REPLAY BEFORE YOU REPLY: SECOND-LOOK PROMPTING FOR UNIFIED MULTIMODAL ENCODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Solving multimodal tasks requires Vision-Language Models (VLMs) to form a unified understanding of the image and the question text. However, under causal attention in current VLMs, a token can attend only to what precedes it. Whichever modality comes first is therefore encoded under partial observation of the input, uninformed by the rest of the multimodal context. We introduce Second-Look Prompting (SLP), a training-free method that repeats the image–question pair in the input. In the repeated copy, every visual and textual token can attend to the complete input, yielding a unified multimodal encoding of the pair. Moreover, the added visual tokens contribute computation beyond re-exposing the image—even a blank image in their place can improve reasoning. Across 12 benchmarks, SLP improves the overall average score of every evaluated model, spanning nine open-source models (3B–72B) and three proprietary frontier models, with gains reaching 5.6 points for Gemini 3 Flash on MATH-Vision. Our analysis reveals two complementary mechanisms: the question guides where the second image attends, and the added visual tokens act as a prefill scratchpad whose hidden states retain answer-relevant information from the image and question. These findings highlight input presentation as a simple yet effective way to improve VLM performance.

1 INTRODUCTION

Vision-Language Models (VLMs) (Bai et al., 2025; Gemma Team et al., 2025; Li et al., 2025) answer questions about images by encoding the image and the question as a single causal token sequence. Prior work improves their reasoning through generation-time strategies (Wei et al., 2022; Wang et al., 2023; Zhang et al., 2024); we instead ask how the input presentation shapes representations during prefill. Under causal attention, a single-pass presentation in either order allows only one-way cross-modal interaction: the earlier modality cannot attend to the later one. In the standard image-before-question prompt, denoted $\langle \mathcal{V}, \mathcal{T} \rangle$, visual tokens therefore cannot attend to the question. Here, $\mathcal{V}$ denotes the image block and $\mathcal{T}$ the complete question text, including instructions and options. We refer to this constraint as the *cross-modal partial-observation problem*.

Reversing the prompt to $\langle \mathcal{T}, \mathcal{V} \rangle$ merely relocates the partial observation to the question, now encoded without access to the image. Across the tested families and scales, VLM performance remains sensitive to this ordering: $\langle \mathcal{T}, \mathcal{V} \rangle$ yields a lower overall average than $\langle \mathcal{V}, \mathcal{T} \rangle$ for every tested model (Figure 1, left; Appendix C), possibly because the reversed order differs from the training format. Relying on reasoning is also insufficient. In text-only models, rationales re-read the prompt and repetition gains diminish (Leviathan et al., 2025); in VLMs, generated reasoning tokens can likewise attend to both modalities, but they cannot restore the missing attention direction (Section 3.3.3).

We introduce *Second-Look Prompting* (SLP), a training-free presentation establishing unified multimodal encoding within a causal backbone: append the same image and question to the standard prompt, forming $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$. During prefill, the second image attends to both the initial image and the question, yielding a question-guided second look; the repeated question then conditions on this updated visual state alongside the full preceding context. SLP requires no parameter updates, added task information, or model-internal access.

Across 12 reasoning and perception benchmarks, SLP improves the overall average of all 12 models: nine open-source models from 3B to 72B and three proprietary frontier models. Overall gains reach up to 1.7 points (Figure 1, middle; Table 1) and are larger on reasoning tasks on average. Measured

