

A Evaluation

A.1 Hallucination Reduction Index

A.1.1 Metric Design

HRI represents an aggregate improvement metric across five different benchmarks. Simply summing the raw scores from each benchmark would not be a reasonable or rigorous approach, as the metrics are not directly comparable. Therefore, we calculate the improvement ratio for each benchmark based on its potential improvement range, effectively converting the raw metric gains into an additive proportion of improvement. Furthermore, we employ a conservative aggregation method to avoid overestimating the effectiveness of our approach.

Let $a_i, i \in \{1, 2, 3, 4, 5\}$ denotes $F1_{\text{AMB-gen}}, \text{Score}_{\text{MMHal}}, F1_{\text{ObjectHal}}, LV_{\text{score}}, F1_{\text{AMB-dis}}$ respectively, namely the results on each benchmark, superscript “base” represents performances of the baseline model and “ref” represents the set reference performances. Then HRI is calculated as:

$$\text{HRI} = 2 \times \sum_{i=1}^5 \frac{a_i - a_i^{\text{base}}}{a_i^{\text{ref}} - a_i^{\text{base}}} \quad (1)$$

A.1.2 Main Results

For 7B model, we set the reference performances as **OVIP**_{2ep}, so it comes:

$$\text{HRI} = 2 \times \left(\frac{a_1 - 65.01}{67.12 - 65.01} + \frac{a_2 - 1.90}{2.65 - 1.90} + \frac{a_3 - 72.40}{74.18 - 72.40} + \frac{a_4 - 57.20}{60.90 - 57.20} + \frac{a_5 - 85.5}{87.4 - 85.5} \right)$$

For 13B model, we also use **OVIP**_{2ep} as the reference performances except for the ObjectHal benchmark which almost all methods fail to improve. We set the reference performance of ObjectHal to 79.0.

$$\text{HRI} = 2 \times \left(\frac{a_1 - 65.99}{68.98 - 65.99} + \frac{a_2 - 2.24}{2.57 - 2.24} + \frac{a_3 - 76.73}{79.00 - 76.73} + \frac{a_4 - 62.60}{67.90 - 62.60} + \frac{a_5 - 89.1}{90.2 - 89.1} \right)$$

A.1.3 Ablation Study: Loss Functions

There is no method surpassing other methods significantly, so we consider the best performance on the benchmark as its reference peerformance.

$$\text{HRI} = 2 \times \left(\frac{a_1 - 65.01}{68.57 - 65.01} + \frac{a_2 - 1.90}{2.70 - 1.90} + \frac{a_3 - 72.40}{74.14 - 72.40} + \frac{a_4 - 57.20}{64.10 - 57.20} + \frac{a_5 - 85.5}{87.20 - 85.5} \right)$$

A.1.4 Ablation Study: Online and Offline

Same as Main Results.

$$\text{HRI} = 2 \times \left(\frac{a_1 - 65.01}{67.12 - 65.01} + \frac{a_2 - 1.90}{2.65 - 1.90} + \frac{a_3 - 72.40}{74.18 - 72.40} + \frac{a_4 - 57.20}{60.90 - 57.20} + \frac{a_5 - 85.5}{87.4 - 85.5} \right)$$

A.1.5 Further Study

Same as Main Results.

$$\text{HRI} = 2 \times \left(\frac{a_1 - 65.01}{67.12 - 65.01} + \frac{a_2 - 1.90}{2.65 - 1.90} + \frac{a_3 - 72.40}{74.18 - 72.40} + \frac{a_4 - 57.20}{60.90 - 57.20} + \frac{a_5 - 85.5}{87.4 - 85.5} \right)$$

A.2 Bad Cases

A.2.1 MMHal

Shown in Figure 1, the original evaluation protocol utilizes the text-only gpt-4-turbo-2024-04-09 to evaluate the model response, which has no access to the input image and can only infer from the given image contents and ground truth, so it will lead to many incorrect judgments. We replace it with gpt-4o-2024-05-13, which yields more accurate assessments.


