FGAIF: Aligning Large Vision-Language Models with Fine-grained AI Feedback

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) have demonstrated proficiency in tackling a variety of visual-language tasks. However, current LVLMs suffer from misalignment between text and image modalities which causes three kinds of hallucination problems, i.e., object existence, object attribute, and object relationship. To tackle this issue, existing methods mainly utilize Reinforcement Learning (RL) to align modalities in LVLMs. However, they still suffer from three main limitations: (1) General feedback can not indicate the hallucination type contained in the response; (2) Sparse rewards only give the sequence-level reward for the whole response; and (3)Annotation cost is time-consuming and labor-intensive. To handle these limitations, we propose an innovative method to align modalities in LVLMs through **F**ine-**G**rained **A**rtificial **I**ntelligence **F**eedback (**FGAIF**), which mainly consists of three steps: AI-based Feedback Collection, Fine-grained Reward Model Training, and Reinforcement Learning with Fine-grained Reward. Specifically, We first utilize AI tools to predict the types of hallucination for each segment in the response and obtain a collection of fine-grained feedback. Then, based on the collected reward data, three specialized reward models are trained to produce dense rewards. Finally, a novel fine-grained feedback module is integrated into the Proximal Policy Optimization (PPO) algorithm. Extensive experiments are conducted on hallucination and general benchmarks, demonstrating the superior performance of our proposed method. Notably, compared with previous models trained with the RL-based aligning method, our proposed method is effective even with fewer parameters.

1 Introduction

Large Language Models (LLMs) like GPT-3 (Brown et al., 2020) and ChatGPT (OpenAI, 2022) have showcased remarkable abilities in language processing. However, their ability to handle

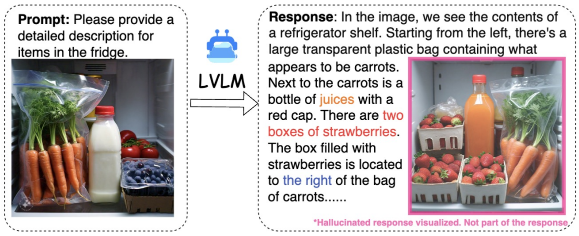

Figure 1: Illustration of the hallucination in the response generated by the LVLM. We illustrate all three kinds of hallucinations in this figure, where orange fonts denote object existence hallucinations, red fonts denote object attribute hallucinations, and blue fonts for object relation hallucinations.

multimodal inputs combining both visual and textual data remains inadequate. This limitation has drawn research attention to Large Vision-Language Models (LVLMs) which achieve massive success in various vision and language tasks (e.g. Visual Question Answering (Antol et al., 2015) and Image Captioning (Lin et al., 2014)).

Although LVLMs have achieved significant success in tasks requiring visual-textual understandings, the challenge of misalignment between vision and language modalities (Sun et al., 2023) has not been solved, leading to "hallucination" in generated textual responses (Jing et al., 2023). As shown in Figure 1, there are three kinds of hallucinations in the context of LVLMs, including (1) Object Existence Hallucination, where non-existent objects are mistakenly referenced; (2) Object Attribute Hallucination, involving inaccuracies in the depiction of object attributes like color, shape, and size; and (3) Object Relationship Hallucination, where the descriptions inaccurately portray the interactions or spatial relationships between objects, leading to misrepresentations of their positions, interactions, and actions involving two or more objects (Jing et al., 2023; Zhai et al., 2023). Therefore, mitigating the hallucinations and generating faithful responses are key to building practical applications of LVLMs.

Hallucinations in LVLMs stem from their inclination to lean on common sense or stereotypical knowledge ingrained in the textual data used for training and frequently ignore the visual information presented (Cui et al., 2023), where the specific details contained in the input images (Zhou et al., 2024) are greatly overlooked. Such discrepancies are largely caused by the misalignment between textual and visual modalities (i.e., modality misalignment problem). To tackle this kind of misalignment problem, most existing methodologies rely on Reinforcement Learning (RL) (Ziegler et al., 2019; Sun et al., 2023; Li et al., 2023a; Zhou et al., 2024). For example, LLaVA-RLHF (Sun et al., 2023) aims to first gather human preferences and then incorporate these preferences into the reinforcement learning process for fine-tuning Language Models.

Despite their great success, the existing modality alignment method still suffers from three limitations: (1) General Feedback. Only broad and general feedback is generated by the reward model employed in current methodologies, and hallucination of specific types like objects and relations is not contained, making it challenging to precisely identify and correct inaccuracies in the generated content in the training stage. (2) Sparse Rewards. During the modality alignment training process, sequence-level feedback is gathered by current methodologies for the entirety of long responses, which is a kind of sparse training signal and is suitable to the task requiring the generation of long-form text. Moreover, sequence-level feedback tends to overlook the detailed hallucinations that may occur within individual segments of the response. (3) High Annotation Costs. Prevailing methods primarily utilize rewards based on human annotations, which is time-consuming and labor-intensive. Thus, scalability is another constraint for existing methods requiring massive accurate feedback.

To mitigate above-mentioned limitations, we propose to align modalities in large vision-language models with Fine-Grained AI Feedback (FGAIF), an innovative approach to refine large vision-language models via fine-tuning. In particular, our method mainly consists of three steps: AI-based feedback collection, fine-grained reward model training, and reinforcement learning with fine-grained rewards. The AI-based feedback collection step provides three kinds of segment-level (i.e., sub-sentence-level) hallucination labels based on AI feedback. We train three reward models that can produce fine-grained rewards, i.e., multiple types and segment-level rewards, using the collected fine-grained reward data, in the second step. The last step integrates novel fine-grained feedback into the Proximal Policy Optimization (PPO) algorithm to further fine-tune the LVLM.

Our contribution can be summarized as follows: 1) We propose a novel fine-grained artificial intelligence-based hallucination labeling method, which can detect three types of hallucinations (i.e., object existence, object attribute, and object relation) in terms of sub-sentence level and eliminate the need for manual annotation. 2) To the best of our knowledge, we are the first to provide multiple types and segment-level feedback towards modalities alignment in LVLMs, which can mitigate three kinds of hallucination in LVLMs. 3) We conduct comprehensive experiments on several hallucination benchmarks and one general benchmark. The experimental results demonstrate the effectiveness of FGAIF. In addition, the ablation study shows the necessity of each module in FGAIF.

