generated texts on out-of-domain images. Although the ITM$_{Hard}$ can be seen as an analogy to the object hallucination problem (plausible but not correct) in a global and discriminative way, it has a negligible effect on reducing hallucination for downstream generative tasks.

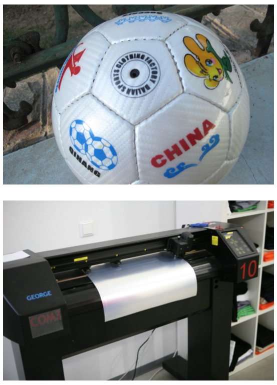
5.3.4 Image-Conditioned Language Modeling

Various image-conditioned language modeling losses have been proposed in the VLP research, in the form of masked language modeling (MLM) (Sun et al., 2019; Tan and Bansal, 2019; Su et al., 2020), text infilling (Dai et al., 2022; Wang et al., 2022a), prefix LM (Wang et al., 2022b), and causal LM (Hu et al., 2022). This is one of the most crucial pre-training losses to activate the cross-modal text generation ability for the VLP model.

COCO Caption
“Several boats docked at a floating dock at a marina.”,
“Several boats sitting on a docking station on the water.”,
“A bunch of speedboats near a harbor with flags from all over the world.”, etc.

Visual Genome Caption
“A dock in a city.”, “Long silver dock in water.”
“Very blue, calm water in marina.”, “The water is calm.”
“A dock is floating on the water.”, “Row of docked boats.”, etc.

Figure 2: Comparison of ground truth captions in COCO and Visual Genome datasets for the same image.



Ground Truth: “A soccer ball is next to a wall.”, “A soccer ball that is placed on the ground.”, etc.

ViT-L/14 w/o VG: “A close up of a soccer ball on a **table**.”

ViT-L/14 w/ VG: “A close up of a soccer ball on the ground.”

Ground Truth: “A large black printer seems to have a piece of paper in it sideways.”, “A large printer with paper coming out of it”, etc.

ViT-L/14 w/o VG: “A pair of **scissors** sitting on top of a piece of paper.”

ViT-L/14 w/ VG: “A large black machine.”

Figure 3: Comparison of generated captions with or without the image-conditioned language modeling pre-training on the VG dataset before finetuning.

We first examine the causal LM loss, which is exactly the same loss as the image captioning loss, but used in the pre-training on a much larger scale. Surprisingly, as shown in Table 3 (f), although pre-training on the VG Caption does not improve previous standard metrics like CIDEr, it helps to reduce object hallucination by a large margin when compared to (a).

There are two reasons behind this performance lift. Firstly, as described in Figure 2, for each image, VG contains more captions than COCO.

Each caption in VG is much shorter and only describes one specific aspect of the image, unlike the global descriptions in COCO. Therefore, pre-training on VG and then finetuning on COCO is a fine-to-coarse process. It enables models to first accurately describe different parts of an image and connect these clues together at a higher viewing point. Secondly, due to the nature of the short length of VG captions, the model becomes slightly more cautious. On average, after adding VG data in the pre-training, there are 0.08 and 0.24 fewer objects generated in each caption on COCO and NoCaps, respectively. This observation aligns with the sentence simplification method proposed by Biten et al. (2022), which simplifies sentences to augment data and further mitigate object hallucination. Figure 3 illustrates VG's effects on generated samples. The model is more faithful but more likely to lack some details when it is not confident.

For CC3M, we observe a leap in all metrics. It improves the general image translation ability of the model, which can be seen as large-scale data augmentation. This indicates that seeing a sufficient amount of data and co-occurrence of various objects during pre-training help to mitigate object hallucination to some extent. However, data augmentation may not be the key to drastically tackle object hallucination. As discussed in Section 4, object hallucination still happens frequently even if the model is pre-trained on large-scale data. Therefore, we believe that enhancing the controllability of vision-conditioned text generation would be a promising future direction. More case studies are exhibited in Appendix B.

6 Object Masked Language Modeling

Based on the findings in Section 5, we propose a simple yet effective pre-training objective to mitigate object hallucination by improving object-level image-text alignment. It is named Object Masked Language Modeling (ObjMLM). As shown in Figure 4, ObjMLM can be seen as a variant of the MLM loss by masking all the objects in the text that appear in the image. For each sentence, we mask the object words and phrases as defined in the object category lists of both COCO and NoCaps by performing exact matching. Similar to the whole word masking (Cui et al., 2021), we conduct whole object masking so that there will be only one [MASK] token to replace each object.