	Question: How many people are there in the image? Model response: There are four people in the image: a man, a woman, and two children.
	Ground truth: There are four people in the image.
	gpt-4-turbo-2024-04-09: The LMM's response claims that there are four people in the image, which matches the standard human-generated answer. <i>However, the LMM further specifies the composition of these four people as "a man, a woman, and two children." This additional detail is not supported by the image contents provided,</i> which only mention a "Girl" and does not specify any other individuals or their roles/ages. Therefore, the LMM's response includes information that is not present or implied in the image, leading to a hallucination. Explanation: The response is informative in terms of the number of people, but it hallucinates details about their specific identities (man, woman, two children) which are not mentioned in the image contents. Rating: 1, somewhat informative, with hallucination.
	gpt-4o-2024-05-13: The LMM's response correctly identifies the number of people in the image, which is four. It also provides additional information by specifying the likely roles of the individuals (a man, a woman, and two children). This additional detail is accurate and does not introduce any false claims or hallucinations. The response is informative and provides a good analysis without any hallucination. Rating: 6, very informative with good analysis or reasoning, no hallucination

Figure 1: Text-only LLM can not correctly judge the response.



	Model description The image features a dog running on a sandy beach near the water. The dog is positioned towards the center of the scene, with its legs spread out as it moves across the sand. The beach appears to be relatively empty, <i>with no other people or animals visible in the image.</i> The sky above the beach is cloudy, adding a sense of depth to the scene.
	Annotated correct & wrong words Truth: sky, cloud, sea, beach, dog, mountain Hallucination: lounge, umbrella, tree, person, seagull
	Model description The image captures a woman running on a sandy beach. She is wearing a white top and appears to be enjoying her run. The woman's shadow is visible on the sand, <i>indicating that the sun is shining and casting light on the scene.</i> The beach setting provides a serene and picturesque backdrop for her activity.
	Annotated correct & wrong words Truth: woman, beach Hallucination: sky, sun, sea, dog, lounge

Figure 2: Rule-based extraction will lead to misjudgments to some extent.


	OPA-DPO: The image features a stop sign prominently displayed against a backdrop of a blue wall. The stop sign is positioned on the left side of the frame, drawing attention to it as the main subject. The blue wall behind the sign adds a contrasting element to the scene, creating visual interest. The overall composition of the image suggests it might be capturing a moment in an urban or suburban setting where traffic regulations are important.
	OVIP: The image features a man walking down a street, passing by a stop sign. The stop sign is positioned on the left side of the scene, and the man is walking towards it. The street appears to be empty, with <i>no other people or vehicles visible in the image.</i> The man is wearing a striped shirt, and his pants are rolled up. The scene is set against a blue wall, which adds a vibrant touch to the otherwise quiet street. Failing to mention the main entity in the generated description is also a form of hallucination.

Figure 3: OPA-DPO fails to mention the man, a deficiency that is captured by the “Cover” metric but often overlooked in previous evaluations. “vehicles” is incorrectly identified as a hallucination word.

A.2.2 AMBER-generative & ObjectHal

AMBER uses an automatic method for detecting the hallucination entity, which primarily relies on the pre-defined hallucination words. ObjectHal introduces LLM to extract the mentioned entities, its metrics are basically the same with AMBER.

Figure 2 illustrates several cases of misjudgment in AMBER. Since the score is determined solely by the presence of specific predefined words rather than the actual semantic correctness, the hallucination rate (Chair score) is often overestimated. **Moreover, this issue becomes more pronounced as the diversity and informativeness of model responses increases.**

Many methods achieve great improvements in the Chair score (entity-wise hallucination rate), but often at the cost of a significant decrease in the cover rate (completeness and informativeness). Figure 3 provides an example of this information deficit phenomenon, which should also be considered in the evaluation of model performance.

B Experiments

B.1 Loss Functions

Base image loss $\mathcal{L}_{\text{Image}}^{\text{base}}$ is similar to DPO loss which replace the response pair with the image pair:

$$\mathcal{L}_{\text{Image}}^{\text{base}}(\mathcal{I}^+, \mathcal{I}^-; \mathcal{Q}, \mathcal{A}^+) = \log \sigma \left(\beta \cdot \left[\log \frac{\pi_{\theta}(\mathcal{A}^+ | \mathcal{I}^+, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^+ | \mathcal{I}^+, \mathcal{Q})} - \log \frac{\pi_{\theta}(\mathcal{A}^+ | \mathcal{I}^-, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^+ | \mathcal{I}^-, \mathcal{Q})} \right] \right)$$

