Mitigating Hallucinations in Large Vision-Language Models via DPO: On-Policy Data Hold the Key
Supplementary Material

The Appendix is organized as follows:

- In Chapter A, we offer a comprehensive description of our implementation details, complementing the information presented in Chapter 5.1. The training details and hyperparameter settings are reported in Chapter A.1, while the GPT-4V prompt and corresponding examples are provided in Chapter A.2.
- In Chapter B, we supply some additional experimental results. The helpfulness-related benchmark evaluations are conducted in Chapter B.1. The extended experiments on LLaVA-OneVision [63] are introduced in Chapter B.2, and additional ablation studies on the hyperparameter choosing are presented in Chapter B.3.
- In Chapter C, we provide additional analytical examples to complement the example presented in Figure 6.

Algorithm 1 OPA-DPO Training

```
Phase 1 Training: On-Policy Alignment
Require: Initial policy $\pi_\theta$, datasets $\mathcal{D} = \{\mathbf{x}, \mathbf{m}, \mathbf{y}_{\mathrm{GT}}, \mathbf{y}_{\mathrm{Rev}}\}^N$
for SFT epochs do
    for $\{\mathbf{x}, \mathbf{m}, \mathbf{y}_{\mathrm{GT}}, \mathbf{y}_{\mathrm{Rev}}\}^{M_1} \sim \mathcal{D}$ do
        Calculate loss in Eq. (4) for $\pi_\theta(\mathbf{y}_{\mathrm{GT}}|\mathbf{x}, \mathbf{m})$ and $\pi_\theta(\mathbf{y}_{\mathrm{Rev}}|\mathbf{x}, \mathbf{m})$
        Update $\pi_\theta$
    end for
end for
return OPA policy $\pi_{\mathrm{OPA}} = \pi_\theta^{\mathrm{final}}$
Phase 2 Training: OPA-DPO
Require: Initial policy $\pi_{\theta'} = \pi_{\mathrm{OPA}}$; hyperparameters $\beta, \gamma_1, \gamma_2, \delta$; datasets
    $\mathcal{D} = \{\mathbf{x}, \mathbf{m}, \mathbf{y}_{\mathrm{Gen}}, \mathbf{y}_{\mathrm{GT}}, \mathbf{y}_{\mathrm{Rev}}, S_{\mathrm{hal}}, S_{\mathrm{img}}\}^N$
for OPA-DPO epochs do
    for $\{\mathbf{x}, \mathbf{m}, \mathbf{y}_{\mathrm{Gen}}, \mathbf{y}_{\mathrm{GT}}, \mathbf{y}_{\mathrm{Rev}}, S_{\mathrm{hal}}, S_{\mathrm{img}}\}^{M_2} \sim \mathcal{D}$ do
        Calculate loss in Eq. (7) with $\{\mathbf{x}, \mathbf{m}, \mathbf{y}_{\mathrm{Gen}}, \mathbf{y}_{\mathrm{GT}}, \mathbf{y}_{\mathrm{Rev}}, S_{\mathrm{hal}}\}^{M_2}$
        Produce distorted image $\mathbf{m}' = \mathbf{m} \odot \text{pixel\_mask}$
        Calculate loss in Eq. (8) with $\{\mathbf{x}, \mathbf{m}, \mathbf{m}', \mathbf{y}_{\mathrm{GT}}, \mathbf{y}_{\mathrm{Rev}}, S_{\mathrm{img}}\}^{M_2}$
        Calculate loss in Eq. (9) with $\{\mathbf{x}, \mathbf{m}, \mathbf{y}_{\mathrm{GT}}, \mathbf{y}_{\mathrm{Rev}}\}^{M_2}$
        Combine the losses as in Eq. (10)
        Update $\pi_{\theta'}$
    end for
end for
return OPA-DPO policy $\pi_{\mathrm{OPA}}^{\mathrm{DPO}} = \pi_{\theta'}^{\mathrm{final}}$
```

A. Implementation Details

A.1. Training Details.

Importantly, we emphasize that OPA-DPO does not depend on detailed hyperparameter tuning for different base models or training datasets. In our experiments, we apply OPA-DPO to two LVLMs with varying parameter sizes: LLAVA-v1.5-7B and LLAVA-v1.5-13B. We maintain consistent hyperparameter settings for OPA-DPO across different models and training datasets.

As shown in Figure 3, the initial step in OPA-DPO involves instructing the model (slated for training) to generate responses $\mathbf{y}_{\mathrm{Gen}}$ based on pre-collected images $\mathbf{m}$ and prompts $\mathbf{x}$. Notably, we employ a combination of ”top-k” and ”top-p” sampling methods to select tokens with relatively high sampling probabilities according to the initial policy, thereby revealing the intrinsic hallucinations of the policy itself. For token sampling, we set $\mathrm{topk} = 30$, $\mathrm{topp} = 0.95$, and use a temperature of 1.0.

Following that, GPT-4V is tasked with identifying hallucinations by evaluating the generated responses at the sentence level. Each sentence in a response is assigned a hallucination severity score, S_{hal}, on a scale from one to four, indicating the severity of any hallucination present. As we introduced in Eq. 7, this score is incorporated into hallucination-weighted policy updating, with the corresponding mapping between scores and weights provided in Table 6. Additionally, GPT-4V is required to categorize sentences with incorrect description as either *image recognition errors* or *language comprehension errors*. The classification results S_{img} is utilized for image-weighted policy updating, as defined in Eq. (8). Table 7 outlines the mapping between these classifications and their respective updating weights. Note that both S_{hal} and S_{img} are evaluated at the sentence level, ensuring that W_{hal} and W_{img} are assigned the same value for each token within a sentence. Lastly but most importantly, GPT-4V is also instructed to make minimal revisions to any erroneous sentences, and the aggregate of these revised sentences is denoted as $\mathbf{y}_{\mathrm{Rev}}$. Please refer to Chapter A.2 for detailed prompt and example. In our implementation, we utilize the GPT-4V version from 2024-02-15, with the generation temperature set to 0.

