Detecting and Mitigating Hallucination in Large Vision Language Models via Fine-Grained AI Feedback

Anonymous submission

A Details of Generating AI feedback from LVLMs

A.1 Prompts

Figure 1 showcases the prompt, complemented by object annotations, for fine-grained AI feedback on detailed description task with GPT-4. Figure 2 illustrates the prompt designed to gather fine-grained AI feedback on visual complex reasoning task with GPT-4V. Table 9 shows the prompt to rewrtie hallucinatory response into non-hallucinatory one.

A.2 Hallucination Severity Score

Hallucination Severity Score is a metric we propose to evaluate severity of hallucinations in model response. Hallucination Severity Scores range from 0 to 3, with definitions provided in Table 10.

B Details of Hallucination Detection Model

In this section, we present further details regarding the hallucination detection model. This includes results on the pilot experiment for the backbone, experiments comparing different ratios of hallucinated to non-hallucinated data, and detailed results of fine-grained hallucination detection.

B.1 Pilot Experiment of Hallucination Detection Model

To choose a strong backbone for our hallucination detection model, we evaluate various models, including LLaVA-1.5-13b, Yi-VL-34b, and InternVL-Chat-Plus-v1.2-40b, on our fine-grained hallucination detection dataset. The results are detailed in Table 1. The analysis indicates that InternVL exhibits the best performance.

Table 1: Results of different base model on Fine-Grained Hallucination Detection.

Detection Model	Accuracy	Precision	Recall	F1
LLaVA-1.5	0.8000	0.8493	0.7654	0.8052
Yi-VL	0.8467	0.8537	0.8642	0.8588
InternVL-Chat-Plus	0.8733	0.8780	0.8889	0.8834

B.2 Ratios of Hallucinated to Non-hallucinated Data for Training Detection Model

We test multiple ratios of hallucinated to non-hallucinated data on our fine-grained hallucination detection dataset. Results are reported in Table 3. According to the results, the final ratio is set to 1:1.2 of hallucinated to non-hallucinated data.

We note that at a 1:1 ratio, the model is prone to falsely identifying non-hallucinated segments as hallucinated. Conversely, at a 1:1.4 ratio, it is inclined to classify hallucinated segments as non-hallucinated. Empirically, it is observed that a lower ratio results in a more lenient hallucination detection, whereas a higher ratio leads to increased strictness in hallucination detection.

B.3 Detailed Results of Fine-Grained Hallucination Detection

We show detailed results in Table 4 and Table 5 on the proposed fine-grained hallucination detection dataset, especially for the Multi part.

We observe that the performance in the 2-shot setting is significantly influenced by the prompt choice. Consequently, after testing numerous prompts, we report the best results for both the LLaVA-1.6-34b 2-shot and GPT-4V 2-shot.

C Details of Hallucination Mitigation Model

C.1 Metrics

In AMBER, we use metrics as follows:

1. CHAIR: This metric is commonly used for evaluating hallucinations and measures the frequency of hallucinatory objects appearing in the responses.

It is calculated using the formula:

$$\text{CHAIR}(R) = 1 - \frac{\text{len}(R'_{\text{obj}} \cap A_{\text{obj}})}{\text{len}(R'_{\text{obj}})}$$

where R'_{obj} is the set of objects mentioned in the model's response that are also in the annotated object list, and A_{obj} is the set of objects that actually exist in the image.

2. Cover: This metric measures the object coverage of responses, quantifying the proportion of objects mentioned in the response relative to the objects identified in the annotated list A_{obj}.

Table 2: Detailed evaluation results for different models on MMHal-Bench.

Model	Overall Score ↑	Hallucination Rate ↓	Score in Each Question Type ↑							
			Attribute	Adversarial	Comparison	Counting	Relation	Environment	Holistic	Other
LLaVA-RLHF	2.53	0.57	3.33	2.67	1.75	2.25	2.33	3.25	2.25	2.42
Qwen-VL-Chat	2.89	0.43	3.58	3.42	2.67	3.0	2.75	3.58	2.0	2.17
Silkie	3.01	0.41	4.42	3.42	2.0	2.83	3.17	4.0	1.75	2.5
Qwen w/ HSA-DPO	3.07	0.34	4.0	3.33	2.25	2.67	3.0	4.0	2.67	2.67
LLaVA w/ HSA-DPO	2.61	0.48	3.33	3.42	3.33	1.92	2.08	3.58	1.83	1.42

Table 3: Effect of data ratio between hallucinated and non-hallucinated samples on Fine-Grained Hallucination Detection.

