Supplementary Materials of Instruction-Oriented Preference Alignment for Enhancing Multi-Modal Comprehension Capability of MLLMs

1. More Implementation Details

1.1. Training Configuration

For model optimization, we leverage Direct Preference Optimization (DPO) with Low-Rank Adaptation (LoRA). Specifically, we use the AdamW optimizer with a cosine learning rate scheduler, starting from the initial learning rate of 5e-6. The total batch size is set to 8 and the gradient accumulation step is 8, resulting in an effective batch size of 64. All the models are trained for 1 epoch in the comparisons.

1.2. Prompt Design

We detail the prompts used in the automated preference construction. Specifically, these comprises the prompt I^{fb} for feedback generation (Equation 3), the prompt I^{rev} for response refinement (Equation 4), and the prompt I^{ver} for instruction-oriented verification (Equation 5). The prompts are displayed in Table 2, 3 and 4 respectively.

2. Qualitative Analysis

As illustrated in Figure 1, we analyze how each stage of our automated preference construction contributes to the construction of high-quality preference pairs. The revision stage improves the sampled responses by correcting factual errors and expanding the depth of the responses. Through critical feedback generation and refinement, potential flaws have the opportunity to be corrected and the responses are enriched with more valuable information. The verification stage then quantifies instruction fulfillment capacity of the responses. Responses which pass the verification are identified as preferred due to their stronger intrinsic alignment with instruction objectives.

As demonstrated in Figure 2, we present comparative examples to illustrate the improvements achieved by our approach over the baseline model. The visual comparisons reveal that IPA-aligned responses exhibit enhanced capability in addressing the intrinsic demands of instructions. These examples validate that our approach enables the model to better interpret and fulfill the requirements embedded in diverse instructions, moving toward more comprehensive understanding capability.

3. More Experiments

3.1. Experiments with More Advanced Models

We also conduct experiments using larger and more advanced models, InternVL3-14B and Qwen2.5VL-72B, where the same dataset as in the main manuscript is used for training these models. As shown in Table 1, despite the preference dataset being constructed with a relatively weaker and smaller model (Qwen2VL-7B), it brings consistent gains to significantly larger models, demonstrating strong generalizability across model scales and architectures. While the improvements are more moderate due to the already strong baselines, our method still yields meaningful enhancements in overall comprehension ability. We believe that using stronger models for data collection can yield further improvements.

Table 1. Experiments with more advanced models. The training dataset is the same as in the main manuscript.

Model	**Hallucination**			**General VQA**					**Text**
	HallBench$_{avg}$	POPE	MMHal	MMMU$_{val}$	MMStar	MMVet$_{turbo}$	MME$_P$	LLaVABench	OCRBench
InternVL3-14B†	55.7	89.5	3.8	63.1	68.8	73.0	1740.5	83.0	87.6
+ IPA (ours)	56.0 +0.3	89.6 +0.1	3.9 +0.1	63.3 +0.2	69.1 +0.3	75.3 +2.3	1742.3 +1.8	89.9 +6.9	88.0+0.4
Qwen2.5VL-72B†	55.3	86.4	4.3	68.1	71.6	74.4	1721.3	97.6	88.8
+ IPA (ours)	58.4 +3.1	86.9 +0.5	4.4 +0.1	69.0 +0.9	71.5 -0.1	74.6 +0.2	1730.8 +9.5	102.2 +4.6	89.2 +0.4

Table 2. Prompt I^{fb} for feedback generation.

You are a evaluation model tasked with providing a detailed critical analysis of the given instruction-response pair.
Please evaluate the response based on the criteria below:

Accuracy of Image Content Description
- Assess whether all descriptions of the image content in the response are accurate.
- Identify any instances of hallucinations, where the response describes elements inconsistent with the image.
- If any inaccuracies or hallucinations are found, provide a corrected description.

Correctness of Knowledge
- Verify the factual accuracy of any commonsense statements or knowledge presented in the response.
- If any incorrect knowledge or misconceptions are identified, provide the correct information.

Validity of Reasoning and Inferences
- Evaluate the logical consistency of the reasoning processes and inferences made in the response.
- Ensure that conclusions are appropriately derived from correct premises.
- If any flawed reasoning or incorrect inferences are detected, provide a revised reasoning or inference.

Verifiability and Expression of Uncertainty
- Determine if the response includes statements that are difficult to verify or falsify.
- Check whether the response clearly indicates uncertainty for such statements when appropriate.
- If unverifiable statements are present without indicated uncertainty, suggest how to express the uncertainty clearly.

Adherence to Instructions
- Note any deviations from the instructions or omissions of important requirements.
- If the response deviates from the instructions or omits key elements, provide a corrected version that fully adheres to the instructions.

You should offer feedback detailing all identified issues, followed by a comprehensive evaluation and actionable steps for improvement with direct corrections.

Output Format:
Accuracy of Image Content Description
...