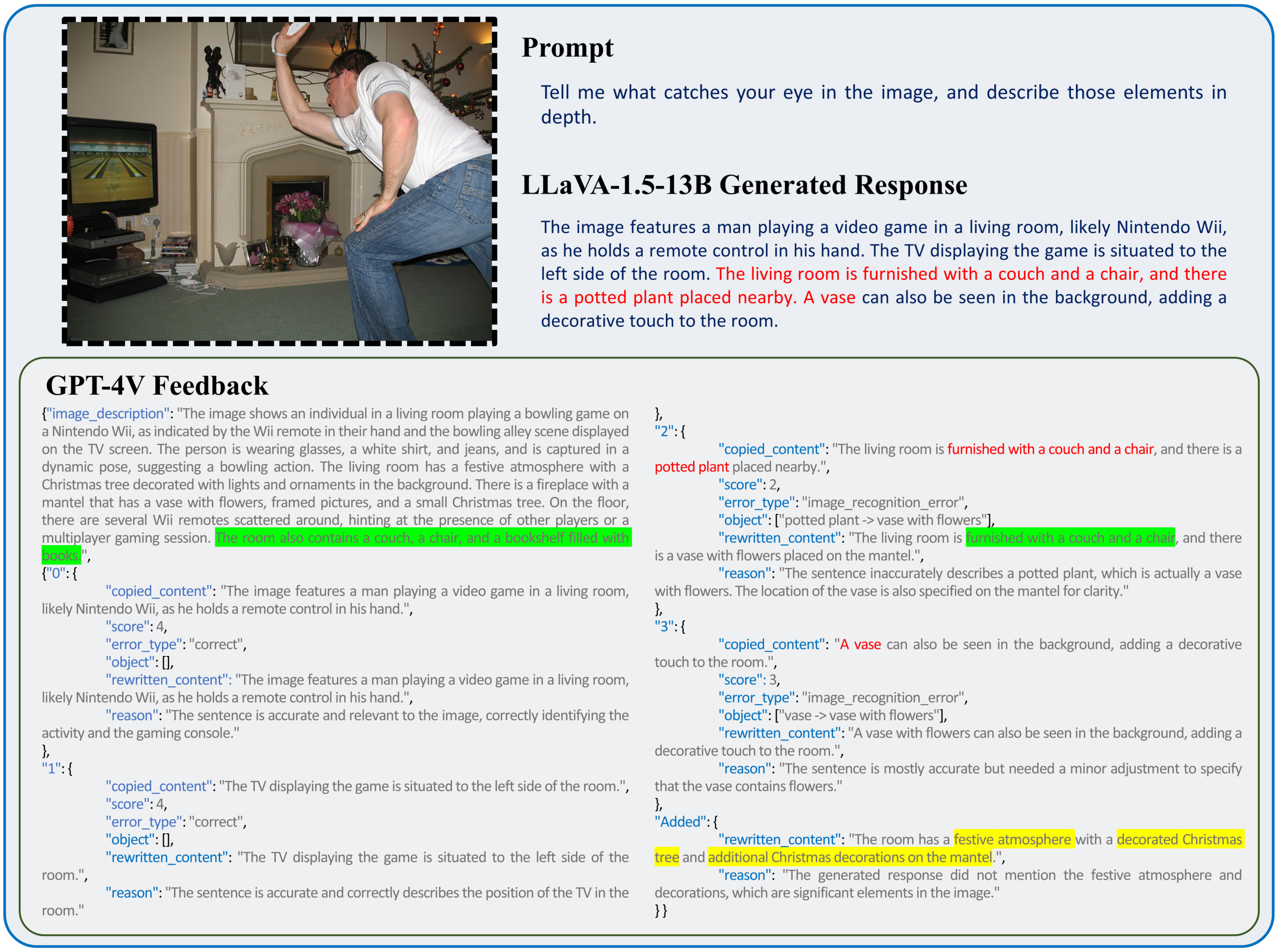
Table 6. GPT-4V assigned hallucination scores and the corresponding update weights for language correction loss, as described in Eq. (7).

Hallucination Severity	Score from GPT-4V (S_{hal})	Updating Weight (W_{hal})
Not at all	4	2.5
Minor	3	2.0
Major	2	1.5
Totally	1	1.0

Table 7. GPT-4V labeled error types and the corresponding update weights for image focus loss, as described in Eq. (8).

Label from GPT-4V (S_{img})	Updating Weight (W_{img})
correct	1.0
language_comprehension_error	1.0
image_recognition_error	3.0

After completing the data collection, we proceed with a two-phase training for the initial models as detailed in Algo-



Prompt

Tell me what catches your eye in the image, and describe those elements in depth.

LLaVA-1.5-13B Generated Response

The image features a man playing a video game in a living room, likely Nintendo Wii, as he holds a remote control in his hand. The TV displaying the game is situated to the left side of the room. The living room is furnished with a couch and a chair, and there is a potted plant placed nearby. A vase can also be seen in the background, adding a decorative touch to the room.

GPT-4V Feedback

```
{"image_description": "The image shows an individual in a living room playing a bowling game on
a Nintendo Wii, as indicated by the Wii remote in their hand and the bowling alley scene displayed
on the TV screen. The person is wearing glasses, a white shirt, and jeans, and is captured in a
dynamic pose, suggesting a bowling action. The living room has a festive atmosphere with a
Christmas tree decorated with lights and ornaments in the background. There is a fireplace with a
mantel that has a vase with flowers, framed pictures, and a small Christmas tree. On the floor,
there are several Wii remotes scattered around, hinting at the presence of other players or a
multiplayer gaming session. The room also contains a couch, a chair, and a bookshelf filled with
books.",
{"0": {
        "copied_content": "The image features a man playing a video game in a living room,
likely Nintendo Wii, as he holds a remote control in his hand.",
        "score": 4,
        "error_type": "correct",
        "object": [],
        "rewritten_content": "The image features a man playing a video game in a living room,
likely Nintendo Wii, as he holds a remote control in his hand.",
        "reason": "The sentence is accurate and relevant to the image, correctly identifying the
activity and the gaming console."
},
"1": {
        "copied_content": "The TV displaying the game is situated to the left side of the room.",
        "score": 4,
        "error_type": "correct",
        "object": [],
        "rewritten_content": "The TV displaying the game is situated to the left side of the
room.",
        "reason": "The sentence is accurate and correctly describes the position of the TV in the
room."
},
"2": {
        "copied_content": "The living room is furnished with a couch and a chair, and there is a
potted plant placed nearby.",
        "score": 2,
        "error_type": "image_recognition_error",
        "object": ["potted plant -> vase with flowers"],
        "rewritten_content": "The living room is furnished with a couch and a chair, and there
is a vase with flowers placed on the mantel.",
        "reason": "The sentence inaccurately describes a potted plant, which is actually a vase
with flowers. The location of the vase is also specified on the mantel for clarity."
},
"3": {
        "copied_content": "A vase can also be seen in the background, adding a decorative
touch to the room.",
        "score": 3,
        "error_type": "image_recognition_error",
        "object": ["vase -> vase with flowers"],
        "rewritten_content": "A vase with flowers can also be seen in the background, adding a
decorative touch to the room.",
        "reason": "The sentence is mostly accurate but needed a minor adjustment to specify
that the vase contains flowers."
},
"Added": {
        "rewritten_content": "The room has a festive atmosphere with a decorated Christmas
tree and additional Christmas decorations on the mantel.",
        "reason": "The generated response did not mention the festive atmosphere and
decorations, which are significant elements in the image."
}}
```