2 Related Work

Large Vision-Language Model The recent pivot of the multimodal learning community towards LVLMs has been largely inspired by the effective pretraining approaches seen in LLMs and Vision Foundation Models (VFMs). At the heart of modern advanced LVLMs lie three fundamental components: a text encoder, an image encoder, and a cross-modal alignment module (Rohrbach et al., 2018). The text encoder typically manifests as a language model, with notable examples being LLaMA (Touvron et al., 2023) and Vicuna (Chiang et al., 2023), whereas the image encoder usually borrows from VFMs like ViT (Dosovitskiy et al., 2021). The critical role of the cross-modal alignment module is to fuse the visual and textual domains, thereby enabling the text encoder to grasp visual semantics more effectively. LVLMs generally undergo a multi-stage training approach to master visual comprehension (Gong et al., 2023; Zhu et al., 2023; Liu et al., 2023b,c; Ye et al., 2023; Dai et al., 2023). For example, Liu et al. (2023c) initially pre-trains the model by aligning image features with the word embeddings from a pre-trained LLM, followed by fine-tuning on specific language-image instruction datasets. To boost training efficiency, LVLMs often employ techniques like freezing parameters in the LLM or VFM components and utilize efficient fine-tuning methods such as LoRA (Hu et al., 2022a).

Despite their significant progress, LVLMs still face challenges with hallucinations, which can severely affect their performance on various vision-language tasks (Rohrbach et al., 2018).

Hallucinations in LVLMs Motivated the hallucination in LLMs, more researchers shifted research attention to hallucination in LVLMs. Hallucination in the context of LVLMs is the inconsistent content between the generated response and the input image. To evaluate the hallucination in LVLMs, some work devised metrics to measure the hallucination in the response, such as FaithScore (Jing et al., 2023), CHAIR (Rohrbach et al., 2018), POPE (Li et al., 2023d), and NOPE (Lovenia et al., 2023). Recently, there have been works to mitigate hallucinations in LVLMs utilizing various technologies, such as decoding approaches (Leng et al., 2023; Huang et al., 2023), post-processing (Zhou et al., 2023; Yin et al., 2023), and construction of the higher-quality dataset (Liu et al., 2023a; Li et al., 2023c). To address the challenge of aligning image and text modalities within LVLMs and to mitigate the issue of hallucination, existing strategies offer partial solutions but lack direct guidance for modality alignment. Therefore, some research efforts (Li et al., 2023b; Yu et al., 2023; Zhou et al., 2024) have embraced the use of reinforcement learning for direct modality alignment. For example, Sun et al. (2023) developed the LLaVA-RLHF model, harnessing human-annotated preference data to minimize hallucinations in LLaVA.

Motivated by the fine-grained RL (Wu et al., 2023) and AI-based RL (Lee et al., 2023; Bai et al., 2022) methods, we propose to align modalities in LVLMs with fine-grained AI feedback. Different from existing work which needs human annotation and only provides coarse-grained feedback, our method provides fine-grained rewards and learns from AI automatic feedback.

3 Problem Formulation

Suppose we have a set of N images $\{I_i\}_{i=1}^{N}$ and the corresponding prompts $\{P_i\}_{i=1}^{N}$. Next, we omit the index of I_i and P_i for simplicity. Then we feed the prompt P and image I into an LVLM $\mathcal{M}$ and get the sampled response as $R = \mathcal{M}(I, P|\Theta_m)$, where R is the response for (I, P). Θ_M refers to the parameters of LVLM $\mathcal{M}$. Next, we resort to another AI-based method $\mathcal{A}$ to identify three kinds of hallucination (i.e., object existence, object attribute, and object relation) in the generated response and **train three reward models** as $F^o, F^a, F^r = \mathcal{A}(R, I, P), \mathcal{R}^{o/a/r}(R, I, P|\Theta_{o/a/r}) \rightarrow F^{o/a/r}$, where $F^{o/a/r} = \{f_1^{o/a/r}, \cdots, f_s^{o/a/r}\}$ denotes the object existence/attribute/relation hallucination labels. $\Theta_{o/a/r}$ is the parameters of the reward model $\mathcal{R}^{o/a/r}$. $f_j^{o/a/r}$ is the label which means whether the j-th sub-sentence in the response contains the object existence/attribute/relation hallucination. $\mathcal{R}^{o/a/r}$ denotes reward models which aim to detect object existence/attribute/relation hallucinations.

Finally, we utilize well-trained reward models and a set of N_f images $\{I_i^f\}_{i=1}^{N_f}$ and the corresponding prompts $\{P_i^f\}_{i=1}^{N_f}$. to **fine-tune the LVLM** to make it generate faithful responses as $\hat{R} = \mathcal{M}(I^f, P^f, |\Theta_f, \mathcal{R}^o, \mathcal{R}^a, \mathcal{R}^r)$, where Θ_f is final optimized parameters of the LVLM $\mathcal{M}$. We also omit the index in this equation. N_f is the size of data for finetuning LVLMs.

4 Methodology

In this section, we detail the proposed FGAIF, which consists of three steps: AI-based feedback collection, fine-grained reward model training, and reinforcement learning with fine-grained rewards.

4.1 AI-based Feedback Collection

In our method, we explore a reward function informed by multiple detailed reward models for aligning modalities in LVLMs. These models (1) provide rewards at frequent intervals (namely, for sub-sentence of the generated content) and (2) assign rewards according to various categories of hallucinations. Each category of hallucination is evaluated by a distinct reward model. Therefore, in this stage, to train the reward model that can detect the hallucination, we collect the reward dataset first. Different from the most existing work which collects coarse-grained reward data via human feedback to refine VLMs, we collect fine-grained reward data by automatic AI model (left of Figure 2).

