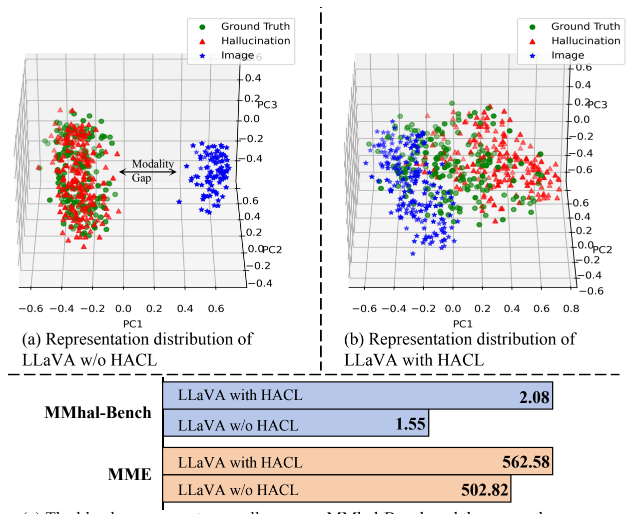
arXiv:2312.06968v4 [cs.CV] 24 Feb 2024

Hallucination Augmented Contrastive Learning for Multimodal Large Language Model

Chaoya Jiang[1] Haiyang Xu[2*] Mengfan Dong[1] Jiaxing Chen[1] Wei Ye[1*]
Ming Yan[2] Qinghao Ye[2] Ji Zhang[2] Fei Huang[2] Shikun Zhang[1]
[1]National Engineering Research Center for Software Engineering, Peking University
[2]Alibaba Group
{jiangchaoya, wye}@pku.edu.cn, shuofeng.xhy@alibaba-inc.com

Abstract

Multi-modal large language models (MLLMs) have been shown to efficiently integrate natural language with visual information to handle multi-modal tasks. However, MLLMs still face a fundamental limitation of hallucinations, where they tend to generate erroneous or fabricated information. In this paper, we address hallucinations in MLLMs from a novel perspective of representation learning. We first analyzed the representation distribution of textual and visual tokens in MLLM, revealing two important findings: 1) there is a significant gap between textual and visual representations, indicating unsatisfactory cross-modal representation alignment; 2) representations of texts that contain and do not contain hallucinations are entangled, making it challenging to distinguish them. These two observations inspire us with a simple yet effective method to mitigate hallucinations. Specifically, we introduce contrastive learning into MLLMs and use text with hallucination as hard negative examples, naturally bringing representations of non-hallucinative text and visual samples closer while pushing way representations of non-hallucinating and hallucinative text. We evaluate our method quantitatively and qualitatively, showing its effectiveness in reducing hallucination occurrences and improving performance across multiple benchmarks. On the MMhal-Bench benchmark, our method obtains a 34.66% /29.5% improvement over the baseline MiniGPT-4/LLaVA. Our code is available on https://github.com/X-PLUG/mPLUG-HalOwl/tree/main/hacl.



(a) Representation distribution of LLaVA w/o HACL

(b) Representation distribution of LLaVA with HACL

(c) The blue bar represents overall score on MMhal-Bench and the orange bar represents the overall score on MME. Noted biger numbers mean better results.

Figure 1. Subfigure (a) and subfigure (b) show the distributions of the last token's representations yielded by LLM for visual or textual token sequences. Blue icons represent images, green icons represent ground truth captions, and red ones represent hallucinative captions generated by GPT-4. HACL refers to our proposed method, Hallucination Augmented Contrastive Learning. In subfigure (a), textual and visual representations have cross-model semantic gaps, while non-hallucinating and hallucinative text representations are mixed. This phenomenon is alleviated by HACL, as shown in subfigure (b). Subfigure (c) shows the empirical results of the hallucination evaluation benchmark MMhal-Bench [44] and the model performance evaluation metric MME [12].

1. Introduction

Large Language Models (LLMs) like GPT-3 [4], LLaMA [45, 46], and GPT-4 [39] have received significant attention for their remarkable text understanding and generation capabilities. Recently, GPT-4V[1] [38] has demonstrated impressive multi-modal abilities in tasks such as image caption and visual question answering, shedding light on the vision-language domain and attracting widespread research interests. Consequently, a new category of models, known as Multi-modal Large Language Models (MLLMs) [2, 10, 27, 33, 49–51, 55], has emerged, aiming to enhance

*Corresponding author

[1]https://openai.com/research/gpt-4v-system-card

LLMs with the capacity to comprehend and handle visual information. To integrate natural language with other modalities, MLLMs incorporate a learnable interface between pre-trained visual encoders and LLMs. Such interface includes learnable query tokens [10, 27, 51, 55] or a projection-based linear model [32, 33] that extracts and integrates information from visual modalities. MLLMs learn this interface to generate answers for multimodal instructions, resulting in remarkable performance in many multimodal tasks.

However, a fundamental limitation of MLLMs is their tendency to produce erroneous or fabricated information that doesn't match the provided visual input, known as hallucination [28, 31, 44, 47]. In this paper, we aim to tackle the issue from the perspective of representation learning. We first check the distribution of textual and visual tokens within the representation space of LLMs (Vicuna [54] in our experiments), in which visual representations are projected by the learned interface. As shown in Figure 1, we have two primary findings:

- *A significant modality gap remains between the textual and visual tokens despite visual projection;*
- *Representations of texts that contain and do not contain hallucinations are entangled, making it challenging to differentiate them.*

These preliminary observations indicate that the current learned interfaces are not effective enough to map visual representations into the textual representation space of LLMs. As a result, it is difficult for MLLMs to discriminate between texts containing minor errors at the level of objects or attributes and those manifesting typical hallucinative expressions. This issue potentially heightens the tendency for MLLMs to generate more hallucinations. Therefore, exploring more effective approaches to align visual representations with LLMs' textual representation space to address hallucinations is crucial.

