RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback

Tianyu Yu[1] Yuan Yao[2*] Haoye Zhang[1] Taiwen He[1] Yifeng Han[1]
Ganqu Cui[1] Jinyi Hu[1] Zhiyuan Liu[1*] Hai-Tao Zheng[134*] Maosong Sun[1]

[1]Tsinghua University [2]National University of Singapore
[3]Shenzhen International Graduate School, Tsinghua University
[4]Pengcheng Laboratory, Shenzhen, China

yiranytianyu@gmail.com yaoyuanthu@gmail.com

https://rlhf-v.github.io

Abstract

Multimodal Large Language Models (MLLMs) have recently demonstrated impressive capabilities in multimodal understanding, reasoning, and interaction. However, existing MLLMs prevalently suffer from serious hallucination problems, generating text that is not factually grounded in associated images. The problem makes existing MLLMs untrustworthy and thus impractical in real-world (especially high-stakes) applications. To address the challenge, we present RLHF-V, which enhances MLLM trustworthiness via behavior alignment from fine-grained correctional human feedback. Specifically, RLHF-V collects human preference in the form of segment-level corrections on hallucinations, and performs dense direct preference optimization over the human feedback. Comprehensive experiments on five benchmarks in both automatic and human evaluation show that, RLHF-V can enable substantially more trustworthy MLLM behaviors with promising data and computation efficiency. Remarkably, using 1.4k annotated data samples, RLHF-V significantly reduces the hallucination rate of the base MLLM by 34.8%, outperforming the concurrent LLaVA-RLHF trained on 10k annotated data. The final model achieves state-of-the-art performance in trustworthiness among open-source MLLMs, and shows better robustness than GPT-4V in preventing hallucinations aroused from over-generalization.

1. Introduction

The recent success of Multimodal Large Language Models (MLLMs) marks a significant milestone in AI research [2, 4, 11, 12, 19, 27, 29, 42, 51]. By connecting visual signals and Large Language Models (LLMs), MLLMs show unprecedented capabilities in multimodal understanding, reasoning, and interaction [28, 29, 44]. The models are typically pre-trained on large-scale image-text data to learn the foundational multimodal knowledge and capabilities [2, 4, 12, 19]. To steer the model behavior, most MLLMs are further fine-tuned with instruction tuning (also known as supervised fine-tuning), which supervises models to clone the behavior from demonstration data, enabling MLLMs to engage users in various open-world tasks [4, 11, 25, 27, 47].

However, current MLLM behaviors are not well aligned with human preferences. A glaring issue is their tendency to produce *hallucinations* — responses that are not factually grounded in the associated images [21, 25, 29, 37]. This typically includes descriptions of non-existing visual contents and errors in descriptions. As shown in Figure 1, current MLLMs can hallucinate about objects, attributes, numbers, positions, actions, etc. Quantitatively, our human evaluation shows that the problem is prevalent among state-of-the-art MLLMs, where even the most advanced GPT-4V [29] contains obvious hallucinations in 45.9% responses. The problem makes existing MLLMs untrustworthy and thus impractical in real-world (especially high-stakes) applications, such as guiding visually impaired individuals [29] or autonomous driving systems [43].

We argue that the problem arises from the lack of positive/negative human feedback in instruction-tuned models, making it challenging to learn the precise behavior boundaries to exclude hallucination. To address the problem, we propose RLHF-V, a novel framework that aligns MLLM behavior by learning from human feedback. However, simply applying traditional Reinforcement Learning from Hu-

*Corresponding authors

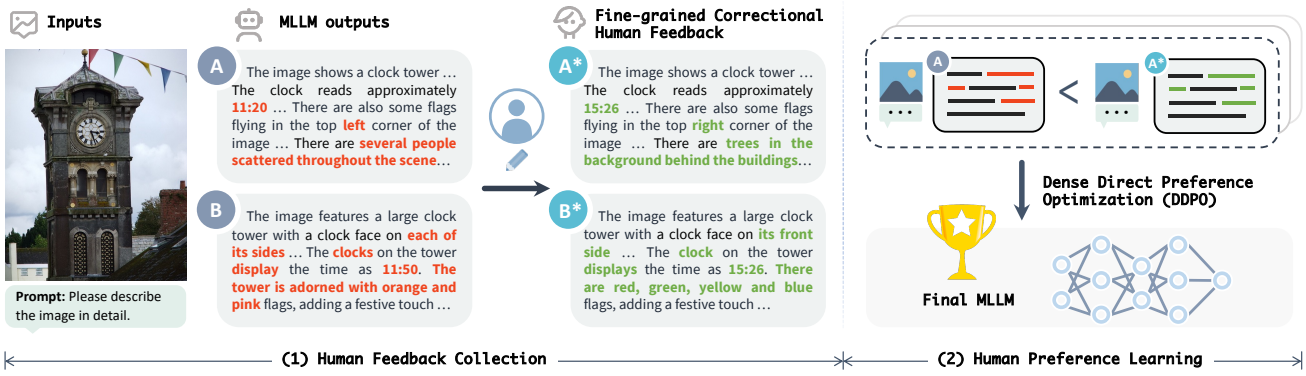

Figure 1. The RLHF-V framework for MLLM behavior alignment from human feedback. (1) Given the input image and prompt, we obtain outputs from MLLMs and collect human feedback in the form of fine-grained segment-level **corrections** on **hallucinations**. (2) During human preference learning, we perform dense direct preference optimization over the fine-grained correctional human feedback.