To achieve this, we first sample responses from the backbone LVLM as depicted in Section 3. Inspired by the existing fine-grained evaluation work (Jing et al., 2023; Min et al., 2023), we devise a fine-grained AI-based feedback collection method. In particular, we utilize AI models to annotate three kinds of hallucinations (i.e., object

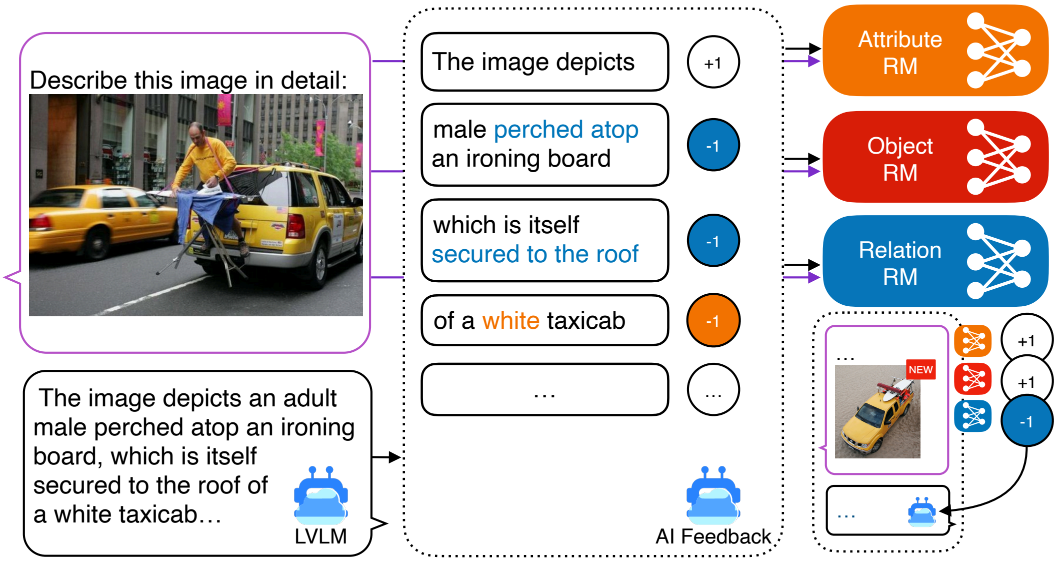

Figure 2: The illustration of our proposed FGAIF, which consists of three steps: AI-based feedback collection, fine-grained reward model training, and reinforcement learning with fine-grained rewards.

existence hallucination, object attribute hallucination, and object relationship hallucination) on the sub-sentence level for the response. In particular, to get the hallucination labels for each sub-sentence, we first split the response from the LVLM into sub-sentences as follows,

$$(s_1, \cdots, s_n) = \mathrm{SPLIT}(R), \tag{1}$$

where s_i is the i-th sub-sentence of the response. Thereafter, to accurately annotate three kinds of hallucination in the sub-sentence, we extract three kinds of atomic facts (Jing et al., 2023): object existence, object attribute, and object relationship atomic facts, from the sub-sentence, using ChatGPT as follows,

$$\begin{aligned} \{\{a_1^o, \cdots, a_{n^o}^o\}, \{a_1^a, \cdots, a_{n^a}^a\}, \{a_1^r, \cdots, a_{n^r}^r\}\} \quad (2) \\ = \mathrm{ChatGPT}(\mathrm{P}_s(s, \{s_i\}_{i=1}^n)), \end{aligned}$$

where a_i^o, a_i^a and a_i^r denote the i-th object existence, object attribute, and object relation types of atomic fact derived from the sub-sentence, respectively. And $n^{o/a/r}$ is the total number of object existence/attribute/relation atomic facts for the sub-sentence. Here we omit the index j of the sub-sentence for simplicity. Atomic fact is the minimal information unit and we show some examples in Appendix A. $\mathrm{P}_s(\cdot)$ is a prompt that can instruct ChatGPT to generate three kinds of atomic facts, and corresponding details can be found in Appendix A.

Thereafter, to get the label of each type of hallucination for each sub-sentence, we need to verify whether the atomic fact is consistent with the input image. We utilize superior LLaVA 1.5 (Liu et al., 2023b) to annotate the object existence hallucination, attribute hallucination, and relationship hallucination. Specifically, we feed LLaVA 1.5 with the image, the atomic fact, and the prompt, which can instruct LLaVA 1.5 to identify the consistency between atomic facts and the input image as follows,

$$f_{a_i}^{o/a/r} = LLaVA(\mathrm{P}_{con}(I, a_i^{o/a/r})), \tag{3}$$

where $f_{a_i}^o \in \{0,1\}$, $f_{a_i}^a \in \{0,1\}$ and $f_{a_i}^r \in \{0,1\}$ denote the hallucination label of i-th atomic fact in the sub-sentence in terms of object existence, object attribute, and object relationship types of atomic facts, respectively. $f_{a_i}^{o/a/r}$ is set to 1 when the output of LLaVA 1.5 indicates that the input image and the atomic fact are inconsistent (i.e., the corresponding atomic fact is a hallucination), otherwise, it is set to 0. $\mathrm{P}_{con}(\cdot)$ is the prompt that can be used to prompt the LLaVA 1.5 to annotate hallucination and it is shown in Applendix A.

Finally, we can aggregate the hallucination labels of atomic facts for each sub-sentence and then get the fine-grained sub-sentence-level hallucination labels as $f^{o/a/r} = sgn(\sum_i f_{a_i}^{o/a/r})$, where $f^{o/a/r}$ is the hallucination label for the sub-sentence in terms of object existence/attribute/relation. $sgn(\cdot)$ is the sign function. In addition, if there is not any atomic fact in a sub-sentence, the corresponding label $f^{o/a/r}$ is set to 2.

The reason why we use LVLM to verify the consistency between atomic fact and image even if the LVLM may also introduce hallucination: Our method converts the AI labeling task into a discriminative task that usually generates a short response, and this kind of task tends not to generate hallucination, which has been demonstrated in existing work (Jing et al., 2023; Min et al., 2023). Therefore, our AI-based feedback collection method can reduce the hallucination as much as possible.