Inspired by the findings above, we propose hallucination-augmented cross-modal contrastive learning (HACL), which enhances the alignment between visual and textual representations to alleviate hallucinations. Texts with hallucination are used as hard negative examples for image anchors, naturally pulling closer representations of non-hallucinating text and visual samples while pushing way representations of non-hallucinating and hallucinative text. Specifically, we separately feed the visual and textual token sequences into LLMs to obtain global representations for each modality, which is used for contrastive learning. We generate hallucinative image captions with GPT-4 [39]. These hallucinative texts contain partial object attribute errors or introduce additional non-existent information compared to the original image captions. As shown in Figure 1 (b), introducing HACL into LLaVA [33] forces the visual representation closer to the text representation and makes the correct and hallucinated text representations more distinguishable. This effective alignment helps to prevent the generation of hallucinations. Our experiments also show that equipping MLLMs with HACL not only reduces the occurrence of hallucinations but also yields improvements across multiple benchmark evaluations. As shown in Subfigure 1 (c), when equipped with HACL, LLaVA achieves a 29% increase in overall score on the MMhal-Bench benchmark [44], as well as an 11% improvement on the MME [12] benchmark. In conclusion, this paper makes the following contributions:

- We underline a significant cross-modal semantic gap between visual and textual representations and an unexpected representation tangling among text containing and not containing hallucinations in MLLMs. These findings expose the inadequacies of current methodologies in efficiently bridging the gap between visual and textual representations.
- Based on these insights, we propose a simple yet effective method named Hallucination Augmented Cross-Modal Contrastive Learning (HACL). Introducing contrastive learning into MLLMs and using hallucinative text as hard negative samples yields a better cross-modal and more hallucinations-distinguishable representation space.
- Our experiments show that equipping MLLMs with HACL not only minigates hallucinations but also effectively improve the performance across multiple benchmark evaluations.

2. Related Work

Multi-Modal Large Language Foundation Models. The successful application of Large Language Models (LLMs) has paved the way for developing several approaches aiming to augment the perceptual capacities of LLMs with additional modalities, all within a unified model. There are three primary methods for constructing multi-modal large language foundational models, each showing promise for robust zero-shot generalization capabilities in the vision-language domain. For instance, Flamingo [1] is a forerunner in this area, using a frozen vision encoder and a large language model equipped with gated cross-attention for cross-modality alignment. In contrast, PaLM-E [11] integrates extracted visual features directly through linear layers into the pre-trained PaLM [9] model, which boasts 520 billion parameters, thereby leading to robust performance across numerous real-world applications. This approach has been broadly adopted by models such as LLaVA [33], Shikra [7], etc. One significant limitation of this method, however, is the creation of lengthy visual sequences. To address this, BLIP-2 [27], drawing inspiration from DETR [5], developed a Q-former to reduce the sequence length of visual features efficiently. This design has been mirrored by Kosmos-1 [17], mPLUG-Owl [51], and MiniGPT-4 [55].

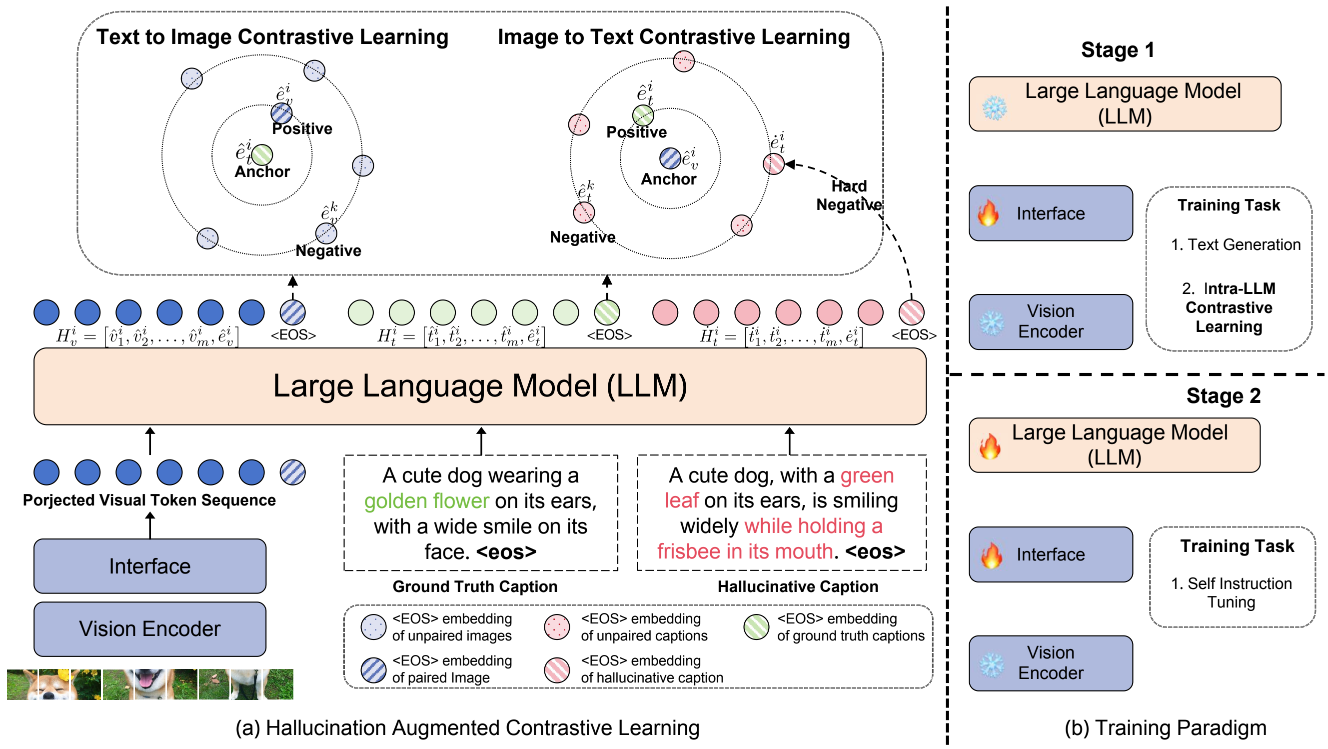


Figure 2. Subfigure (a) illustrates the proposed HACL. In this framework, we employ GPT-4 [39] to generate the hallucinative captions as the the hard negative samples in the image-to-text contrastive learning. Subfigure (b) shows the training paradigm of HACL.