man Feedback (RLHF) is fraught with two key challenges: (1) *Annotation ambiguity*. Helpful and engaging responses about rich image content are typically long and complex, making it usually non-obvious to decide which response is preferable. As shown in Figure 1 (responses A and B), annotators usually face dilemmas when presenting responses with respective advantages and flaws. Besides, even if labeled with a clear preference, the optimal response remains unknown (e.g., the exact time of the clock). (2) *Learning efficiency*. The coarse-grained ranking feedback makes it difficult to accurately allocate credit to the desirable behaviors. Considering the linguistic complexity and variance of responses, the desirable behavior often requires a large amount of labeled data to learn [10, 31, 37]. Moreover, misallocation of credit to the non-robust bias correlated with the data usually leads to reward hacking and behavior degeneration problems [5, 39].

RLHF-V addresses these challenges by introducing two key innovations: (1) At the data level, we propose to collect human feedback in the form of fine-grained segment-level corrections. As shown in Figure 1, we ask human annotators to directly correct the hallucinated segments from model responses, providing a clear, dense, and fine-grained human preference, as well as optimal responses. This strategy also avoids linguistic variance and non-robust bias, ensuring that the feedback is accurately allocated to the desirable behaviors, thereby enhancing learning efficiency and preventing reward hacking problems. (2) At the method level, we propose dense direct preference optimization (DDPO), a new variant of DPO [33] that addresses the traditional RLHF objective in an equivalent simple and efficient supervised fashion. DDPO directly optimizes the policy model against dense and fine-grained segment-level preference, where the hallucinated segments receive stronger feedback to be factually grounded.

Comprehensive experiments on five benchmarks show that, RLHF-V can substantially enhance the trustworthiness of MLLMs with promising data and computation efficiency. Using 1.4k preference data, RLHF-V significantly reduces the object hallucination rate of the base MLLM by 34.8%, surpassing the concurrent LLaVA-RLHF [37] trained on 10k preference data. We also show that RLHF-V achieves better robustness than the strong GPT-4V [29] in preventing hallucinations aroused from over-generalization.

The contribution of this work can be summarized as threefold: (1) We present RLHF-V, a novel framework that aligns MLLM behavior through fine-grained correctional human feedback. (2) We collect high-quality human preference data to provide human-aligned learning signals for MLLMs. (3) We conduct comprehensive experiments to demonstrate the effectiveness of the proposed framework, achieving state-of-the-art performance in trustworthiness among open-source MLLMs.

2. Human Preference Collection

The goal of human preference data is to distinguish human-preferred high-quality responses from inferior ones, providing human-aligned learning signals to steer the MLLM behaviors. We first provide an analysis of underlying factors of human preference data, based on which we motivate the human preference collection procedure of RLHF-V.

Human Preference Data: Underlying Factors and Challenges. Given the input x (including the image and the prompt), denote the difference between a preferred output y_w and an inferior output y_l as Y. The difference Y can be essentially decomposed into three factors:

$$Y = Y_p + Y_s + Y_n, \tag{1}$$

where Y_p is the truly preferred behavior such as being trustworthy and helpful, Y_s denotes the shallow non-robust bias correlated with the data but unrelated to human judgment

(e.g., y_w contains more usage of specific words), and Y_n is the random noise factor denoting the linguistic variance of natural language (e.g., different ways of expressing the same meaning). Y_p is the factor we want to learn from the difference Y, while fitting to Y_s can lead to reward hacking problems and thus should be avoided. The linguistic variance Y_n does not bias the preference learning but makes the learning more difficult, demanding more labeled data to learn to the preferred factor Y_p, and thus should also be avoided if possible.

The common RLHF practices in LLMs collect human preference Y in the form of ranking labels, indicating the overall relative quality of responses [30, 31, 39]. According to the above analysis, the practice faces several key challenges: (1) *Annotation ambiguity.* It can be non-obvious to annotate which response is superior using an overall ranking label due to the fine-grained nature of Y_p, especially for complex responses. As shown in Figure 1, annotators usually cannot agree on assigning an overall ranking to different responses with respective advantages and flaws. We observe the issue leads to unsatisfactory annotation quality of existing RLHF data. Moreover, even if labeled with a clear preference, the optimal responses for the questions typically remain unknown. (2) *Learning efficiency.* During reinforcement learning, it can be challenging and data-demanding to precisely allocate the sparse and coarse-grained credit from Y through the linguistic variance Y_n to the preferred behavior Y_p. Misallocation to the non-robust bias factor Y_s will lead models to collapse to exploit trivial rewards [5, 39].

Fine-grained Correctional Human Preference Collection. To address the challenges, we propose to collect fine-grained human preferences in the form of segment-level corrections. As shown in Figure 1, given a flawed output y_l from MLLMs, we ask human annotators to directly correct the hallucinated segments, resulting in a factually optimal output y_w. The annotation simultaneously yields a segment-level incremental preference pair (y_w, y_l). The simple procedure effectively addresses the challenges: (1) The annotation of incremental correction in segments is clearer and more operable for human labelers. (2) The dense and fine-grained feedback is directly allocated to the preferred behavior Y_p, excluding the linguistic variance Y_n and the non-robust bias Y_s, therefore improving learning efficiency and preventing reward hacking problems. In experiments, we find that the procedure greatly improves the annotation quality and data efficiency, enabling our model to surpass concurrent models trained on an order of magnitude more labeled preference data (see Section 4.3).