4.2 Fine-grained Reward Model Training

As mentioned before, the existing LVLMs mainly suffer from three aspects of hallucinations, i.e., object existence, object attribute, and object relation. Based on the process above, we can get three kinds of hallucination labels for each sample. Thereafter, we train three reward models corresponding to each kind of hallucination (middle of Figure 2). Specifically, we first split the input of the reward model into tokens and get the index of the last token of each sub-sentence for the subsequent hallucination prediction as follows,

$$\begin{cases} T = \mathrm{Tokenizer}([P, I, R]), \\ \{ind_1, \cdots, ind_n\} = \mathrm{Search}([P, I, R, T]), \end{cases} \tag{4}$$

where ind_i is the index of the last token of the i-th sub-sentence. n is the total number of sub-sentences and T is the tokens for the input R (response), P (prompt) and I (image). Seach is a function that can get the index of the last token for each sub-sentence.

Finally, we can utilize the above-recognized indices to train reward models which is able to detect various kinds of hallucinations in the sub-sentence of response. In particular, we first feed the tokens above into the reward model backbones as follows,

$$\mathbf{F}^o = \mathrm{RM}^o(T), \mathbf{F}^a = \mathrm{RM}^a(T), \mathbf{F}^r = \mathrm{RM}^r(T). \tag{5}$$

Then, we connect the output from reward models, corresponding to the last token, with an MLP classifier. Thereafter, we can predict the hallucination label with the classifier. The above process can be formulated as follows,

$$\hat{f}_j^{o/a/r} = \mathrm{MLP}_{o/a/r}(\mathbf{F}_{ind_j}^{o/a/r}), \tag{6}$$

where $\mathbf{F}_{ind_j}^{o/a/r}$ is the feature vector of the last token for the j-th sub-sentence. $\hat{f}_j^o$, $\hat{f}_j^a$ and $\hat{f}_j^r$ are the predicted labels. To equip the three reward models with hallucination detection ability and give further rewards for reinforcement learning, we train the three reward models with a cross-entropy loss as $\mathcal{L}_{o/a/r} = \sum_{j=1}^n CE(f_j^{o/a/r}, \hat{f}_j^{o/a/r})/n$, where $CE(\cdot)$ is the cross-entropy function and $\mathcal{L}_o$, $\mathcal{L}_a$ and $\mathcal{L}_r$ are loss functions for different reward models (i.e., object existence, object attribute, and object relation).

4.3 Reinforcement Learning with Fine-grained Reward

Fine-tuning language models with reinforcement learning is an effective approach to align modalities in LVLMs. To make LVLMs generate more faithful responses rather than hallucinated responses, we also resort to reinforcement learning to further fine-tune LVLMs with the fine-grained reward (right of Figure 2). Specifically, we first segment the generated response from the LVLM into K sub-sentences $(s^1, \cdots, s^K)$. Then we get all kinds of rewards for each sub-sentence based on the well-trained reward model by cross-entropy loss. We define r_o^i, r_a^i, and r_r^i as the object existence, object attribute, and object relation rewards for the j-th sub-sentence. Then we have a combined reward function for each token as $r_t = -\sum_{l\in\{o,a,r\}} \sum_{i=1}^K (\mathbb{I}(t = T_i) w_l r_l^i)$, where T_i is the timestep for the last token of s^i. $\mathbb{I}(\cdot)$ is the indicator function. $w_l \in \mathbb{R}$ is a weight assigned to rewards. Thereafter, we utilize the PPO algorithm to train the policy model (i.e., the LVLM) following the existing work (Sun et al., 2023).

5 Experiment

5.1 Experimental Details

To ensure a fair and equitable comparison, we utilized same base model with the LLaVA-RLHF model whose network architecture is LLaVA$_{7B}$. In addition, we also adopt the same architecture (i.e., LLaVA$_{13B}$) with LLaVA-RLHF for the reward model. We compared our method with these models that used the same model backbone as ours (i.e., **LLaVA**$_{7B}$ (Liu et al., 2023c) and **LLaVA-RLHF**$_{7B}$). We also introduced some methods with the same backbone architecture but a larger model size (i.e., **LLaVA**$_{13B}$ and **LLaVA-RLHF**$_{13B}$). Besides, we further incorporated more advanced LVLMs for comparison, i.e., MiniGPT-4$_{7B}$ (Zhu et al., 2023), mPLUG-Owl$_{7B}$ (Ye et al., 2023), InstructBLIP$_{7B}$ (Dai et al., 2023), and

Table 1: POPE evaluation benchmark. Accuracy denotes the accuracy of predictions. “Yes” represents the probability of the model outputting a positive answer. ↑ denotes that the larger the value, the better the performance. The bold font denotes the best performance among our model and baselines with the same backbone architecture (LLaVA). The underlined font denotes the second-best performance among our model and baselines with the same backbone architecture.

Model	POPE							
	Random		Popular		Adversarial		Overall	
	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	F1↑	Yes
MiniGPT-4$_{7B}$	79.7	80.2	69.7	73.0	65.2	70.4	74.5	60.8
mPLUG-Owl$_{7B}$	54.0	68.4	50.9	66.9	50.7	66.8	67.2	97.6
InstructBLIP$_{7B}$	88.6	89.3	79.7	80.2	65.2	70.4	80.0	59.0
InstructBLIP$_{13B}$	88.7	89.3	81.4	83.5	74.4	78.5	83.7	62.2
LLaVA$_{7B}$	50.4	66.6	49.9	66.4	49.7	66.3	66.4	99.2
LLaVA$_{13B}$	73.7	78.8	73.6	78.2	67.2	74.4	77.1	73.7
LLaVA-RLHF$_{7B}$	84.8	83.3	83.3	81.8	80.7	79.5	81.5	41.8
LLaVA-RLHF$_{13B}$	85.2	83.5	83.9	81.8	**82.3**	80.5	81.9	39.0
FGAIF$_{7B}$	**87.0**	**86.7**	**84.0**	**83.7**	79.6	79.9	**83.4**	48.3

Table 2: Evaluation results for different LLMs on MMHal-Bench and LLaVA-Bench. “Over” and “Hal” denotes “Overall Score” and “Hallucination Rate”, respectively. “Con”, “De” and “Com” denote “Conversation”, “Detailed Description”, and “Complex Question”.

