Appendix

A. Prompt Templates

A.1. Prompt Templates for the Three Inference Strategies

We list the exact prompt templates used for the three inference modes. The placeholder `[QUESTION]` is replaced by the actual question (e.g., Where is the dog?).

Direct Strategy.

> `[QUESTION]` The final answer **MUST BE** in `<answer> </answer>` tags. Put your final answer in `<answer></answer>`, e.g., `<answer>blink</answer>`.

Reasoning Strategy.

> `[QUESTION]` You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process **MUST BE** enclosed within `<think> </think>` tags. The final answer **MUST BE** in `<answer> </answer>` tags. Put your final answer in `<answer></answer>`, e.g., `<answer>blink</answer>`.

Region-Guided Strategy.

> `[QUESTION]` Focus on areas in the image related to the question, then think and respond. The final answer **MUST BE** in `<answer> </answer>` tags. Put your final answer in `<answer></answer>`, e.g., `<answer>blink</answer>`.

A.2. Evaluation Prompt Template

We provide the exact evaluation prompt used in all experiments.

> You are an intelligent evaluator. Your task is to evaluate the prediction answer based on the question and the reference answer list along two independent dimensions: Correctness and Irrelevance.
>
> Evaluation Instructions:
>
> 1) Correctness (A): Evaluate whether the prediction semantically answers the question correctly. - A=2: Fully correct — The prediction answer does not conflict with the reference answer - A=1: Partially correct or uncertain — The answer is partially correct, or the information is not sufficient to determine correctness - A=0: Incorrect — If the prediction answer does not contain any of the reference answers
>
> 2) Irrelevance (I): Evaluate how much irrelevant content is present. - I=0: No irrelevant content — The answer is concise and fully focused on the question. - I=1: Slightly irrelevant content — The answer includes brief extra comments, filler text, or small unnecessary additions. - I=2: Strongly irrelevant content — The answer contains long unrelated explanations or digressions.
>
> Question: `<question>`
> Reference Answer: `<reference_answers>`
> Prediction Answer: `<model_prediction>`
>
> You may provide reasoning steps in your evaluation. However, after your reasoning, you MUST provide the final result in the following EXACT format, and DO NOT include any extra words beyond this final format:
>
> Correctness: A=0/1/2
> Irrelevance: I=0/1/2
>
> Now give your evaluation:
> Output:

B. Additional Experimental Results

To better understand how our response-stage attention enhancement influences model performance, we conduct an ablation study on the HaloQuest dataset by varying two key hyperparameters: α (enhancement strength) and ρ (region selection ratio).

α controls how strongly attention is amplified on question-relevant image regions—larger values produce stronger boosting. Meanwhile, ρ determines the proportion of visual regions selected for enhancement, where higher values include more regions.

The α parameter controls the strength of attention amplification applied to question-relevant image regions—larger values produce stronger boosting. Meanwhile, ρ determines the proportion of visual regions selected for enhancement, where higher values include more regions. Tables 1 and 2 report the accuracy of Qwen2.5-VL-3B and Qwen2.5-VL-7B under different settings of these two hyperparameters,

allowing us to examine how amplification intensity and region selection jointly affect downstream performance.

α \ ρ	**0.50**	**0.55**	**0.60**	**0.65**	**0.70**
1.5	**64.3%**	61.5%	63.4%	63.5%	62.7%
2.0	63.2%	61.9%	61.5%	61.2%	62.4%
2.5	62.2%	61.9%	62.2%	61.0%	61.9%

Table 1. Accuracy of Qwen2.5-VL-3B on the HaloQuest dataset under different combinations of α and ρ. Here, α controls the strength of attention amplification on question-relevant image regions (higher values apply stronger enhancement), while ρ controls the proportion of regions selected for enhancement (larger values select more regions).

α \ ρ	**0.50**	**0.55**	**0.60**	**0.65**	**0.70**
1.5	**73.8%**	72.3%	73.2%	72.3%	71.7%
2.0	71.6%	71.5%	72.5%	72.1%	71.6%
2.5	71.5%	72.3%	72.3%	72.3%	71.1%

Table 2. Accuracy of Qwen2.5-VL-7B on the HaloQuest dataset under varying α and ρ values. The hyperparameter α determines the intensity of attention boosting applied to the question-relevant visual regions, whereas ρ specifies the proportion of regions chosen for amplification. Larger values indicate stronger enhancement and broader region selection, respectively.