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

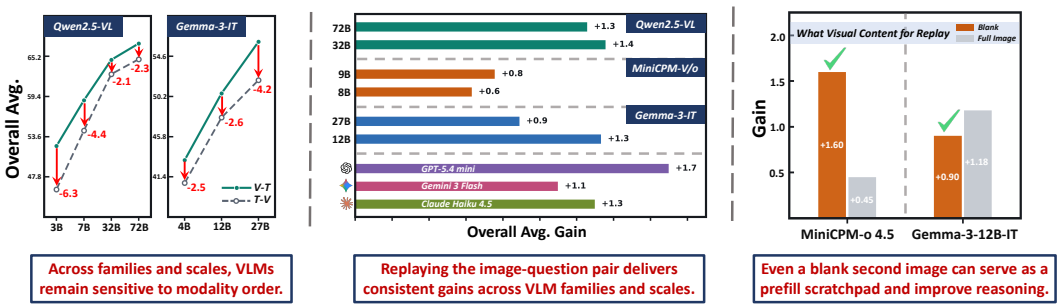


Figure 1: **Motivation, broad gains, and scratchpad evidence for Second-Look Prompting.** Left: across multiple scales in both the Qwen2.5-VL and Gemma-3-IT families, reversing the standard $\langle \mathcal{V}, \mathcal{T} \rangle$ order to $\langle \mathcal{T}, \mathcal{V} \rangle$ lowers the overall average by 2.1–6.3 points, showing that VLMs remain sensitive to input order. Middle: relative to $\langle \mathcal{V}, \mathcal{T} \rangle$, SLP consistently improves the overall average of representative open-source and proprietary VLMs. Right: averaged over seven reasoning benchmarks, the bars compare inserting a blank image, defined as $\langle \mathcal{V}, \mathcal{T}, \square, \mathcal{T} \rangle$, or replaying the same image, $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$, both relative to text-only replay, $\langle \mathcal{V}, \mathcal{T}, \mathcal{T} \rangle$. The positive blank-image gains suggest that added visual positions can support reasoning without adding task-relevant pixels.

end-to-end latency on the nine open-source models changes by at most ± 0.16 seconds on family average (Section 3.5). A continued-training study also shows that the model can learn to better use the second look: with matched training data and steps, LLaVA-OneVision-7B-SI gains 1.97 points under SLP versus 0.55 under the standard prompt (Section 3.4).

Three controlled studies examine how the second look supports answering. **First, the question actively guides the second look.** Once it can attend to the question, the second image directs cross-image attention toward question-relevant regions (Section 3.3.1). **Second, added visual tokens can act as a prefill scratchpad.** Their hidden states retain information from the image and question that the model can later use to answer, giving the second look a computational role beyond revisiting the image. For models that effectively write information into a blank second image, placing one between the repeated questions improves reasoning over text-only replay and can even outperform full-image replay (Figure 1, right; Section 3.3.2). **Third, reasoning does not resolve cross-modal partial observation.** Generated text cannot reorganize visual states formed during prefill; empirically, SLP gains remain similar under direct-answer and default prompts, while text-only repetition gains shrink (Section 3.3.3).

Together, these results show that input presentation alone can shape prefill representations and improve VLM performance. This complements work on generation-time reasoning (Zhang et al., 2024; Zhou et al., 2024; Cao et al., 2024) and input order (Tan et al., 2024; Wardle & Sušnjak, 2025; Wei et al., 2026). We will release our code.

2 From Causal Attention to Second-Look Prompting

Figure 2 shows how input presentation shapes cross-modal attention under a causal mask. In the standard prompt $\langle \mathcal{V}, \mathcal{T} \rangle$, visual tokens precede the question, so their hidden states are computed under partial observation of the input and are never revised. Later generated tokens can read both modalities but only add text states; they cannot produce a question-conditioned visual representation. Creating one requires visual tokens that follow the question, which only the input can provide.