Model	MMHal-Bench					LLaVA-Bench			
	Over↑	Hal ↓	Object↑	Attribute↑	Relation↑	Con↑	De↑	Com↑	Full↑
MiniGPT-4$_{7B}$	3.39	0.24	3.0	2.54	3.67	80.5	74.5	81.6	78.9
mPLUG-Owl$_{7B}$	2.49	0.43	0.33	2.58	1.5	78.7	46.0	47.4	57.5
InstructBLIP$_{7B}$	2.10	0.58	2.08	2.67	2.17	95.4	96.3	99.1	97.0
InstructBLIP$_{13B}$	2.14	0.58	1.75	2.82	2.5	90.9	91.7	109.3	97.2
LLaVA$_{7B}$	1.55	0.76	0.00	1.25	2.00	75.1	75.4	92.3	81.0
LLaVA$_{13B}$	1.11	0.84	0.00	1.13	1.5	87.2	74.3	92.9	84.9
LLaVA-RLHF$_{7B}$	2.04	0.68	1.83	2.42	2.25	93.0	79.0	109.5	94.1
LLaVA-RLHF$_{13B}$	2.53	0.57	2.67	2.79	2.33	93.9	82.5	**110.1**	95.6
FGAIF$_{7B}$	**3.09**	**0.36**	**3.58**	**3.21**	**3.33**	**98.2**	**93.6**	110.0	**100.1**

InstructBLIP$_{13B}$.

To verify the effectiveness of our proposed FGAIF, we compare our method with baselines on several benchmarks, including **QA-based hallucination benchmarks** POPE (Li et al., 2023d) and MMHal-Bench (Sun et al., 2023), **hallucination metrics** CHAIR (Rohrbach et al., 2018) and Faith-Score (Jing et al., 2023), and the **general** benchmark LLaVA-Bench (Liu et al., 2023c). More detailed setups for dataset and model training are shown in Appendix B.

5.2 On Model Comparison

The results on **QA-based hallucination benchmarks** (i.e., POPE and MMHal-Bench) are summarized in Table 1 and Table 2. From this table, we have several observations. (1) LLaVA$_{7B}$ and InstructBLIP$_{7B}$ performs worse than LLaVA$_{13B}$ and InstructBLIP$_{13B}$ on most cases, respectively. Compared with LLaVA$_{13B}$, LLaVA$_{7B}$ has a strong hallucination problem, especially its over-confident problem on POPE. This indicates the importance of model size. (2) LLaVA-RLHF$_{7B}$ is better than LLaVA$_{7B}$, which indicates the superiority of further fine-tuning with human feedback. Notably, LLaVA-RLHF$_{7B}$ even has a better performance compared to LLaVA$_{13B}$, even though the latter has specifically more parameters. (3) Our model consistently performs better than the previous advanced in terms of all metrics and testing sets. This verifies that fine-grained artificial intelligence feedback also can be beneficial for hallucination mitigation in LVLMs. (4) Our FGAIF surpasses LLaVA-RLHF$_{7B}$ across all metrics. This implies the advantage of fine-grained artificial intelligence feedback compared to human feedback. (5) To fur-

Table 3: Results of CHAIR and FaithScore on LVLMs.

Model	CHAIR		FaithScore		Length
	CHAIR$_I$↓	CHAIR$_S$↓	F-Score ↑	F-Score$_S$↑	
MiniGPT-4$_{7B}$	9.4	17.4	63.9	61.8	245.1
mPLUG-Owl$_{7B}$	6.2	9.5	85.6	65.7	75.2
InstructBLIP$_{7B}$	2.4	3.8	93.6	80.0	45.6
InstructBLIP$_{13B}$	2.7	4.0	94.1	80.8	46.3
LLaVA$_{7B}$	9.1	22.0	88.9	72.3	216.0
LLaVA$_{13B}$	10.3	19.8	87.9	68.3	121.0
LLaVA-RLHF$_{7B}$	4.6	7.0	89.3	71.1	58.8
LLaVA-RLHF$_{13B}$	7.7	20.3	89.7	73.8	413.8
FGAIF$_{7B}$	**3.9**	**6.2**	**91.2**	**74.7**	60.2

Table 4: Ablation study of our FGAIF. The best results are highlighted in boldface. "Over" and "Hal" denotes "Overall Score" and "Hallucination Rate", respectively.

Model	CHAIR		FaithScore		POPE	MMHal-Bench	
	CHAIR$_I$ ↓	CHAIR$_S$ ↓	F-Score ↑	F-Score$_S$ ↑	F1 ↑	Over ↑	Hal ↓
FGAIF$_{7B}$	**3.9**	**6.2**	**91.2**	**74.7**	**83.4**	**3.09**	**0.36**
w/o-Obj	4.7	6.8	89.9	73.1	81.5	2.31	0.56
w/o-Att	4.1	6.3	90.3	73.7	82.4	2.56	0.45
w/o-Rel	4.2	6.4	90.3	73.4	82.6	2.64	0.44
w/o-AIF	4.8	7.0	89.1	72.8	81.0	1.76	0.67
w-Coarse	4.7	7.0	89.5	72.1	81.4	2.41	0.60

ther understand the performance of our FGAIF, we split the MMHal-Bench into three classes based on the original dataset: a) object existence (class "adversarial object"), b) object attribute (classes "object attribute" and "counting"), and c) object relation (class "spatial relation"). We observe that our method consistently achieves the best performance across all question categories.

We further show the performance of our FGAIF and baselines on **hallucination metrics** CHAIR and FaithScore in Table 3. InstructBLIP$_{7B}$ and InstructBLIP$_{13B}$ achieve the best performance in CHAIR and FaithScore metrics. The potential reason is that these two models tend to generate short answers and these two metrics just measure the precision of faithfulness but do not contain recall of faithfulness. Despite this, our FGAIF still outperforms the RLHF-based baseline (i.e., LLaVA-RLHF$_{7B}$) whose answers are shorter than FGAIF, which verifies the superiority of our method.