Figure 7. Example of feedback from GPT-4V. Hallucinated parts in the base-model generated responses are marked in red, missing details are highlighted in yellow. Note that the feedback from GPT-4V also contain hallucinations, as highlighted in green.

rithm 1. The first phase training (line 1-7) termed On-Policy Alignment (OPA), involves performing a 2-epoch LoRA-SFT on both ground-truth responses and GPT-4V revised responses. The entire backbone model, including the vision encoder and multimodal connection layers, is wrapped with LoRA modules. We employ a cosine learning rate schedule beginning at 2e-5 with a batch size of 128. The LoRA rank is set to 256, and LoRA alpha is set to 512. The updated policy from this phase is denoted as π_{OPA}, which serves as the initial (reference) policy for the subsequent OPA-DPO training. The second phase of training (lines 8-18) uses the same LoRA module as in phase 1, extending over 4 additional epochs with a batch size of 32 and a cosine learning rate starting at 1e-6. In our equations, we set $\beta = 0.1$ in Eqs. (7)(8)(9), $\delta = 0$ in Eq. (9), and $\gamma_1 = 0.2, \gamma_2 = 1.0$ in Eq. (10). For the distorted images $\mathbf{m}'$ in Eq. (8), we randomly mask 30% of pixels, assigning the masked areas the average pixel values. For ablation studies on the relative hyperparameters, please refer to Chapter B.3.

A.2. Prompts for GPT-4V

To obtain fine-grained feedback from GPT-4V, we crafted a detailed prompt, as shown in the TextBox on the following page. To establish a one-to-one correspondence between the revised and original responses, we instruct GPT-4V to first copy the generated sentence before proceeding with assessment and revision. Additionally, we request that GPT-4V provide the rationale behind its assigned score or revision. It is important to note that GPT-4V may itself produce hallucinations, which can affect the reliability of its feedback. An example is provided in Figure 7.

GPT-4V Prompt for Fine-Grained Sentence-Level Revision of Generated Responses

Your role is as a discerning assistant tasked with evaluating and refining responses for multimodal tasks. Upon being presented with a question that requires the interpretation of both text and images, you will receive two distinct responses. The first is crafted by our sophisticated multimodal model, while the second represents an approximate ideal answer—it may be incomplete or incorrect. You will also be provided with the images pertinent to the question.

Your objective is to meticulously assess these responses. You are to enhance the model-generated response by making precise, minimal modifications that bring it into closer alignment with both the image and the approximate ideal answer. Your revisions should preserve the integrity of the original response as much as possible.

Be mindful that the approximate ideal response may not contain all the necessary information to fully address the question or may include mistakes. In such cases, you must carefully evaluate the accuracy of the model-generated response by consulting the image, which serves as the primary reference.

Your analysis should prioritize the information provided in the image to ascertain the accuracy and completeness of the model-generated response. The ultimate goal is to ensure that the final response is both accurate in relation to the images and as informative as possible while remaining true to the content originally produced by the model.

Your task involves meticulous scrutiny of the generated response to a multimodal task, sentence by sentence. Here's how you should approach the revision process:

Evaluate each sentence within the generated response.

- If a sentence is both accurate and relevant to the task, it should remain unchanged.
- If you encounter a sentence that is only partially correct, carefully adjust the erroneous or incomplete segments to improve its precision. Ensure that these modifications are minimal and directly address the inaccuracies.
- If you find any sentences that contain hallucinations or extraneous information, these must be either rephrased or replaced entirely. Use the image and the approximate ideal response as your sources for correction, aiming to retain the essence of the original content when possible.

You are to present your output in a structured JSON format. Begin with the key "image_description" where a comprehensive description of the provided images should be articulated. Following this, evaluate the generated response sentence by sentence. For each sentence, craft a JSON object that contains the original sentence, your refined version, and a brief commentary explaining your revisions. The format is as follows:

1. "copied_content": Copy and paste the original sentence as it appears in the generated response.
2. "score": Provide a score between 1 and 4, reflecting the sentence's accuracy and relevance to the image and question:
   - 4 for a sentence that is completely accurate and relevant, aligning perfectly with the image information and the approximate ideal answer, requiring no adjustments.
   - 3 for a sentence that is largely correct but needs minor tweaks, like an accurate object described with an incorrect count or size.
   - 2 for a sentence with substantial issues requiring significant changes, such as incorrect object recognition or incorrect relationships between objects.
   - 1 for a sentence that is completely irrelevant or incorrect, with no relation to the image or the question at hand.
3. "error_type": Specify the type of error detected in the sentence:
   - "correct" if the sentence is accurate or requires only minor adjustments, applicable only to a score of 4.
   - "image_recognition_error" when the error arises from an incorrect interpretation of the visual content, like mistaking an apple for a pear.
   - "language_comprehension_error" when the image is correctly understood, but the language used is incorrect, such as placing the Eiffel Tower in Berlin instead of Paris.
4. "object": List any objects that are hallucinated or misidentified, and provide the correct identification. Leave this field empty if there are no hallucinations or misidentifications.
   - For instance, if the sentence inaccurately identifies a cat sleeping on a table as a dog standing on a blanket, the "object" should be ["dog -> cat", "standing -> sleeping", "blanket -> table"].
5. "rewritten_content": Present the corrected sentence after applying necessary adjustments, considering all information from the image captions and the approximate ideal answer.
6. "reason": Explain the rationale for the given score, the identified error type, and any modifications made. This should include the reasoning behind changes and the decision to maintain certain parts of the original sentence.

If the rewritten sentences still lack essential information necessary for answering the given questions, add the missing part to the "Added" section and incorporate that missing information minimally. Only do this if absolutely necessary.

You should never bring other hallucinations into the rewritten parts. Only do the modifications when you are one hundred percent sure that the original sentence is incorrect or irrelevant. Please note that the rewritten sentence should retain as much of the generated response as possible. All unnecessary changes should be minimized.

B. Additional Experiments

B.1. Helpfulness Benchmark Evaluations.

To demonstrate that the exceptional performance of OPA-DPO on hallucination-related metrics does not result in a decline in helpfulness-related metrics, we evaluated the performance of various RLHF/RLAIF-based algorithms designed to enhance LVLMs on the LLaVA-Bench [5], as shown in Table 8. The results indicate that the OPA-DPO trained model performs at an upper-middle level. With the exception of LLaVA-RLHF, the performance of each algorithm on the LLaVA-Benchmark shows minimal variation. However, LLaVA-RLHF is significantly less effective than other algorithms in hallucination-related metrics.

