Mitigating Open-Vocabulary Caption Hallucinations

Anonymous ACL submission

Abstract

While recent years have seen rapid progress in image-conditioned text generation, image captioning still suffers from the fundamental issue of hallucinations, namely, the generation of spurious details that cannot be inferred from the given image. Existing methods largely use closed-vocabulary object lists to mitigate or evaluate hallucinations in image captioning, ignoring the long-tailed nature of hallucinations that occur in practice. To this end, we propose a framework for addressing hallucinations in image captioning in the open-vocabulary setting. Our framework includes a new benchmark, *OpenCHAIR*, that leverages generative foundation models to evaluate open-vocabulary object hallucinations for image captioning, surpassing the popular and similarly-sized CHAIR benchmark in both diversity and accuracy. Furthermore, to mitigate open-vocabulary hallucinations without using a closed object list, we propose *MOCHa*, an approach harnessing advancements in reinforcement learning. Our multi-objective reward function explicitly targets the trade-off between fidelity and adequacy in generations without requiring any strong supervision. *MOCHa* improves a large variety of image captioning models, as captured by our *OpenCHAIR* benchmark and other existing metrics. We will release our code and models.

1 Introduction

Image captioning, the task of generating text that describes an image, is one of the most fundamental machine learning tasks combining vision and language. Unfortunately, *hallucinations* plague the current state-of-the-art (SOTA), making it less usable for practical tasks that require confidence in the factual correctness of generated captions. Consider, for instance, the image in Figure 1. SOTA image captioning models can generate text that is highly semantically related to its associated imagery, but also contains spurious details ("*skateboard*"). Such hallucinated spurious details either damage user confidence or lead to uncritical acceptance of fallacious (and even potentially dangerous) generated content (Chong et al., 2022; McGowan et al., 2023; Chong et al., 2023).

BLIP-2
A group of people jumping on a skateboard.

BLIP-2 + MOCHa
Several people jumping up and down a flight of stairs.

Figure 1: Hallucinated details (shown as *highlighted text*) are prevalent in the outputs of modern image captioning models, such as the above generation sampled from BLIP2 (Li et al., 2023a). By considering hallucinations in the open-vocabulary setting, we can both quantify and mitigate their effects, illustrated by the improvement provided by our RL-based *MOCHa* framework (**+MOCHa**).

Hallucinations may take a variety of forms in text. However, prior work addressing hallucinations in image captioning has largely focused on detecting or mitigating hallucinations by using closed-vocabulary object lists. While this simplifies the problem under consideration, it fails to capture the diversity of hallucinations observed in modern image captioning models. Thus, we propose a framework for both quantifying and mitigating hallucinations in the open-vocabulary setting.

While established benchmarks and metrics for quantifying hallucinations in captioning models exist for closed-vocabulary object sets, they do not exist (to our knowledge) in an open-vocabulary setup. Accordingly, we introduce *OpenCHAIR*, a new benchmark for quantifying object hallucinations in an open-vocabulary setting. We construct our benchmark using text-to-image models and large language models (LLMs) for generating data and performing evaluation. This allows for capturing and accurately quantifying a wide variety of object hallucination types without being limited

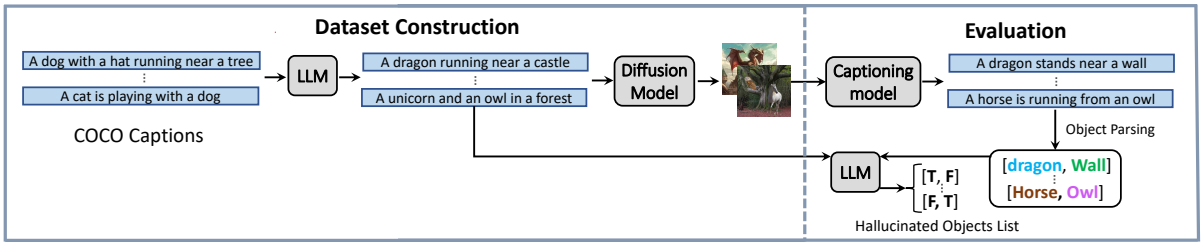


Figure 2: **The *OpenCHAIR* Benchmark**. We illustrate the construction of the *OpenCHAIR* benchmark via an LLM and text-to-image generation model, and its usage for evaluating image captioning models. We first use captions from MS-COCO as seeds to generate diverse synthetic captions. Using syntactic parsing and filtering heuristics, we select for captions containing various open-vocabulary objects. We then generate images corresponding to these captions, producing our benchmark of images linked with object annotations. To evaluate a captioning model, we run it on this benchmark and compare predicted and GT object categories.

to a fixed set of categories. Moreover, our open-vocabulary evaluation method considers free-text predictions without referencing a fixed synonym list. Our evaluations show that this outperforms the CHAIR closed-vocabulary metric (Rohrbach et al., 2018) at capturing performance over diverse hallucinations, providing a complementary measure to CHAIR's evaluation over eighty common object types on natural images.

Equipped with this metric, we turn to hallucination mitigation. A major cause for hallucinations in image captioning are deficiencies in the standard language modeling (LM) objective. The *token-level* language modeling objective does not directly optimize the *sequence-level* quality of generated text, and *factual groundedness* is inherently a sequence-level property of text. Yet, many prior works that directly optimize hallucinations in image captioning limit their scope to a fixed set of possible object tokens, e.g. objects in MS-COCO (Biten et al., 2021; Liu et al., 2022; Petryk et al., 2023), which is incompatible with an open-vocabulary setting.

To mitigate hallucinations without using a closed-vocabulary object list, we introduce *MOCHa*, a **M**ulti-Objective reinforcement learning (RL) based approach for Mitigating **O**pen-vocabulary **C**aption **Ha**llucinations. We observe that RL applied to caption fidelity alone fails to preserve the semantic adequacy (i.e. descriptiveness) of output text, while optimizing for the latter does not enforce factually grounded text. Our key insight is that these two goals can be jointly optimized at the sequence-level by applying RL with a multi-objective reward function. Furthermore, we perform this optimization fully automatically by leveraging SOTA text-based learned metrics, without requiring direct supervision. By considering hallucinations in an open setting, we are able to improve performance across diverse hallucination types, as demonstrated by our *OpenCHAIR* benchmark as well as other metrics. Moreover, we show that our approach can be flexibly applied to a variety of captioning architectures and sizes.

Explicitly stated, our key contributions are: (i) *OpenCHAIR*, a benchmark for open-vocabulary object hallucinations in image captioning. (ii) *MOCHa*, a framework for optimizing a wide array of VLMs to produce high-quality factually-grounded output. (iii) Experiments showing the advantage of *OpenCHAIR* for measuring hallucinations in the open setting, and of *MOCHa* for reducing them.

2 The *OpenCHAIR* Benchmark

To measure object hallucination in the open-vocabulary setting, we propose the *OpenCHAIR* (OCH) benchmark, consisting of ∼5K images illustrating diverse object types in context, accompanied by an evaluation procedure to measure object hallucinations in captioning models. Following existing works (Minderer et al., 2022; Bravo et al., 2023; Chatterjee et al., 2024), we consider our benchmark to be *open-vocabulary* as it contains diverse and uncommon items reflecting the unlimited distribution found in the real world, as well as having the ability to perform evaluation against arbitrary strings. *OpenCHAIR* modifies the previous object hallucination metric CHAIR (Rohrbach et al., 2018), by relaxing its strong reliance on the object annotations in the MS-COCO dataset, which constitute only 80 common object types. We control the diversity of object types in our benchmark by leveraging generative models to produce synthetic caption-image pairs, providing a complementary measure to CHAIR's evaluation

of a closed set of 80 commmon objects over natural images. The use of synthetic images for this purpose is further motivated by prior works which show that models training on synthetic image data may generalize to favorable performance on real images (Tian et al., 2024), as well as the recent growth in usage of synthetic data in general (Sun et al., 2024; Betker et al., 2023). We provide an overview of *OpenCHAIR* below; further implementation details are provided in the appendix.

In order to create a new benchmark that enables measuring the hallucination rate of arbitrary objects, while still maintaining high quality ground-truth captions, we use the pipeline illustrated in Figure 2. We first prompt the LLM Llama-2 (Touvron et al., 2023) with few-shot examples of image captions from MS-COCO, having it generate captions with a similar style but containing diverse details (and in particular, objects that are likely not contained in the closed set of MS-COCO object labels). We then parse these synthetic captions with a syntactic parsing model, identify nouns with high concreteness scores (Brysbaert et al., 2014) (as these generally represent concrete objects), and balance the generated captions among object types to cover a wide array of objects. Subsequently, we utilize the text-to-image diffusion model Stable Diffusion XL (Podell et al., 2023) to generate images from these newly formed captions. This process results in a dataset that consists of synthetic images with corresponding captions including diverse, open-vocabulary objects. While this approach naturally scales to any number of desired image-caption pairs, we generate 5K such pairs (the same order of items found in the widely-used MS-COCO Karpathy test split) and perform manual filtering to assure each pair's alignment and general quality. In total, we removed a small minority (3%) of generated image-caption pairs. Figure 3 shows examples of image-captions pairs from *OpenCHAIR*.