Data ratio	Accuracy	Precision	Recall	F1
1:1.2	0.8733	0.8780	0.8889	0.8834
1:1.0	0.8467	0.8295	0.9012	0.8639
1:1.4	0.8467	0.9143	0.7901	0.8477

Table 4: Detailed results of Fine-Grained Hallucination Detection (Binary).

Model	P	R	ACC	F1
GPT-4V 2-shot	59.7	98.7	63.3	74.4
LLaVA-1.6-34B 2-shot	55.5	100	56.67	71.4
Ours	87.8	88.8	87.3	88.2

It is calculated using the formula:

$$\text{Cover}(R) = \frac{\text{len}(R'_{\text{obj}} \cap A_{\text{obj}})}{\text{len}(A_{\text{obj}})}$$

3. Hal: This metric represents the proportion of hallucinated responses. A response is considered to contain hallucinations if $\text{CHAIR}(R) \neq 0$.

It is calculated using the formula:

$$\text{Hal}(R) = \begin{cases} 1 & \text{if CHAIR}(R) \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

4. Cog: This metric evaluates whether the hallucinations in model responses are similar to those in human cognition. It evaluates the likelihood of LVLMs generating objects from a set of hallucinatory target objects H_{obj}.

using the formula:

$$\text{Cog}(R) = \frac{\text{len}(R'_{\text{obj}} \cap H_{\text{obj}})}{\text{len}(R'_{\text{obj}})}$$

In MMHal-Bench, we use metrics including:

1.Overall Score: this benchmark consists of 96 image-question pairs, ranging in 8 question categories × 12 object topics. And the overall score is the average score of 8 categories scores which are rated from 0 to 6.

2.Response-level hallucination rate: this benchmark employs GPT-4 to compare model response with human response and object labels to ascertain the presence of hallucinations, and then calculates the response-level hallucination rate out of 96 questions.

Table 5: Detailed results of Fine-Grained Hallucination Detection (Multi).

Model	Obj.	Attr.	Rela.	ACC
GPT-4V 2-shot	47.5	54.5	78.9	40.6
LLaVA-1.6-34B 2-shot	85.0	63.6	15.7	36.7
Ours	92.5	31.8	47.3	74.3

Table 6: Detailed results on LLaVA Bench in the wild.

Model	Conv.	Detail	Comp.	Overall
Qwen-VL-Chat	70.3	76.6	87.8	79.8
Silkie	83.3	76.0	90.6	84.9
LLaVA-1.5-Chat	74.7	51.1	83.7	72.7
Qwen w/ HSA-DPO	79.2	79.5	86.0	82.4
LLaVA w/ HSA-DPO	76.0	71.8	88.2	80.5

In LLaVA-Bench in the wild, we use score metric.

LLaVA-Bench in the wild consists of 60 image-question pairs to assessing multimodal conversation, detailed description and complex reasoning capabilities. The capabilities are quantified in the form of scores, which are provides via GPT-4 comparing model response against reference response.

C.2 Detailed Results on Hallucination Mitigation

In Table 2 and 6, we report the results of different question type in MMHal-Bench and LLaVA Bench in the wild, respectively. In table 7, in addition to the F1 score, we include accuracy, precision, recall and Yes Ratio on the adversarial part of POPE.

From the detailed results, we note that HSA-DPO enables both LLaVA-1.5-13b and Qwen-VL-Chat-7b to effectively reduce hallucinations, as evidenced on the MMHal-Bench and POPE, while preserving their multi-modal capabilities, as shown by the LLaVA Bench in the wild results. More-

Table 7: Detailed results on POPE adversarial.

Model	Acc	P	R	F1	Yes Ratio
Qwen-VL-Chat	82.7	82.2	83.6	82.9	50.9
Silkie	83.9	92.1	74.1	82.1	41.3
LLaVA-1.5	85.5	91.0	78.8	84.5	43.3
Qwen w/ HSA-DPO	83.3	85.2	80.7	82.9	47.3
LLaVA w/ HSA-DPO	84.0	80.2	90.1	84.9	56.2

Table 8: Hallucination severity scores rated by GPT-4V.