Table 8. Comparison of RLAIF/RLHF-based algorithms for enhancing LVLMs on LLaVA-Bench.

Algorithm	LLaVA-Bench			
	Conv.↑	Detail↑	Comp.↑	All↑
LLaVA-Instruct-1.5-7B [5, 6]	84.1	74.4	89.8	83.0
+ LLaVA-RLHF [21]	84.1	75.3	106.8	88.9
+ HA-DPO [18]	80.7	74.5	88.4	81.4
+ POVID [16]	84.9	77.3	90.3	84.3
+ RLAIF-V [20]	75.8	83.7	90.7	83.5
+ OPA-DPO (ours)	82.1	79.5	87.9	83.2
LLaVA-Instruct-1.5-13B [5, 6]	79.6	77.3	91.4	82.9
+ LLaVA-RLHF [21]	93.1	76.2	105.6	91.8
+ RLHF-V [15]	93.1	75.3	91.6	86.7
+ HSA-DPO [19]	76.0	71.8	88.2	80.5
+ OPA-DPO (ours)	87.1	78.3	90.7	85.5

B.2. Experiments on LLaVA-OneVision.

To validate the generalizability of OPA-DPO, we implement it on more advanced LVLMs, LLaVA-OneVision [63] (in Table 9). This series of models is built upon Qwen2 [45] and employs a completely different visual information encoding mechanism in comparison with LLaVA-Instruct-v1.5. We train the models using 2.4k data samples with the same hyperparameter settings as introduced in Chapter A. Notably, LLaVA-OneVision tends to generate excessive redundant content, with a more severe hallucination phenomenon. OPA-DPO proves to be particularly effective on LLaVA-OneVision.

Table 9. Evaluation of hallucination metrics for OPA-DPO applied to LLaVA-OneVision [63], trained with 2.4k samples.

Model size	Algo.	Object-Hal		MMHal-Bench		AMBER			
		CHAIRs↓	CHAIRi↓	Score↑	HalRate↓	CHAIR↓	Cover↑	HalRate↓	Cog↓
7B	**Initial**	44.67	8.64	2.98	0.31	8.5	**72.9**	73.3	9.9
	DPO	34.67	7.08	3.42	0.25	7.6	71.1	68.3	8.6
	OPA	22.67	6.20	3.60	0.25	6.0	63.5	32.7	3.5
	OPA-DPO	**15.33**	**3.72**	**3.69**	**0.23**	**4.7**	61.2	**25.0**	**2.6**
0.5B	**Initial**	34.33	9.81	2.53	0.51	9.2	**72.8**	73.7	9.2
	DPO	30.33	7.84	2.62	0.50	8.9	71.0	71.2	9.6
	OPA	26.33	7.71	2.70	0.48	8.2	67.0	53.1	6.3
	OPA-DPO	**21.00**	**5.93**	**2.72**	**0.48**	**7.0**	62.6	**39.9**	**4.5**

Table 10. Ablation studies on the mask ratio of the distorted image and the term coefficient γ_1 in the image focus mechanism.

Model Size	Mask Ratio	IF Coef γ_1	AMBER			MMHal-Bench		Object Hal	
			Cover↑	HalRate↓	repeat	Score↑	HalRate↓	CHAIRs↓	CHAIRi↓
7B	0.1	0.2	46.1	12.5	0.3%	2.60	0.49	14.67	4.28
	0.3	0.2	47.9	11.6	0.6%	2.83	0.45	13.00	4.25
	0.5	0.2	46.3	11.6	0.3%	2.73	0.47	14.67	4.18
	0.7	0.2	45.6	12.2	0.2%	2.69	0.47	13.33	4.45
	0.3	0.1	46.2	12.3	0.6%	2.70	0.47	14.67	4.05
	0.3	0.5	44.3	11.1	5.3%	2.79	0.45	14.67	4.32
	0.3	1.0	43.5	9.3	20.8%	2.26	0.59	9.67	2.98
13B	0.1	0.2	48.3	13.9	0.4%	2.84	0.45	19.00	6.16
	0.3	0.2	48.3	12.8	0.8%	3.07	0.39	16.33	5.48
	0.5	0.2	48.2	13.4	0.8%	2.99	0.41	18.00	5.41
	0.7	0.2	48.3	13.9	0.4%	2.84	0.45	19.00	6.16
	0.3	0.1	48.3	13.0	0.2%	2.95	0.42	18.33	5.89
	0.3	0.5	48.1	12.3	2.6%	2.97	0.44	16.67	5.35
	0.3	1.0	46.1	10.3	7.9%	2.72	0.44	16.33	5.00

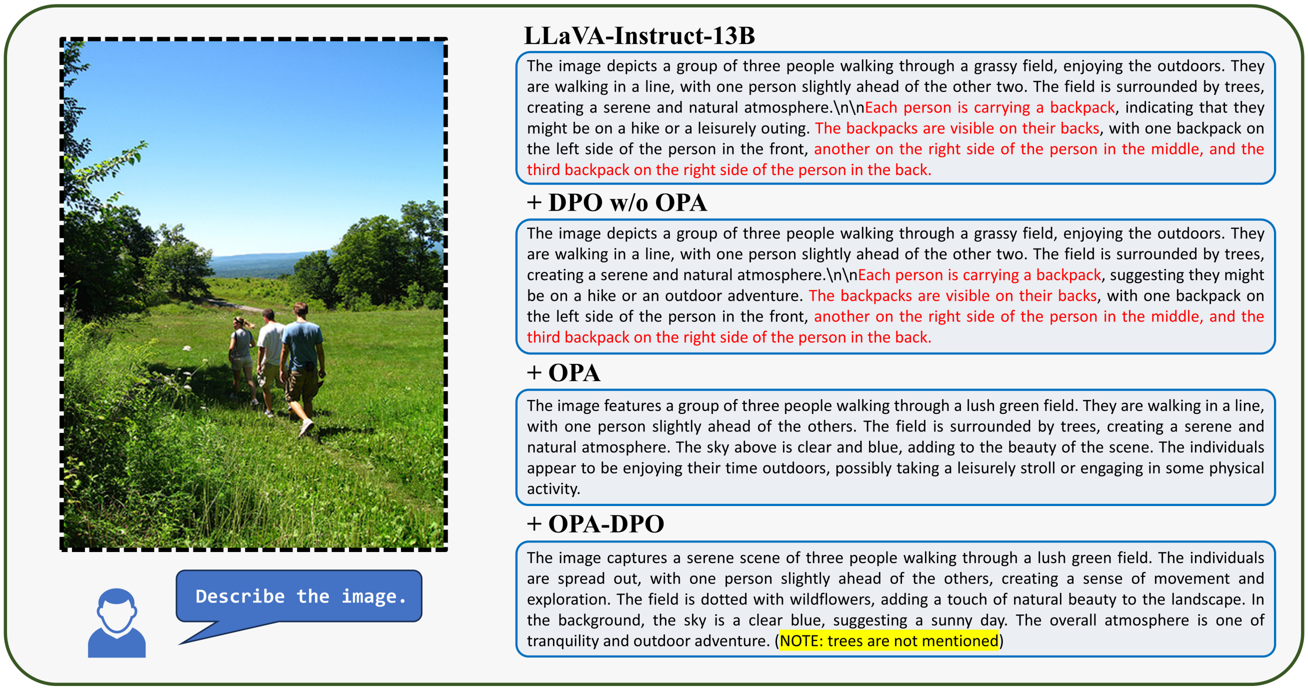
B.3. Additional Ablation Studies.