In practice, we obtain a total of 1.4k prompts as input from existing instruction tuning dataset [47] and image description prompts generated by GPT-4, and get the responses from Muffin [47] for human annotation. The responses after annotation contain 64.4 words and 2.65 corrected segments on average. We observe that the corrections are diverse in hallucination types, including objects (41.2%), positions (20.3%), numbers (16.5%), attributes (10.0%), actions (5.3%) and miscellaneous types (6.8%).

3. Method

We introduce the RLHF-V approach that learns the fine-grained correctional human feedback by dense direct preference optimization. In addition, we also mitigate existing sources of hallucination in MLLM training by addressing the vision-language mismatch problem.

3.1. Dense Direct Preference Optimization

To leverage the dense and fine-grained human feedback, we present DDPO, a new variant of direct preference optimization [33] for directly optimizing the MLLM policy against dense human preference. The prevalent RLHF approaches involve fitting a reward model on the preference data, and then training the critique, policy and value models to maximize the reward without deviating too far from the reference model [10, 31, 39]. This procedure requires training multiple LLMs with extensive sampling and training, which suffers from complex procedures and high computation cost.

Direct Preference Optimization (DPO) [33] solves this reinforcement learning objective in a simpler equivalent supervised fashion. Here we briefly introduce the DPO method, and refer readers to the original paper for more details. The key observation of DPO is that the reward function $r(x, y)$ can be analytically expressed by its optimal policy model $\pi_*(y|x)$ and reference model $\pi_{\text{ref}}(y|x)$, and therefore we can directly optimize the policy model under proper forms on the preference data. Specifically, the reward model $r(x, y)$ can be represented as:

$$r(x, y) = \beta \log \frac{\pi_*(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x), \tag{2}$$

where β is a constant and $Z(x)$ is the partition function. Based on this observation, the policy model can be directly optimized on the human feedback data:

$$\begin{aligned} \mathcal{L} &= -\mathbb{E}_{(x,y_w,y_l)}\left[\log \sigma(r(x, y_w) - r(x, y_l))\right] \\ &= -\mathbb{E}_{(x,y_w,y_l)}\left[\log \sigma(\beta \log \frac{\pi_*(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_*(y_l|x)}{\pi_{\text{ref}}(y_l|x)})\right], \end{aligned} \tag{3}$$

where the reference model $\pi_{\text{ref}}(y|x)$ is usually implemented by an instruction-tuned base model we want to improve, and is kept fixed during DPO training. Only the policy model $\pi_*(y|x)$ is updated. We note that DPO is more simple, efficient and stable in aligning MLLM behaviors compared with traditional RLHF approaches.

Leveraging dense and fine-grained segment-level feedback requires the model to evaluate the reward of segment-level actions. However, DPO is designed for learning preference in the form of overall response ranking labels:

$$\log \pi(y|x) = \sum_{y_i \in y} \log p(y_i|x, y_{<i}), \quad (4)$$

where y_i is the i-th token of the response y. We argue that compared with unchanged segments y_u, corrected segments y_c more directly reveal human judgment in hallucination, and thus should contribute more to the overall action evaluation. Therefore, we propose to score the response as a weighted aggregation of the fine-grained segments:[1]

$$\log \pi(y|x) = \frac{1}{N}\Big[\sum_{y_i \in y_u} \log p(y_i|x, y_{<i}) + \gamma \sum_{y_i \in y_c} \log p(y_i|x, y_{<i})\Big], \quad (5)$$

where $\gamma > 1$ is a weighting hyperprameter, and larger γ means more contribution from the corrected segments. $N = |y_u| + \gamma|y_c|$ is a normalizing factor, preventing longer responses from getting higher scores. In this way, corrected segments are highlighted to receive stronger human preference feedback to be factually grounded. In experiments, we find that DDPO can better exploit the fine-grained human feedback, leading to more trustworthy responses.

3.2. Mitigating Hallucination from VL Mismatch

DDPO reduces hallucination by learning from human feedback. From another cause-and-effect view, we examine the mainstream MLLM training paradigm, and identify sources of hallucinations in training MLLMs. Based on the observations, we motivate a more trustworthy training recipe.

In general, current MLLMs learn multimodal capabilities in a supervised learning paradigm, where the model outputs are supervised against the ground-truth text associated with the image. In such a paradigm, hallucinations can be introduced by mismatches between images and text data. In practice, the mismatch can come from: (1) low-quality text in pre-training and instruction tuning data, and (2) careless image augmentation during training. We specify the issues and solutions in the following.

Addressing Low-quality Text Influence. Current pre-training data of MLLMs are automatically crawled from the Web [6, 7, 35], which inevitably suffers from severe noise in the text even after extensive post-processing. Supervising MLLMs against such data is essentially teaching them to hallucinate (e.g., describing elements not present in the image, or producing inconsistent descriptions with the image). Similarly, most existing visual instruction tuning datasets are generated by ChatGPT/GPT-4 according to intermediate text annotations [25, 27, 47], which inevitably introduces hallucination into instruction data. While it can be difficult to repair existing pre-training and instruction-tuning data, we find that the influence can be countered by simply post-training MLLMs on high-quality visual question-answering datasets. Intuitively, human-labeled datasets can provide accurate learning signals to calibrate model behaviors from hallucinations, and also enhance instruction-following capabilities. In our experiments, we find that simply fine-tuning the model on VQAv2 [14] can significantly reduce the hallucination rate (see Section 4.3).