Minigating Hallucination for MLLMs. To address the issue of hallucination in MLLMs, researchers have developed various methods, which can be broadly categorized into two lines. The first line [30, 47] involve limiting the length of instruction data, which typically leads to a reduction in hallucination. For instance, LRV-Instruction[30] takes an intuitive approach by constraining the text length of instructions and constructing counterfactual instructions. However, this may result in less detailed descriptions from the fine-tuned model. The second line utilizes additional artificial data or tools to modify hallucinations in the model's output. For example, LLaVA-RLHF [44] employs manually annotated data as reward signals to guide the model in generating less hallucinative responses. Although effective, this approach requires extra manual annotation data. In this paper, we propose a method from the perspective of representation learning. We introduce hallucinative captions as hard negative samples in contrastive learning, aiming to narrow the gap between visual representations and correct textual representations, while pushing away from hallucinative textual representations. This approach effectively addresses the issue of hallucination and also enhances the model's visual understanding capability.

3. Method

The learnable interface of MLLMs plays a vital role in bridging diverse modalities and mapping visual representations to the representation space of LLMs. Our goal is to refine this interface to facilitate better matching of visual representations with the ground truth text in the representation space, while also increasing the distance between them and hallucinative text. To accomplish this, we propose a new approach called Hallucination Augmented Cross-modal Contrastive Learning (HACL). This approach is inspired by contrastive learning, which is a well-established technique in the fields of representation learning [37] and self-supervised learning [8, 16, 21, 41]. In the following subsection, we first introduce how to incorporate cross-modal contrastive learning during training. Next, we describe how to boost contrastive learning through additional generated hallucinative captions. Finally, we introduce the hallucination augmented contrastive learning training paradigm.

3.1. Cross-modal Contrastive Learning

As shown in Figure 2 (a), our approach can be applied to any MLLMs that maps or abstracts visual information to the textual representation space through an learnable interface. Formally, we assume that the MLLM consists of a vision encoder denoted as $\mathbf{V}_\theta$, a learnable interface denoted as $\mathbf{F}_\alpha$, and a decoder-only based Large Language Model denoted $\mathbf{L}_\beta$ where θ, α, β represent the parameters of each module. Additionally, we also have an unsupervised pre-training dataset, containing N image-text pairs, denoted as $D = \{I_i, T_i\}, i \in [1, 2, \ldots, N]$.

Assuming an image I_i is processed by the vision encoder $\mathbf{V}_\theta$ and the learnable interface $\mathbf{F}_\alpha$, it is transformed into a visual token sequence of length m. Since most LLMs are decoder-only models, in order to obtain the representations that can capture global semantic information. We pass a $< EOS >$ token through an embedding layer $\mathbf{L}_\beta$ to obtain the vector representation $e \in \mathbf{R}^D$ and append it to this sequence. Thus, the new visual token sequence becomes $S_v^i = [v_1^i, v_2^i, \ldots, v_m^i, e_v^i]$, where $v_k^i \in \mathbb{R}^D, k \in [1, 2, /dots, m]$. Similarly, for the caption paired with this image, we also append an <EOS> token to the text token sequence and pass it through the embedding layer of the LLM to obtain the text embedding sequence $S_t^i = [t_1^i, t_2^i, \ldots, t_n^i, e_t^i]$, where $t_k^i \in \mathbb{R}^D, k \in [1, 2, /dots, n]$. Subsequently, the visual embedding sequence S_v and the text embedding sequence S_v are individually passed through the LLM $\mathbf{L}_\beta$ to obtain the final output from the last layer of $\mathbf{L}_\beta$ as following:

$$H_t^i = \mathbf{L}_\beta \left(S_t^i \right) \tag{1}$$

$$H_v^i = \mathbf{L}_\beta \left(S_v^i \right) \tag{2}$$

where $H_v^i = [\hat{v}_1^i, \hat{v}_2^i, \ldots, \hat{v}_m^i, \hat{e}_v^i]$ and $H_t^i = [\hat{t}_1^i, \hat{t}_2^i, \ldots, \hat{t}_n^i, \hat{e}_t^i]$. Afterwards, we obtain the global representation $\hat{e}_v^i$ that captures the overall semantic information of the image I_i, as well as the global representation $\hat{e}_t^i$ that captures the overall semantic information of the ground truth caption T_i.

Afterwards, similar to many existing methods in the field of vision-language pretraining [3, 18–20, 25, 26, 48, 53], we introduce the following contrastive learning strategy. Assuming a batch size of B during the training process, we compute the text-to-image contrastive learning loss as follows:

$$\mathcal{L}_{CL}^t = -\sum_{i=1:B} \frac{1}{B} \, log \left[\frac{f\left(\hat{e}_t^i, \hat{e}_v^i\right)}{f\left(\hat{e}_t^i, \hat{e}_v^i\right) + \sum_{k \neq i} f\left(\hat{e}_t^i, \hat{e}_v^k\right)} \right] \tag{3}$$

where $f\left(\hat{e}_t^i, \hat{e}_v^i\right)$ measures the distance between $\hat{e}_t^i$ and $\hat{e}_v^i$ in a semantic space. Similar, the image-to-text contrastive learning loss as follows:

$$\mathcal{L}_{CL}^v = -\sum_{i=1:B} \frac{1}{B} \, log \left[\frac{f\left(\hat{e}_v^i, \hat{e}_t^i\right)}{f\left(\hat{e}_v^i, \hat{e}_t^i\right) + \sum_{k \neq i} f\left(\hat{e}_v^i, \hat{e}_t^k\right)} \right] \tag{4}$$

3.2. Improving Contrastive Learning with Hallucinative Captions

We propose to improve the effectiveness of contrastive learning by introducing hard negative samples which mimic the hallucinative text generated by MLLMs.