Captioning models may predict free-text objects semantically matching the ground-truth while taking a different surface form (e.g. *chihuahua* vs. *dog*). To capture this in the open-vocabulary setting (rather than using a fixed list of synonyms as done in CHAIR), we evaluate captioning models as follows: After predicting a caption for each image in the *OpenCHAIR* dataset, we parse them to identify objects as described above. For each extracted object o, we compare it to the ground-truth synthetic caption c by prompting an LLM, asking it whether an image with caption c contains the object o and using its answers to count hallucinations. Following CHAIR, we calculate the hallucination rate as n_h/n_{tot}, where n_h is the number of hallucinated objects (*no* answers) and n_{tot} is the total number of objects considered. Figure 4 illustrates the difference between *OpenCHAIR* evaluation and the closed-vocabulary CHAIR metric.

"A green emerald is perched on a rock in a cave." *"A group of mushrooms in the forest."* *"A dog dressed as a human with a wig and eyeglasses."*

Figure 3: ***OpenCHAIR* Examples**. We show examples of images from the *OpenCHAIR* benchmark along with their accompanying ground-truth captions, illustrating its diverse coverage of object types. Long captions are truncated due to space considerations.

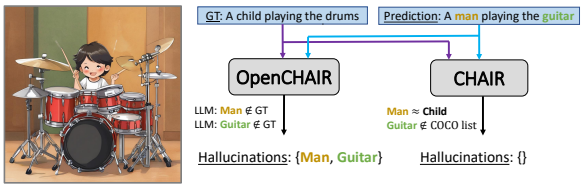


Figure 4: ***OpenCHAIR* vs. CHAIR**. In the above the predicted object *guitar* would not be counted by CHAIR since it is not in its fixed vocabulary, while *man* would not be classified as a hallucination since it is defined by CHAIR as a synonym of *child*. In contrast, *OpenCHAIR*'s LLM classifies both as hallucinations.

3 The *MOCHa* Framework

To mitigate captioning hallucinations in the open-vocabulary setting, we propose *MOCHa*, an RL-based pipeline using SOTA methods for stable reinforcement along with a carefully designed reward function that jointly optimizes for caption fidelity and semantic adequacy. Figure 5 presents it. We turn to describe the learning procedure and objectives used in *MOCHa*. We start with preliminaries, then describe the reward function that *MOCHa* optimizes (Section 3.1), and finally present the RL algorithm used for optimization (Section 3.2).

Preliminaries. In general, RL views a model as an *agent* that interacts with the external *environment* and receives a *reward*, learning to optimize for this reward via exploring the environment (Sutton and Barto, 2018). In the case of image captioning, this model is a VLM operating in an environment of images and reference captions (Rennie et al., 2017).

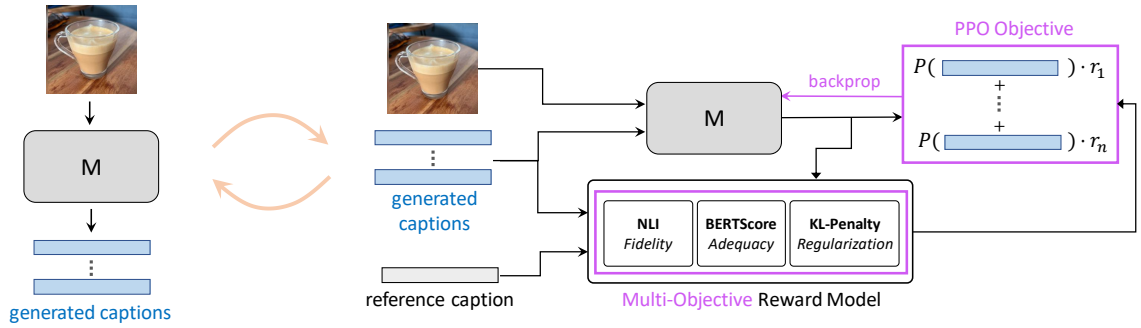


Figure 5: **MOCHa scheme**. The algorithm iteratively collects a minibatch of data from an image captioning model M (left side) and then applies an optimization step to the captioning model (right side). The multi-objective reward reinforces M to produce captions closer to the high-scoring captions and further from the low-scoring captions.

During training, the agent generates a caption by sampling from its own predicted distribution as shown in Figure 5 (left), receiving a reward based on an estimate of the caption quality. After collecting a full batch of rewards, a RL optimization step is applied as shown in Figure 5 (right), and this process repeats iteratively until convergence.

We use the following notation: Let T and I be the sets of possible texts and images, with joint distribution X. Given image $i \in I$, an image captioning model M with weights θ induces a conditional probability distribution $\pi_\theta(\cdot|\cdot)$ over generated captions $\hat{c} \in T$ conditioned on images $i \in I$. In the RL context, we refer to π_θ as the *policy*. A *reward function* $r : T \times T \times I \to \mathbb{R}$ assigns *reward* (or score) $r(\hat{c}; c, i)$ to generated caption $\hat{c}$ relative to ground-truth caption c and image i.

3.1 Reward Function

We wish to optimize for the competing objectives of output fidelity (low hallucination rate) and adequacy (including sufficient details to describe the input image), as optimizing for one of these alone causes the other to deteriorate (as shown in our ablations). We also wish to preserve other desired generation properties such as fluency and diversity. To achieve this, we design a reward function combining multiple objectives as follows:

Fidelity Objective. (r_f). To measure output fidelity to the input image, we use the GT reference captions as a proxy, checking for logical consistency via a pretrained Natural Language Inference (NLI) model. This outputs the probability $\overline{p}(\hat{c}, c)$ that the generated text $\hat{c}$ logically contradicts c, serving as a strong signal for fidelity, as details which contradict ground-truth information about the image are guaranteed to be hallucinations. We scale to the range $[-1, 1]$ by using $r_f(\hat{c}; c) := 1 - 2\overline{p}(\hat{c}, c)$ as the fidelity reward. We implement this with BART (Lewis et al., 2019) fine-tuned on the MNLI dataset (Williams et al., 2018). We average values over all reference captions.

Adequacy Objective. (r_a). To measure adequacy (whether the output caption contains sufficient detail), we use BERTScore (Zhang et al., 2019), a pretrained model measuring text quality relative to ground-truth references. We calculate its F1 value, scaled scale to be approximately in the range $[-1, 1]$ as described in the appendix.

KL Regularization. Following prior work (Jaques et al., 2017, 2019; Ziegler et al., 2020; Stiennon et al., 2020; Ouyang et al., 2022), we add a Kullback–Leibler (KL) divergence penalty to the reward model which constrains the agent to stay close to its initial policy π_0. This serves to prevent mode collapse (i.e. preserving diversity of outputs) and adversarial policies which over-optimize the reward function. The KL penalty adds a term proportional to $K(\hat{c}; i) := -\log(\pi_\theta(\hat{c}|i)/\pi_0(\hat{c}|i))$ to the reward, which limits the agent from excessively distancing itself from the initial policy.

Combined Objective. Our total reward function takes the form $r(\hat{c}; c, i) := \alpha \cdot r_f(\hat{c}; c) + (1 - \alpha) \cdot r_a(\hat{c}; c) + \beta K(\hat{c}; i)$, where $\alpha \in [0, 1]$ and $\beta > 0$ control the trade-off between objectives.

3.2 Learning Procedure

To optimize for caption generations that satisfy the desired properties (described above in Section 3.1), we adopt the Proximal Policy Optimization (PPO) RL algorithm (Schulman et al., 2017), which has been used by recent works on text generation as discussed in Section 5. This is a *policy gradient* algorithm, meaning that it optimizes the parameters θ in order to (approximately) maximize the expected reward $L(\theta) = E_{i,c\sim X,\hat{c}\sim\pi_\theta(\hat{c}|i)}[r(\hat{c}; c, i)]$. PPO extends the REINFORCE algorithm (Sutton and Barto, 2018), also known as SCST in the context

of image captioning (Rennie et al., 2017), by using a clipped surrogate objective to avoid instabilities.

4 Experiments and Results

OpenCHAIR Analysis. We analyze the utility of *OpenCHAIR* by comparing its distribution of objects to the existing closed-vocabulary CHAIR metric, as well as by performing a human evaluation to compare their correlations to human judgements of hallucinations.