We further evaluate how our attention-enhancement strategy influences performance on Hallusion Bench, which includes two subsets of questions: VS and VD. VS questions correspond to vision-supplement cases, where the model can often answer correctly even without relying heavily on the image. In contrast, VD questions require vision-dependent reasoning, where accurate answers depend on attending to image content. Given this distinction, we expect that reducing image attention may benefit VS performance, while increasing image attention should favor VD performance. Tables 3 and 4 present results under different values of α and ρ, confirming this contrasting behavior across the two subsets.

α \ ρ	**0.50**	**0.55**	**0.65**	**0.70**
2.5	56.4%	54.4%	53.9%	55.3%
2.0	**57.5%**	56.4%	57.5%	57.2%

Table 3. Performance of Qwen2.5-VL-3B on the VS subset of Hallusion Bench under different (α, topR) settings with Rimg=0.5. VS questions generally benefit from weaker image-attention enhancement.

α \ ρ	**0.50**	**0.55**	**0.65**	**0.70**
2.5	45.4%	44.8%	45.5%	**46.5%**
2.0	44.7%	45.2%	44.0%	45.5%

Table 4. Performance of Qwen2.5-VL-3B on the VD subset of Hallusion Bench under different (α, topR) settings with Rimg=0.5. VD questions require visual grounding and thus benefit from stronger image-attention enhancement.

C. Case Studies

Below we present additional qualitative cases. **Response (base)** refers to the outputs produced by the original base model, while **Response (ours)** denotes the outputs generated after applying our train-free attention-enhancement method on top of the same base model. These examples illustrate how our method helps the model better focus on the image regions relevant to the queried content.

Question:What color is the man's scarf?
Response (base) :<think>The man in the image is wearing a blue scarf around his neck. The scarf appears to be a light blue or grayish color, matching the overall color scheme of his outfit.</think><answer>The man's scarf is blue.</answer>
Response (ours) :<think>The image shows a man wearing a blue coat with a white collar. There is no visible scarf in the image, so it cannot be determined what color the scarf is.</think><answer>The man's scarf is not visible in the image.</answer>

Figure 1. Case 1. **Response (base)** shows the base model output, while **Response (ours)** shows the output after applying our train-free attention-enhancement method.

Question:Is her necklace gold or silver?
Response (base) :<think>To determine whether the mermaid's necklace is gold or silver, I need to carefully examine the image for any clues. The necklace appears to be made of small, round, golden beads that match the color of the water and the coins floating around. This suggests that the necklace is likely gold.</think><answer>gold</answer>
Response (ours) :<think>\nThe image shows a mermaid sitting on a rock in the water, reading a book. There is no visible necklace in the image, so it is not possible to determine if the necklace is gold or silver based solely on this picture.</think><answer>\nThe image does not show any necklace, so it is not possible to determine if the necklace is gold or silver.</answer>

Figure 2. Case 2. Comparison between **Response (base)** and **Response (ours)**.

Question:What color of boots is the worker with the red hat wearing?
Response (base) :<think>The image shows toy figures dressed in construction gear, including hard hats and safety vests. One of the figures has a red hat, and I need to determine the color of their boots. The boots appear to be blue.</think><answer>blue</answer>
Response (ours) :<think>The image shows several workers, each with different colored hats. The worker with the red hat is not visible in this particular view. However, based on the description provided, it seems that the worker with the red hat is not present in the image.</think><answer>The worker with the red hat is not visible in this particular view. Therefore, it is not possible to determine the color of their boots from the given information.</answer>

Figure 3. Case 3. Comparison between **Response (base)** and **Response (ours)**.

Question:Is the person holding the photo crying?
Response (base) :<think>To determine if the person holding the photo is crying, I need to consider several factors:1. The image is sepia-toned, which suggests it might be old or vintage.2. The person's hands appear to be shaking slightly, which could indicate nervousness or emotion.3. The act of holding a photo close to the face often signifies a strong emotional connection.4. There are no visible tears or other clear signs of crying on the person's face.Given these observations, while there's a possibility of emotional distress, the lack of visible tears and the shaking hands suggest that the person might not be actively crying at this moment. They could be holding the photo with a sense of nostalgia or sadness.</think><answer>No, the person holding the photo is not crying.</answer>
Response (ours) :<think>The image shows a person holding a photograph of a woman with long hair, but there is no visible indication of the person's emotional state or whether they are crying. The focus is on the photograph itself rather than the person's facial expression or body language.</think><answer>I don't know if the person holding the photo is crying.</answer>

Figure 4. Case 4. Comparison between **Response (base)** and **Response (ours)**.