Figure 3. This figure showcases a range of hallucinative captions generated by GPT-4. The hallucinative text is highlighted in red.

Generation of Hallucinative Captions In order to do this, we utilize GPT-4 [39] to incorporate some elements into the ground truth captions that are either inconsistent with the image content or completely absent from it. As shown in Figure 3, these hallucinations can be coarse-grained, focusing on the presence of objects, or fine-grained, focusing on specific attributes such as quantity, properties, or locations. Here is our prompt to GPT-4:

Hallucination in Large-scale Visual Language Models (LVLMs) refers to cases where these models generate descriptions introducing elements that are inconsistent with the content or completely absent from a provided image. These hallucinations can be coarse-grained, focusing on the mere existence of objects, or fine-grained, focusing on more specific attributes or characteristics such as quantity, properties, and locations. Your task is to revise a given caption to create a mirrored version that closely aligns with the original's content and length but incorporates elements of hallucination. The first step involves identifying the objects involved and their associated attributes within the given caption. Subsequently, combine this insight with the details concerning hallucinations provided above to complete your task.

To improve the generation of more appropriate hallucinative captions, we also provide some contextual examples for GPT-4. Please check our appendix for more details.

Hallucination Augmented Contrastive Learning Assuming that we have generated an hallucinative caption $\dot{T}_i$ based on the original caption T_i for the image I_i , and obtained the global representation $\dot{e}_t^i$ of the hallucinative caption using the approach described in subsection 3.1, we can treat it as a negative sample in the image-text contrastive learning. Therefore, the new formula for the image-to-text contrastive learning becomes:

$$\mathcal{L}_{CL}^{v} = - \sum_{i=1:B+1} \frac{1}{B+1} log \left[\frac{f\left(\hat{e}_v^i, \hat{e}_t^i\right)}{f\left(\hat{e}_v^i, \hat{e}_t^i\right) + f\left(\hat{e}_v^i, \check{e}_t^i\right) + \sum_{k \neq i} f\left(\hat{e}_v^i, \hat{e}_t^k\right)} \right] \tag{5}$$

For the text-to-image contrastive learning, we have not made changes and have maintained consistency with the content presented in subsection 3.1.

3.3. Training Paradigm

As shown in Figure 2 (b) which demonstrates how HACL is introduced during the training process of MLLMs. Typically, we incorporate HACL into the first-stage pretraining of the model to optimize the interface $\mathbf{F}_\alpha$ better. Therefore, suppose the loss function of text generation task is denoted as $\mathcal{L}_G$ and the optimization object of the first stage can be defined as follow:

$$\mathcal{O}_\alpha = \arg\min_\alpha \mathcal{L}_G + \left(\mathcal{L}_{CL}^v + \mathcal{L}_{CL}^t\right)/2 \tag{6}$$

In the second stage, we follow the same approach as other methods and fine-tune the model using only instructional data.

4. Experiments

4.1. Implementation

We validated the effectiveness of our method by applying it to four different models: miniGPT-4 [55], LLaVA [33] and LLaVA-1.5 [32].

Data sets For MiniGPT-4, the pre-training phase utilized significantly large datasets such as LAION[42] (115 million), Conceptual Captions [6] (CC3M/CC12M), and others. However, generating hallucinative captions for such enormous datasets is very costly. As a result, for MiniGPT-4, we randomly sampled about 10 million data, representing 10% of the total, and didn't use hallucinative captions for contrastive learning for the remaining data during training. Moreover, we discovered that regardless of not using hallucinative captions for enhancement, our model still significantly enhances models such as MiniGPT-4 [55]. On the other hand, for the LLaVA [33] and LLaVA1.5 [32], which used subsets of LAION/CC/SBU datasets with roughly 558K data, we generated hallucinative captions for every training datum.

Training Settings We followed the original approach for MiniGPT-4 [55] and retrained it using the complete pre-training dataset, about 10M data included hallucinative captions. For LLaVA [33] and LLaVA 1.5 [32], we used the complete pre-training dataset introduced HACL during the first stage of pre-training. We keeping the same hyperparameter settings for all above models. Our experiments were conducted using 16 NVIDIA A100 GPUs with 80G of memory. Due to the increased memory usage during MLLMs training (which includes model and gradient data), the batch size during contrastive learning was affected. To address this, we used a queue of size 16,384, similar to the approaches used for ALBEF [25] and MOCO [8], to store more negative samples. We used Deepspeed [**?**] for LLaVA and LLaVA 1.5, with a batch size of 64 and 32 on a single GPU, respectively. For MiniGPT-4, the batch size was 8.

4.2. Effectiveness of HACL on Mitigating Hallucination

To verify the efficacy of our proposed method in addressing hallucination issues, we leveraged two widely used benchmark evaluation datasets that evaluate the presence of hallucinations in models. These datasets included MMHal-Bench [44] and POPE [28]. MMHal-Bench offers a comprehensive evaluation of models that encompasses multiple perspectives, such as attributes, relations, and counting. On the other hand, POPE particularly focuses on hallucinations related to objects. We employed both datasets to measure the effectiveness of our method in addressing hallucination across various scenarios.

Evaluation on MMHal-Bench For the MMHal-Bench [44]. We apply our method to iniGPT-4 [55], LLaVA [33], LLaVA1.5 [32] and compare the results with other recent vision-language models, including MKosmos-2 [40], IDEFICS [22], InstructBLIP [10], and anther LLaVA-RLHF [44]. Following [44], we use GPT-4 to evaluate the overall score and hallucination rate of different MLLMs. Table 1 demonstrates a significant improvement in the overall performance of MMHal-Bench after applying our method to LLaVA [33], MiniGPT-4[55], and LLaVA1.5[32]. Notably, MiniGPT-4-HACL exhibited considerable performance gain over MiniGPT-4 [55]. Moreover, compared with LLaVA-RLHF[44], a recently proposed method that uses human feedback and reinforcement learning to address hallucinations, LLaVA-HACL showed an even more significant improvement.