Symmetrical image loss $\mathcal{L}_{\text{Image-Sym}}$ considers the negative image and the negative response a correct pair, then calculate Image loss using negative response and image as the positive one:

$$\begin{aligned} \mathcal{L}_{\text{Image-Sym}}(\mathcal{I}^+, \mathcal{I}^-, \mathcal{A}^+, \mathcal{A}^-; \mathcal{Q}) &= \mathcal{L}_{\text{Image}}(\mathcal{I}^+, \mathcal{I}^-; \mathcal{Q}, \mathcal{A}^+) + \mathcal{L}_{\text{Image}}(\mathcal{I}^-, \mathcal{I}^+; \mathcal{Q}, \mathcal{A}^-) \\ &= -\log \sigma \left(\beta_1 \cdot \left[\log \frac{\pi_{\theta}(\mathcal{A}^+ | \mathcal{I}^+, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^+ | \mathcal{I}^+, \mathcal{Q})} - \log \frac{\pi_{\theta}(\mathcal{A}^+ | \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^+ | \mathcal{Q})} \right] \right. \\ &\quad \left. + \beta_2 \cdot \left[\log \frac{\pi_{\theta}(\mathcal{A}^+ | \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^+ | \mathcal{Q})} - \log \frac{\pi_{\theta}(\mathcal{A}^+ | \mathcal{I}^-, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^+ | \mathcal{I}^-, \mathcal{Q})} \right] \right) \\ &\quad -\log \sigma \left(\beta_1 \cdot \left[\log \frac{\pi_{\theta}(\mathcal{A}^- | \mathcal{I}^-, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^- | \mathcal{I}^-, \mathcal{Q})} - \log \frac{\pi_{\theta}(\mathcal{A}^- | \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^- | \mathcal{Q})} \right] \right. \\ &\quad \left. + \beta_2 \cdot \left[\log \frac{\pi_{\theta}(\mathcal{A}^- | \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^- | \mathcal{Q})} - \log \frac{\pi_{\theta}(\mathcal{A}^- | \mathcal{I}^+, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^- | \mathcal{I}^+, \mathcal{Q})} \right] \right) \end{aligned}$$

Anchor loss $\mathcal{L}_{\text{Anchor}}$ directly enforces the probability of positive response to be higher for intuitively better optimization results.

$$\mathcal{L}_{\text{Anchor}}(\mathcal{A}^+, \mathcal{A}^-; \mathcal{Q}, \mathcal{I}^+) = -\log \sigma \left(\beta \cdot \log \frac{\pi_{\theta}(\mathcal{A}^+ | \mathcal{I}^+, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^+ | \mathcal{I}^+, \mathcal{Q})} \right)$$

Bi-directional anchor loss $\mathcal{L}_{\text{Bi-Anchor}}$ not only exerts supervision on the positive response, but it also makes the negative response probability to be lower.

$$\mathcal{L}_{\text{Bi-Anchor}}(\mathcal{A}^+, \mathcal{A}^-; \mathcal{Q}, \mathcal{I}^+) = -\log \sigma \left(\beta \cdot \log \frac{\pi_{\theta}(\mathcal{A}^+ | \mathcal{I}^+, \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^+ | \mathcal{I}^+, \mathcal{Q})} \right) + \log \sigma \left(\beta \cdot \log \frac{\pi_{\theta}(\mathcal{A}^- | \mathcal{Q})}{\pi_{\text{ref}}(\mathcal{A}^- | \mathcal{Q})} \right)$$

B.2 Settings

By default, we use the following settings:

Software infrastructure. In our implementation, we deploy the non-training LLM and diffusion models as services using FastAPI. During training, the system interacts with these services via API calls to obtain feedback, image prompts, and the paths to generated images.

Table 1: OViP pseudocode

Algorithm 1 Algorithm of OViP

Input: training dataset $\mathcal{D} = \{(\mathcal{I}^+, \mathcal{Q}, \mathcal{A}^*)\}$;
target model π ; reward model G_r ; prompt generator G_{diff} ; diffusion model diff

Initialize: experience buffer $\mathcal{B} \leftarrow \emptyset$

Output: optimized model π

for each $(\mathcal{I}^+, \mathcal{Q}, \mathcal{A}^*) \in \mathcal{D}$ **do**

 Sample candidate responses $\{\mathcal{A}^i\}_{i=1}^k \sim \pi(\cdot | \mathcal{I}^+, \mathcal{Q})$

 Compute reward scores: $r^i = G_r(\mathcal{A}^i, \mathcal{A}^*)$

 Compute standard deviation σ_r of $\{r^i\}$

 Initialize temporary pair list $\mathcal{T} \leftarrow \emptyset$

while $\exists (\mathcal{A}^+, \mathcal{A}^-)$ satisfying:

$|r^+ - r^-| > \max(\delta, 2\sigma_r), r^+ > \tau_{\text{pos}}, r^- < \tau_{\text{neg}}$ **do**

 Add $(\mathcal{A}^+, \mathcal{A}^-)$ to \mathcal{T} and remove from candidate pool

end while

if $\mathcal{T} = \emptyset$ and $\min_i r^i < \tau_{\text{neg}}$ **then**

 Let \mathcal{A}^- be the lowest-scoring response

 Add $(\mathcal{A}^*, \mathcal{A}^-)$ to \mathcal{T}

endif

for each $(\mathcal{A}^+, \mathcal{A}^-) \in \mathcal{T}$ **do**

 Generate prompt: $\mathcal{T}^- = G_{\text{diff}}(\mathcal{A}^+, \mathcal{A}^-)$

 Synthesize image: $\mathcal{I}^- = \text{diff}(\mathcal{T}^-)$

 Add $(\mathcal{I}^+, \mathcal{I}^-, \mathcal{Q}, \mathcal{A}^+, \mathcal{A}^-)$ to buffer \mathcal{B}

end for

if $|\mathcal{B}| \geq N$ **then**

 Sample N samples from \mathcal{B} for training

 Compute total loss: $\mathcal{L}_{\text{OViP}}$

 Update $\pi \leftarrow \pi - \eta \nabla_{\pi} \mathcal{L}_{\text{OViP}}$

endif

end for

Models. The LLM we use for judging response and providing image-generation prompt is Qwen-2.5-7b-instruct (<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>). The diffusion model for image generation is FLUX.1-dev (<https://huggingface.co/black-forest-labs/FLUX.1-dev>).

Sampling and Filter. The score is between 0 and 10, which 10 means a perfect response and 0 means a totally incorrect response. We sample 16 responses for one query and set the lower-bound margin δ to 3. Moreover, the quality criterion coefficients $\tau_{\text{pos}} = \tau_{\text{neg}} = 5$, which means the score of positive response should be at least 6 and negative response be at most 4. The **temperature** of the LLM scorer is 0.1.

Image Generation. For image prompt generation, we set the model’s **temperature** as 0.1, **top_p** as 0.9, and **max_new_tokens** as 128. We generate a 384×384 image given the prompt with **num_inference_steps**=40 and **guidance_scale**=7.5.

Training. We list training setups in Section ??.

We perform ablation and further study using LLaVA-1.5-7B. The following describes the relevant experimental settings.

B.2.1 Ablation on Loss Functions

We fine-tune the model for one epoch using data generated by the model itself immediately before training, following the OViP data construction pipeline.

For *iterative training*, we first fine-tune the base model on the original dataset using DPO to obtain a stronger initialization. We then sample and filter 4,730 instances as the second-stage contrastive dataset, which remains fixed across all variants. To improve supervision quality, model responses are annotated using DeepSeek-V3 for more accurate reward estimation.

B.2.2 Ablation on Online Learning

Although online methods can continuously improve when trained with another epoch, we conduct the experiment with one epoch for both online and offline methods.

B.2.3 Further Study

We save several checkpoints during training and evaluate them to gain a complete understanding of the training process.

We select 227 instances from the original OPA-DPO dataset—distinct from the training set—for analyzing changes in token-wise log-probability and output quality distributions.

For token-wise log-probability analysis, we treat the original negative samples from the dataset as out-of-distribution (OOD) responses and compute their log-probabilities. For in-distribution (IND) responses, we perform 16 response samplings per query using our model with a temperature of 0.2, and compute the average log-probability across these samples.

To examine output distribution shifts, we sample 16 responses per query with a higher temperature of 1.2, assign scores to each response, and analyze the distribution of these scores.

For comparison about different image generation strategy, we implement two other representative methods. 1. Random Cropping (R.Crop). We randomly crop 020% area from the image to form the negative image. This method is from mDPO. 2. Offline construction. We use an LLM to generate the image description with some inaccordance with the positive image, then we use the diffusion model to generate the image according to the description. The prompt is at Table 5.

C Algorithm

The pseudocode is at Table 1.

D Efficiency and Time Consuming

OViP training takes approximately 17 hours on 7× A800 (40G) GPUs. Among them, 4 GPUs are allocated for VLM training, 1 GPU for LLM deployment, and 2 GPUs for diffusion model deployment. We divide each training step into six stages: sampling (response generation), scoring (response evaluation), description (image prompt construction), negative image (counterfactual image generation), forward (model inference), and post-processing. Figure 4 illustrates the proportion of time spent on each stage, where post-processing refers to the period after forward propagation and before the next training step begins, including gradient accumulation, backpropagation, optimizer updates, and other related operations.

