

Appendix

001 A. Prompt Templates

002 A.1. Prompt Templates for the Three Inference 003 Strategies

004 We list the exact prompt templates used for the three infer-
005 ence modes. The placeholder [QUESTION] is replaced by
006 the actual question (e.g., Where is the dog?).

007 Direct Strategy.

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

008 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** en-
closed 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>.

009 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>.

010 A.2. Evaluation Prompt Template

011 We provide the exact evaluation prompt used in all experi-
012 ments.

013 You are an intelligent evaluator. Your task is to evalu-

ate the prediction answer based on the question and
the reference answer list along two independent di-
mensions: 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 con-
flict 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 in-
cludes brief extra comments, filler text, or small un-
necessary 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 evalua-
tion. However, after your reasoning, you **MUST** pro-
vide the final result in the following EXACT format,
and **DO NOT** include any extra words beyond this fi-
nal format:

Correctness: A=0/1/2

Irrelevance: I=0/1/2

Now give your evaluation:

Output:

014

B. Additional Experimental Results

015

To better understand how our response-stage attention en-
hancement 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).

016

017

018

019

020

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.

021

022

023

024

025

The α parameter controls the strength of attention ampli-
fication 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,

026

027

028

029

030

031

032

033
034

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

$\alpha \backslash \rho$	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).

$\alpha \backslash \rho$	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.

035
036
037
038
039
040
041
042
043
044
045
046

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.

$\alpha \backslash \rho$	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.

047

$\alpha \backslash \rho$	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

048

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.

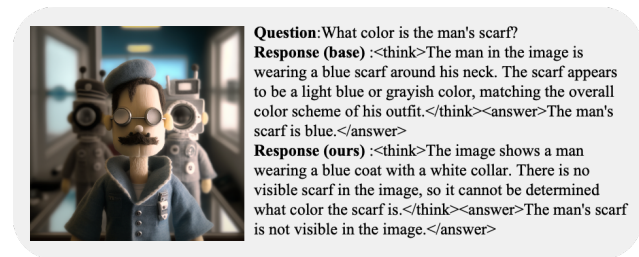
049
050
051
052
053
054
055

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.

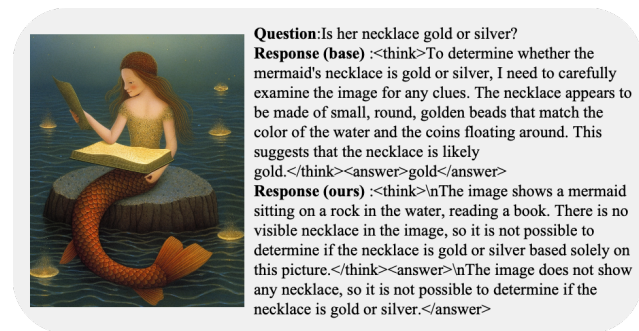


Figure 2. Case 2. Comparison between **Response (base)** and **Response (ours)**.

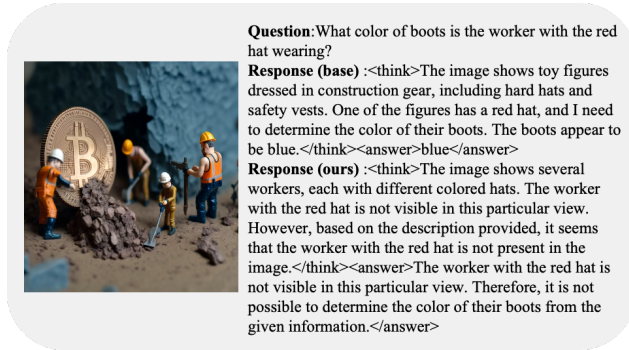


Figure 3. Case 3. Comparison between **Response (base)** and **Response (ours)**.

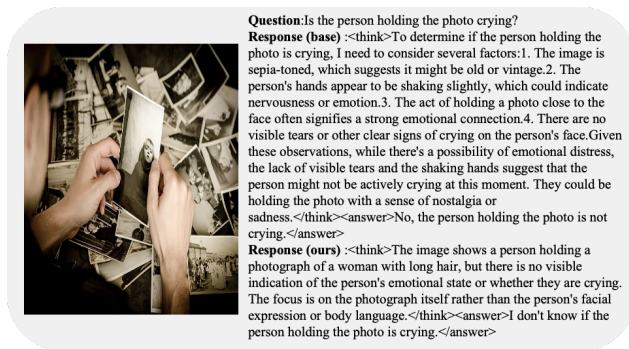


Figure 4. Case 4. Comparison between **Response (base)** and **Response (ours)**.