In the first column of Table 1 and in Figure 14 (appendix), we show the difference in the number of unique object types found in CHAIR and *OpenCHAIR*, which both contain approximately the same number of images (∼5K). The open-vocabulary design of *OpenCHAIR* enables a significantly larger coverage of object types; in particular, the 2.4K unique object types in *OpenCHAIR* reflect an approximately 30-fold increase relative to the 80 object types found in CHAIR. Furthermore, we find that 53% of object types appear at most three times, and 22% appear only once, illustrating *OpenCHAIR*'s coverage of the long tail of uncommon objects. This is also reflected qualitatively, as the closed-vocabulary benchmark is missing many common object types, including daily objects like *shoe* and *guitar* (see the left image in Figure 6 for a visual example). In contrast, our benchmark includes diverse object types, such as: *pearl, tiger, sand, tricycle, corkscrew, toy, charcoal, text, pine-cone, grandfather, chocolate, wheelchair, wand, etc.* A large sample of additional objects (those not included in CHAIR) can be found in `openchair_objects.txt`. Another source of confusion is its synonym list (*e.g.*, see Figure 4).

We show that *OpenCHAIR* evaluations are grounded in human intuitions via a manual evaluation, comparing its performance to that of CHAIR. For each benchmark (*OpenCHAIR* and CHAIR), we generate captions for a random subset of its dataset and manually check object-level decisions (predicted as existing or hallucinated) for over 400 random objects. Results using various captioning models are found in Table 1. As the presence of hallucinations is highly imbalanced (the large majority of predicted objects are not hallucinated), we report balanced accuracy. We provide further details in the appendix, including full confusion matrices.

Surprisingly, although operating over a much more diverse scope, *OpenCHAIR* achieves higher accuracy than CHAIR. We identify that this stems

	# Obj Types	Balanced Accuracy			
		BLIP2	BLIP-L	GIT-B	OFA-L
CH	80*	0.844	0.774	0.899	0.810
OCH	**2400**	**0.945**	**0.944**	**0.943**	**0.930**

Table 1: **Human Evaluation of *OpenCHAIR* and CHAIR**. We perform a manual evaluation of *OpenCHAIR* and CHAIR object-level predictions, as described in Section 4. As seen above, *OpenCHAIR* covers a much larger variety of unique object types while also outperforming CHAIR in per-object predictive accuracy (of whether the given object is present or hallucinated). *CHAIR includes also a synonym list.

Figure 6: **CHAIR Limitations.** The left image exhibits CHAIR's limited vocabulary. Out of all objects predicted by BLIP2, *Scissors* is the only object CHAIR considers during the evaluation. The right image illustrates a limitation stemming from CHAIR's use of a fixed list of synonyms to coarsely aggregate different, semantically similar objects. Hallucinations that occur within the same synonym group are considered as a correct detection; in this example both *Goose* and *Duck* are defined as synonyms of *Bird* even though the image does not display a duck (but rather a goose).

from CHAIR's heavy reliance on coarse synonym lists, as seen in Figure 6 (right). By assessing whether pairs of object names match using a knowledgeable LLM, *OpenCHAIR* performs finer-grained hallucination measurements and achieves superior accuracy even in the more general open-vocabulary setting. We note that this reflects a trade-off between true and false positives, as predicted objects may not be found in *OpenCHAIR* ground-truth lists despite being present in the accompanying images, due to the limited descriptive capacity of text used to generate images. See more details in the Appendix (Tables 3 and 4).

As *OpenCHAIR* was produced by automatic generation followed by manual filtering, we investigate the effect of the small proportion of erroneous data removed (3%) on performance. Table 12 (appendix) shows that it only marginally impacts the resulting *OpenCHAIR* score, validating the high quality of its automatic generation mechanism.

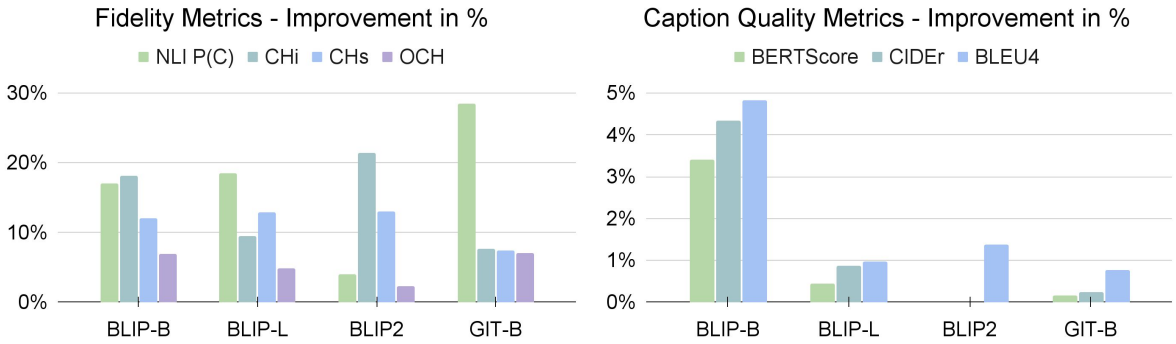


Figure 7: **Reducing Hallucinations While Maintaining Caption Quality**. We show the relative improvement of state-of-the-art VLM models when optimized using *MOCHa* optimization on the COCO Caption Karpathy test set. CH and OCH refer to Chair and *OpenCHAIR* respectively. All results are generated by using their officially provided checkpoints and hyperparameters. Full numeric results are provided in the appendix.

B	*A man in a suit and tie standing by another man in a suit and tie*	*A person taking a tray of apples out of an oven*	*A man sitting on a couch talking on a cell phone*
B+M	*A man in a military uniform talking to a man in a suit and tie*	*A person taking a pan of food out of an oven*	*A man sitting on a couch using a laptop computer*

Figure 8: **Qualitative results** of *MOCHa* applied to an image captioning model (BLIP-Large), along with baseline results without optimization (noted as B+M, B, respectively). We show captions (over COCO) produced from each model using beam search decoding with five beams. Hallucinated details are highlighted. The results illustrate that *MOCHa* encourages captions with high fidelity to the input image (avoiding hallucinations), while preserving a satisfying level of detail.

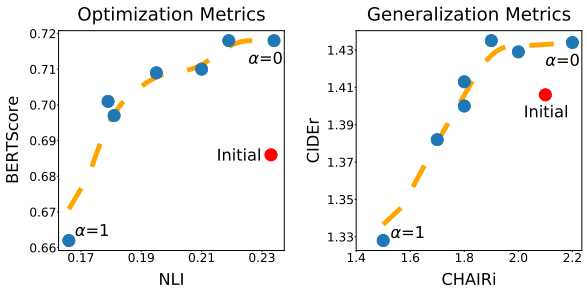


Figure 9: **Fidelity-Adequacy graphs** for pretrained ("initial") and *MOCHa*-optimized BLIP models. As seen above, varying the reward weighting α adjusts the trade-off between caption fidelity (x-axis) and adequacy (y-axis), with intermediate values outperforming the initial model ("Initial"). This holds both for metrics we directly optimize (left) and additional metrics (right), illustrating the generalization ability of our approach.

***MOCHa* Implementation Details.** We test image captioning with *MOCHa* on various SOTA image captioning models of varying architectures and across various sizes. In particular, we test BLIP (Li et al., 2022a), BLIP-2 (Li et al., 2023a) and GIT (Wang et al., 2022). Following standard practice in RL-based image captioning, we use models that have first been fine-tuned on with a standard language modeling loss on the captioning dataset, and then applying PPO reinforcement with our reward function ($\alpha = 0.5$). See the appendix for model checkpoints, parameter counts, and further training settings and hyperparameters.

We test our method on the MS-COCO (Lin et al., 2015) captioning benchmark, using the data split of Karpathy and Fei-Fei (Karpathy and Fei-Fei, 2015) (113K items for training, 5K for evaluation). We report standard captioning metrics along with CHAIR (Rohrbach et al., 2018) and *OpenCHAIR* over generated captions (beam search decoding with 5 beams). We also provide NLI ($\overline{p}$) and BERTScore values, directly optimized by *MOCHa*, as described in Section 3.1. In the appendix, we provide results on additional captioning datasets and metrics to further demonstrate generalization.

***MOCHa* Results.** Figure 7 presents quantitative results of image captioning models on MS-COCO showing the relative improvement of optimizing the baseline SOTA captioning models with *MOCHa*. As shown there, *MOCHa* improves measures of hallucinations in image captioning while preserving or even enhancing standard measures of caption quality. We note that this is despite the fact that the trade-off between these qualities may degrade one or the other when using a sub-optimal reward weighting (see ablations below). Figure 8 provides qualitative examples, illustrating that the *MOCHa*-optimized model generates captions consistent with the image while preserving a satisfying level of detail, consistent with our numeric results.