As an important component of OPA-DPO, the image focus mechanism (see Chapter 4.2) involves two hyperparameters that require tuning: the term coefficient γ_1 in Eq. 10 and the mask ratio of the distorted image. We found that setting γ_1 to 0.2 and randomly masking 30% of the pixels is optimal for both the LLaVA-1.5-7B and LLaVA-1.5-13B models. In contrast, the pioneering algorithm mDPO [27], which first utilized this mechanism, opts to set γ_1 to 1.0 and employs a variable masking ratio of 0-20% of pixels randomly. We find that the mask ratio has a slight impact on the model's performance, whereas the term coefficient γ_1 has a more significant effect. In particular, setting the coefficient too high results in excellent performance in metrics related to hallucination rate, but at the cost of being overly conservative and severely lacking in explanatory detail. Additionally, the model tends to repeat its last sentence or words and fails to generate an EOS token when using greedy sampling. As a compromise, we set $\gamma_1 = 0.2$. Ablation studies supporting our findings are presented in Table 10.

C. Additional Qualitative Examples

Image Descriptions. As introduced in Chapter 5, OPA-DPO is particularly effective in preventing hallucinations by adopting a somewhat conservative strategy that avoids uncertain assertions. Such strategy significantly enhances the credibility of the responses but may lead to the omission of some ambiguous details, necessitating a trade-off. In addition to the case presented in Chapter 5.5, we offer further examples involving image detail descriptions, as illustrated in Figures 8, 9, 10, and 11. In these cases, the initial model's output contained numerous hallucinations and flawed reasoning. This issue persisted even after training with naive DPO without OPA. However, after applying OPA to 4.8k samples, hallucinations were nearly eliminated, with only minor instances remaining. The subsequent implementation of OPA-DPO completely resolved these issues, although some details from the original description were omitted. It is important to note that the omitted details are often not central to the image's main information and do not cause the overall description to deviate.

False Premise Queries. Another interesting phenomenon we observed in our experiments is that, LVLMs consistently experience hallucinations when presented with queries based on false premises. These queries contain objects or details that do not exist in the image or are irrelevant to it. For example, the LVLM is asked to describe the girl's outfit given a picture of a basketball. As demonstrated in Figures 12, 13, and 14, the base model consistently produces absurd responses to nonsensical questions due to linguistic inertia. The application of DPO without OPA does not generally modify these responses. Furthermore, utilizing the OPA operation in isolation is sometimes insufficient to address the issue. However, when both methods are combined, through training with OPA-DPO, the model is able to discern false premises in queries or prompts and provide reasoned responses.



Describe the image.

LLaVA-Instruct-13B

The image depicts a group of three people walking through a grassy field, enjoying the outdoors. They are walking in a line, with one person slightly ahead of the other two. The field is surrounded by trees, creating a serene and natural atmosphere.\n\nEach person is carrying a backpack, indicating that they might be on a hike or a leisurely outing. The backpacks are visible on their backs, with one backpack on the left side of the person in the front, another on the right side of the person in the middle, and the third backpack on the right side of the person in the back.

+ DPO w/o OPA

The image depicts a group of three people walking through a grassy field, enjoying the outdoors. They are walking in a line, with one person slightly ahead of the other two. The field is surrounded by trees, creating a serene and natural atmosphere.\n\nEach person is carrying a backpack, suggesting they might be on a hike or an outdoor adventure. The backpacks are visible on their backs, with one backpack on the left side of the person in the front, another on the right side of the person in the middle, and the third backpack on the right side of the person in the back.

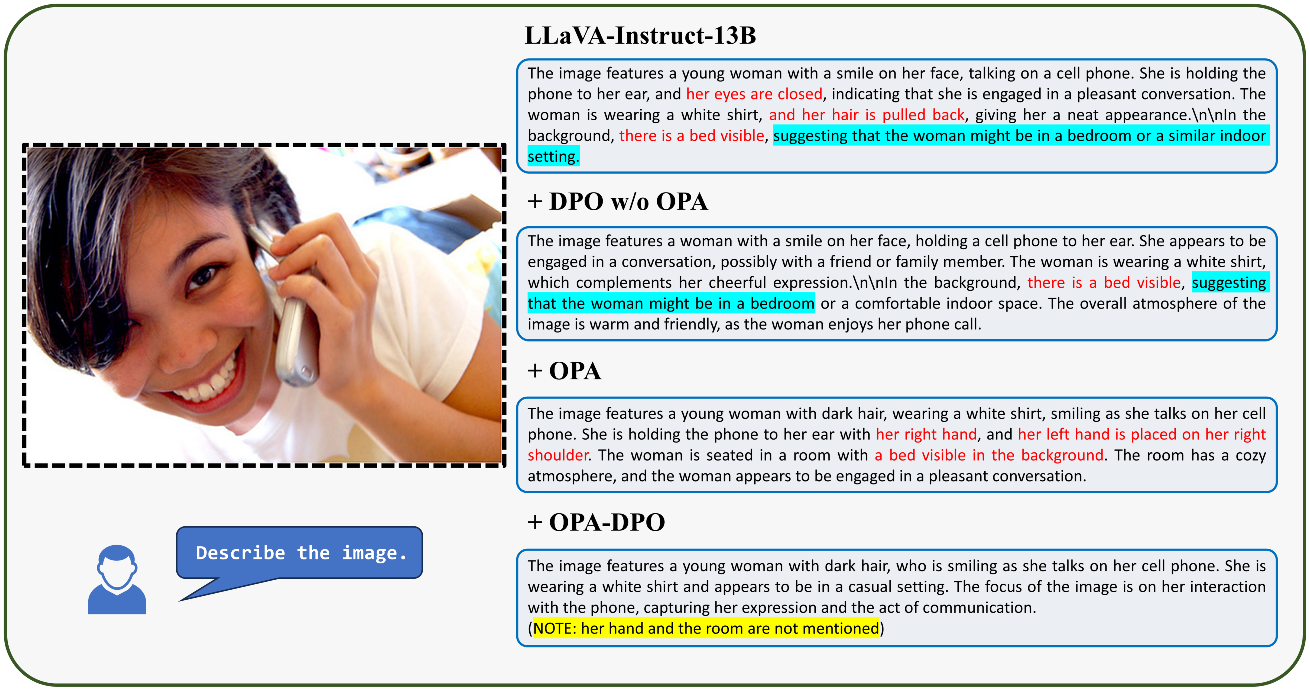
+ OPA

The image features a group of three people walking through a lush green field. They are walking in a line, with one person slightly ahead of the others. The field is surrounded by trees, creating a serene and natural atmosphere. The sky above is clear and blue, adding to the beauty of the scene. The individuals appear to be enjoying their time outdoors, possibly taking a leisurely stroll or engaging in some physical activity.