	Avg.	Sum	Avg. length
LLaVA-1.5-13b	0.796	42.99	579.2
LLaVA w/ HSA-DPO	0.602	33.50	466.8

over, our approach yields superior hallucination mitigation compared to models trained with extensive AI feedback or human-labelled datasets.

Task for the Rewrite Expert:
You are tasked with revising a model response. The original response may contain inaccuracies or "hallucinations"—details that do not exist in the actual image. Your objective is to correct these inaccuracies based on the information provided in the tags. The tags indicate the type of hallucination:
<object> for errors related to objects
<attribute> for incorrect descriptions of attributes
<relationship> for inaccuracies in the depiction of relationships between elements in the image
If marked as NO HALLUCINATION, it signifies that there are no inaccuracies present.
Output Format:
Provide a corrected version of the description that accurately reflects the actual contents of the image, addressing any hallucinations as guided by the tagged descriptions.
Original Question:
${Question}
Original Response:
${Response with Detected Hallcinations}
Instructions:
Analyze the original textual response in conjunction with the provided tags and reasons. Correct any elements of the response that contain hallucinations, making the necessary adjustments to ensure accuracy concerning the details of the image.
Now, please proceed to provide the revised response based on the tagged information and the actual image content. Remember that the tags may not always be correct, so use your judgment when making revisions.

Table 9: The prompt we used to rewrite hallucinatory response. ${Question} denotes question, ${Response with Detected Hallcinations} denotes hallucinatory response from model with detected hallucinations and reasons.

- None (0 point): No hallucination.
- Minor (1 point): The hallucination concerns a minor detail and does not significantly affect the overall portrayal of the scene. These hallucinations might include slight color variations, minor inaccuracies in object size, or subtle features that do not draw much attention within the context of the image.
- Moderate (2 points): The hallucination involves a noticeable detail that is incorrect within the context of the scene, yet the overall comprehension of the scene is maintained. This typically includes the misplacement of objects that are not the focal point of the scene or incorrect attributes of secondary elements that do not fundamentally change the scene's narrative or function.
- Major (3 points): The hallucination introduces a significant error or an entirely fabricated element that fundamentally alters the viewer's understanding of the scene. This could involve the description of a non-existent central object, a drastic misrepresentation of a key relationship or scene activity, or an erroneous depiction that leads to a completely different interpretation of the scene.

Table 10: The prompt we used to define Hallucination Severity Scores with GPT-4 and GPT-4V.

D Implementation Details

In this section, we provide more implementation details to enhance reproducibility.

For AI feedback collection, we collect 6k fine-grained AI feedback from proprietary models, which costs $600 and takes 4.5 hours with 3 parallel API request processes.

For the detection model training, we train InternVL-Chat-Plus-v1.2 for 2 epochs with LoRA where the lora rank is 128 and lora alpha is 256. Learning rate is 1e-5 and warm up ratio is 0.03 and total batch size is 16. It takes less than 5 hours on 4 Nvidia A100 GPUs in total. The pipeline for generating 16k preference datapoints takes 5 hours with 4 Nvidia A100 GPUs.

For HSA-DPO, we train LLaVA-1.5-13b for 2 epochs with LoRA where lora rank is 128 and lora alpha is 256. Learning rate is 2e-6 and total batch size is 32 and the DPO hyperparameter β is 0.1. It takes less than 5 hours on 4 A100 GPUs with 8k data in total. Note that we freeze projector for LLaVA-1.5-13b in HSA-DPO .

We train Qwen-VL-Chat-7b for 2 epochs with LoRA where lora rank is 128 and lora alpha is 256. Learning rate is 5e-6 and total batch size is 32 and the DPO hyperparameter β is 0.1. It takes around 3.5 hours on 4 A100 GPUs with 6k data in total.

Code, data, and model weight will be released soon.