What incomplete variants leave out. Repetition and repositioning change this picture only partially. First, intra-modal *repetition* ($\langle \mathcal{V}, \mathcal{T}, \mathcal{T} \rangle$ and $\langle \mathcal{V}, \mathcal{V}, \mathcal{T} \rangle$; Figure 2b) expands context within a single modality: repeating text extends question processing (Xu et al., 2024; Springer et al., 2025; Leviathan et al., 2025), while repeating the image allows $\mathcal{V}_2$ to attend to $\mathcal{V}_1$. However, because all visual tokens still precede $\mathcal{T}$, they cannot attend to the question, so cross-modal interaction remains one-way. Second, *repositioning* ($\langle \mathcal{V}, \mathcal{T}, \mathcal{V} \rangle$; Figure 2c) places an image copy after the question, allowing $\mathcal{V}_2$ to attend to both $\mathcal{V}_1$ and $\mathcal{T}_1$ during prefill (step 1). However, $\mathcal{T}_1$ precedes $\mathcal{V}_2$ and cannot attend to this question-conditioned visual representation.

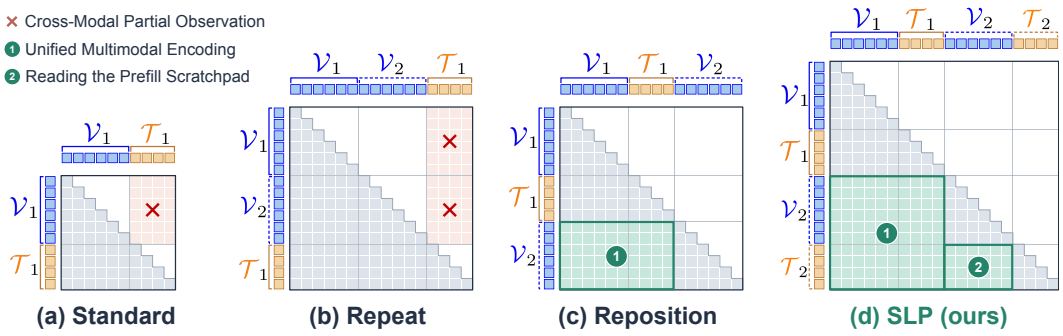


Figure 2: **Causal attention across input presentations.** Rows are querying tokens and columns are keys; gray and green cells are accessible, blank and red cells are masked. Red crosses mark visual blocks that cannot attend to the question; green cells mark the two steps of SLP described in the text (step 1: question-guided visual encoding; step 2: reading the prefill scratchpad via $\mathcal{T}_2$). Within-block attention is schematic and may differ by model.

Unified multimodal encoding via SLP. Second-Look Prompting repeats both modalities:

$$\langle \mathcal{V}, \mathcal{T} \rangle \quad \longrightarrow \quad \langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle := \langle \mathcal{V}_1, \mathcal{T}_1, \mathcal{V}_2, \mathcal{T}_2 \rangle.$$

By resolving the partial-observation constraint, $\mathcal{V}_2$ can selectively integrate visual and textual context during prefill. We call $\mathcal{V}_2$ a multimodal *prefill scratchpad*. Appending $\mathcal{T}_2$ then provides text token positions that read this scratchpad together with the full prompt history (step 2) without disturbing the familiar $\langle \mathcal{V}, \mathcal{T} \rangle$ prefix. In the repeated copy, each modality is thus encoded with the other in context. We examine how this scratchpad retrieves information under question guidance in Section 3.3.1 (Finding 1) and evaluate what it retains and computes in Section 3.3.2 (Finding 2).

3 Experiments

3.1 Experimental Setup

Models. We evaluate SLP on a diverse set of VLM families, covering Qwen2.5-VL (Bai et al., 2025) at four scales (3B, 7B, 32B, and 72B), Gemma-3-IT (Gemma Team et al., 2025) at three scales (4B, 12B, and 27B), and MiniCPM-V 4.5 (8B) and MiniCPM-o 4.5 (9B) (Cui et al., 2026; Yu et al., 2026; Yao et al., 2025), jointly abbreviated as MiniCPM-V/o 4.5. All open-source models are the official instruction-tuned checkpoints. We also include three proprietary models: GPT-5.4 mini, Gemini 3 Flash, and Claude Haiku 4.5.