Mitigating Untrustworthy Image Augmentation. The vision-language mismatch can also come from the image domain. Data augmentation is widely adopted to improve the data diversity and model robustness in various multi-modal models [11, 19, 32, 41, 47]. However, we note that such augmentation must be performed with care in training MLLMs. The key problem is that some image augmentation operations can significantly change the semantics of images, which may make the augmented image inconsistent with the associated text. For example, during augmentation, random cropping can make the objects mentioned in the text absent from the image. This can make the model describe non-existing objects, with wrong numbers, and in wrong positions. In our model training, we exclude image cropping in data augmentation, which improves the trustworthiness of MLLMs (see Section 4.3).

4. Experiments

In this section, we empirically investigate the effectiveness of RLHF-V in aligning MLLM behaviors. In addition to evaluating the trustworthiness and helpfulness of conversation, we also analyze the data efficiency and scalability as well as the robustness.

4.1. Experimental Settings

We first introduce the experimental settings, including evaluation, baselines, and implementation details.

Evaluation. We evaluate the models from two perspectives, including trustworthiness reflecting the hallucination degree, and helpfulness reflecting the general interaction quality. Similar to [37], we find binary classification evaluation (i.e., answering yes/no) [13, 21] cannot well reflect the MLLM behaviors in open-ended long-form interactions. We thus adopt benchmarks that directly evaluate the long-form responses, which are more closely related to the practical usage scenarios of MLLMs. For trustworthiness, we perform evaluation on three benchmarks:

(1) **Object HalBench** [34] is a widely adopted benchmark for assessing object hallucination in detailed image descriptions. It compares the objects in the model output with object labels exhaustively annotated for COCO images [23] to detect object hallucination. We report the response-level hallucination rate (i.e., the percentage of responses that have hallucinations), as well as the mention-level hallucination rate (i.e., the percentage of hallucinated object mentions among all object mentions).

[1] For denotation simplicity, without confusion we also use y_u and y_c to denote the set of tokens in unchanged and corrected segments respectively.

Model	Object HalBench ↓		MHumanEval ↓				MMHal-Bench		LLaVA Bench			VQAv2
	Resp.	Mention	Object	Position	Number	All	Info.	Resp.↓	Conv.	Detail	Comp.	testdev
LLaVA [27]	63.0	29.5	46.6	21.2	19.9	80.8	31.9	70.8	85.4	74.3	96.3	-
Muffin [47]	50.5	24.5	33.6	16.4	26.0	74.7	33.4	68.8	89.3	**79.7**	97.7	-
LRV [25]	32.3	22.3	43.2	11.6	19.2	82.9	22.2	78.1	61.7	47.3	55.0	-
LLaVA-RLHF [37]	38.1	18.9	37.7	17.8	18.5	72.6	39.9	65.6	**93.8**	74.3	**111.4**	-
InstructBLIP [11]	25.9	14.3	30.8	15.1	17.1	63.7	29.5	64.4	83.2	67.6	90.6	-
Qwen-VL-Chat [4]	43.8	20.0	34.9	16.4	15.8	61.0	38.5	**52.1**	81.9	77.1	92.3	79.5
LLaVA 1.5 [26]	46.3	22.6	30.8	17.8	17.1	61.0	39.2	**52.1**	81.6	75.5	95.2	**80.0**
RLHF-V	**12.2**	**7.5**	**21.9**	**7.5**	**14.4**	**55.5**	**40.0**	**52.1**	93.1	75.3	91.6	**80.0**
GPT-4V [29]	13.6	7.3	22.6	12.3	11.0	45.9	47.6	31.3	96.0	102.5	106.7	77.2*

Table 1. Main experimental results on hallucination. We report hallucination rates in different granularities, including response-level (Resp.) and mention-level (Mention), and response-level hallucination rates in different types. We also show scores on informativeness (Info.), multimodal conversation (Conv.), detailed description (Detail), and complex reasoning (Comp.). * denotes zero-shot results on VQAv2.[2] The best and second best open-source results are shown in **bold** and underlined respectively.

(2) **MMHal-Bench** [37] evaluates hallucinations and response informativeness. It employs GPT-4 to compare model output with human response and several object labels to decide the scores. In experiments, we find that GPT-4 cannot reliably detect hallucinations due to the incompleteness of MMHal-Bench text annotations. We therefore only report the informativeness score from GPT-4, and assess response-level hallucination rate by human evaluation.

(3) **MHumanEval.** The above evaluations are either limited to common object hallucination or dominated by short-form question answering (i.e., questions that can be sufficiently answered by a few words). To provide a more reliable and comprehensive evaluation over diverse hallucination types, we present MHumanEval benchmark, which covers both long-form image descriptions, and short-form questions. The benchmark contains 146 samples collected from Object HalBench (50) and MMHal-Bench (96). Given model responses, we ask human annotators to label the hallucinated segments and hallucination types of the segments, including objects, positions, numbers and others. We report the response-level hallucination rate on these types.

For helpfulness, we adopt two benchmarks: (1) **LLaVA Bench** [27] is a widely adopted benchmark for assessing multimodal conversation, detailed description and complex reasoning capabilities. It scores model output against reference response via GPT-4. (2) **VQAv2** [14] is a popular dataset for short-form visual question answering.