Correctness of Knowledge
...

Validity of Reasoning and Inferences
...

Verifiability and Expression of Uncertainty
...

Adherence to Instructions
...

Overall Assessment
...

Guidance for Improvement
...

Input:
<Instruction>{`instruction`}</Instruction>

<Response>{`response`}</Response>

Table 3. Prompt I^{rev} for response refinement.

Given an instruction-response pair related to an image, an assistant has provided feedback identifying issues in the response. This feedback enables the generation of an improved response.

So your task is to derive a revised response by:
Response + Feedback $\rightarrow$ Revised Response

Output Format:
Revised Response
...

Input:
<Instruction>{`instruction`}</Instruction>

<Response>{`response`}</Response>

<Feedback>{`feedback`}</Feedback>

Table 4. Prompt I^{ver} for instruction-oriented verification.

You are an evaluation model. Given an instruction-response pair related to an image, your task is to make evaluation by assessing the alignment between the response and a provided reference response.

You can assess the alignment by following the criteria below:
- Can the reference answer be inferred from the response, i.e., Response $\implies$ Reference?
- Are there critical details from the reference missing in the response?
- Are there contradictions between the response and the reference?
- Does the response adhere to the instruction?

Note:
- It is permissible for the response to contain additional information if it is related to the instruction and not contradictory to the reference.
- If this is a multiple-choice/yes-no question, does the response clearly provide the answer?

You should indicate whether the response pass the evaluation ("Yes" or "No") concerning all the criteria.

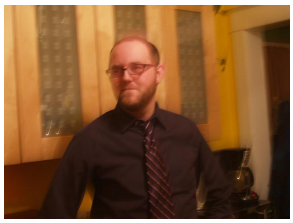
Output Format:
Assessment
...

Verification
(Yes or No)

Input:
<Instruction>{`instruction`}</Instruction>

<Response>{`response`}</Response>

<Reference>{`reference response`}</Reference>



Instruction:
Is this man posing for this photo?

Reference Response:
Yes.

Preferred

Sampling:
It's difficult to determine with certainty, but the man appears to be posing for the photo.

Feedback:
The response is mostly accurate. However, it should explicitly state the answer 'yes' and reference details from the image to support this conclusion.

Refinement:
The man is posing for the photo. He is standing in a room with a neutral expression, and his posture suggests that he is aware of the camera. He is wearing a dark shirt and a striped tie.

Verification:
Pass

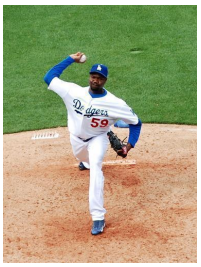
Dispreferred

Sampling:
The man is not posing for the photo; he appears to be in motion, possibly walking or moving, which results in a blurred image.

Feedback:
The response should directly address the instruction by answering the question "Is this man posing for this photo?" with a clear "No" and provide the reasoning that the man was in motion when the photo was taken, resulting in the blurred effect.

Refinement:
No, the man is not posing for this photo. He appears to be in motion, possibly walking or moving, which results in a blurred image. The photo captures a candid moment rather than a posed one.

Verification:
Fail



Instruction:
How many different colors are on the man's uniform?

Reference Response:
3.

Preferred

Sampling:
The man's uniform has three different colors: white, blue, and red.

Feedback:
The response is accurate, correct, valid, and adheres to the instructions. More details could be added to provide a clearer and more comprehensive description.

Refinement:
The man's uniform has three different colors: white, blue, and red. Specifically, the uniform includes white pants, blue sleeves, and a blue cap with a red logo. The red color is also present in the number 59 on the jersey. Therefore, there are three different colors on the man's uniform.

Verification:
Pass

Dispreferred

Sampling:
The man's uniform consists of the following colors: blue, white, red, black. There are 4 different colors on the man's uniform.

Feedback:
The response is accurate, correct, valid, and adheres to the instructions. There are no issues to address.

Refinement:
The man's uniform consists of the following colors: blue, white, red, black. There are 4 different colors on the man's uniform.

Verification:
Fail

Instruction:
Do the pillow and the bed have the same color?

Reference Response:
No, the bed is blue and the pillow is white.

Preferred Response

Sampling :
Yes, the pillow and the bed are the same color. Both are white.

Feedback:
The response contains inaccuracies in describing the colors of the pillow and the bed. It is important to carefully observe and accurately report the details in the image.
To improve, carefully observe the image and accurately describe the colors of the pillow and the bed. Ensure that the description matches the visual details present in the image.

Refinement:
No, the pillow and the bed do not have the same color. The bed has a white base with a blue runner, while the pillow is white with a blue stripe.

Verification:
Pass

Dispreferred Response

Sampling :
Yes, the pillow and the bed are the same color. Both are white.

Feedback:
No corrections or improvements are needed for this response. It effectively addresses the instruction and provides a clear and accurate description of the image content.