Compare the results shown in lines (h) and (i) of Table 3, by plugging ObjMLM into an existing VLP setting, the CHAIR$_s$ score is reduced by 17.4%. This is a non-trivial improvement without introducing more pre-training data. To further validate ObjMLM's effectiveness, we replace it by the standard MLM loss with a 15% masking rate. However, it only reduces CHAIR$_s$ by 1.7%, which is not significant. We conjecture that ObjMLM adds a constraint that indirectly controls the model to only generate objects that are visible in the input image. Additionally, ObjMLM enhances the model's recognition ability when describing the spatial relationship between objects, which is a common scenario that causes hallucinations frequently.

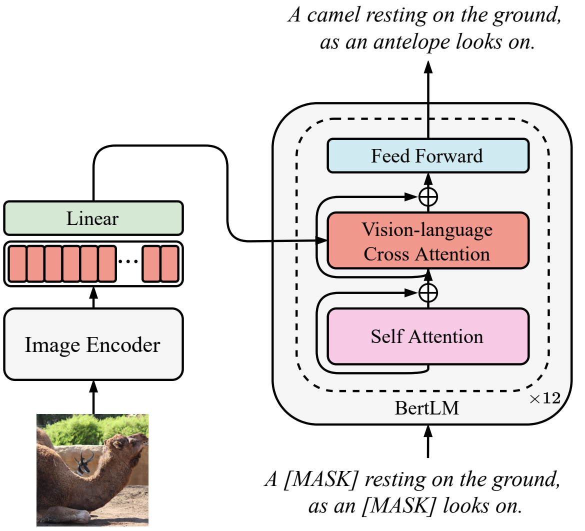

A camel resting on the ground, as an antelope looks on.
Feed Forward
Linear
Vision-language Cross Attention
Image Encoder
Self Attention
×12
BertLM
A [MASK] resting on the ground, as an [MASK] looks on.

Figure 4: An overview of the model architecture and the training of our proposed ObjMLM. We use the same architecture as described in Section 5 to show the effectiveness of ObjMLM. Here, the image encoder can be one of the region-based, grid-based, or patch-based variants as described in Section 5.2. For ObjMLM, we use the ViT-L/14.

7 Conclusion

This paper systematically studies the objection hallucination phenomenon in VLP models, which is a severe problem but neglected in contemporary VLP works. We find that recent large VLP models still hallucinate frequently. Moreover, the widely used SCST method harms the faithfulness of generated sentences in image captioning, even if it improves previous standard metrics. Furthermore, we discover that image encoding matters and the patch-based input with smaller resolution helps mitigate object hallucination. Finally, we ablate commonly used VLP losses and show that token-level image-text alignment and controllability of the generation are crucial. We further propose a new loss named ObjMLM, which reduces object hallucination by 17.4% for an existing VLP setting. We believe our findings are beneficial for future work to build more responsible VLP models.

Limitations

We understand that the hallucination problem is a big research topic and it is not just limited to object hallucination. In this paper, we focus on the investigation and mitigation of object hallucination, leaving other types of hallucination in VLP for future work. Another limitation is that for the discussion of recent VLP models in Section 4, we only study those whose pre-trained checkpoints are publicly available. For the non-released ones, we cannot pre-train them by ourselves due to the lack of large-scale GPU power and private pre-training datasets.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. *International Conference on Computer Vision*, pages 8947–8956.

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 Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*.

Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. 2022. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1381–1390.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR.

Nadav Cohen and Amnon Shashua. 2017. Inductive bias of deep convolutional networks through pooling geometry. In *International Conference on Learning Representations*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Enabling multimodal generation on CLIP via vision-language knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2383–2395, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. 2021. Cogview: Mastering text-to-image generation via transformers. In *NeurIPS*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

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

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, and L. Wang. 2022. Scaling up vision-language pretraining for image captioning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17959–17968, Los Alamitos, CA, USA. IEEE Computer Society.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21798–21809. Curran Associates, Inc.

Andrej Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:664–676.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*.

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. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73.

Alina Kuznetsova, Hassan Rom, Neil Gordon Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The open images dataset v4. *International Journal of Computer Vision*, 128:1956–1981.

Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020b. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online. Association for Computational Linguistics.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021b. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2592–2607, Online. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020c. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Junyang Lin, Rui Men, An Yang, Chan Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, J. Zhang, Jianwei Zhang, Xu Zou, Zhikang Li, Xiao Qing Deng, Jie Liu, Jinbao Xue, Huiling Zhou, Jianxin Ma, Jin Yu, Yong Li, Wei Lin, Jingren Zhou, J ie Tang, and Hongxia Yang. 2021. M6: A chinese multimodal pretrainer. *ArXiv*, abs/2103.00823.

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 *ECCV*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7219–7228.