+ OPA-DPO

The image captures a serene scene of three people walking through a lush green field. The individuals are spread out, with one person slightly ahead of the others, creating a sense of movement and exploration. The field is dotted with wildflowers, adding a touch of natural beauty to the landscape. In the background, the sky is a clear blue, suggesting a sunny day. The overall atmosphere is one of tranquility and outdoor adventure. (NOTE: trees are not mentioned)

Figure 8. Qualitative results of different models. Hallucinated parts are marked in red, and missing details are highlighted in yellow.



Describe the image.

LLaVA-Instruct-13B

The image features a young woman with a smile on her face, talking on a cell phone. She is holding the phone to her ear, and her eyes are closed, indicating that she is engaged in a pleasant conversation. The woman is wearing a white shirt, and her hair is pulled back, giving her a neat appearance.\n\nIn the background, there is a bed visible, suggesting that the woman might be in a bedroom or a similar indoor setting.

+ DPO w/o OPA

The image features a woman with a smile on her face, holding a cell phone to her ear. She appears to be engaged in a conversation, possibly with a friend or family member. The woman is wearing a white shirt, which complements her cheerful expression.\n\nIn the background, there is a bed visible, suggesting that the woman might be in a bedroom or a comfortable indoor space. The overall atmosphere of the image is warm and friendly, as the woman enjoys her phone call.

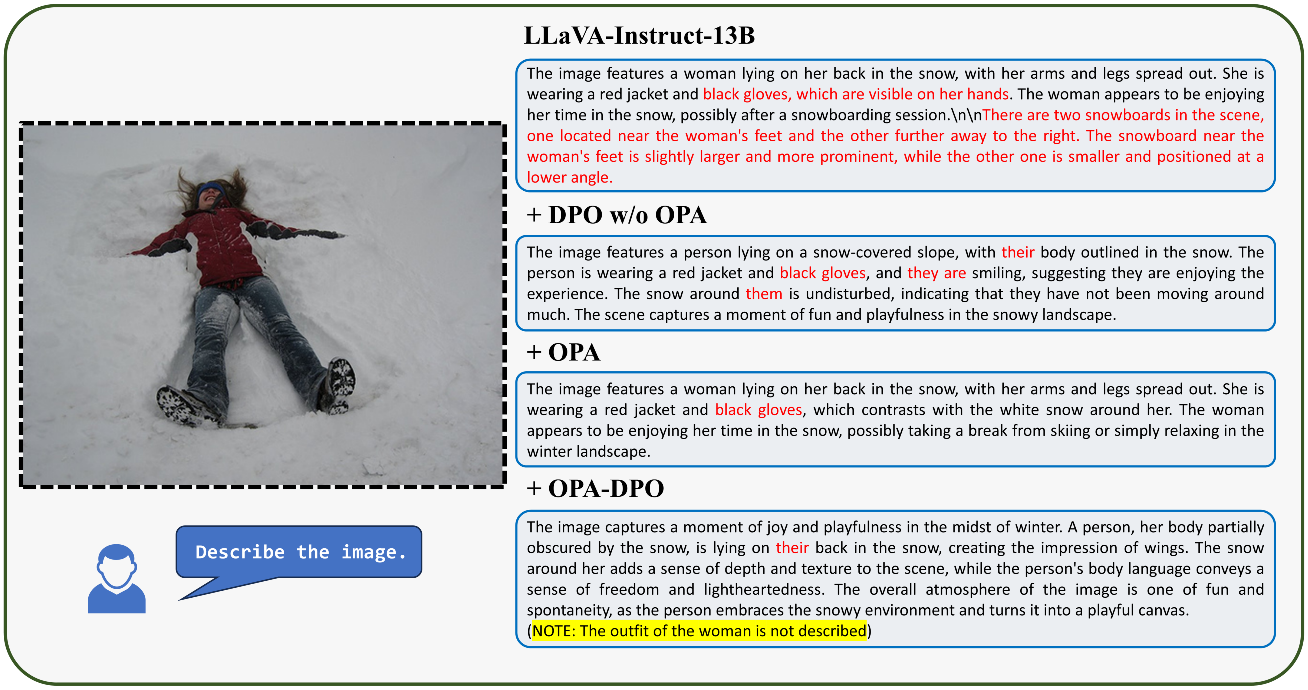
+ OPA

The image features a young woman with dark hair, wearing a white shirt, smiling as she talks on her cell phone. She is holding the phone to her ear with her right hand, and her left hand is placed on her right shoulder. The woman is seated in a room with a bed visible in the background. The room has a cozy atmosphere, and the woman appears to be engaged in a pleasant conversation.

+ OPA-DPO

The image features a young woman with dark hair, who is smiling as she talks on her cell phone. She is wearing a white shirt and appears to be in a casual setting. The focus of the image is on her interaction with the phone, capturing her expression and the act of communication.
(NOTE: her hand and the room are not mentioned)

Figure 9. Qualitative results of different models. Hallucinated parts are marked in red, flawed reasoning is highlighted in blue, and missing details are highlighted in yellow.



LLaVA-Instruct-13B

The image features a woman lying on her back in the snow, with her arms and legs spread out. She is wearing a red jacket and black gloves, which are visible on her hands. The woman appears to be enjoying her time in the snow, possibly after a snowboarding session.\n\nThere are two snowboards in the scene, one located near the woman's feet and the other further away to the right. The snowboard near the woman's feet is slightly larger and more prominent, while the other one is smaller and positioned at a lower angle.

+ DPO w/o OPA

The image features a person lying on a snow-covered slope, with their body outlined in the snow. The person is wearing a red jacket and black gloves, and they are smiling, suggesting they are enjoying the experience. The snow around them is undisturbed, indicating that they have not been moving around much. The scene captures a moment of fun and playfulness in the snowy landscape.