Benchmarks. We group benchmarks into reasoning-oriented and perception-oriented suites. The reasoning suite includes DynaMath (Zou et al., 2025), MATH-Vision (Wang et al., 2024a), LogicVista (Xiao et al., 2024), VisualPuzzles (Song et al., 2026), MMMU (Yue et al., 2024), MathVista (Lu et al., 2024), and We-Math (Qiao et al., 2025), while the perception suite includes AI2D (Kembhavi et al., 2016), OCRBench (Liu et al., 2024c), SEED-Bench-2-Plus (Li et al., 2024), MMBench (Liu et al., 2024b), and MMStar (Chen et al., 2024a). MathVista is evaluated on its `testmini` split and the other benchmarks on their standard evaluation splits; due to inference cost, each proprietary model is evaluated on up to 1,000 samples per benchmark.

Protocol. Unless otherwise stated, experiments use the default benchmark prompts that permit intermediate reasoning. Model-specific thinking settings and decoding parameters are given in Appendix A. Within each comparison, these settings and the prompt template are fixed. We report benchmark scores on a 0–100 scale and take unweighted means across benchmarks: seven for reasoning, five for perception, and all 12 for the overall average.

3.2 Main Results

SLP improves open-source VLMs across families and scales. Using $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$ instead of $\langle \mathcal{V}, \mathcal{T} \rangle$ improves the overall average of all nine open-source models, spanning 3B to 72B, by $+0.8$ points on average. The gains are generally larger on the reasoning suite: SLP raises the reasoning

Table 1: Main results. Shaded rows add SLP to the model above; deltas on the averages are relative to that baseline and computed before rounding; bold marks only the higher suite or overall average within each pair.