In addition, Table 2 shows the comprehensive performance comparison of our FGAIF and the baseline methods on the **general benchmark** LLaVA-Bench. From this table, we observed that most models perform worst on the "Detail" (i.e., detailed description) subset and perform best on the "Complex" (i.e., complex questions) subset. This may be due to the reason that the "Detail" (i.e., detailed description) subset has more stringent requirements for faithfulness because all the content of the response is required to be an accurate description of the input image. On the contrary, the "Complex" (i.e., complex questions) subset often explores the extended content of an image, sometimes leading to open-ended discussions. Therefore, the demand for strict consistency with the image isn't as critical. In addition, we found that the RLHF can boost the LVLM's performance on the whole LLaVA-Bench from 81.0 (LLaVA$_{7B}$) to 94.1 (LLaVA-RLHF$_{7B}$). Furthermore, our FGAIF can bring more performance gain in terms of the "Conv" subset, "Detail", "Complex" subset, and full set), compared with LLaVA-RLHF$_{7B}$. This further indicates the advance of our method.

5.3 On Ablation Study

To verify the effect of each component in our FGAIF, we devise the following variant methods for ablation study: 1) w/o-Obj: To demonstrate the effect of the object hallucination feedback, we remove the object existence reward model in this method; 2) w/o-Att: To show the necessity of the

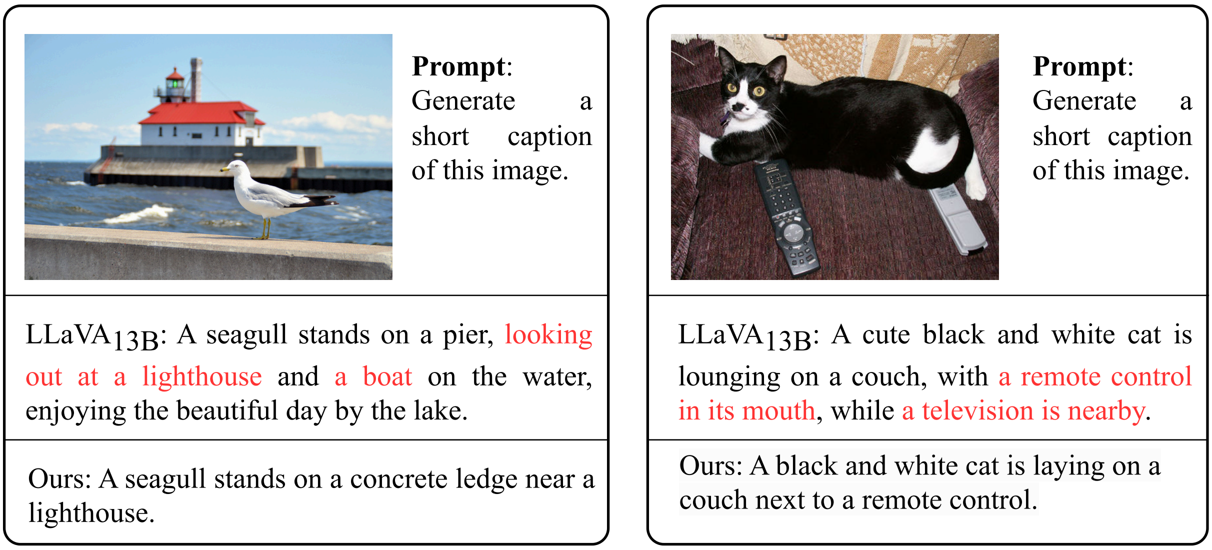

Figure 3: Comparison between the response generated by our method FGAIF and the baseline LLaVA$_{13B}$ on two testing samples. The red fonts denote the generated hallucinations.

attribute hallucination feedback, we remove the object attribute reward model in this method; 3) w/o-Rel: To demonstrate the effect of the relation hallucination feedback, we remove the object relation reward model in this method; 4) w/o-AIF: To show the benefit of using reinforcement learning from fine-grained artificial intelligence feedback, we remove all the reinforcement learning components in this variant; 5) w-Coarse: To verify the advance of the fine-grained feedback compared with the traditional coarse-grained uni reward model, we replace the three fine-grained reward models with one reward model which also is trained with AI annotated data and the training phrase is the same as the previous work (Sun et al., 2023).

Table 4 shows the ablation study results of our FGAIF on several hallucination benchmarks. From this table, we have the following observations. 1) w/o-RLAIF performs terribly compared with FGAIF. It confirms the necessity of using RLAIF for modality alignment and hallucination mitigation in LVLMs. 2) FGAIF consistently outperforms w/o-Obj, w/o-Att, and w/o-Rel, across different evaluation metrics. This is reasonable because each reward model can provide feedback for one kind of hallucination. 3) FGAIF surpasses w-Coarse, denoting that the fine-grained reward models are more essential to align modalities in LVLMs compared with the traditional coarse-grained uni reward model.

5.4 On Case Study

To get an intuitive understanding of the hallucination mitigation capability of our model, we show two testing results of our method and LLaVA$_{13B}$ in Figure 3. Looking into the generated responses of the first sample, we can learn that by incorporating our fine-grained artificial intelligence feedback, our FGAIF is able to generate the faithful description for the input visual image, while the baseline cannot (e.g., the baseline generates “A seagull looking out at a lighthouse” and “a boat on the water” mistakenly). This intuitively demonstrates the necessity of considering the fine-grained feedback in reinforcement learning. A similar result can be found in the second sample.

6 Conclusion

In this paper, we devise an innovative method for refining large vision-language models through Fine-Grained Artificial Intelligence Feedback (FGAIF), which mainly consists of three steps: AI-based feedback collection, fine-grained reward model training, and reinforcement learning with fine-grained rewards. The experimental results on hallucination and general benchmarks show the superiority of our method. The ablation study shows the necessity of each component in our method. In the future, we plan to incorporate more reward models in our method, such as soundness and fluency, which could provide more feedback during the model training stage.

Limitations

Our method enables the collection of feedback through AI, achieving the goal of reducing hallucinations in LVLMs. However, a challenge remains: During the feedback collection process, AI might introduce erroneous information. Some AI-generated feedback may contain imperceptible errors or inaccuracies, which can affect the model's performance.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *IEEE International Conference on Computer Vision*, pages 2425–2433. IEEE Computer Society.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073.

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, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey 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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems*.