Chih-Yao Ma, Yannis Kalantidis, Ghassan AlRegib, Peter Vajda, Marcus Rohrbach, and Zsolt Kira. 2020. Learning to generate grounded image captions without localization supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

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.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

John Pavlopoulos, Vasiliki Kougia, and Ion Androutsopoulos. 2019. A survey on biomedical image captioning. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 26–36, Minneapolis, Minnesota. Association for Computational Linguistics.

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 *ICML*.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2022. How much can CLIP benefit vision-and-language tasks? In *International Conference on Learning Representations*.

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

Gunnar A. Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, João Carreira, Phil Blunsom, and Andrew Zisserman. 2020. Visual grounding in video for unsupervised word translation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10847–10856.

Dan Su, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Read before generate! faithful long form question answering with machine reading. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 744–756.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin P. Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7463–7472.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *ECCV*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022b. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*.

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.

Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2022. FILIP: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021a. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3208–3216.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021b. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3995–4007, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18123–18133.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021a. Vinvl: Revisiting visual representations in vision-language models. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5575–5584.

Wenqiao Zhang, Haochen Shi, Siliang Tang, Jun Xiao, Qiang Yu, and Yueting Zhuang. 2021b. Consensus graph representation learning for better grounded image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3394–3402.

Wenqiao Zhang, Xin Eric Wang, Siliang Tang, Haizhou Shi, Haochen Shi, Jun Xiao, Yueting Zhuang, and William Yang Wang. 2020. Relational graph learning for grounded video description generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 3807–3828, New York, NY, USA. Association for Computing Machinery.

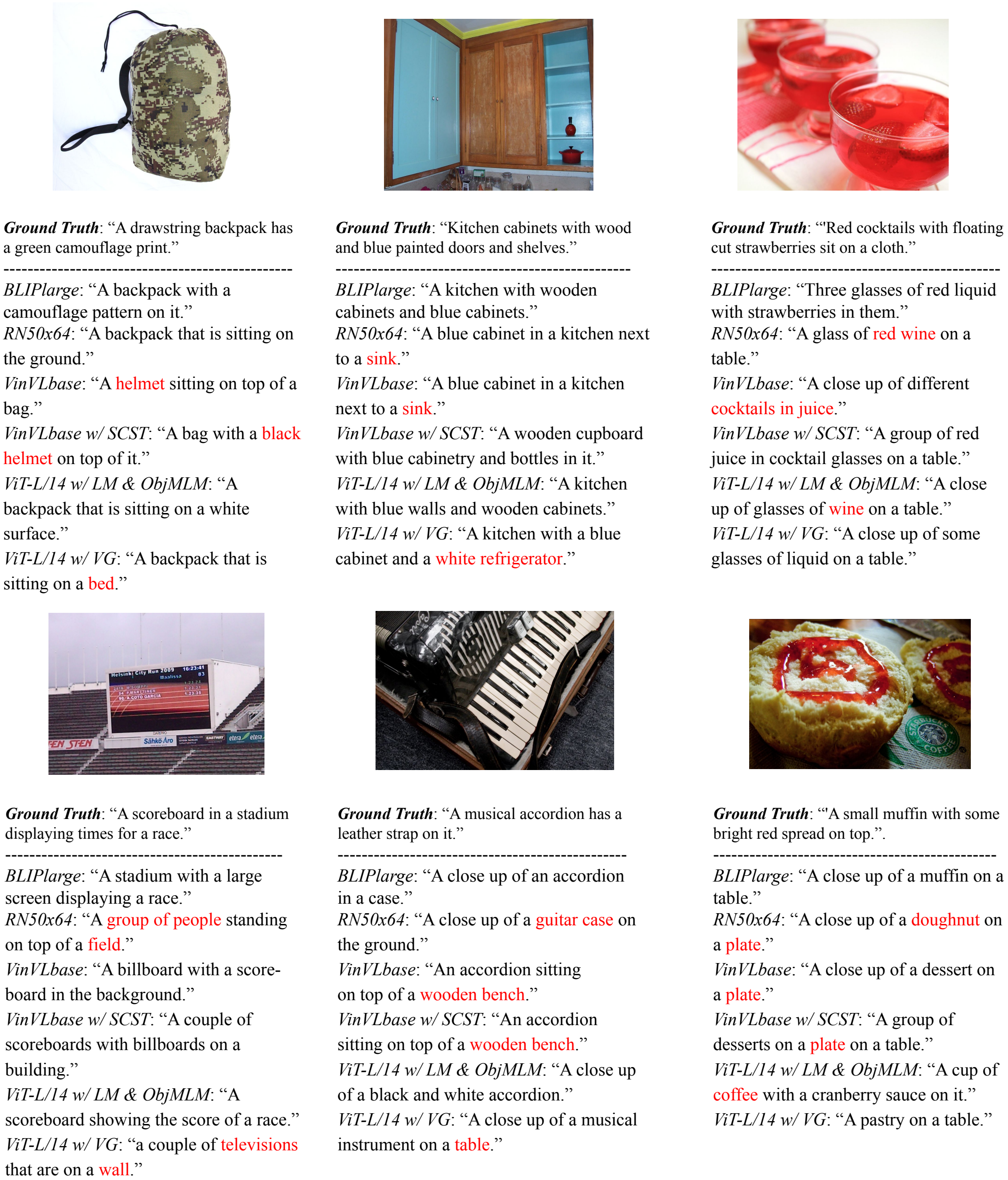
Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. *ArXiv*, abs/1909.11059.

Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4153–4163.

A Implementation Details

Our experiments are implemented in the PyTorch framework (Paszke et al., 2019). For both pre-training and finetuning, we use 8 Nvidia V100 GPUs. As mentioned in Section 5.1, we use the official CLIP checkpoints provided on GitHub. For the text decoder BertLM, we initialize model weights from the bert-base-uncased checkpoint with 110M parameters. For the finetuning on COCO Caption, we use a batch size of 512 and train the models with the AdamW optimizer (Loshchilov and Hutter, 2019) for 10 epochs with a learning rate of 5×10^{-5} and a weight decay of 1×10^{-2}. The learning rate is decayed linearly after each epoch with a rate of 0.85. For the pre-training of text generation losses (LM and ObjMLM), we keep the same hyper-parameters with a learning rate warmup within the first epoch. For ITC and ITM losses, we increase the batch size to 1024 as they tend to have a better performance with more negative samples.

B Additional Case Studies



Ground Truth: “A drawstring backpack has a green camouflage print.”

BLIPlarge: “A backpack with a camouflage pattern on it.”
RN50x64: “A backpack that is sitting on the ground.”
VinVLbase: “A helmet sitting on top of a bag.”
VinVLbase w/ SCST: “A bag with a black helmet on top of it.”
ViT-L/14 w/ LM & ObjMLM: “A backpack that is sitting on a white surface.”
ViT-L/14 w/ VG: “A backpack that is sitting on a bed.”

Ground Truth: “Kitchen cabinets with wood and blue painted doors and shelves.”

BLIPlarge: “A kitchen with wooden cabinets and blue cabinets.”
RN50x64: “A blue cabinet in a kitchen next to a sink.”
VinVLbase: “A blue cabinet in a kitchen next to a sink.”
VinVLbase w/ SCST: “A wooden cupboard with blue cabinetry and bottles in it.”
ViT-L/14 w/ LM & ObjMLM: “A kitchen with blue walls and wooden cabinets.”
ViT-L/14 w/ VG: “A kitchen with a blue cabinet and a white refrigerator.”

Ground Truth: “‘Red cocktails with floating cut strawberries sit on a cloth.”

BLIPlarge: “Three glasses of red liquid with strawberries in them.”
RN50x64: “A glass of red wine on a table.”
VinVLbase: “A close up of different cocktails in juice.”
VinVLbase w/ SCST: “A group of red juice in cocktail glasses on a table.”
ViT-L/14 w/ LM & ObjMLM: “A close up of glasses of wine on a table.”
ViT-L/14 w/ VG: “A close up of some glasses of liquid on a table.”

Ground Truth: “A scoreboard in a stadium displaying times for a race.”

BLIPlarge: “A stadium with a large screen displaying a race.”
RN50x64: “A group of people standing on top of a field.”
VinVLbase: “A billboard with a score-board in the background.”
VinVLbase w/ SCST: “A couple of scoreboards with billboards on a building.”
ViT-L/14 w/ LM & ObjMLM: “A scoreboard showing the score of a race.”
ViT-L/14 w/ VG: “a couple of televisions that are on a wall.”

Ground Truth: “A musical accordion has a leather strap on it.”

BLIPlarge: “A close up of an accordion in a case.”
RN50x64: “A close up of a guitar case on the ground.”
VinVLbase: “An accordion sitting on top of a wooden bench.”
VinVLbase w/ SCST: “An accordion sitting on top of a wooden bench.”
ViT-L/14 w/ LM & ObjMLM: “A close up of a black and white accordion.”
ViT-L/14 w/ VG: “A close up of a musical instrument on a table.”

Ground Truth: “‘A small muffin with some bright red spread on top.”.

BLIPlarge: “A close up of a muffin on a table.”
RN50x64: “A close up of a doughnut on a plate.”
VinVLbase: “A close up of a dessert on a plate.”
VinVLbase w/ SCST: “A group of desserts on a plate on a table.”
ViT-L/14 w/ LM & ObjMLM: “A cup of coffee with a cranberry sauce on it.”
ViT-L/14 w/ VG: “A pastry on a table.”

Figure 5: More cases of generated captions from different models, where the hallucinated objects are marked in red.