Excluding post-processing, the most time-consuming component is the sampling stage, similar to reinforcement learning. This is because it requires autoregressive generation of 16 responses, one token at a time. The second most expensive stage is negative image generation. To reduce latency, we parallelize this process by assigning two diffusion models to handle image generation requests from four sampling subprocesses.

Additionally, since the experience buffer is implemented independently in our system, repeated sampling by one subprocess may block others due to synchronization constraints. This can indirectly slow down the forward and post-processing stages as some processes await completion.

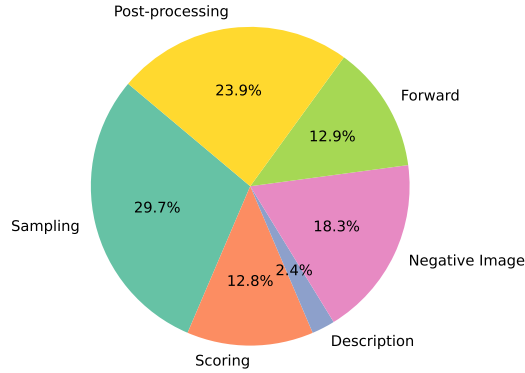


Figure 4: Time consumption for each stage during training.

E Limitations

This work introduces an online training framework that integrates dual contrastive learning across vision and language. While our loss function follows the DPO formulation, we do not explore how existing reinforcement learning algorithms—such as PPO or GRPO—could be effectively combined with image-level contrastive objectives. In terms of evaluation, although we identify and discuss several limitations of prior protocols and propose improved metrics and procedures, the current benchmarks still fall short of fully capturing model capability. We manually identified a subset of erroneous cases through inspection, but did not conduct a comprehensive correction. Lastly, our data filtering strategy during sampling has not been carefully tuned, and a more refined design could potentially lead to better training dynamics and model performance.

F Broader Impacts

This work focuses on improving the factual reliability of vision-language models by reducing hallucination. While it does not directly engage with societal applications, it contributes to the broader goal of building more trustworthy and robust AI systems. Although the method itself does not pose obvious risks, we note that even improved generation quality does not eliminate the possibility of misuse, such as producing misleading content. Responsible deployment and proper safeguards remain necessary when integrating such models into real-world applications.

G Prompts for Judgment and Negative Image Generation

Prompt for Quality Judgment

Task

Your role is as a discerning assistant tasked with evaluating model responses for multimodal tasks (though you have no access with the image). 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. Your objective is to meticulously and precisely assess the model-generated response (the former) based on the provided reference answer (the latter).

- Here's how you should approach the assessment process:

1. The quality of the response depends on its accuracy and the degree of adherence to the correct answer. Therefore, if the response is much more detailed than the reference answer, it should not be considered a very good response (although it may still be considered a good one).

2. Directly provide the score of the response, with a full score of 10. Your response should follow this format: "Score: [x]\n", where "[x]" represents the score you give, and "\n" is a line break.

3. Please do not provide additional reasoning, just give the score directly.

Question

{question}

Response

{response}

Correct answer

{answer}

Table 2: The prompt for judging model response given the answer.

Prompt for image generation-Part 1

Task

Given an unknown image-related question, a correct answer, and an inaccurate response, carefully analyze the differences between the response and the answer. Then, provide a brief description of the image so that it aligns with the correct answer and differs from the incorrect responses. In other words, infer the content of the image.

Example

[Example 1]

<Question>

What is on the sandwich?

<Answer>

The sandwich has tomatoes and lettuce on it.

<Response>

The sandwich has a slice of egg and tomato on it.

<Output Description>

A sandwich with only tomatoes and lettuce on it.

Explanation: The answer mentions lettuce and tomato, while the incorrect response mentions tomato and egg. So there is no egg on the sandwich.

Prompt for image generation-Part 2

[Example 2]

****<Question>****

Can you point out the details that make this image unique?

****<Answer>****

In the image, there is a plate with a slice of pizza topped with tomatoes, herbs, and cheese. The distinctive detail about the image is that the pizza is missing two slices, leaving just one slice remaining on the plate. This suggests that someone has already started enjoying the pizza. The slice appears to be well-cooked and freshly served, creating an appetizing and mouthwatering scene for the viewer.