Model	Quality		Hallucination			
			Closed		Open	
	B@4↑	C↑	CH_i↓	CH_s↓	OCH ↓	$\bar{p}$ ↓
BLIP	41.5	138.4	2.3	3.5	19.2	0.244
BLIP+L	5.5	0.0	12.1	35.4	31.8	0.321
BLIP+T	41.3	137.4	**1.9**	**2.8**	19.2	0.241
BLIP+M	**41.9**	**139.6**	2.1	3.1	**18.3**	**0.206**
BLIP-2	43.4	**144.3**	1.7	2.6	17.0	0.207
BLIP-2+L	5.7	0.0	12.1	33.6	28.4	0.259
BLIP-2+T	43.3	143.5	**1.3**	**2.0**	17.0	0.206
BLIP-2+M	**44.0**	**144.3**	1.4	2.3	**16.6**	**0.199**

Table 2: **Comparison To Prior Works**. Measured for BLIP-Large and BLIP-2. +L/T/M refer to LURE, TLC-A, and *MOCHa* respectively. B@4, C, CH, OCH, and $\bar{p}$ denote BLEU-4, CIDEr, CHAIR, OpenCHAIR, and NLI p(contr.) metrics respectively. All metrics are measured over MS-COCO test set, except for OCH which is measured over our OpenCHAIR benchmark.

Our quantitative results show that *MOCHa* improves performance over base captioning models by most measures, across model architectures and sizes – not only among metrics that we directly optimize but also among non-optimized metrics, measuring general caption quality (e.g. CIDEr), closed-vocabulary hallucinations (CHAIR) and open-vocabulary hallucinations (*OpenCHAIR*). Along with our qualitative observations, this justifies our holistic approach to reducing hallucinations without restriction to a closed object list.

***MOCHa* Comparisons.** In Table 2 we compare *MOCHa* to LURE (Zhou et al., 2024) and TLC-A (Petryk et al., 2023), current SOTA methods addressing VLM hallucinations, applied to the same pretrained BLIP and BLIP-2 models. LURE fails in the pure image captioning setting as its training procedure encourages long-form, highly detailed outputs. While these are in-distribution for instruction-tuned VLMs, they represent an increase in hallucinations relative to concise captions, as well as an extreme deviation from the reference texts; thus it degrades performance across metrics when applied to captioning models such as BLIP and BLIP-2. Regarding TLC-A, as it targets the objects in the closed-vocabulary object list of CHAIR, it shows an expected advantage in this metric, but does not improve the open-vocabulary hallucination rate (measured by *OpenCHAIR*) and even degrades other measures of caption quality, contrasting with the overall improvement shown by our method. More details and results are provided in Appendix B.3, B.4 and C.4.

A number of prior works have proposed dedicated methods for reduced-hallucination image captioning, often using data modification or building multi-component pipelines applied to older vision-language backbones. In Table 8 (appendix), we provide a comparison between these methods and SOTA foundation VLMs applied as-is, reproducing results for the dedicated methods UD-L (Biten et al., 2021), CIIC (Liu et al., 2022), and COS-NET (Li et al., 2022b). We find SOTA VLMs outperform these methods across all metrics, motivating our focus on optimization applied on top of modern foundation models.

Ablations. We ablate the components of our reward function, finding that optimizing for fidelity alone degrades general caption quality, while optimizing for adequacy alone fails to improve hallucinations. This is seen in Figure 9 where extreme values of α (0 or 1) correspond to the edges of the curves. Adjusting the parameter α controlling the trade-off between objectives traces a *Pareto frontier* which outperforms the base model, showing that joint optimization of these objectives has a synergistic effect. The effects of each reward function component are also illustrated qualitatively in Figure 15 (appendix); removing r_f from the reward function leads to increased hallucinations, and removing r_a leads to captions that do not contain sufficient details. We provide full numeric results in the appendix, as well as ablating the effect of our chosen RL algorithm and of the KL-Penalty in our reward.

5 Related Work

We provide a short summary of related works here, with an extended discussion of their methods and differences from our work in the appendix.

Measuring VLM Hallucinations. Several works have proposed holistic measures of generated text fidelity with respect to an input image using embedding similarities or learned metrics; such methods (the "Similarity Based" metrics of Figure 10) include CLIPScore and variants (Hessel et al., 2022; Shi et al., 2022), Semantic Fidelity (Agarwal et al., 2020), VIFIDEL (Madhyastha et al., 2019), and FAIer (Wang et al., 2021). While these metrics may correlate with the presence of hallucinations, they are less interpretable as they do not provide a discrete count of hallucinations in a predicted caption. By contrast, the POPE metric (Li et al., 2023b)

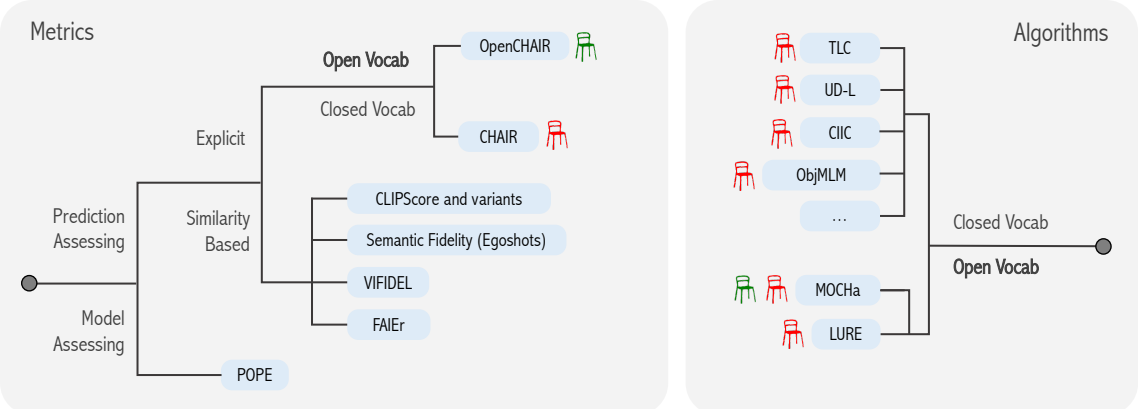


Figure 10: **VLM Caption Hallucination Taxonomy**. We illustrate metrics (left) and algorithms (right) for quantifying and mitigating hallucinations in image-conditioned text generation. We propose an explicit metric for measuring open-vocabulary hallucinations (*OpenCHAIR*) and an open-vocabulary hallucination mitigation algorithm (*MOCHa*). We mark each algorithm with the automatic hallucination rate metric with which it is evaluated (Green – *OpenCHAIR*, Red – CHAIR). Further details are provided in Section 5.

compares ground-truth objects with a model's answers when asked if each object is present; this is open-vocabulary but differs from our setting as it does not score predicted captions but rather assesses a VQA model's general knowledge (indicated as "Model Assessing" in Figure 10(left)).

Reducing VLM Hallucinations. Various methods for mitigating hallucinations in image captioning have been proposed (see Figure 10 (right)). Until recently, research on mitigating hallucinations in captions has largely considered object (noun) hallucinations, typically confined to a closed vocabulary, for instance, objects defined in MS-COCO. Such works include UD-L (Biten et al., 2021), CIIC (Liu et al., 2022), TLC (Petryk et al., 2023), ObjMLM (Dai et al., 2023), and Woodpecker (Yin et al., 2023). Unlike these works, we mitigate hallucinations in the more challenging open-vocabulary setting. The contemporary work LURE (Zhou et al., 2024) proposes a method for the open setting, but their proposed approach (complementary to ours) was not evaluated automatically in an open vocabulary setting due to the lack of an existing benchmark. Figure 10 illustrates which explicit hallucination metric was used to evaluate each algorithm.

As instruction-following VLMs rapidly develop, multiple concurrent works have considered hallucinations in related tasks such as visual question-answering (VQA), applying RL-based methods adopted from research on LLMs (Gunjal et al., 2023; Sun et al., 2023a,b). These methods, which do not directly target our task, also require laborious human annotation to train a supervised reward model to penalize hallucinations, while our approach does not require any explicit supervision.

Deep RL for VLM Text Generation. Deep RL has been widely applied to text generation tasks and specifically for optimizing classical image-captioning metrics (Rennie et al., 2017; Stefanini et al., 2022). Another more recent development is the rise of deep RL for LLMs, which commonly uses the Reinforcement Learning from Human Feedback (RLHF) framework, which requires manual human preference annotation for training a reward model (Ziegler et al., 2020; Stiennon et al., 2020; Ouyang et al., 2022). Beyond LLMs, RLHF has been recently applied to aligning multimodal models with human preferences (Abramson et al., 2022). While such methods succeed in optimizing sequence-level properties, they often suffer from increased hallucinations as a side-effect of optimizing for human preferences or standard NLG sequence-level metrics (as illustrated in Appendix C.4).