Evaluation on POPE In addition, we obtained consistent results using MMHal-Bench [44] in the POPE evaluation benchmark [28]. Table 2 shows that miniGPT-4-HACL and LLaVA-HACL both demonstrated significant improvements compared to the original model. Of particular note, the average F1 score of LLaVA-HACL increased by 17.8% compared to LLaVA [33], while the Yes ratio decreased from 99.55 to 48.25. Furthermore, by applying our method to LLaVA1.5 [32], LLaVA1.5-HACL easily achieved SOTA on this benchmark. Noted that LLaVA1.5 [32] is a high-performing model with a low likelihood of generate hallucination, surpassing MiniGPT-4 [55] and LLaVA [33]. This model's impressive benchmark scores make it a valuable

Method	Overall Score ↑	Hallucination Rate ↓	Score in Each Question Type ↑							
			Attribute	Adversarial	Comparison	Counting	Relation	Environment	Holistic	Other
Kosmos-2 [40]	1.69	0.68	2.00	0.25	1.42	1.67	1.67	2.67	**2.50**	1.33
IDEFICS$_{9B}$ [22]	1.89	0.64	1.58	0.75	2.75	1.83	1.83	2.50	2.17	1.67
IDEFICS$_{80B}$ [22]	2.05	0.61	2.33	1.25	2.00	2.50	1.50	3.33	2.33	1.17
InstructBLIP$_{7B}$ [10]	2.10	0.58	3.42	2.08	1.33	1.92	2.17	3.67	1.17	1.08
InstructBLIP$_{13B}$ [10]	2.14	0.58	2.75	1.75	1.25	**2.08**	**2.50**	**4.08**	1.50	1.17
LLaVA-RLHF$_{7B}$ [44]	2.05	0.68	2.92	1.83	**2.42**	1.92	2.25	2.25	1.75	1.08
LLaVA$_{7B}$ [33]	1.55	0.76	1.33	0.00	1.83	1.17	2.00	2.58	1.67	1.83
LLaVA$_{7B}$-HACL [33]	2.08 (↑ 0.53)	0.62 (↓ 0.15)	2.94	2.01	2.27	1.64	2.35	2.14	1.67	1.63
miniGPT-4$_{7B}$ [33]	1.39	0.71	0.75	1.83	2.16	0.91	1.25	1.33	0.91	1.91
miniGPT-4$_{7B}$-HACL	1.80 (↑ 0.31)	0.65 (↓ 0.06)	1.22	1.85	2.23	1.74	2.13	2.48	1.03	1.58
LLaVA1.5$_{7B}$ [33]	2.08	0.52	2.75	2.00	2.33	**2.08**	1.50	1.91	1.91	2.16
LLaVA1.5$_{7B}$-HACL	**2.13** (↑ 0.05)	**0.50** (↓ 0.02)	**2.95**	**2.15**	2.29	1.97	1.53	1.98	2.02	**2.19**

Table 1. Evaluation results for different MLLMs on MMHal-Bench.

Datasets	Metrics	Shikra [7]	InstructBLIP [10]	MM-GPT [14]	mPLUG-Owl [51]	MiniGPT-4 [55] w/ HACL		LLaVA [33] w/ HACL		LLaVA1.5 [32] w/ HACL	
Random	Accuracy (↑)	86.90	88.57	50.10	53.97	54.64	80.49 (↑ 25.84)	50.97	82.16 (↑ 31.18)	88.17	88.59 (↑ 0.42)
	Precision (↑)	94.40	84.09	50.05	52.07	57.92	94.32 (↑ 36.39)	50.19	87.30 (↑ 37.11)	97.68	98.62 (↑ 0.93
	Recall (↑)	79.27	95.13	100.00	99.60	34.65	75.34 (↑ 40.69)	99.13	76.53 (↓ 22.59)	78.93	80.60 (↑ 1.66)
	F1-Score (↑)	86.19	89.27	66.71	68.39	43.35	83.82 (↑ 40.46)	66.71	81.56 (↑ 14.85)	87.31	88.70 (↑ 1.39)
	Yes (→ 50%)	43.26	56.57	98.90	95.63	31.32	44.33 (↑ 13.01)	99.90	45.19 (↓ 54.71)	41.64	44.43 (↑ 2.78)
Popular	Accuracy (↑)	83.97	82.77	50.00	50.90	56.67	78.32 (↑ 21.64)	49.87	79.32 (↑ 29.44)	87.46	87.94 (↑ 0.48)
	Precision (↑)	87.55	76.27	50.00	50.46	58.69	79.23 (↑ 20.54)	49.93	80.34 (↑ 30.41)	95.17	97.23 (↑ 2.06)
	Recall (↑)	79.20	95.13	100.00	99.40	44.74	74.54 (↑ 29.80)	99.27	76.60 (↓ 22.67)	78.93	79.31 (↑ 0.37)
	F1-Score (↑)	83.16	84.66	66.67	66.94	50.74	76.85 (↑ 26.11)	66.44	78.43 (↑ 11.99)	86.29	87.36 (↑ 1.07)
	Yes (→ 50%)	45.23	62.37	100.00	98.57	62.20	45.23 (↓ 16.97)	99.40	47.64 (↓ 51.76)	41.46	45.03 (↑ 3.57)
Adversarial	Accuracy (↑)	83.10	72.10	50.00	50.67	54.50	71.32 (↑ 16.82)	49.70	74.47 (↑ 24.77)	85.93	86.54 (↑ 0.61)
	Precision (↑)	85.60	65.13	50.00	50.34	57.21	70.53 (↑ 13.32)	49.85	73.55 (↑ 23.70)	91.78	93.01 (↑ 1.23)
	Recall (↑)	79.60	95.13	100.00	99.33	41.45	73.45 (↑ 32.00)	99.07	76.40 (↓ 22.67)	78.93	79.52 (↑ 0.59)
	F1-Score (↑)	82.49	77.32	66.67	66.82	48.07	71.96 (↑ 23.89)	66.32	74.95 (↑ 8.63)	84.87	85.73 (↑ 0.86)
	Yes (→ 50%)	46.50	73.03	100.00	98.67	38.32	48.23 (↑ 9.91)	99.37	51.93 (↓ 47.44)	43.00	46.33 (↑ 3.33)

Table 2. Object hallucination benchmark using POPE [28] evaluation pipeline . "Yes" signifies the likelihood of the model producing a positive response.

foundation to build upon.