Baselines. We compare our model with state-of-the-art baselines. (1) **General baselines.** We adopt Qwen-VL-Chat [4], LLaVA [27], LLaVA 1.5 [26], Muffin [47], and InstructBLIP [11] as representative general baselines. These models are mostly pre-trained on large-scale multi-modal data, and fine-tuned on high-quality instruction data, achieving strong performance across various multimodal tasks. (2) **Baselines tailored for hallucination problems.** LRV [25] is fine-tuned on 400k instruction data generated by GPT-4, and mitigates hallucination by limiting the response length. The concurrent LLaVA-RLHF [37] employs the strong 13B Vicuna v1.5 [50] (fine-tuned from LLaMA-2 [39]) as LLM backbone. It trains the reward model on 10k human-labeled preference data, and performs proximal policy optimization [36] on 72k factually augmented data. (3) **Commercial Baseline.** We also include GPT-4V [29] as a strong reference, evaluating the gap between the open-source models and state-of-the-art commercial models.

Implementation Details. We implement the RLHF-V framework based on Muffin [47]. The model uses BEiT-3 [41] as the visual module, and 13B Vicuna v1.0 [9] (fine-tuned from LLaMA [38]) as the LLM backbone. The hyper-parameter β is 0.5, and the weighting coefficient γ is 5. We train the model with DDPO for 7 epochs, with image resolution 448, learning rate 5e-7 and batch size 32. The training of RLHF-V is computationally efficient, which takes less than 1 hour on 8 A100 GPUs in total.

4.2. Main Results

The main experimental results are reported in Table 1, from which we observe that: (1) RLHF-V achieves state-of-the-art performance in trustworthiness among open-source models, outperforming strong general models and models tailored for hallucination. The framework significantly reduces the hallucination rate of the base model Muffin by 75.8% relative points for common objects on Object HalBench, and by 34.8% for overall objects on MHumanEval. The improvement is consistent in different granularities including response-level and mention-level hallucinations, and different hallucination types including objects, posi-

[2]Due to limited instruction-following capability, most MLLMs need to be specifically fine-tuned to produce short-form VQA answers, and therefore cannot achieve reasonable zero-shot performance on VQAv2.

Model	Living Room book, person, bed chair, couch, remote			Kitchen bottle, bowl, cup person, chair, knife			Bathroom toilet, sink, bottle toothbrush, person, cup			Street person, car, motorcycle traffic light, handbag, truck			$\overline{\Delta}$
	H_a	H_s	Δ	H_a	H_s	Δ	H_a	H_s	Δ	H_a	H_s	Δ	
LLaVA-1.5 [26]	25.2	41.8	+16.6	18.9	23.9	+5.0	22.4	30.4	+8.0	20.6	28.0	+7.4	+9.2
LLaVA-RLHF [37]	23.7	34.5	+10.8	13.1	17.4	+4.3	18.2	19.5	+1.4	18.3	22.7	+4.4	+5.2
QWEN-VL [4]	24.5	34.5	+10.0	16.4	20.8	+4.4	21.6	17.5	**-4.1**	22.5	32.0	+9.5	+5.0
RLHF-V	**5.5**	**8.0**	**+2.5**	**3.8**	**5.9**	**+2.1**	**4.1**	**4.0**	-0.1	**2.3**	**4.6**	**+2.3**	**+1.7**
GPT-4V [29]	8.2	19.4	+11.2	4.6	5.7	+1.1	5.9	13.3	+7.5	4.2	4.6	+0.4	+5.0

Table 2. Experimental results of hallucination from over-generalization on Object HalBench. For each scene, we report the hallucination rate of the top 10 frequent objects on average on the full benchmark (H_a) and under the scene (H_s). Top 6 frequent objects are listed for each scene for brevity. Δ: hallucination rate difference, $\overline{\Delta}$: average difference across the scenes.

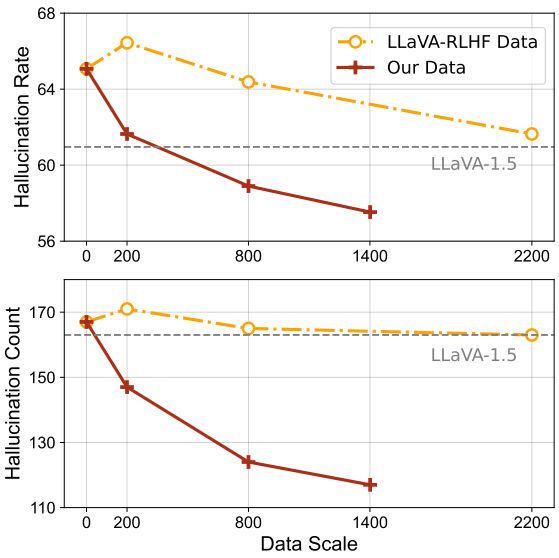

Figure 2. Hallucination rate and number on MHumanEval (all types) with respect to the amount of preference data. We report the results of different models trained on different RLHF data.

tions, and numbers. The reduction is more significant on the more challenging long-form answers on Object HalBench and MHumanEval. The results show that RLHF-V can effectively learn from fine-grained correctional human feedback to enable more trustworthy MLLM behaviors. (2) RLHF-V achieves promising performance in response helpfulness, where the results on MMHalBench, LLaVA Bench and VQAv2 are strong and comparable to the base model. This shows that RLHF-V can enhance the trustworthiness of MLLMs without sacrificing their helpfulness.