6 Conclusion

We have shown the significance of operating in an open-vocabulary setting to effectively quantify and mitigate caption hallucinations. These are explicitly measured by our *OpenCHAIR* benchmark, and our *MOCHa* framework allows for optimizing captioning models to reduce such hallucinations while preserving caption quality. This reduction is demonstrated on our benchmark and other existing metrics. Our method and benchmark may be applied flexibly to a variety of model sizes and architectures, which we foresee providing a framework for future work on hallucination-aware captioning.

7 Limitations

While *OpenCHAIR* provides diverse coverage of object types, it does not directly measure non-object hallucinations (e.g. hallucinated attributes or relations between entities), which are also targeted by sequence-level approaches such as our *MOCHa* optimization. We have focused on objects as a natural extension of the existing closed-vocabulary object hallucination benchmark CHAIR, and due to the fact that extracting and comparing objects from image captions is a relatively well-defined task. Future work may consider extending our OpenCHAIR concept to non-objects, specifically, constructing a robust benchmark for evaluating hallucinations on the attribute-, relation-, predicate-level, or of other types, utilizing elements of our methodology such as open-vocabulary LLM evaluation. Furthermore, we acknowledge that captioning models may show different performance on the synthetic images found in *OpenCHAIR* relative to natural images, although we have found it to correlate empirically to other hallucinations metrics and human intuition.

We emphasize that our work does not solve the hallucination problem completely, although it presents a significant step towards this goal. Note also that we have focused in this work on the image captioning domain, while modern VLMs are often applied to diverse tasks such as VQA and visual instruction-following for which hallucinations also pose a significant challenge. We hope that our proposed strategy will pave the way for future research on hallucination reduction in all of these domains, in which open-vocabulary approaches also present significant promise.

8 Ethics Statement

This work focuses on measuring and mitigating hallucinations in visual-language models (VLMs). As such it is expected to increase the reliability of VLMs and the ability to measure their performance, which is important when using them in real world systems. This is expected to have a positive impact on the use of VLMs in the society. However, we do recognize that the foundation models used in the *OpenCHAIR* construction and evaluation pipeline and those used to calculate the *MOCHa* reward function could propagate biases. We anticipate further research into such biases before relying on our work beyond the research environment.

References

Josh Abramson, Arun Ahuja, Federico Carnevale, Petko Georgiev, Alex Goldin, Alden Hung, Jessica Landon, Jirka Lhotka, Timothy Lillicrap, Alistair Muldal, et al. 2022. Improving multimodal interactive agents with reinforcement learning from human feedback. *arXiv preprint arXiv:2211.11602*.

Pranav Agarwal, Alejandro Betancourt, Vana Panagiotou, and Natalia Díaz-Rodríguez. 2020. Egoshots, an ego-vision life-logging dataset and semantic fidelity metric to evaluate diversity in image captioning models. *arXiv preprint arXiv:2003.11743*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Ali Furkan Biten, Lluis Gomez, and Dimosthenis Karatzas. 2021. Let there be a clock on the beach: Reducing object hallucination in image captioning.

Maria A Bravo, Sudhanshu Mittal, Simon Ging, and Thomas Brox. 2023. Open-vocabulary attribute detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7041–7050.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Dibyadip Chatterjee, Fadime Sener, Shugao Ma, and Angela Yao. 2024. Opening the vocabulary of egocentric actions. *Advances in Neural Information Processing Systems*, 36.

Leah Chong, Ayush Raina, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. 2023. The evolution and impact of human confidence in artificial intelligence and in themselves on ai-assisted decision-making in design. *Journal of Mechanical Design*, 145(3):031401.

Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. 2022. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of ai advice. *Computers in Human Behavior*, 127:107018.

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In *European Chapter of the Association for Computational Linguistics*, pages 2136–2148.

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

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. Clipscore: A reference-free evaluation metric for image captioning.

Jack Hessel, David Mimno, and Lillian Lee. 2018. Quantifying the visual concreteness of words and topics in multimodal datasets. *arXiv preprint arXiv:1804.06786*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

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

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2019. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*.

Natasha Jaques, Shixiang Gu, Dzmitry Bahdanau, José Miguel Hernández-Lobato, Richard E Turner, and Douglas Eck. 2017. Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control. In *International Conference on Machine Learning*, pages 1645–1654. PMLR.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

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

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

Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. 2022b. Comprehending and ordering semantics for image captioning.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context.

Bing Liu, Dong Wang, Xu Yang, Yong Zhou, Rui Yao, Zhiwen Shao, and Jiaqi Zhao. 2022. Show, deconfound and tell: Image captioning with causal inference. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18020–18029.

Pranava Madhyastha, Josiah Wang, and Lucia Specia. 2019. VIFIDEL: Evaluating the visual fidelity of image descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6539–6550, Florence, Italy. Association for Computational Linguistics.

Alessia McGowan, Yunlai Gui, Matthew Dobbs, Sophia Shuster, Matthew Cotter, Alexandria Selloni, Marianne Goodman, Agrima Srivastava, Guillermo A Cecchi, and Cheryl M Corcoran. 2023. Chatgpt and bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Research*, 326:115334.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. 2022. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Suzanne Petryk, Spencer Whitehead, Joseph E. Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. 2023. Simple token-level confidence improves caption correctness.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *CoRR*, abs/1809.02156.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. 2022. Emscore: Evaluating video captioning via coarse-grained and fine-grained embedding matching.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2022. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559.

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

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. 2024. Journeydb: A benchmark for generative image understanding. *Advances in Neural Information Processing Systems*, 36.

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. 2023a. Aligning large multimodal models with factually augmented rlhf.

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023b. Salmon: Self-alignment with principle-following reward models.

Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*, second edition. The MIT Press.

Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. 2024. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36.

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

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Sijin Wang, Ziwei Yao, Ruiping Wang, Zhongqin Wu, and Xilin Chen. 2021. Faier: Fidelity and adequacy ensured image caption evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14050–14059.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Zhenlin Xu, Yi Zhu, Tiffany Deng, Abhay Mittal, Yanbei Chen, Manchen Wang, Paolo Favaro, Joseph Tighe, and Davide Modolo. 2023. Challenges of zero-shot recognition with vision-language models: Granularity and correctness. *arXiv preprint arXiv:2306.16048*.

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

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. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *ICLR*.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences.

A Interactive Visualization

For additional qualitative results, we refer the reader to the interactive visualization tool provided at `index.html`.

We provide image captioning results using BLIP-Large with and without *MOCHa* for 350 randomly selected test images from MS-COCO (Lin et al., 2015) and Flickr30K (Young et al., 2014).

To visually emphasize the hallucination rate in the predictions, for each model we calculate the NLI contradiction probability[1] between the top beam and a ground-truth caption (which is depicted below the image), and report the difference in the contradiction probability between the two models. Samples are ordered via n-gram similarity between the predictions of both models, listing the most different predictions first, allowing for better emphasizing items with evident differences first. This is calculated by considering the top 5 beams of BLIP as reference texts and the top 5 beams of BLIP+*MOCHa* as candidate sentences; we then compute the average BLEU (Papineni et al., 2002) score between each candidate and all references.

B Additional Details

B.1 *MOCHa* Implementation Details

As discussed in Rennie et al. (Rennie et al., 2017), we reduce variance in gradient estimates by shifting the reward function to have zero mean; we apply this to the reward function before adding the KL penalty. To perform this shifting, we subtract the sample mean of this reward (without KL penalty) among all predictions for a given image in a mini-batch.

During each training iteration, we build mini-batches by selecting 10 images and then generating 10 predictions per image (hence 100 image-prediction pairs total). We use nucleus sampling (Holtzman et al., 2019) with $p = 0.9$ and temperature $t = 1.2$, and we cap generations to be at most 40 tokens. We apply PPO reinforcement with clipping parameter $\epsilon = 0.2$. For our reward function, we use coefficients $\alpha = 0.5$ and $\beta \in [0.004, 0.06]$ (depending on the model optimized).

During *MOCHa* training, we freeze the image encoder of all models, training the text encoder components alone. For BLIP-Large and BLIP-Base we use gradient clipping of 5, learning rate of 1e-6 and 4 PPO steps in each iteration. BLIP-2 is trained with low rank adapters (LoRA) over the keys and values of the decoder attention layers (Hu et al., 2021) with a learning rate of 1e-6. GIT-base is trained with a learning rate of 1e-5 with 4 PPO steps and gradient clipping of 5.

All model checkpoints are taken from the Hugging Face Model Hub[2]):

- `salesforce/blip-image-captioning-large`
- `salesforce/blip-image-captioning-base`
- `salesforce/blip2-opt-2.7b-coco`
- `microsoft/git-base-coco`

We train these models for the following number of iterations: 350 for BLIP-B, 1200 for BLIP-L, 3400 for BLIP-2, and 600 for GIT-B.