4.3. Effectiveness of HACL on Visual Comprehension

HACL has shown effectiveness in solving the issue of hallucination. Nevertheless, we intend to explore the influence of HACL on the model's abilities of visual comprehension and generation. To achieve this objective, we carried out assessments on common benchmarks, such as Visual Question Answering (VQA) [15, 36, 43] after incorporating HACL into the MLLMs. Furthermore, as MLLMs possess robust zero-shot capabilities, traditional evaluation metrics often fail to provide a detailed assessment of their abilities. Additionally, their inability to match the given answer correctly exacerbates significant robustness issues. To mitigate these challenges, the research community introduced a series of benchmarks. These benchmarks aim to systematically structure and evaluate complex multi-modal tasks from various perspectives. Therefore, we also evaluated the model's performance on recently designed MLLM-focused Multi-modal Benchmarks including MME [12], MMBench [34], MM-Vet [52], SEED-Bench [23].

Results on Benchmark Tasks Our evaluation includes six popular benchmarks, as summarized in Table 3. We applied the HACL to three baselins: MiniGPT-4, LLaVA, and LLaVA1.5, and compared their performance to other State-of-the-Art (SOTA) MLLMs such as BLIP2[27], InstructBLIP [10], Shikra [7], and Qwen-VL-Chat [2]. Our experimental results show that our approach successfully enhances the performance of original models across a range of VQA datasets. Notably, LLaVA-HACL outperforms LLaVA [33] in terms of consistency and accuracy across all VQA datasets. Additionally, when compared to LLaVA1.5 [32], LLaVA1.5-HACL achieves better results in General VQA benchmarks abd zero-shot VQA tasks, implying that MLLMs may not only mitigate hallucinations but also improve correlations between visual and textual information, which further refines the generalization ability of models.

MLLM-oriented Multi-modal Benchmarks. We applied HACL to MiniGPT-4 [55], LLaVA [33], LLaVA1.5 [32] and evaluate them on five recently popular multi-modal benchmarks in a zero-shot manner. For a fair comparison, we select models with similar language model sizes, particularly those from the LLaMA [45] family, and detail their differences in the vision encoder. The results of our evaluation are listed in Table 4. We discovered that after implementing HACL, all three models exhibited improvements across multiple benchmarks. Notably, for LLaVA and MiniGPT-4, the enhancement was particularly evident on the MME [12] benchmark. For instance, after implementing HACL, LLaVA's MME score improved from 581.67 to 653.94. These results indicate that our methodology can not only reduce the instances of model hallucination but also enhance the model's visual comprehension capabilities.

Method	#Params	General VQA		General VQA (Zero-shot)		
		VQAv2	GQA	VizWizQA	TextVQA	SciQA (IMG)
BLIP-2 [27]	8.2B	65.0	41.0	19.6	42.5	61.0
InstructBLIP [10]	8.2B	-	49.2	34.5	50.1†	60.5
Unified-IO$_{XL}$ [35]	2.9B	77.9	-	57.4‡	-	-
PaLM-E-12B [11]	12B	76.2	-	-	-	-
Shikra [7]	7.2B	77.4	-	-	-	-
Qwen-VL-Chat [2]	9.6B	78.2	57.5	38.9	61.5‡	**68.2**
LLaVA [32]	7.2B	71.3	41.3	36.7	50.2†	61.5
LLaVA-HACL [32]	7.2B	73.3	42.5	37.4	52.2†	62.4
MiniGPT-4 [32]	7.2B	65.2	30.8	30.2	52.3†	58.4
MiniGPT-4-HACL [32]	7.2B	68.9	32.3	31.7	54.2†	60.3
LLaVA1.5 [32]	7.2B	78.5	62.0	50.0	58.2†	66.8
LLaVA1.5-HACL [32]	7.2B	**79.1**	**62.5**	**50.5**	59.8†	67.3

Table 3. **Performance comparison on visual question answering.** For VQA, accuracy is reported. Note that specialists are fine-tuned on each individual dataset. † denotes OCR inputs are utilized. ‡ indicates the model has trained on the dataset. We gray out those specialists' methods which are individually fine-tuned on the dataset as well as those fine-tuned results of generalists.

Method	Vision Encoder	Language Model	MME	MMBench	MM-Vet	SEED-Bench
BLIP-2 [27]	ViT-g (1.3B)	Vicuna (7B)	1293.84	-	22.4	46.4
mPLUG-Owl [51]	ViT-L (0.3B)	LLaMA (7B)	967.34	46.6	-	34.0
InstructBLIP [10]	ViT-g (1.3B)	Vicuna (7B)	1212.82	36.0	26.2	53.4
LLaMA-Adapter-v2 [13]	ViT-L (0.3B)	LLaMA (7B)	1328.40	39.5	31.4	32.7
Otter [24]	ViT-L (0.3B)	LLaMA (7B)	1292.26	48.3	24.6	32.9
Qwen-VL-Chat [2]	ViT-G (1.9B)	Qwen (7B)	1487.58	60.6	-	58.2
LLaVA [33]	ViT-L (0.3B)	Vicuna (7B)	502.82	36.2	28.1	33.5
LLaVA-HACL [33]	ViT-L (0.3B)	Vicuna (7B)	562.58	37.8	28.4	33.9
MiniGPT-4 [55]	ViT-g (1.3B)	Vicuna (7B)	581.67	23.0	22.1	42.8
MiniGPT-4-HACL [55]	ViT-g (1.3B)	Vicuna (7B)	653.94	24.5	23.8	42.5
LLaVA-1.5 [32]	ViT-L (0.3B)	Vicuna (7B)	1510.70	64.3	**30.5**	58.6
LLaVA-1.5-HACL [32]	ViT-L (0.3B)	Vicuna (7B)	**1530.10**	**64.5**	30.4	**58.9**

Table 4. **Zero-shot multi-modal evaluation on multi-modal benchmarks** including MME [12], MMBench [34], MM-Vet [52], SEED-Bench [23]. The overall scores are reported for evaluation. For MMBench, we report test results.