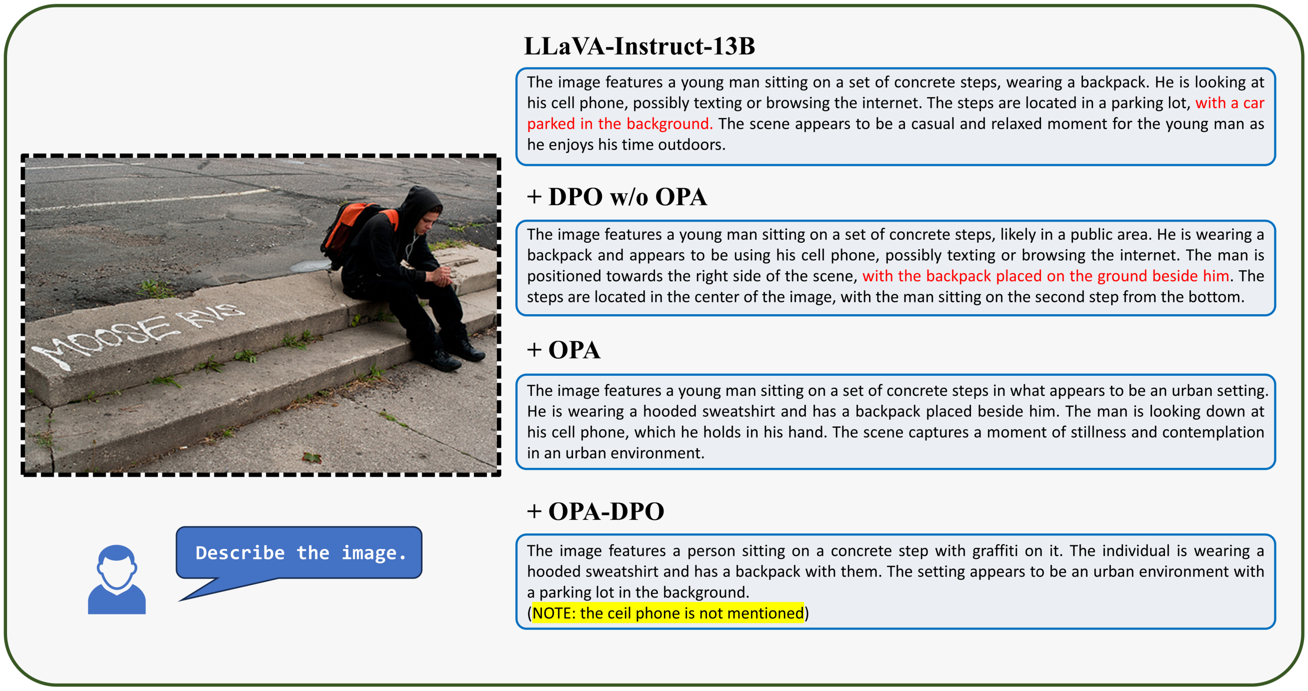
+ OPA

The image features a woman lying on her back in the snow, with her arms and legs spread out. She is wearing a red jacket and black gloves, which contrasts with the white snow around her. The woman appears to be enjoying her time in the snow, possibly taking a break from skiing or simply relaxing in the winter landscape.

+ OPA-DPO

The image captures a moment of joy and playfulness in the midst of winter. A person, her body partially obscured by the snow, is lying on their back in the snow, creating the impression of wings. The snow around her adds a sense of depth and texture to the scene, while the person's body language conveys a sense of freedom and lightheartedness. The overall atmosphere of the image is one of fun and spontaneity, as the person embraces the snowy environment and turns it into a playful canvas.
(NOTE: The outfit of the woman is not described)

Figure 10. Qualitative results of different models. Hallucinated parts are marked in red, and missing details are highlighted in yellow.



LLaVA-Instruct-13B

The image features a young man sitting on a set of concrete steps, wearing a backpack. He is looking at his cell phone, possibly texting or browsing the internet. The steps are located in a parking lot, with a car parked in the background. The scene appears to be a casual and relaxed moment for the young man as he enjoys his time outdoors.

+ DPO w/o OPA

The image features a young man sitting on a set of concrete steps, likely in a public area. He is wearing a backpack and appears to be using his cell phone, possibly texting or browsing the internet. The man is positioned towards the right side of the scene, with the backpack placed on the ground beside him. The steps are located in the center of the image, with the man sitting on the second step from the bottom.

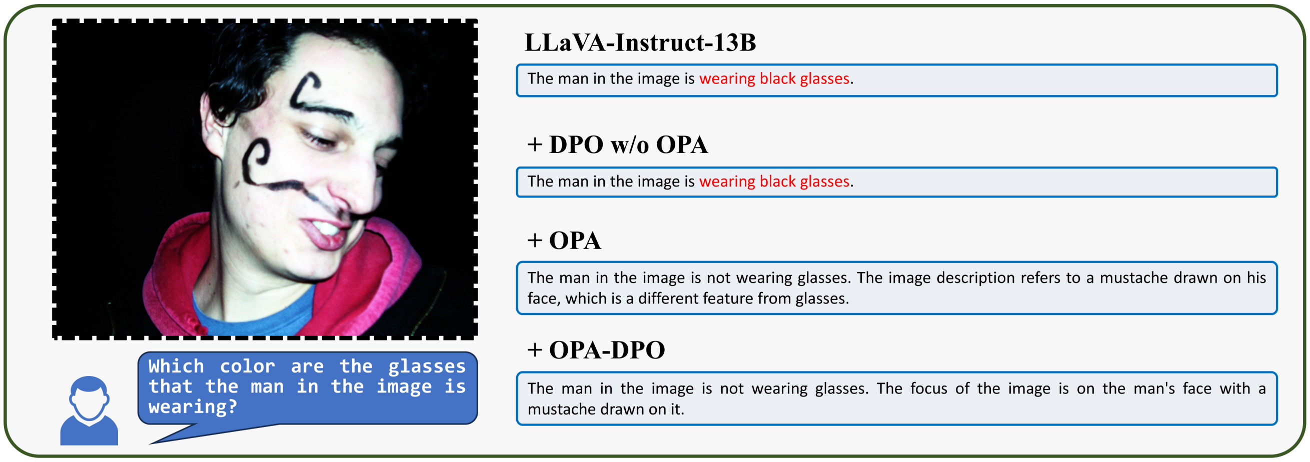
+ OPA

The image features a young man sitting on a set of concrete steps in what appears to be an urban setting. He is wearing a hooded sweatshirt and has a backpack placed beside him. The man is looking down at his cell phone, which he holds in his hand. The scene captures a moment of stillness and contemplation in an urban environment.