Model	Reasoning								Perception						*Avg.*
	Dyna Math	MATH-Vision	Logic Vista	Visual Puzzles	MMMU	Math Vista	We-Math	**Avg.**	AI2D	OCR Bench	SEED-Bench-2-Plus	MM Bench	MMStar	**Avg.**	
Proprietary															
GPT-5.4 mini	80.1	48.7	63.5	46.2	70.4	68.9	88.2	66.6	84.3	79.5	72.2	84.0	65.2	77.0	70.9
+SLP	80.3	50.2	63.5	46.0	71.6	74.3	89.6	**67.9** (+1.4)	86.8	80.1	72.6	87.6	68.8	**79.2** (+2.1)	**72.6** (+1.7)
Gemini 3 Flash	85.3	76.8	81.4	71.0	83.9	87.5	95.2	83.0	88.5	90.3	76.6	93.2	79.9	85.7	84.1
+SLP	84.5	82.4	83.7	73.1	84.4	87.4	96.3	**84.5** (+1.5)	89.3	90.1	75.7	93.6	82.1	**86.2** (+0.5)	**85.2** (+1.1)
Claude Haiku 4.5	71.3	53.7	57.7	43.6	68.2	73.4	86.4	64.9	79.9	71.6	69.1	81.2	63.2	73.0	68.3
+SLP	72.4	54.4	58.4	44.6	69.7	75.1	87.5	**66.0** (+1.1)	80.6	72.4	69.2	84.0	66.0	**74.4** (+1.4)	**69.5** (+1.3)
Open-Source															
Qwen2.5-VL-72B	66.7	38.6	57.5	41.6	67.7	73.5	48.8	56.3	88.6	88.6	73.7	88.2	71.0	82.0	67.0
+SLP	69.0	40.5	59.7	45.0	66.7	75.4	50.5	**58.1** (+1.8)	88.8	89.1	73.7	89.7	71.5	**82.6** (+0.6)	**68.3** (+1.3)
Qwen2.5-VL-32B	67.0	37.3	54.4	38.2	65.8	72.7	46.9	54.6	83.6	85.5	72.6	85.5	66.8	78.8	64.7
+SLP	67.5	38.8	57.7	40.8	67.7	74.9	49.4	**56.7** (+2.1)	84.4	85.8	71.8	86.1	67.7	**79.2** (+0.4)	**66.1** (+1.4)
Qwen2.5-VL-7B	54.2	25.4	46.5	32.9	53.8	66.6	35.3	45.0	85.1	88.3	70.9	83.2	63.9	**78.3**	58.8
+SLP	54.6	26.4	46.5	32.8	55.0	67.0	37.5	**45.7** (+0.7)	84.6	88.3	70.7	82.4	61.5	77.5 (-0.8)	**58.9** (+0.1)
Qwen2.5-VL-3B	43.6	21.7	36.0	28.5	48.4	62.1	21.6	37.4	81.5	82.3	68.5	77.1	55.7	73.0	52.3
+SLP	43.5	21.3	40.0	28.8	50.9	61.2	24.3	**38.6** (+1.1)	81.8	82.9	68.9	77.1	55.6	**73.2** (+0.2)	**53.0** (+0.8)
MiniCPM-o 4.5	61.6	46.0	51.2	16.4	69.2	72.1	57.5	53.4	80.8	81.8	65.4	84.9	66.6	75.9	62.8
+SLP	62.5	46.1	54.1	15.3	70.8	74.2	57.0	**54.3** (+0.9)	81.4	82.3	65.7	85.5	67.7	**76.5** (+0.6)	**63.6** (+0.8)
MiniCPM-V 4.5	63.1	42.2	53.9	20.5	66.5	73.4	50.1	52.8	80.5	80.8	67.4	86.2	70.5	77.1	62.9
+SLP	63.7	40.6	55.0	20.6	67.2	75.8	51.1	**53.4** (+0.6)	82.5	80.9	67.7	86.6	71.0	**77.7** (+0.7)	**63.6** (+0.6)
Gemma-3-27B-IT	50.6	36.7	47.2	15.2	57.6	60.1	42.0	44.2	83.3	75.3	70.1	79.5	56.5	72.9	56.2
+SLP	48.1	37.2	45.9	16.4	59.0	62.1	46.3	**45.0** (+0.8)	84.6	74.6	70.2	81.2	59.2	**73.9** (+1.0)	**57.1** (+0.9)
Gemma-3-12B-IT	30.8	31.2	36.9	16.4	51.6	54.1	35.3	36.6	80.3	72.9	67.2	75.2	54.7	70.1	50.5
+SLP	37.4	31.4	37.1	16.2	53.0	54.9	35.9	**38.0** (+1.4)	82.2	71.1	67.4	79.2	56.8	**71.3** (+1.3)	**51.9** (+1.3)
Gemma-3-4B-IT	25.9	22.2	30.7	13.8	42.0	43.8	25.7	29.1	73.5	67.9	61.5	66.1	45.9	63.0	43.2
+SLP	22.6	23.9	27.3	16.0	41.9	43.0	29.7	**29.2** (+0.1)	74.7	65.6	61.5	69.1	46.3	**63.5** (+0.5)	**43.5** (+0.2)

averages of Qwen2.5-VL-32B and -72B by 2.1 and 1.8 points, against 0.4 and 0.6 on perception. The reasoning-suite average improves on all nine models and the perception average on eight; only Qwen2.5-VL-7B's perception average declines (-0.8).

SLP also improves proprietary VLMs. SLP improves the overall averages of GPT-5.4 mini, Gemini 3 Flash, and Claude Haiku 4.5 by $+1.7$, $+1.1$, and $+1.3$ points, and all three improve on both suites. A paired bootstrap over the 12 benchmarks gives 95% confidence intervals excluding zero for the overall gains of nine of the 12 models, including all three proprietary models (Appendix B).

3.3 Dissecting the Second Look

We ask three questions about the second look: where the second image attends during prefill (Section 3.3.1), what role its hidden states play in answering (Section 3.3.2), and whether generated reasoning can substitute for prefill replay (Section 3.3.3).