4.4. Ablation Study

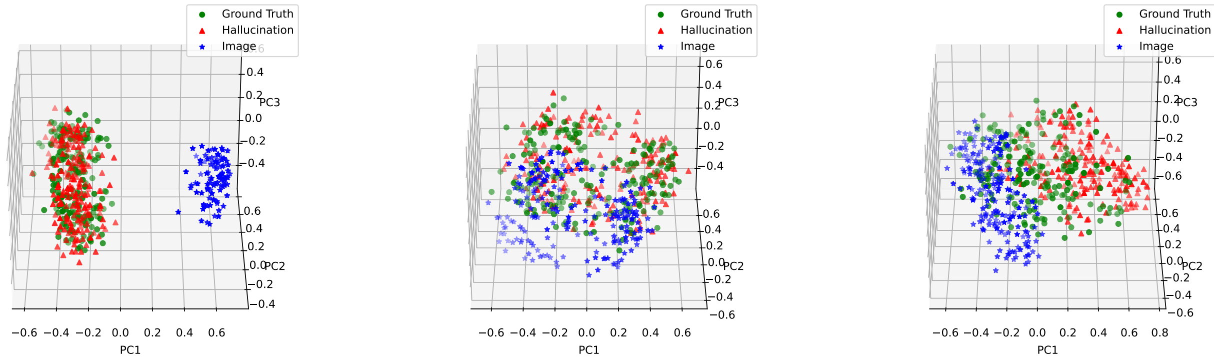
Model	w/ CL	w/ HC	POPE	MMHal	VQA	MME
LLaVA	×	×	66.48	1.55	71.32	502.82
LLaVA	✓	×	69.23	1.67	72.98	549.04
LLaVA	✓	✓	78.31	2.08	73.30	562.58
MiniGP-4	×	×	47.38	1.39	65.2	581.67
MiniGP-4	✓	×	53.54	1.45	67.6	633.21
MiniGP-4	✓	✓	77.54	1.80	68.9	653.94
LLaVA1.5	×	×	86.15	2.08	78.5	1510.70
LLaVA1.5	✓	×	86.31	2.09	78.7	1523.84
LLaVA1.5	✓	✓	87.26	2.13	79.1	1530.10

Table 5. The result of ablations for the impact of hallucinative captions. We report the text-dev score results of POPE[28], MMhal-Bench [44], VQA and MME. w/ CL refers to training MLLMs with Contrastive Learning for MLLMs, w/ HC refers to utilize hallucinative captions to enhance the contrastive learning. For POPE, we report the average F1 score.

model	+VE	+LLM	POPE	MMHal	VQA	MME
LLaVA-HACL	✓	×	63.42	1.43	65.0	324.50
LLaVA-HACL	×	✓	78.53	2.08	74.2	580.32
LLaVA-HACL	×	×	78.31	2.08	73.3	562.58
LLaVA1.5-HACL	✓	×	68.89	1.53	69.6	459.34
LLaVA1.5-HACL	×	✓	87.23	2.13	79.4	1542.48
LLaVA1.5-HACL	×	×	87.26	2.13	79.1	1530.10

Table 6. Results of models under different training paradigms. "+VE" denotes training the Visual Encoder during Stage 1 pretraining, while "+LLM" indicates training the LLM during Stage 1 pretraining.

Impact of Hallucinative Captions To validate the effectiveness of using hallucinative captions as hard negative samples in contrastive learning for resolving hallucinations, we conducted the following experiments: In the Stage 1 pre-training phase, we did not introduce any additional hal-



(a) w/o HACL (b) w/ CL (c) w/ HACL

Figure 4. This figure illustrates the visualization of various data distributions. The blue icons represent visual data extracted from images, the green icons denote ground truth caption data, and the red icons signify hallucinative caption data. The label "w/o HACL " represents the data distribution obtained from the original model output without employing our proposed method. On the other hand, "w/ CL" indicates the data distribution resulting from the model output utilizing contrastive learning. Lastly, "w/ HACL " indicates the data distribution generated by the model output using our proposed method.

lucinative captions, and the contrastive learning loss was calculated solely based on the equations 3 and 9 discussed in subsection 3.1 of our paper. We conducted experiments on MLLMs including LLaVA [33], MiniGPT-4 [55], and LLaVA1.5 [32], and reported the results on benchmarks such as POPE and MMHal-Bench. Additionally, we also reported results on the MME and VQA benchmarks. As illustrated in the Table 5, absent the facilitation from hallucinative captions, the models displayed moderate improvements on hallucination benchmarks such as MMhal-Bench, yet these improvements were somewhat constrained. However, the subsequent inclusion of hallucinative captions resulted in a marked enhancement on the same hallucination benchmark, thus affirming the potency of the hallucinative captions. Furthermore, we observed analogous improvements in the model's performance on both MME and VQA. Our hypothesis asserts that hallucinative captions aid MLLMs in diverting the visual representation from hallucinations and other textual inaccuracies. This action helps avoid instances of hallucination. Furthermore, contrastive learning supports the model by aligning the semantics of image-text, which ultimately enhances the model's effectiveness.