B.2 *OpenCHAIR* Implementation details

Generating Diverse Captions We start by parsing all objects in MS-COCO's human-annotated captions by first identifying nouns via syntactic parsing[3]. We then filter these for highly concrete nouns, by using the values recorded by Hessel et al. (Hessel et al., 2018) with threshold 4.5. We used these objects, coupled with their corresponding captions, to prompt an instruction-tuned LLM[4] to rephrase the captions with different objects. We used stochastic sampling with top-p of 0.9 and temperature of 0.6 for this LLM generation. While this stage increases the object diversity, we notice that the output still includes many common objects that have a significant overlap with those in MS-COCO. To overcome this issue, we filter out all captions that do not include rare objects, defining an object as rare if its appearance frequency in the dataset is in the lowest 10th percentile. The remaining captions are used as few-shot examples for a LLM[5] (base, not instruction-tuned) to generate new captions, to further increase diversity. We used 10 few shot example for each generated caption, and text is generated using sampling with temperature 0.8. We generate 2,000 captions from the LLM and feed them as prompts to the text-to-image generation model Stable Diffusion XL (Podell et al., 2023),

[1] Using the same pretrained NLI model described in the main paper.

[2] https://www.huggingface.co/models

[3] Using the *en_core_web_md* pipeline from the SpaCy (Honnibal and Montani, 2017) library.

[4] `meta-llama/Llama-2-70b-chat-hf (4-bit quant.)`

[5] `meta-llama/Llama-2-13b`

which generates a single image for each caption. For image generation, we use 40 sampling steps and guidance scale of 10. We also employ negative prompting using the prompt *"unclear, deformed, out of image, disfigured, body out of frame"* to encourage generation of clear objects in the output images.

Evaluation on the *OpenCHAIR* Benchmark Evaluating a captioning model on *OpenCHAIR* is performed as follows: First, all the objects in the caption generated by the captioning model are extracted using the parsing method described in the previous paragraph. For each detected object, an LLM[4] is prompted to determine whether the object is in the GT caption or not using the prompt: *"<s>[INST] An image has the following caption: "⟨input caption⟩". Does the image contain the following object? "⟨input object⟩". Answer yes/no/unsure. The answer is: [/INST]"* . We use greedy decoding for this stage. Objects for which the LLM answers "no" are counted as hallucinations and objects for which the LLM answers "yes" are counted as existing objects. We ignore objects that receive any other response, and report that the amount of such objects are <2% of the total objects considered. Finally, the *OpenCHAIR* hallucination rate is calculated as $OCH := n_h/(n_h + n_e)$, where n_h is the number of hallucinated objects and n_e is the number of existing objects. We note that we added a short list of objects to ignore: *['painting', 'drawing', 'photo', 'picture', 'portrait', 'photograph']*. Since the prefix of the prediction tends to have the following form: *"A photograph of..."*, *"A picture of..."*, these words are identified as concrete objects and then classified as hallucinations by the LLM (as they dont appear in the GT caption), hence should be ignored.

B.3 LURE Comparison

To evaluate LURE (Zhou et al., 2024) in our setup, we followed the authors' instructions[6] and applied their pre-trained model (`YiyangAiLab/LURE`, over MiniGPT-4 with VICUNA-13b) to our predicted captions. Both BLIP-L's and BLIP-2's predictions (with beam search decoding, 5 beams) were supplied to LURE's revisor along with the probabilities of each predicted token for the highest scor-

[5]Reference ground truth captions: *Painting of oranges, a bowl, candle, and a pitcher* (left) and *A giraffe grazing on a tree in the wilderness with other wildlife* (right).

[6]`https://github.com/YiyangZhou/LURE/blob/main/README.md`

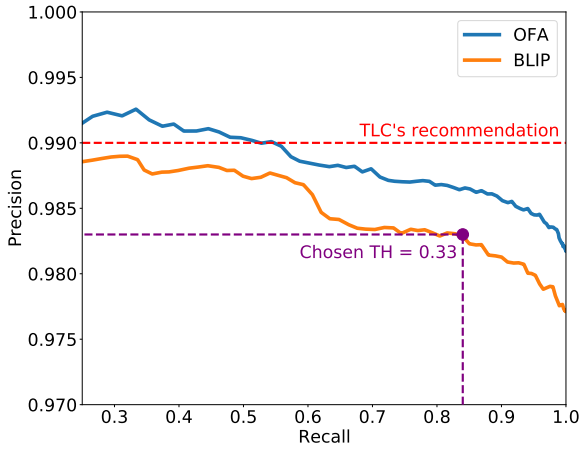


Figure 11: **Precision-recall curve for selecting TLC-A threshold.** As detailed in (Petryk et al., 2023), we compute a precision-recall curve over the predicted object confidences. As illustrated above, the 99% precision threshold recommended by Petryk et al. (Petryk et al., 2023) cannot be achieved by BLIP-Large on the COCO Karpathy validation set. Hence, in our setting we must adjust the threshold to find a reasonable balance between precision and recall.

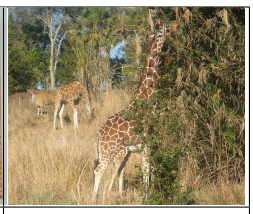
		
$\emptyset$	*a painting of oranges and a silver pitcher on a table*	*two giraffes eating leaves from a tree*
$-r_{kl}$	*a painting of some items*	*some giraffes in the field*
r	*a painting of a pitcher, oranges, and a candle on a table*	*a giraffe eating leaves from a tree in a field*

Figure 12: **Ablating the KL-penalty reward**. Above we show captions sampled from various models: the initial model (BLIP-Large) before optimization ($\emptyset$), the model with *MOCHa* optimization applied and KL penalty ablated ($-r_{kl}$), and an optimized model with our full reward function (r). As is seen above, while the base model outputs various hallucinations (e.g. *a silver pitcher*), the model optimized without KL penalty outputs generic texts without adequate detail, due to over-optimization of the fidelity objective. Optimizing with the full reward function yields captions that are both descriptive and consistent with the input condition.

BLIP2	Pred = 'E'	Pred = 'H'
GT = 'E'	332	42
GT = 'H'	0	54

BLIP-L	Pred = 'E'	Pred = 'H'
GT = 'E'	353	44
GT = 'H'	0	31

GIT-B	Pred = 'E'	Pred = 'H'
GT = 'E'	325	36
GT = 'H'	1	66

OFA-L	Pred = 'E'	Pred = 'H'
GT = 'E'	336	45
GT = 'H'	1	46

Table 3: **Human Evaluation of *OpenCHAIR* Benchmark**. The tables illustrate a correlation measurement between *OpenCHAIR*'s automatic hallucination annotations (Pred) and manual human hallucination annotations (GT). 'E', 'H' stand for 'object **E**xists', 'object **H**allucinated', respectively. BLIP2, BLIP-L, GIT-B and OFA-L stand for BLIP2-2.7b, BLIP-Large, GIT-Base, OFA-Large, all fine-tuned for image-captioning over COCO.

BLIP2	Pred = 'E'	Pred = 'H'
GT = 'E'	416	3
GT = 'H'	4	5

BLIP-L	Pred = 'E'	Pred = 'H'
GT = 'E'	413	2
GT = 'H'	4	9

GIT-B	Pred = 'E'	Pred = 'H'
GT = 'E'	412	1
GT = 'H'	3	12

OFA-L	Pred = 'E'	Pred = 'H'
GT = 'E'	418	2
GT = 'H'	3	5

Table 4: **Human Evaluation of CHAIR Benchmark**. The tables illustrates a correlation measurement between CHAIR's automatic hallucination annotations (Pred) and manual human hallucination annotations (GT). 'E', 'H' stand for 'object **E**xists', 'object **H**allucinated', respectively. BLIP2, BLIP-L, GIT-B and OFA-L stand for BLIP2-2.7b, BLIP-Large, GIT-Base, OFA-Large, all fine-tuned for image-captioning over COCO.

ing beam. After additional parsing, LURE revised BLIP-L's and BLIP-2's sentences, which we then evaluated with various metrics. An example of this procedure is given below:

- Original caption from BLIP-L, masked by LURE's algorithm: *a woman standing in a kitchen with [IDK]*
- LURE instruction: *According to the picture, remove the information that does not exist in the following description:*
- LURE output revision: *In the image, a smiling woman is standing in a kitchen with various food items spread out on a table. She seems to be posing and enjoying the scene, perhaps taken during a cooking session. The table is filled with food, including several pieces of sushi placed near the center, along with some pasta on the left side, cucumbers near the right side, and a couple of apples towards the back.*

B.4 TLC-A Comparison

In order to compare our method to TLC-A (Petryk et al., 2023), we received code from its authors and implemented it in our setup. TLC-A is a decoding-time method applied to auto-regressive captioning models, and in our setting we apply it to different models (e.g. BLIP-Large) than those tested by Petryk et al (e.g. OFA). Of particular note is that TLC-A requires selecting a threshold confidence value, which is used in the decoding phase to re-rank generated beams according to the confidence assigned to COCO object tokens. Petryk et al. recommend calibrating this threshold using the COCO validation set to achieve a precision level of at least 99%; however, in our experiments we find that this value cannot be achieved by the models we consider without sacrificing most of the recall, as illustrated in Figure 11. Therefore, we instead use the COCO validation set to select the best-performing threshold with respect to the CHAIR metric, as shown in Table 5. The selected confidence threshold is 0.33 and it achieves a precision of 98.3% and a recall of 84% over the validation set.