3.3.1 Where Does the Second Image Look?

> FINDING 1 **The question can guide where second-image tokens attend.** *With the full input in context, second-image attention responds to what is asked.*

Figure 1 (left) shows that VLM performance is sensitive to the order of the image and question. This motivates comparing where to place a single image replay: before the question ($\langle \mathcal{V}, \mathcal{V}, \mathcal{T} \rangle$) or after it ($\langle \mathcal{V}, \mathcal{T}, \mathcal{V} \rangle$). Both formats contain identical inputs, differing only in the position of $\mathcal{V}_2$ (Table 2).

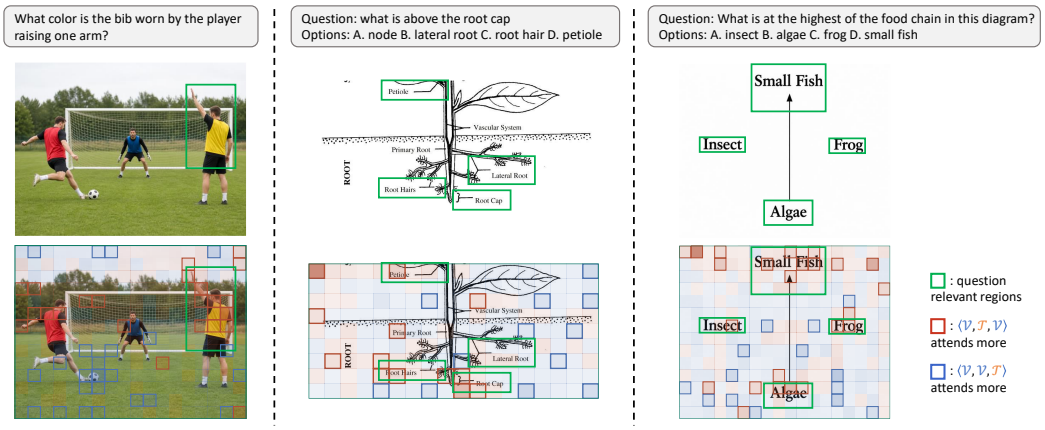


Figure 3: Three examples of attention from the second image to the first. Attention is normalized within the first image before averaging over second-image patch queries; we visualize the difference ($\langle \mathcal{V}, \mathcal{T}, \mathcal{V} \rangle - \langle \mathcal{V}, \mathcal{V}, \mathcal{T} \rangle$). Green boxes mark question-relevant regions. Red and blue outlines mark the smallest set of patches accounting for 50% of the total absolute difference (red: $\langle \mathcal{V}, \mathcal{T}, \mathcal{V} \rangle$ higher; blue: $\langle \mathcal{V}, \mathcal{V}, \mathcal{T} \rangle$ higher).

Post-question placement achieves higher gains in two of the three model families and on average (+0.5 vs. +0.4), despite a small reversal on MiniCPM-V/o 4.5. Repeating only the question ($\langle \mathcal{V}, \mathcal{T}, \mathcal{T} \rangle$) yields a smaller average gain of +0.3 points, although Gemma-3-IT favors question repetition over image repositioning. To examine how access to the question affects visual attention, we inspect cross-image attention in the language decoder.

Table 2: Gains over $\langle \mathcal{V}, \mathcal{T} \rangle$ under default prompts (points), averaged over the 12 benchmarks. Family rows average over their models; the last row weights the three families equally.

Family	$\langle \mathcal{V}, \mathcal{V}, \mathcal{T} \rangle$	$\langle \mathcal{V}, \mathcal{T}, \mathcal{T} \rangle$	$\langle \mathcal{V}, \mathcal{T}, \mathcal{V} \rangle$	$\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$
Qwen2.5-VL	+0.4	+0.1	+0.8	**+0.9**
Gemma-3-IT	+0.2	+0.5	+0.3	**+0.8**
MiniCPM-V/o 4.5	+0.5	+0.3	+0.4	**+0.7**
Average	+0.4	+0.3	+0.5	**+0.8**