****<Response>****

In the image, there is a slice of pizza on a plate with tomatoes and cheese. The pizza appears to be homemade and has been cut into two pieces. The tomatoes are sliced in half, revealing their juicy interior. The cheese on top of the pizza is melted, creating a delicious-looking dish. Additionally, there is a fork nearby, suggesting that someone might be planning to enjoy this pizza soon.

****<Output Description>****

A plate with a one-third remaining piece of pizza, topped with herbs, cheese, and tomatoes; someone has finished eating and left.

****Explanation****: The answer mentions that only one-third of the pizza remains and that someone has just finished eating and left, which is inconsistent with the response. Therefore, the image should include these two features.

[Example 3]

****<Question>****

Bird or cow?

****<Answer>****

Bird

****<Response>****

The bird in the image is a small, brown and white bird with a distinctive head shape and coloration. It is not a cow. The bird is perched on a branch, which is situated in front of a white building.

****<Output Description>****

A big, blue bird perched on a branch in front of a black building.

****Explanation****: Both the answer and the response mention the bird, but the response is more detailed. So the description should be contrastive to the features of the bird in the response.

Requirements

- The description should be brief but precise.
- If both the answer and the response are long, focus on describing the one or two most significant differences.
- Do not provide any analysis or explanation; only describe the image.
- A common approach is to describe what is present in the image and what is missing.

****<Question>****

{question}

****<Answer>****

{answer}

****<Response>****

{response}

****<Output Description>****

Table 3: The prompt for image generation instruction.

Prompt for image distortion-Part 1

Task

Given an unknown image-related question, a correct answer, and an inaccurate response, carefully analyze the differences between the response and the answer. Then, provide a brief description of the image so that it aligns with the correct answer and differs from the incorrect responses. In other words, infer the content of the image.

Example

[Example 1]

****<Question>****

What is on the sandwich?

****<Answer>****

The sandwich has tomatoes and lettuce on it.

****<Response>****

The sandwich has a slice of egg and tomato on it.

****<Output Description>****

A sandwich with only tomatoes and lettuce on it.

****Explanation****: The answer mentions lettuce and tomato, while the incorrect response mentions tomato and egg. So there is no egg on the sandwich.

[Example 2]

****<Question>****

Can you point out the details that make this image unique?

****<Answer>****

In the image, there is a plate with a slice of pizza topped with tomatoes, herbs, and cheese. The distinctive detail about the image is that the pizza is missing two slices, leaving just one slice remaining on the plate. This suggests that someone has already started enjoying the pizza. The slice appears to be well-cooked and freshly served, creating an appetizing and mouthwatering scene for the viewer.

****<Response>****

In the image, there is a slice of pizza on a plate with tomatoes and cheese. The pizza appears to be homemade and has been cut into two pieces. The tomatoes are sliced in half, revealing their juicy interior. The cheese on top of the pizza is melted, creating a delicious-looking dish. Additionally, there is a fork nearby, suggesting that someone might be planning to enjoy this pizza soon.

****<Output Description>****

A plate with a one-third remaining piece of pizza, topped with herbs, cheese, and tomatoes; someone has finished eating and left.

****Explanation****: The answer mentions that only one-third of the pizza remains and that someone has just finished eating and left, which is inconsistent with the response. Therefore, the image should include these two features.

Table 4: The prompt for distorted image generation instruction.

Prompt for image distortion-Part 2
<p>[Example 3]</p> <p>**<Question>** Bird or cow?</p> <p>**<Answer>** Bird</p> <p>**<Response>** The bird in the image is a small, brown and white bird with a distinctive head shape and coloration. It is not a cow. The bird is perched on a branch, which is situated in front of a white building.</p> <p>**<Output Description>** A big, blue bird perched on a branch in front of a black building.</p> <p>**Explanation**: Both the answer and the response mention the bird, but the response is more detailed. So the description should be contrastive to the features of the bird in the response.</p> <p># Requirements</p> <ul style="list-style-type: none"> - The description should be brief but precise. - If both the answer and the response are long, focus on describing the one or two most significant differences. - Do not provide any analysis or explanation; only describe the image. - A common approach is to describe what is present in the image and what is missing. <p>**<Question>** question</p> <p>**<Answer>** answer</p> <p>**<Response>** response</p> <p>**<Output Description>**</p>

Table 5: The prompt for distorted image generation instruction.