Discussion on Training Paradigm We have observed that certain Multimodal Language-and-Vision Models (MLLMs) may not freeze the activity of either the visual encoder or the Language-and-Vision Models (LLMs) during the initial stage of pretraining. To assess the impact of our methodology under such distinct training paradigms, we independently tested models where either the Visual Encoder or the LLMs were active during the first pretraining phase. These tests were conducted on two platforms: LLaVA and LLaVA1.5 and subsequently evaluated against multiple benchmark standards. As illustrated in Table 6, the models experienced a significant performance decline when LLMs are activated. We hypothesize that this downturn could be linked to low-quality data in the first pretraining stage and the introduction of additional contrast learning tasks, both of which affect the LLMs' representation distribution. This culminates in the catastrophic forgetting of the LLMs. Conversely, initiating the Visual Encoder led to a modest performance boost. This might be attributed to the fact that the target parameters our model can optimize extend beyond the learnable interface and incorporate the visual encoder as well. This expanded scope paves the way for a more successful alignment of visual and text representations within the MLLMs.

4.5. Visualization

The objective of our research is the introduction of HACL, to further enhance the visual representation output of our interface. The aim is to closely align the output to the correct textual representation within the representation space of Language Models (LLMs) and, at the same time, distance it from hallucinative and other incorrect textual representations. To substantiate our objective, we randomly selected 200 image-text pairs from the COCO [29] val2017 dataset. Using GPT-4, we generated hallucination samples and subsequently reduced these samples using the hidden state representation of the last token through LLMs for visualization purposes. The data distribution under three conditions: without employing HACL, instigating cross-modal contrastive learning but without the use of hallucination-enhanced samples, and usage of hallucination-enhanced sample contrast learning was visualized respectively. The MLLM utilized in our study was LLaVA. As illustrated in Figure 4 (a), a substantial modality gap is observable in the data distribution without contrast learning. In Figure 4 (b), after applying contrast learning, although the modal gap decreased, a differentiation in the distribution of hallucination samples and ground truth samples was unattainable. In Figure 4 (c), with the application of hallucination augmentation in contrast learning, not only did the modal gap decrease, but the hallucination sample distribution was also significantly distanced.

5. Conclusion

This paper addresses the issue of hallucinations in Multi-modal Large Language Models (MLLMs) and proposes a method called Hallucination Augmented Contrastive Learning (HACL) to improve the alignment between visual and textual representations. By using contrastive learning on projected text and visual token sequences, and incorporating hallucinative captions as hard negative samples, HACL effectively reduces the occurrence of hallucinations. Experimental results demonstrate that incorporating HACL enhances the performance of MLLMs and significantly reduces the occurrence of hallucinations in benchmark evaluations.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *ArXiv*, abs/2308.12966, 2023. 1, 6, 7

[3] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022. 4

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. 1

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 5

[7] Ke Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *ArXiv*, abs/2306.15195, 2023. 2, 6, 7

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3, 5

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113, 2022. 2

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500, 2023. 1, 2, 5, 6, 7

[11] Danny Driess, F. Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Ho Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Peter R. Florence. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, 2023. 2, 7

[12] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 1, 2, 6, 7

[13] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, W. Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Jiao Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. *ArXiv*, abs/2304.15010, 2023. 7

[14] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping

Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023. 6

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[17] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. *ArXiv*, abs/2302.14045, 2023. 2

[18] Chaoya Jiang, Haiyang Xu, Chenliang Li, Ming Yan, Wei Ye, Shikun Zhang, Bin Bi, and Songfang Huang. TRIPS: Efficient vision-and-language pre-training with text-relevant image patch selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4084–4096, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 4

[19] Chaoya Jiang, Haiyang Xu, Wei Ye, Qinghao Ye, Chenliang Li, Ming Yan, Bin Bi, Shikun Zhang, Fei Huang, and Songfang Huang. Bus: Efficient and effective vision-language pretraining with bottom-up patch summarization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2900–2910, 2023.

[20] Chaoya Jiang, Haiyang Xu, Wei Ye, Qinghao Ye, Chenliang Li, Ming Yan, Bin Bi, Shikun Zhang, Fei Huang, and Ji Zhang. Copa: Efficient vision-language pre-training through collaborative object-and patch-text alignment. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4480–4491, 2023. 4

[21] Chaoya Jiang, Wei Ye, Haiyang Xu, Miang yan, Shikun Zhang, Jie Zhang, and Fei Huang. Vision language pretraining by contrastive learning with cross-modal similarity regulation. In *Annual Meeting of the Association for Computational Linguistics*, 2023. 3

[22] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open webscale filtered dataset of interleaved image-text documents, 2023. 5, 6

[23] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 6, 7

[24] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *ArXiv*, abs/2305.03726, 2023. 7

[25] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 4, 5

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 4

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023. 1, 2, 6, 7

[28] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji rong Wen. Evaluating object hallucination in large vision-language models. *ArXiv*, abs/2305.10355, 2023. 2, 5, 6, 7

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 8

[30] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 3

[31] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning, 2023. 2

[32] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *ArXiv*, abs/2310.03744, 2023. 2, 5, 6, 7, 8

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023. 1, 2, 5, 6, 7, 8

[34] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 6, 7

[35] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *ArXiv*, abs/2206.08916, 2022. 7

[36] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 6

[37] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[38] OpenAI. Gpt-4v(ision) system card. 2023. 1

[39] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 1, 2, 3, 4

[40] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *ArXiv*, abs/2306.14824, 2023. 5, 6

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 5

[43] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 6

[44] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf. *ArXiv*, abs/2309.14525, 2023. 1, 2, 3, 5, 6, 7

[45] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023. 1, 6

[46] Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023. 1

[47] Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiao wen Dong, Weijia Li, Wei Li, Jiaqi Wang, and Conghui He. Vigc: Visual instruction generation and correction. *ArXiv*, abs/2308.12714, 2023. 2, 3

[48] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 4

[49] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-docowl: Modularized multimodal large language model for document understanding. *CoRR*, abs/2307.02499, 2023. 1

[50] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[51] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 1, 2, 6, 7

[52] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 6, 7

[53] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021. 4

[54] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 2

[55] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv*, abs/2304.10592, 2023. 1, 2, 5, 6, 7, 8