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

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v(ision): Bias and interference challenges. *CoRR*, abs/2311.03287.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *CoRR*, abs/2305.04790.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022a. Lora: Low-rank adaptation of large language models. In *The International Conference on Learning Representations*. OpenReview.net.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022b. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *CoRR*, abs/2311.17911.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. FAITHSCORE: evaluating hallucinations in large vision-language models. *CoRR*, abs/2311.01477.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. RLAIF: scaling reinforcement learning from human feedback with AI feedback. *CoRR*, abs/2309.00267.

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

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023a. Silkie: Preference distillation for large visual language models. *CoRR*, abs/2312.10665.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023b. Silkie: Preference distillation for large visual language models. *CoRR*, abs/2312.10665.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023c. M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *CoRR*, abs/2306.04387.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji rong Wen. 2023d. Evaluating object hallucination in large vision-language models. *ArXiv*, abs/2305.10355.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *European Conference on Computer Vision*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. *CoRR*, abs/2304.08485.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *CoRR*, abs/2305.14251.

OpenAI. 2022. Chatgpt blog post. `https://openai.com/blog/chatgpt`.

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

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

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

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *CoRR*, abs/2310.16045.

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

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. 2023. Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *CoRR*, abs/2310.01779.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. Aligning modalities in vision large language models via preference fine-tuning. *CoRR*, abs/2402.11411.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593.

A Prompts

We provide the prompt of annotating the consistency between the image and atomic fact in Figure 4. We also provide the prompt of atomic fact generation in Figure 5. In this prompt, we asked ChatGPT to generate three types of atomic facts: object existence, object attribute, and object relation. To get better performance on atomic fact generation, we added some samples in this prompt. You can refer to these broken-down samples to understand atomic facts.

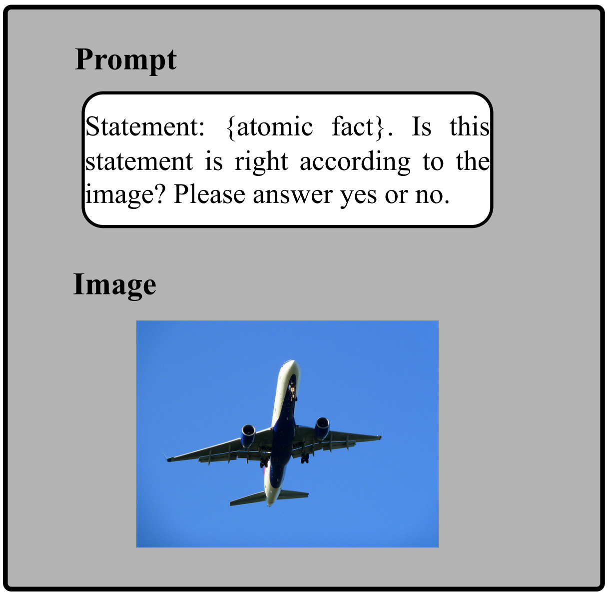

Figure 4: The prompt for verifying the consistency between the image and atomic fact.

B Experimental Settings

All experiments are conducted on a 4 $\times$ A100 80G GPU Server. For the reward model training, we use the Adam optimizer, and the learning rate, batch size, and epoch are set to 2e-5, 4, and 100. For the PPO training, we use the Adam optimizer, and the learning rate, batch size, and epoch are set to 1e-7, 256, and 2. We sample 3,500 and 14,000 examples from the MSCOCO 2014 (Lin et al., 2014) training set for reward model training and LVLM training, respectively. The prompt is set to "Describe this image in detail." for model training and sample. we adopt LoRA (Hu et al., 2022b) for all the reward model training and the LVLM fine-tuning processes.

POPE is a framework specifically designed for assessing object existence hallucinations in LVLMs. Specifically, POPE formulates the evaluation of object hallucination as a binary classification task that prompts LVLMs to output "Yes" or "No", e.g., "Is there a chair in the image?" "Yes" questions can be directly constructed based on objects appearing in the image. The "No" questions are constructed by three distinct sampling settings: random, popular, and adversarial. In the random setting, objects that are not present in the image are selected randomly. For the popular setting, the chosen non-existent objects are those from a pool of objects that appear most frequently in the MSCOCO dataset. In the adversarial setting, the sampling negative objects are often seen together with the objects in the image but are absent in the image under evaluation. This comprehensive approach allows for a nuanced analysis of the model's tendency to hallucinate across different scenarios. Finally, POPE consists of 3,000 samples under the setting of each type of negative sampling and 9,000 samples for the whole dataset.

MMHal-Bench benchmark has been introduced to assess and measure the degree of hallucination in responses by LVLMs. MMHAL-BENCH comprises 96 carefully constructed image-question pairs across eight different question categories and 12 object topics. These pairs are crafted to challenge LVLMs on common points of failure, including 1) Object Attribute, 2) Adversarial Object, 3) Comparison, 4) Counting, 5) Spatial Relation, 6) Environment, 7) Holistic Description, 8) Others. Different with POPE, it can evaluate more fine-grained hallucinations rather than only object existence.

CHAIR is a framework to quantify object hallucination in image captions. This method compares objects generated in captions against the ground truth objects within the images. CHAIR assesses hallucination on two levels: sentence-level and instance-level. The sentence-level score, referred to as CHAIR$_S$, quantifies the proportion of captions that contain hallucinated content, whereas the instance-level score, CHAIR$_I$, measures the frequency of hallucinated objects relative to the total number of objects mentioned by the model. Our evaluation involves a randomly selected subset of 1,000 images from the MSCOCO validation set, allowing for an analysis of our model's performance in minimizing object existence hallucination.

FaithScore is another framework to assess the accuracy and relevance of response generated by LVLMs. This innovative approach focuses on evaluating the consistency of atomic facts within the

Given an answer output by a vision-language model, break down its sub-sentence into independent atomic facts from it.
First extract elements from the answer. Then classify each element into a category (object, attribute, relation).
Finally, generate atomic facts for each element. You can refer to the context of the sub-sentence.
The relation must be the relationship between two objects.
Please note that you only need to output atomic facts. Besides, you must follow the format of examples. Facts are separated directly by periods.
The context is: %s
Please do not output other irrelevant information.