+ OPA-DPO

The image features a person sitting on a concrete step with graffiti on it. The individual is wearing a hooded sweatshirt and has a backpack with them. The setting appears to be an urban environment with a parking lot in the background.
(NOTE: the ceil phone is not mentioned)

Figure 11. Qualitative results of different models. Hallucinated parts are marked in red, and missing details are highlighted in yellow.



Which color are the glasses that the man in the image is wearing?

LLaVA-Instruct-13B

The man in the image is wearing black glasses.

+ DPO w/o OPA

The man in the image is wearing black glasses.

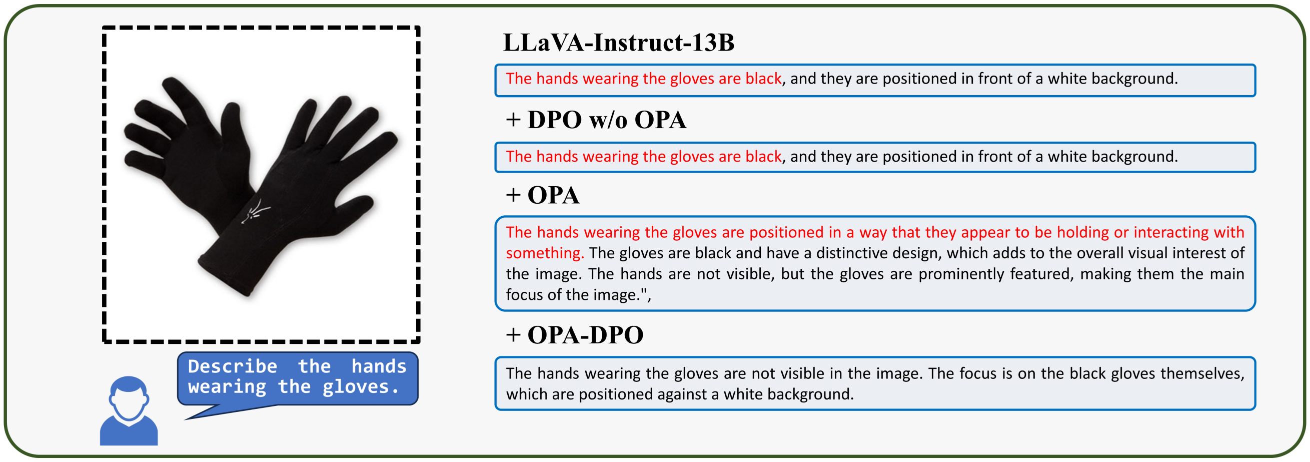
+ OPA

The man in the image is not wearing glasses. The image description refers to a mustache drawn on his face, which is a different feature from glasses.

+ OPA-DPO

The man in the image is not wearing glasses. The focus of the image is on the man's face with a mustache drawn on it.

Figure 12. Qualitative results of different models. Hallucinated parts are marked in red.



Describe the hands wearing the gloves.

LLaVA-Instruct-13B

The hands wearing the gloves are black, and they are positioned in front of a white background.

+ DPO w/o OPA

The hands wearing the gloves are black, and they are positioned in front of a white background.

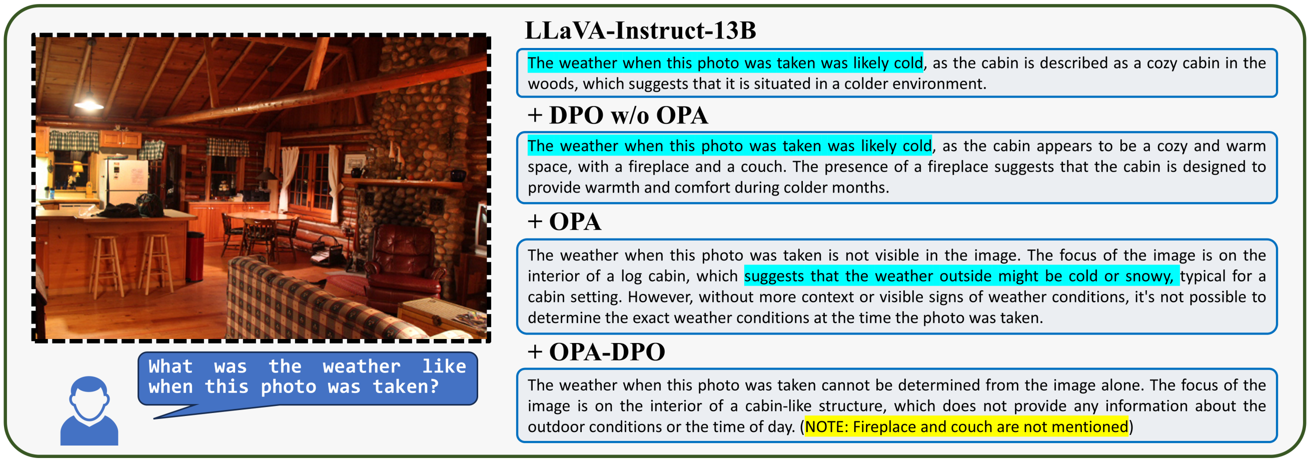
+ OPA

The hands wearing the gloves are positioned in a way that they appear to be holding or interacting with something. The gloves are black and have a distinctive design, which adds to the overall visual interest of the image. The hands are not visible, but the gloves are prominently featured, making them the main focus of the image.",

+ OPA-DPO

The hands wearing the gloves are not visible in the image. The focus is on the black gloves themselves, which are positioned against a white background.

Figure 13. Qualitative results of different models. Hallucinated parts are marked in red.



What was the weather like when this photo was taken?

LLaVA-Instruct-13B

The weather when this photo was taken was likely cold, as the cabin is described as a cozy cabin in the woods, which suggests that it is situated in a colder environment.

+ DPO w/o OPA

The weather when this photo was taken was likely cold, as the cabin appears to be a cozy and warm space, with a fireplace and a couch. The presence of a fireplace suggests that the cabin is designed to provide warmth and comfort during colder months.

+ OPA

The weather when this photo was taken is not visible in the image. The focus of the image is on the interior of a log cabin, which suggests that the weather outside might be cold or snowy, typical for a cabin setting. However, without more context or visible signs of weather conditions, it's not possible to determine the exact weather conditions at the time the photo was taken.

+ OPA-DPO

The weather when this photo was taken cannot be determined from the image alone. The focus of the image is on the interior of a cabin-like structure, which does not provide any information about the outdoor conditions or the time of day. (NOTE: Fireplace and couch are not mentioned)

Figure 14. Qualitative results of different models. Flawed reasoning is highlighted in blue, and missing details are highlighted in yellow.