Task instructions:
You are tasked with evaluating the accuracy of a detailed description of an image against provided regions. If the description includes hallucinations (i.e., details not present in the image):
Please identify hallucinated parts of the description and tag each hallucination at the end of the sentence containing it, not within the sentence.Using one of the following tags:
<object> for object hallucinations, like perceiving physical entities that are not actually present.
<relationship> for relationship hallucinations, like misperception about the relationships between people or position relationship between objects.
<attribute> for attribute hallucinations, like false perceptions about the qualities or characteristics of person or object.
For each identified hallucination, provide a reason why it is considered a hallucination by comparing it with the actual content of the image regions.
Once you have tagged and provided reasons for the hallucinations, assign a score to each tagged hallucination based on its impact on the overall accuracy of the description. Use the following scale:
- None (0 point): There is no hallucination.
- Minor (1 point): The hallucination concerns a minor detail and does not significantly affect the overall portrayal of the scene. These hallucinations might include slight color variations, minor inaccuracies in object size, or subtle features that do not draw much attention within the context of the image.
- Moderate (2 points): The hallucination involves a noticeable detail that is incorrect within the context of the scene, yet the overall comprehension of the scene is maintained. This typically includes the misplacement of objects that are not the focal point of the scene or incorrect attributes of secondary elements that do not fundamentally change the scene's narrative or function.
- Major (3 points): The hallucination introduces a significant error or an entirely fabricated element that fundamentally alters the viewer's understanding of the scene. This could involve the description of a non-existent central object, a drastic misrepresentation of a key relationship or scene activity, or an erroneous depiction that leads to a completely different interpretation of the scene.
After scoring, please rewrite the description, correcting any hallucinated parts to align with the actual content of the image.
Output format: Provide the tagged description and reasons first, and then provide scores for all hallucination tag with a brief justification.Last, give the corrected (rewritten) description.
Image Regions Information - regions
The regions are described in a JSON list, where each element contains the following details:
"w": Width of the region
"h": Height of the region
"phrase": Description of the region
"y": Vertical coordinate of the region in the image
"x": Horizontal coordinate of the region in the image
${regions}
Description - description
Analyze this textual description to match the given image regions. Identify any inaccuracies based on the regions' information.
Description:
${description}
Output Format - Output Template
Tagged Description:
The image features a woman standing on a tennis court, holding a tennis racket in her hand<object>. She is wearing a white shirt and a visor, which is a common attire for playing tennis<object>. The tennis racket is positioned in her right hand, and she appears to be focused on the game<relationship>.The tennis court is surrounded by numerous chairs, arranged in rows and placed at various distances from the court. These chairs are likely meant for spectators or players to rest during breaks in the game. The chairs are all green, creating a sense of depth and dimension within the scene<attribute>.
<object>
1. The woman is not standing on a tennis court, she is in a stadium.
2. The woman is not wearing a white shirt, she is wearing a white tank top.
<attribute>
1. The chairs are blue instead of green.
<relationship>
1. The tennis racket is in her left hand instead of in her right hand
The original description with tagged hallucinations immediately following the holistic score with a brief justification. For example:
Score:
Example:
<object> Tennis court: Major (3 points): The location of the woman is a significant detail in the description, and its incorrect portrayal fundamentally alters the viewer's understanding of the scene.
<object> White shirt: Moderate (2 points): The woman's attire is a noticeable detail, but its incorrect description does not fundamentally change the scene's narrative.
<attribute> Blue chairs: Minor (1 point): The wrong color of chairs is not a significant detail. The color is a less impactful detail compared to the presence or absence of objects.
<relationship> Left hand: Moderate (2 points): The woman's hand for playing is an important detail, but it does not significantly alter the scene's understanding.
The scores and justifications immediately following the rewritten Description. For example:
Rewritten Description:
The image features a woman standing in a stadium, holding a tennis racket in her hand. She is wearing a white tank top and a visor, which is a common attire for playing tennis. The tennis racket is positioned in her left hand and she appears to be focused on the game. The stadium is surrounded by numerous chairs, arranged in rows and placed at various distances from the court. These chairs are likely meant for spectators or players to rest during breaks in the game. The chairs are all blue, creating a sense of depth and dimension within the scene.
###
Now, please proceed with the analysis, tagging, scoring and rewriting of the description as per the instructions and templates provided above.Please tag only when you are very confident and provide the reason for tagging.