You should convert the pronoun into a specific object according to the context.
Please note that you only need to output atomic facts that are in the sub-sentence, the context is only used to help you understand context information such as the object to which the pronoun refers, don't output any content that didn't appear in the given sub-sentence.
Please note that the object is an objective description, not a subjective analysis, such as the atmosphere is not an object.
If the sub-sentence does not contain any object/attribute/relation, leave the corresponding line empty such as Object:

Sub-sentence: A man posing for a selfie in a jacket and bow tie.
Atomic facts:
Object: There is a man. There is a selfie. There is a jacket. There is a bow tie.
Attribute:
Relation: A man is in a jacket. A man is in a bow tie. A man posing for a selfie.

Sub-sentence: The image features a red velvet couch with a cat lying on it.
Atomic facts:
Object: There is a couch. There is a cat.
Attribute: The couch is red. The couch is velvet.
Relation: A cat is lying on a couch.

Sub-sentence: The photo is about a close-up image of a giraffe's head.
Atomic facts:
Object: There is a giraffe's head.
Attribute:
Relation:

Sub-sentence: A horse and several cows feed on hay.
Atomic facts:
Object: There is a horse. There are cows. There is a hay.
Attribute:
Relation: A horse feeds on hay. Cows feed on hay.

Sub-sentence: A red colored dog.
Atomic facts:
Object: There is a dog.
Attribute: The dog is red.
Relation:

Sub-sentence: {sub-sentence}
Atomic facts:

Figure 5: The prompt of atomic fact generation. In this prompt, we asked ChatGPT to generate three kinds of atomic facts: object existence, object attribute, and object relation. To get better performance on atomic fact generation, we added some samples in this prompt.

response against the depicted scenes in the input images. Different from CHAIR, FaithScore can demonstrate the model's hallucination performance in terms of object existence, attribute, and relation. Our evaluation involves a randomly selected subset of 1,000 images from the MSCOCO validation set, allowing for an analysis of our model's performance in mitigating object existence, attribute, and relation hallucination. It also provides an instance-level score F-Score and sentence-level score F-Score$_S$.

LLaVA-Bench is a general benchmark to assess the performance of LVLMs. LLaVA-Bench consists of 90 samples which can be categorized into three categories: detailed description, conversation, and complex question. All the prompts in this benchmark and answers are generated by GPT-4. In the evaluation process, the standard answer and generated response are fed into GPT-4 and GPT-4 then given a rating. Following the existing work (Sun et al., 2023), we also report the relative scores of LVLMs compared to GPT-4.

C Detailed Results

We report the detailed performance on MMHal-Bench and POPE in Table 5 and Table 6.

To understand the performance of our FGAIF, we split the MMHal-Bench into three classes based on the original dataset 1) object existence (class "adversarial object"), 2) object attribute (classes "object attribute" and "counting"), and 3) object relation (class "spatial relation"). From Table 5, we can observe that our method achieves the best performance consistently on all question categories (object existence, object attribute, and object relation), which further demonstrates the effectiveness of our method.

Table 5: Detailed evaluation results for different LMMs on MMHal-Bench. ↓ denotes that the less the value, the better the performance.

LLM	Overall Score↑	Hallucination Rate ↓	Score in Different Question Type Existence	Attribute	Relation
MiniGPT-4 $7B$	3.39	0.24	3.0	2.54	3.67
mPLUG-Owl $7B$	2.49	0.43	0.33	2.58	1.5
$\text{InstructBLIP}_{\text{7B}}$	2.10	0.58	2.08	2.67	2.17
$\text{InstructBLIP}_{\text{13B}}$	2.14	2.75	1.75	2.82	2.5
LLaVA_{7B}	1.55	0.76	0.00	1.25	2.00
$\text{LLaVA}_{\text{13B}}$	1.11	0.84	0.00	1.13	1.5
$\text{LLaVA-RLHF}_{\text{7B}}$	2.04	0.68	1.83	2.42	2.25
$\text{LLaVA-RLHF}_{\text{13B}}$	<u>2.53</u>	<u>0.57</u>	<u>2.67</u>	<u>2.79</u>	<u>2.33</u>
FGAIF_{7B}	**3.09**	**0.36**	**3.58**	**3.21**	**3.33**

Table 6: POPE evaluation benchmark. Accuracy denotes the accuracy of predictions. "Yes" represents the probability of the model outputting a positive answer. ↑ denotes that the larger the value, the better the performance. The bold font denotes the best performance among our model and baselines with the same backbone model. The underlined font denotes the second-best performance among our model and baselines with the same backbone model.

Model	**Random**			**Popular**			**Adversarial**			**Overall**	
	Acc↑	**F1↑**	**Yes**	**Acc↑**	**F1↑**	**Yes**	**Acc↑**	**F1↑**	**Yes**	**F1↑**	**Yes**
MiniGPT-4_{7B}	79.7	80.2	52.5	69.7	73.0	62.2	65.2	70.4	67.8	74.5	60.8
mPLUG-Owl_{7B}	54.0	68.4	95.6	50.9	66.9	98.6	50.7	66.8	98.7	67.2	97.6
InstructBLIP_{7B}	88.6	89.3	56.6	79.7	80.2	52.5	65.2	70.4	67.8	80.0	59.0
$\text{InstructBLIP}_{13B}$	88.7	89.3	55.2	81.4	83.5	62.6	74.4	78.5	69.0	83.7	62.2
LLaVA_{7B}	50.4	66.6	98.8	49.9	66.4	99.4	49.7	66.3	99.4	66.4	99.2
LLaVA_{13B}	73.7	78.8	72.3	73.6	78.2	71.0	67.2	74.4	77.8	77.1	73.7
LLaVA-RLHF_{7B}	84.8	83.3	39.6	83.3	<u>81.8</u>	41.8	<u>80.7</u>	79.5	44.0	81.5	41.8
LLaVA-RLHF_{13B}	<u>85.2</u>	<u>83.5</u>	38.4	<u>83.9</u>	<u>81.8</u>	38.0	**82.3**	<u>80.5</u>	40.5	<u>81.9</u>	39.0
FGAIF_{7B}	**87.0**	**86.7**	45.9	**84.0**	**83.7**	48.1	79.6	<u>79.9</u>	50.9	**83.4**	48.3