4.3. Analysis

In this section, we conduct analyses on the framework considering the following research questions: (1) How does RLHF-V's performance scale with feedback data amount? (2) What is the advantage of fine-grained correctional preference data over traditional overall ranking data? (3) Can RLHF-V's data and method be adopted to enhance the trustworthiness of other MLLMs? (4) How does human feedback alleviate hallucinations intuitively?

Scaling feedback data leads to promising results. We report the hallucination rate and numbers of hallucinated segments on MHumanEval under different amounts of feedback data in Figure 2. We observe that the hallucination rate and number of RLHF-V show a significant and rapid decrease as the data amount grows. This shows that fine-grained correctional human feedback provides effective and efficient learning signals for MLLM behavior alignment. Based on this tendency, we expect better performance can be achieved with an increasing amount of feedback data. We leave this for future work.

Fine-grained correctional human feedback enables better learning efficiency. To quantify the advantage of fine-grained correctional human feedback, we replace our data with the 2.2k human preference data on hallucination from LLaVA-RLHF, which gives overall ranking labels following common RLHF practices. From the experimental results in Figure 2, we observe that model equipped with our data shows a more significant and rapid reduction in hallucination rate and number. Notably, using only 200 preference data, our model achieves comparable hallucination rate to the model that uses an order of magnitude more labeled data from LLaVA-RLHF. The superior data efficiency is due to (1) better data quality since label ambiguity is minimized, and (2) more direct feedback on hallucinated segments, excluding non-robust bias and linguistic variance.

RLHF-V generalizes to enhance other MLLMs. To investigate the generalization capability of the framework, we adopt RLHF-V's data and approach to align the behavior of LLaVA [27], a representative and widely used MLLM. Experimental results show that RLHF-V effectively reduces the hallucination count of LLaVA by 13.8 relative points, as well as the hallucination rate by 5.9

Model	MHumanEval↓				MHB↓	VQAv2
	Obj.	Pos.	Num.	All	Resp.	testdev
Muffin [47]	33.6	16.4	26.0	74.7	68.8	-
RLHF-V	21.9	**7.5**	14.4	**55.5**	**52.1**	**80.0**
w/ vanilla DPO	21.9	11.6	**11.6**	57.5	54.2	**80.0**
w/ IT-VQA only	34.3	17.1	17.1	65.1	58.3	**80.0**
w/ untrust aug.	**18.5**	13.7	14.4	59.6	54.2	77.1

Table 3. Ablation results on different components. MHB: MMHal-Bench, IT-VQA: instruction tuning on VQAv2, untrust aug.: untrustworthy data augmentation.

relative points. We also apply RLHF-V to stronger base models and build the OmniLMM-12B [1] which achieves new SoTA results on multiple hallucination benchmarks. For example, OmniLMM-12B exhibits only 4.5% mention-level hallucination on the Object HalBench. Moreover, OmniLMM-12B also shows leading performance among comparable-sized models on multiple benchmarks (1637 on MME-Perception [13], 71.1 on SeedBench-I [17]). The results demonstrate that RLHF-V is applicable across different MLLMs to improve trustworthiness.

RLHF-V reduces hallucination from correlation and over-generalization. Without proper positive/negative human feedback, MLLMs can over-generalize to produce highly correlated and plausible concepts, which leads to hallucinations. For example, a prevalent hallucination case observed across different MLLMs is claiming the presence of *person* as long as they see an image of *street*. To quantify the problem, we select a set of representative scenes $\{living\ room, kitchen, bathroom, street\}$. For each scene, we identify the corresponding images in COCO by lexically matching the captions with the scene name. Then we obtain the top 10 frequent objects in the scene from the COCO object annotations. We compare the response-level hallucination rate for these objects (1) on average across all test samples, and (2) on samples under the target scene. Models prone to over-generalization will expect a significant increase in the hallucination rate (Δ).

From the experimental results in Table 2, we observe that: (1) All models including GPT-4V show a substantial increase in the hallucination rate, which demonstrates the over-generalization hypothesis. (2) RLHF-V exhibits the smallest change in the hallucination rate, which is even more robust than GPT-4V. The reason for the robustness is that RLHF-V provides crucial positive/negative fine-grained correctional human feedback for MLLMs, which helps to learn clear behavior boundaries between reasonable generalizations and over-generalizations. (3) RLHF-V achieves the lowest hallucination rates for these common objects both on average and under common scenes.

Ablation Study. To investigate the contribution of each component, we perform an ablation study. From the experimental results in Table 3, we can observe that: (1) Learning human feedback with vanilla DPO leads to performance degrades, showing the advantage of DDPO in exploiting the fine-grained human preference. (2) Fine-tuning on VQAv2 leads to a significant reduction in hallucination rates compared with the base model. This reveals the value of traditional human-annotated datasets from a new perspective of hallucination mitigation. (3) Including untrustworthy data augmentation (i.e., image cropping) in training hurts the performance on both hallucination and VQAv2. This shows that careless data augmentation can be a double-edged sword in training MLLMs.

Case Study. To provide an intuitive understanding and comparison of different models, we provide qualitative results in Figure 3. We show cases in two representative scenarios: (1) Short-form QA (i.e., questions that can be sufficiently answered in a few words). Our model typically maintains a good balance between helpfulness and clarity. (2) Long-form QA (i.e., questions that require long text to answer). We observe that MLLMs are significantly more prone to hallucinations in long-form QA, whereas RLHF-V introduces significantly fewer hallucinations while delivering a comparable amount of effective information.