C Additional Results

C.1 Full Quantitative Results

We show in Table 6 the full results, comparing the *MOCHa* optimized models (marked by +M) to the baselines (Figure 7 was prepared using this data).

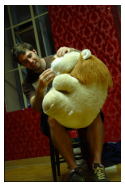
TH	P	R	B@4↑	C↑	CH_i↓	CH_s↓	$\bar{p}$ ↓	BSc ↑
-	-	-	41.5	138.4	2.3	3.5	0.246	0.679
0.10	0.978	0.99	41.4	138.0	2.2	3.38	0.246	0.677
0.21	0.980	0.94	41.4	137.7	2.1	3.14	0.243	0.677
0.33	**0.983**	**0.84**	**41.2**	**137.5**	**1.91**	**2.82**	**0.243**	**0.676**
0.52	0.986	0.61	41.1	136.7	1.97	2.9	0.242	0.675
0.56	0.988	0.55	41.2	136.8	1.94	2.86	0.243	0.675
0.94	1	0.01	41.4	137.7	2.21	3.32	0.247	0.677

Table 5: **Selecting a threshold for TLC-A.** We evaluate TLC-A with different thresholds (as described by Petryk et al. (Petryk et al., 2023)) over the COCO caption Karpathy validation set. In the first row we have BLIP without TLC-A. We indicate the selected threshold which achieves the best CHAIR scores overall **in bold**. B@4, C, CHi, CHs, BSc, $\bar{p}$ denote BLEU-4, CIDEr, CHAIR instance and CHAIR sentence, BERTScore, and NLI p(contr.) metrics respectively. P, R are the precision and recall that each threshold (for predicted object confidences) achieves over the validation set.

Model	B@4↑	C↑	CH_i↓	CH_s↓	OCH ↓	$\bar{p}$ ↓	BSc ↑
BLIP-B	24.8	87.5	2.6	2.8	17.6	0.206	0.557
BLIP-B+M (ours)	**26.0**	**91.3**	**2.2**	**2.5**	**16.4**	**0.176**	**0.576**
BLIP-L	41.5	138.4	2.3	3.5	19.2	0.244	0.679
BLIP-L+M (ours)	**41.9**	**139.6**	**2.1**	**3.1**	**18.3**	**0.206**	**0.682**
BLIP2	43.4	**144.3**	1.7	2.6	17.0	0.207	**0.684**
BLIP2+M (ours)	**44.0**	**144.3**	**1.4**	**2.3**	**16.6**	**0.199**	**0.684**
GIT-B	38.7	128.1	4.2	2.9	24.7	0.284	0.656
GIT-B+M (ours)	**39.0**	**128.4**	**3.9**	**2.7**	**22.9**	**0.221**	**0.657**

Table 6: **Quantitative results** for state-of-the-art VLM models on the COCO Caption Karpathy test set. +M refers to *MOCHa*. BSc and $\bar{p}$ denote BERTScore and NLI contradiction probability rewards. B@4, C, CH, OCH, BSc and $\bar{p}$ denote BLEU-4, CIDEr, CHAIR (i for instance, s for sentence), OpenCHAIR, BERTScore, and NLI p(contr.) metrics respectively. All results are generated by using their officially provided checkpoints and hyperparameters. Best results are shown in **bold**.

LLaVa-RLHF **BLIP-L+*MOCHa***



A man sitting on a chair with a stuffed animal, specijically a teady bear, on his lap

a man sitting on a chair holding a large stuffed animal

Figure 13: **LLaVa-RLHF vs. *MOCHa***. We illustrate that RLHF training does not necessarily solve the hallucination problem of VLM models by showing a generation produced by LLaVa-RLHF (Sun et al., 2023a) compared to BLIP+*MOCHa*. For both models, we use the prompt *"a photography of"* for generation. See Table 10 for a quantitative comparison.

C.2 Comparisons of *OpenCHAIR* and CHAIR

In Tables 3–4 we provide full numeric results for our human evaluation of *OpenCHAIR* and CHAIR across a variety of captioning model predictions, as we discuss in the main paper.

In Figure 14, we illustrate the number of unique object types found in these benchmarks. We note that *OpenCHAIR* contains a much larger diversity of object types, even when considering the full contents of CHAIR's synonym list.

C.3 Additional Ablations

Reward Ablations. In Table 9, we provide numeric results for ablating the fidelity and adequacy terms in our reward function. As discussed in the

[2]Reference ground truth captions: *A car with some surfboards in a field* (left) and *A boy holding umbrella while standing next to livestock* (right).

Model	OCH ↓	B@4↑	C↑	CH_i↓	CH_s↓	$\bar{p}$ ↓	BSc ↑
BLIP-L	0.270	41.5	138.4	2.3	3.5	0.244	0.679
BLIP-L+M	0.259	41.9	139.6	2.1	3.1	0.206	0.682
$-r_f$	0.267	43.0	142.3	2.8	4.4	0.249	0.691
$-r_a$	0.257	41.1	132.9	1.5	2.3	0.174	0.66
$-r_{kl}$	0.241	27.6	98.9	1.4	1.9	0.135	0.62
$-ppo$	0.287	39.4	127.6	2.5	3.76	0.212	0.664

Table 7: **Additional ablation results.** We ablate the effect of the KL penalty reward r_{kl} and the selection of PPO algorithm. As seen above, removing r_{kl} causes the model to over-optimize the fidelity reward ($\bar{p}$), while replacing PPO with the simpler SCST algorithm (described in Section C.3) leads to instabilities that degrade performance across metrics.

Model	B@4↑	M↑	C↑	CH_s↓	CH_i↓
Dedicated					
UD-L+Occ_{XE}	33.9	27.0	110.7	5.9	3.8
UD-L+Occ_{SC}	37.7	28.7	125.2	5.8	3.7
$CIIC_{XE}$	37.3	28.5	119.0	5.3	3.6
$CIIC_{SC}$	40.2	29.5	133.1	7.7	4.5
$COSNet_{XE}$	39.1	29.7	127.4	4.7	3.2
$COSNet_{SC}$	<u>42.0</u>	30.6	<u>141.1</u>	6.8	4.2
End-to-end					
BLIP	41.5	<u>31.1</u>	138.4	<u>3.5</u>	<u>2.3</u>
BLIP-2	**43.4**	**31.7**	**144.3**	**2.6**	**1.7**

Table 8: **Older dedicated methods for reduced-hallucination captioning vs. end-to-end modern VLMs for image captioning**. Results are given on the Karpathy test split of MS-COCO dataset, including closed-vocabulary hallucination metrics as commonly reported by such dedicated methods. B@4, C, M, CH denote BLEU-4, CIDEr, METEOR, and CHAIR metrics respectively. We see that older, dedicated methods with weaker backbones are outperformed by modern VLMs on all metrics, including the smaller BLIP(-Large) and the larger BLIP-2(-2.7B). XE and SC indicate cross-entropy and SCST (RL) optimization respectively. Best and second-best metric values are shown in **bold** and <u>underlined text</u> respectively.

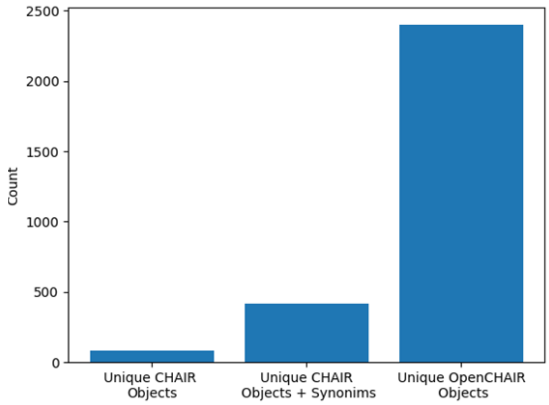


Figure 14: **Object *Type* Coverage, CHAIR vs. *OpenCHAIR*.** We display the object type coverage of CHAIR (over MS-COCO) and *OpenCHAIR*, measured as the number of unique objects. In OPENChair, objects are found using the parsing method described in Section B.2. As can be observed, the proposed benchmark has significantly greater coverage of different objects.

main paper, removing either of these reward terms leads to a degradation with respect to either hallucinations or textual quality, while using both together displays a synergistic effect with hallucinations reduced (as reflected by metrics such as CHAIR) while preserving or even improving caption quality (as reflected by general textual quality metrics such as BLEU-4). We also show a qualitative illustration of these results in Figure 15.