Figure 1: The prompt we used to get fine-grained AI feedback with GPT-4 for Detailed Description. ${regions} denotes input annotations from VG. ${description} denotes input hallucinatory response from model.

Task instructions:
You are tasked with evaluating the accuracy of a complex reasoning of an image provided. The complex reasoning contains a question about the content of the image and a answer to the question. If the answer includes hallucinations (i.e., details not present in the image):
Please identify hallucinated parts of the answer and tag each hallucination at the end of the sentence containing it, not within the sentence. Using one of the following tags:
<object> for object hallucinations, like perceiving physical entities that are not actually present.
<relationship> for relationship hallucinations, like misperception about the relationships between people or position relationship between objects.
<attribute> for attribute hallucinations, like false perceptions about the qualities or characteristics of person or object. For each identified hallucination, provide a reason why it is considered a hallucination by comparing it with the actual content of the image provided.(Notice that you can not tag more than three hallucination sentences)
Once you have tagged and provided reasons for the hallucinations, assign a score to each tagged hallucination based on its impact on the overall accuracy of the description. Use the following scale:
- None (0 point): No hallucination.
- Minor (1 point): The hallucination concerns a minor detail and does not significantly affect the overall portrayal of the scene. These hallucinations might include slight color variations, minor inaccuracies in object size, or subtle features that do not draw much attention within the context of the image.
- Moderate (2 points): The hallucination involves a noticeable detail that is incorrect within the context of the scene, yet the overall comprehension of the scene is maintained. This typically includes the misplacement of objects that are not the focal point of the scene or incorrect attributes of secondary elements that do not fundamentally change the scene's narrative or function.
- Major (3 points): The hallucination introduces a significant error or an entirely fabricated element that fundamentally alters the viewer's understanding of the scene. This could involve the description of a non-existent central object, a drastic misrepresentation of a key relationship or scene activity, or an erroneous depiction that leads to a completely different interpretation of the scene.

After scoring, please rewrite the answer, correcting any hallucinated parts to align with the actual content of the image(Please try to make corrections based on the original answers rather than directly deleting the hallucinations).

Output format: Provide the question, tagged answer and reason first, and then provide scores for all hallucination tag with a brief justification. Last, give the corrected (rewritten) answer.
Complex reasoning - reasoning:
Analyze the question and answer to match the given image. Identify any inaccuracies based on the image's information.
Reasoning:
${Reasoning}
Output Format - Output Template(Please notice that you should present the Reasons and Scores sequentially according to the order in which tags appear in the Tagged Answer.)
Question and Tagged Answer:
Example:
Question: Where is the woman, where is the tennis racket and what is the color of surrounding chairs?
Answer: The woman is standing on the lawn, holding a tennis racket in her hand<object>.The tennis racket is positioned in her right hand, and she appears to be focused on the game<relationship>.The chairs are all green, creating a sense of depth and dimension within the scene<attribute>.

Reasons:
Example:
<object>
The woman is not standing on a lawn, she is in a stadium.
<relationship>
The tennis racket is in her left hand instead of in her right hand.
<attribute>
The chairs are blue instead of green.

Scores:
Example:
<object> Tennis court: Major (3 points): The location of the woman is a significant detail in the description, and its incorrect portrayal fundamentally alters the viewer's understanding of the scene.
<relationship> Left hand: Moderate (2 points): The woman's hand for playing is an important detail, but it does not significantly alter the scene's understanding.
<attribute> Blue chairs: Minor (1 point): The wrong color of chairs is not a significant detail. The color is a less impactful detail compared to the presence or absence of objects.

The scores and justifications immediately following the rewritten answer.
For example:
Rewritten Answer:
Answer: The woman is standing in a stadium, holding a tennis racket in her hand.The tennis racket is positioned in her left hand and she appears to be focused on the game.The chairs are all blue, creating a sense of depth and dimension within the scene.
###
Now, please proceed with the analysis, tagging, scoring and rewriting of the answer as per the instructions and templates provided above.Please tag only when you are very confident and provide the reason for tagging.

Figure 2: The prompt we used to get fine-grained AI feedback with GPT-4V for Complex Reasoning. ${Reasoning} denotes input hallucinatory reasoning from model.