5. Related Work

Multimodal Large Language Models. Recent trends in multimodal learning have witnessed the success of building MLLMs by connecting visual encoders with powerful LLMs [8, 16, 18, 45, 49]. The current MLLM training paradigm typically involves two stages: (1) Pre-training. Models are pretrained on large-scale image-text pairs [4, 11, 19, 42, 47] or interleaved data [2, 3, 16] to learn the semantic mapping between visual and text signals. (2) Instruction Tuning. To enable the model with instruction-following capability, MLLMs are further fine-tuned on visual instruction data, including collections of existing human-annotated datasets [11, 20, 26], and generated data from ChatGPT/GPT-4 [20, 25, 27, 47]. Despite the success, current MLLMs suffer from serious hallucination problems [21, 24, 25, 37]. Notably, even after extensive efforts, GPT-4V has still been found to be prone to hallucinations, making basic factual errors confidently [29].

Reducing Hallucination for MLLMs. Like LLMs [22, 48], MLLMs also exhibits the problem of generating unfaithful contents. Some preliminary efforts have been made to alleviate hallucination problems in MLLMs. LRV [25] generates instruction data with negative responses, and mitigates hallucination by limiting the response length. VIGC [40] iteratively refines the instruction data for better instruction tuning. Woodpecker [46] proposes to post-edit hallucinations by merging the output of MLLMs and a more accurate expert VQA model using GPT-3.5. The post-

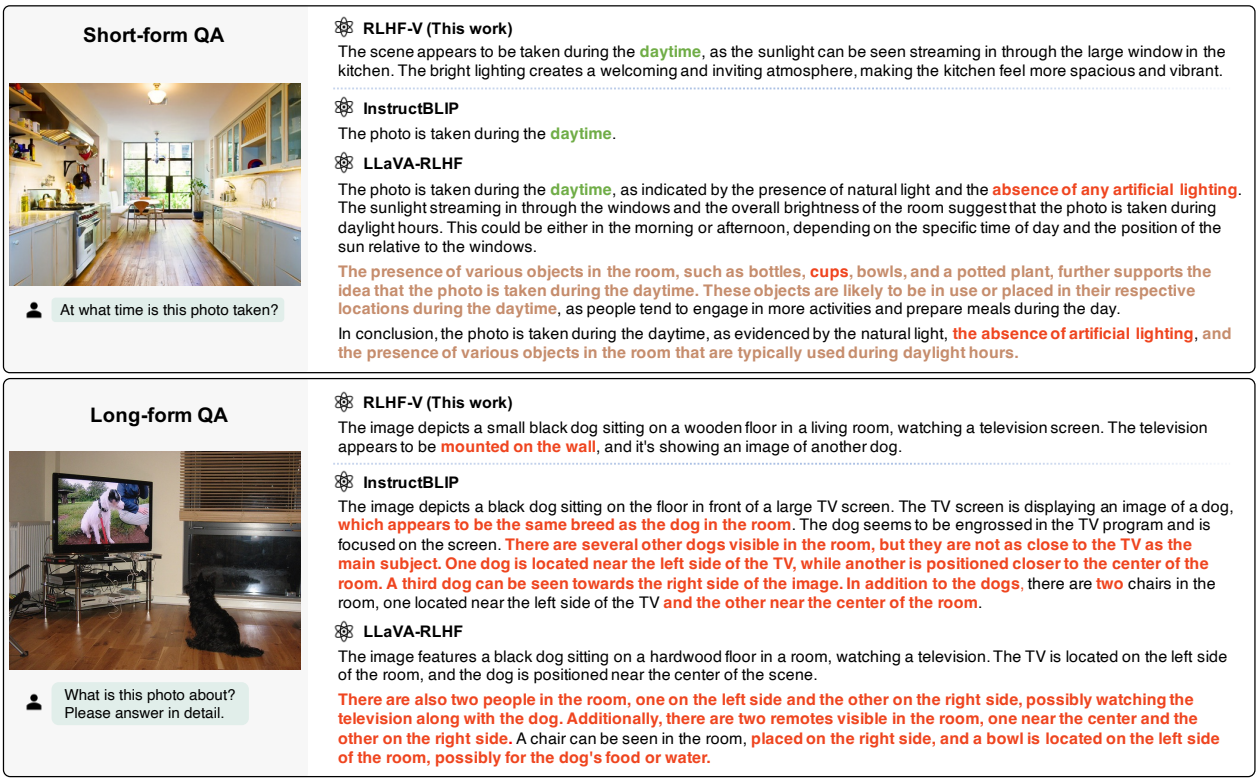

Figure 3. Qualitative results of different models on short-form QA and long-form QA. **Correct answers**, **unreasonable extensions** and **hallucinations** are highlighted in color respectively.

editing procedure involves external tools and LLMs much larger than the target MLLM online in multiple stages, which leads to high inference costs and delays. Gunjal *et al*. [15] distinguishes the inaccurate parts in responses via human annotation, and internally discourages the hallucinated parts by direct preference optimization. However, the positive behaviors for hallucinated parts are unknown, making the human feedback not complete enough to learn the behavior boundary. The concurrent LLaVA-RLHF [37] employs the traditional RLHF approach [31] on MLLMs, and augments the reward model with rich additional text descriptions. It is therefore similarly challenged with label ambiguity, learning efficiency, and complex training. In comparison, RLHF-V presents the first fine-grained correctional human feedback learning framework for behavior alignment, and systematically addresses different hallucination sources in training MLLMs, achieving strong performance in trustworthiness.