We demonstrate the effect of our KL penalty in the reward function by performing *MOCHa* optimization without this term. As can be observed in the fifth row of Table 7, optimization without this penalty improves the NLI-based reward $\bar{p}$ while degrading other measures of text quality (including non-optimized metrics like CIDEr). We hypothesize that allowing the model to freely deviate from its initial distribution encourages it towards a degenerate solution with respect to $\bar{p}$, which may be the easiest reward term to over-optimize in an unconstrained setting. This is also reflected qualitatively as seen in Figure 12. As illustrated in the figure, captions generated by the model trained without the KL penalty ($-r_{kl}$) do not contradict the image, but rather contain generic text (e.g. *a painting with some items*), lacking adequate detail. By contrast, optimizing with the KL penalty reward yields captions that are both descriptive and consistent with the input condition, reflected in the improved scores across metrics in Table 7 and the

Model	B@4↑	C↑	CH_i↓	CH_s↓	$\bar{p}$ ↓	BSc ↑
BLIP	41.5	138.4	2.3	3.5	0.246	0.679
BLIP+M	41.9	139.6	2.1	3.1	0.206	0.682
$-r_f$	43.0	142.3	2.8	4.4	0.249	0.691
$-r_a$	41.1	132.9	1.5	2.3	0.174	0.66

Table 9: **Reward Ablation.** We ablate the effect of the fidelity r_f and adequacy r_a terms in our reward function, finding that using each alone significantly degrades performance with respect to hallucinations or textual quality.

$\emptyset$	*This is a picture of a large old fashioned car that was parked by a group of people*	*People at festival standing around in open field*
$-r_f$	*A car parked in the grass with a surfer standing near it*	*A woman standing next to a herd of animals with an umbrella*
$-r_a$	*Spectators could enjoy the old fashions of the fifties*	*That are some very nice people who are very fun to view them*
r	*A vintage car parked on a field next to people*	*A young man with a large umbrella next to a herd of animals*

Figure 15: **Ablating our multi-objective reward function**. Above we show captions sampled from models with different reward functions. Top row depicts the initial model (before optimization). As can be seen in the table, generations of the base model ($\emptyset$) and the model trained without the fidelity objective ($-r_f$) contain various hallucinations that contradict the image, like stating that the car was *parked by a group of people*, confusing between an ordinary person and a *surfer*, and stating that the boy is a *woman*. In contrast, those from the model without the adequacy objective ($-r_a$) are generic and neutral with respect to the image (without explicitly contradicting it), e.g. the abstract statement about the *spectators enjoying the old fashions of the fifties*. At last, optimizing for both (r) yields captions that are both descriptive and consistent with the input condition, similar to the reference captions[2] that were provided by human annotators.

quality of predictions of the full reward model (r) in Figure 12. This is attributed to the ability of the KL penalty to mitigate over-optimization, which benefits both optimized rewards.

PPO Ablation. We also ablated the selection of RL algorithm, by replacing PPO with the SCST algorithm upon which it is based (noting that SCST is the common name for the REINFORCE algorithm in the context of image captioning) (Sutton and Barto, 2018; Schulman et al., 2017; Rennie et al., 2017). As is seen in Table 7, PPO outperforms SCST across metrics, consistent with prior work on PPO finding that it avoids instabilities during optimization that may allow it to converge to a more optimal solution (Schulman et al., 2017; Ouyang et al., 2022; Ziegler et al., 2020).

C.4 Additional Comparisons

Comparison to Dedicated Models In Table 8 we provide full numeric results for older dedicated models compared to a modern VLM without further optimization, showing that they are outperformed by all metrics.

Comparison to RLHF-Tuned VLMs. LLaVa-RLHF (Sun et al., 2023a) is a concurrent work, which aims to reduce hallucinations in instruction tuned models using factually-grounded RLHF. In Table 10, we provide a quantitative comparison between LLaVa-RLHF and BLIP+*MOCHa* over 100 samples of the OPENChair benchmark. For LLaVa-RLHF decoding we use both stochastic sampling with the default parameters recommended by the authors, as well as greedy sampling (as beam search is not implemented for LLaVa-RLHF). For a fair comparison, we use greedy decoding for BLIP+MOCHa as well. As LLaVa-RLHF tends to generate long paragraphs which follow an image description with subjective commentary, we terminate generation after a single sentence, which usually corresponds to an image caption. The instruction given to LLaVa-RLHF is "describe the image briefly". As seen in the table, our method outperforms LLaVa-RLHF by this measure of open-vocabulary hallucinations. This is further seen in Figure 13, which shows example captioning predictions for these models, illustrating that LLaVa-RLHF may be more prone to hallucinations.

Evaluation over Flickr30K dataset. We perform a zero-shot generalization test by evaluating a *MOCHa*-tuned model on an additional dataset (different from COCO upon which the model was *MOCHa*-tuned). In Table 11 we can see that the model with *MOCHa* fine-tuning shows an improvement in metrics (NLI and BERTScore) that were optimized on the training data from COCO. Furthermore, we see that non-optimized text quality

Model	OCH ↓
LLaVa-RLHF$_S$	0.396
LLaVa-RLHF$_G$	0.401
BLIP-L+M$_G$	**0.360**

Table 10: OPENChair comparison between LLaVa-RLHF and BLIP-L+*MOCHa* over 100 random samples. For LLaVa-RLHF, S stands for stochastic sampling with default parameters, and G stands for greedy decoding (as beam search is not implemented for LLaVa-RLHF). For fair comparison, we also apply greedy decoding to BLIP-L+*MOCHa*.

Model	B@4↑	C↑	$\bar{p}$ ↓	BSc ↑
BLIP	**29.0**	73.2	0.335	0.603
BLIP+M	28.9	**73.6**	**0.296**	**0.607**

Table 11: **Evaluation over Flickr30K dataset.** We perform a zero-shot evaluation of BLIP-Large with and without *MOCHa* (performed on COCO) on an additional dataset. As seen above, improvements to the optimized metrics ($\bar{p}$ and BERTScore) transfer to the new dataset, while other text quality metrics have similar values before and after *MOCHa*-tuning, suggesting that overall text quality is generally preserved.

Model	*MOCHa*'s Improvement (OCH) in % without filtering	with filtering
BLIP-B	4.9%	4.8%
BLIP-L	2.0%	2.3%
BLIP2	7.3%	6.9%
GIT-B	7.0%	7.1%

Table 12: **Performance of *MOCHa* with and without manual filtering**. We compare performance on the *OpenCHAIR* (OCH) benchmark before and after it is manually filtered, as measured by the improvement provided by *MOCHa* on *OpenCHAIR* scores across various models. We observe similar results before and after filtering, corresponding to the relative high quality of the generated data and consistent with the small proportion of data that was removed.

metrics have similar values between both models, suggesting that *MOCHa* tuning generally preserves overall text quality. Supporting this quantitative evaluation, we provide detailed qualitative results on the Flickr30K dataset in the attached visualization tool.

D Extended Discussion of Previous Work

We provide here an extended discussion of related methods, shown in Figure 10.

D.1 Similarity Based Metrics

CLIPScore (Hessel et al., 2022) propose CLIP cross-modal similarity for detecting mismatches between text and images, including hallucinations, and Shi et al. (2022) propose a similar embedding-based metric for video captioning. However, Xu et al. (2023) find that CLIP tends to assign high similarity to texts with minor modifications ("hard negatives") that contradict the corresponding image. The Egoshots Semantic Fidelity metric (Agarwal et al., 2020) and VIFIDEL (Madhyastha et al., 2019) use embedding similarity between object annotations or detections in images and items in predicted captions. FAIEr (Wang et al., 2021) proposes a learned fidelity metric, which must be trained on automatically-generated scene graphs. Unlike these methods, our benchmark provides an explicit measure of hallucinations that can be directly examined (predicted captions on the *OpenCHAIR* benchmark images).

D.2 Closed Vocabulary Algorithms

UD-L (Biten et al., 2021) identifies object hallucinations with bias towards the prior distribution of objects in context found in the training data, and proposes the use of synthetically debiased captions. CIIC (Liu et al., 2022) focuses on captioning models with a closed-vocabulary object detection backbone, inserting components into the object detector and text decoder to reduce spurious correlations. TLC (Petryk et al., 2023) proposes a text decoding method applied to existing captioning models, to avoid generating COCO object tokens if they have insufficient confidence. The more recent work ObjMLM (Dai et al., 2023) proposes masking objects from closed vocabulary lists as a training objective. The concurrent work Woodpecker (Yin et al., 2023) combines closed-vocabulary object detection with LLM-guided decoding to avoid hallucinations in generated text. Unlike these works, our *MOCHa* optimization method does not rely on a closed list of object types.