Refinement:
Yes, the pillow and the bed are the same color. Both are white.

Verification:
Fail

Instruction:
Is this a natural activity for the animal in the picture?

Reference Response:
Yes.

Preferred Response

Sampling :
No, hugging another cat is not a natural behavior for cats. Cats typically exhibit more aggressive or solitary behavior.

Feedback:
The response is partially accurate in describing the image content but is incorrect in its assessment of the natural behavior of cats. It also deviates from the instruction by providing an incorrect answer.
To improve, the response should accurately describe the image content and provide a correct answer to the question. The response should state that hugging another cat can be a natural behavior for cats, especially when they are socialized and comfortable with each other. The response should also adhere to the instruction by providing a correct answer.

Refinement:
Hugging another cat can be a natural behavior for cats, especially when they are socialized and comfortable with each other. Cats can exhibit a variety of behaviors, including social grooming, which can involve hugging.

Verification:
Pass

Dispreferred Response

Sampling :
No, hugging another cat is not a natural behavior for cats. Cats typically exhibit more aggressive or solitary behavior.

Feedback:
The response is accurate and correctly identifies that hugging is not a natural behavior for cats. The explanation is logically sound and aligns with common knowledge about feline behavior.
To enhance the response, consider adding a brief explanation about why cats might display affectionate behaviors, such as through head butting or grooming, which are more typical. This would provide a more comprehensive understanding of feline social behavior

Refinement:
No, hugging another cat is not a natural behavior for cats. While cats are often solitary animals and may exhibit more aggressive or solitary behavior, they can also form close bonds and display affectionate behaviors such as head butting or grooming. These actions are more typical displays of feline social interaction.

Verification:
Fail

Figure 1. Some examples of the automated preference construction process.

Instruction:
What makes it possible to see light shine on the carpet?

Qwen2VL-7B

The light shining on the carpet in the image is coming from the lamps on the side tables.

Qwen2VL-7B + IPA

The light shining on the carpet is due to the presence of multiple table lamps with lit bulbs, which are placed on side tables and coffee tables throughout the room. The light from these lamps reflects off the carpet, creating a well-lit environment. Additionally, there are ceiling lights that contribute to the overall illumination. The combination of these light sources ensures that the carpet is well-lit.

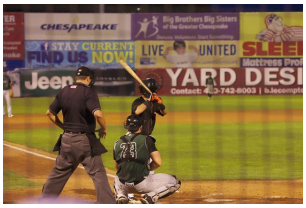
Instruction:
Is the snowboarder on a ski slope?

Qwen2VL-7B

Yes, the snowboarder is on a ski slope.

Qwen2VL-7B + IPA

The snowboarder is on a rail, not a ski slope. The image shows a rail in the foreground, and the snowboarder is performing a trick on it.



Instruction:
What's the purpose of the colorful banners in the outfield?

Qwen2VL-7B

The colorful banners in the outfield serve multiple purposes. They are used to display advertisements for various businesses and organizations, which help fund the baseball field and its operations. These banners also add a vibrant and lively atmosphere to the stadium, enhancing the overall experience for spectators. Additionally, they can serve as a backdrop for photographs and as a way to promote local businesses and community involvement in the sport.

Qwen2VL-7B + IPA

The purpose of the colorful banners in the outfield is to display advertisements for various businesses and organizations. These banners serve as a form of outdoor advertising, promoting the brands and services they represent. The presence of these banners enhances the visibility of the advertisements and helps to attract the attention of spectators and passersby. In the image, the banners include advertisements for Chesapeake, Jeep, Chrysler, Sleepy's Mattress, and Yard Design, among others. These banners are strategically placed to ensure maximum exposure and engagement with the target audience.

Instruction:
What is the relationship of the woman to the parachutist?

Qwen2VL-7B

The woman appears to be watching the parachutist from a safe distance. She is sitting on the grass, waving, and seems to be enjoying the view of the city skyline and the water.

Qwen2VL-7B + IPA

The woman is likely a spectator or friend of the parachutist, as she is waving and watching the parachutist from the grassy hillside. Given the context, it is reasonable to infer that she is familiar with the parachutist, possibly a friend or family member.

Instruction:
What made the slot for the filling?

Qwen2VL-7B

The slot in the sandwich is made from a long, crusty bread roll, typically known as a baguette. The bread is sliced lengthwise to create a pocket that can hold the filling.

Qwen2VL-7B + IPA

The slot for the filling in the sandwich is made by cutting the bread lengthwise, creating a long, open-ended pocket to hold the filling. This is a common method used to prepare sandwiches, especially those with substantial fillings like meats, cheeses, or vegetables. The image shows a sandwich with a generous amount of filling, which is neatly placed inside the bread, demonstrating the effectiveness of this method.

Figure 2. Some examples of the comparison between the baseline model the the model aligned with our approach.