6. Conclusion

Hallucination is a critical problem preventing practical applications of MLLMs in real-world scenarios. In this work, we present RLHF-V, a novel framework that enhances the trustworthiness of MLLMs by behavior alignment from fine-grained correctional human feedback. Comprehensive experimental results show that our model achieves state-of-the-art performance in trustworthiness especially in challenging long-form responses while maintaining strong helpfulness. In future, with the progress of more trustworthy and capable MLLMs, we will explore collecting accurate preferences from MLLMs, which can facilitate large-scale preference learning. Besides, we note that the framework of RLHF-V can potentially also help reduce the hallucinations in LLMs, which we will explore in future.

7. Acknowledgement

This research is supported by National Natural Science Foundation of China (Grant No.62276154), Research Center for Computer Network (Shenzhen) Ministry of Education, the Natural Science Foundation of Guangdong Province (Grant No. 2023A1515012914), Basic Research Fund of Shenzhen City (Grant No. JCYJ20210324120012033 and JSGG20210802154402007), the Major Key Project of PCL for Experiments and Applications (PCL2021A06), and Overseas Cooperation Research Fund of Tsinghua Shenzhen International Graduate School (HW2021008).

References

[1] Large multi-modal models for strong performance and efficient deployment. https://github.com/OpenBMB/OmniLMM. Accessed: 2024-03-05. 7

[2] 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. *NeurIPS*, 35: 23716–23736, 2022. 1, 7

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 7

[4] 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 preprint arXiv:2308.12966*, 2023. 1, 5, 6, 7

[5] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. 2, 3

[6] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. COYO-700M: Image-text pair dataset, 2022. 4

[7] 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 *CVPR*, pages 3558–3568, 2021. 4

[8] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. PaLI: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 7

[9] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality, 2023. 5

[10] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023. 2, 3

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning, 2023. 1, 4, 5, 7

[12] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. PaLM-E: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 1

[13] 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. 4, 7

[14] 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 *CVPR*, pages 6904–6913, 2017. 4, 5

[15] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*, 2023. 8

[16] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 7

[17] 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. 7

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

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 4, 7

[20] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M3IT: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023. 7

[21] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 1, 4, 7

[22] Yinghui Li, Zishan Xu, Shaoshen Chen, Haojing Huang, Yangning Li, Yong Jiang, Zhongli Li, Qingyu Zhou, Hai-Tao Zheng, and Ying Shen. Towards real-world writing assistance: A chinese character checking benchmark with faked and misspelled characters. *CoRR*, abs/2311.11268, 2023. 7

[23] 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 *ECCV*, pages 740–755. Springer, 2014. 4

[24] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusion-Bench: You see what you think? Or you think what you see? An image-context reasoning benchmark challenging for GPT-4V(ision), LLaVA-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2023. 7

[25] 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. 1, 4, 5, 7

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 5, 6, 7

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 4, 5, 6, 7

[28] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 1

[29] OpenAI. GPT-4V(ision) system card. 2023. 1, 2, 5, 6, 7

[30] OpenAI. GPT-4 technical report, 2023. 3

[31] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 35:27730–27744, 2022. 2, 3, 8

[32] 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 *ICML*, pages 8748–8763. PMLR, 2021. 4

[33] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023. 2, 3

[34] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, pages 4035–4045, 2018. 4

[35] 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. *NeurIPS*, 35:25278–25294, 2022. 4

[36] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 5

[37] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented RLHF. *arXiv preprint arXiv:2309.14525*, 2023. 1, 2, 4, 5, 6, 7, 8

[38] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 5

[39] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2, 3, 5

[40] Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. VIGC: Visual instruction generation and correction. *arXiv preprint arXiv:2308.12714*, 2023. 7

[41] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *CVPR*, pages 19175–19186, 2023. 4, 5

[42] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. CogVLM: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1, 7

[43] Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, et al. On the road with GPT-4V (ision): Early explorations of visual-language model on autonomous driving. *arXiv preprint arXiv:2311.05332*, 2023. 1

[44] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of LMMs: Preliminary explorations with GPT-4V(ision). *arXiv preprint arXiv:2309.17421*, 9, 2023. 1

[45] 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. 7

[46] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023. 7

[47] Tianyu Yu, Jinyi Hu, Yuan Yao, Haoye Zhang, Yue Zhao, Chongyi Wang, Shan Wang, Yinxv Pan, Jiao Xue, Dahai Li, et al. Reformulating vision-language foundation models and datasets towards universal multimodal assistants. *arXiv preprint arXiv:2310.00653*, 2023. 1, 3, 4, 5, 7

[48] Tianyu Yu, Chengyue Jiang, Chao Lou, Shen Huang, Xiaobin Wang, Wei Liu, Jiong Cai, Yangning Li, Yinghui Li, Kewei Tu, Hai-Tao Zheng, Ningyu Zhang, Pengjun Xie, Fei Huang, and Yong Jiang. Seqgpt: An out-of-the-box large language model for open domain sequence understanding, 2023. 7

[49] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 7

[50] 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. 5

[51] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1