Question-guided redirection of visual attention. Figure 3 visualizes cross-image attention from $\mathcal{V}_2$ patch queries to $\mathcal{V}_1$ keys on Qwen2.5-VL-32B across three representative tasks (layers 60–63; protocol in Appendix D). In all three cases, placing the image copy after the question shifts attention directly toward question-relevant regions (green boxes). Queried about a player's bib color (left), $\langle \mathcal{V}, \mathcal{T}, \mathcal{V} \rangle$ shifts attention from the kicker and ball to that player; queried about structures above the root cap (middle) or the top of a food web (right), it shifts attention from diagram boundaries and producers to the queried regions. These examples show that access to the question can redirect second-image attention toward relevant evidence in the first image. Under the causal mask, $\mathcal{V}_2$ has the same prefix in $\langle \mathcal{V}, \mathcal{T}, \mathcal{V} \rangle$ and $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$, so these attention maps also hold for SLP.

3.3.2 What Role Does the Second Image Play?

FINDING 2 **Added visual tokens act as a prefill scratchpad, storing answer-relevant information for later tokens to read.** *They need no task-relevant pixels: even a blank image in its place can improve reasoning.*

To test whether the scratchpad's hidden states retain answer-relevant input information, we prefill $\langle \mathcal{V}_1, \mathcal{T}_1, \mathcal{R}, \text{“Answer:”} \rangle$, where the appended block $\mathcal{R}$ reads the image and question. We then predict the next option letter with attention restricted to $\mathcal{R}$ and the forced “Answer:” cue, itself blocked from directly attending to $\mathcal{V}_1$ and $\mathcal{T}_1$ (Figure 4; Appendix E). We call this test the *scratchpad probe*. The *probe gain* is its accuracy gain over a *cue-only baseline*, where only the cue is available for prediction. We test image and text blocks with matched core token counts, except for the question copy (Table 3).

Readout from the image copy recovers most of the gain. With $\mathcal{R} = \mathcal{V}_2$, the readout recovers 81%–103% of the gain of the unrestricted standard prompt (Full), despite having no direct access to $\mathcal{V}_1$ or $\mathcal{T}_1$ (Table 3).

Content-free blocks enable answering. Remarkably, $\mathcal{R}$ need not contain task information: content-free blocks achieve substantial probe gains across all tested models (up to +35.8 on MiniCPM-o 4.5 and +40.9 on Qwen2.5-VL-32B). Blocking $\mathcal{R}$'s access to the earlier context during prefill largely removes these gains, returning accuracy to within 4.6 points of the cue-only baseline (Appendix E). The gains therefore primarily depend on input information written into $\mathcal{R}$, not on the block itself shifting the model's option preferences. The effective block varies by model: on Qwen2.5-VL, white and yellow images yield near-zero gains, whereas noise yields +21.1 to +33.4 points; on MiniCPM-V/o 4.5 and Gemma-3-12B-IT, uniform images yield strong gains. One possible explanation is that Qwen2.5-VL maps a uniform image to identical visual tokens that are hard to perform computation on, whereas the other models add absolute position embeddings that keep these tokens distinct (Appendix E).

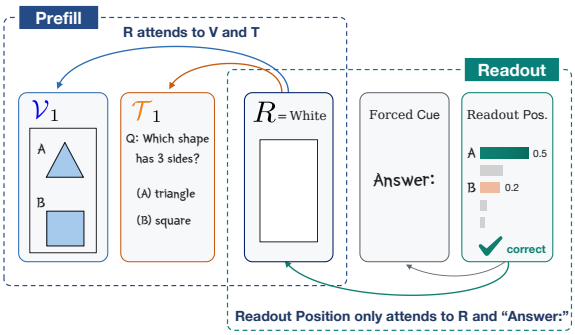


Figure 4: Scratchpad probe with a white image as $\mathcal{R}$. Arrows point from queries to attended blocks.

Table 3: Probe gains over the cue-only baseline. *Full*: the standard prompt $\langle \mathcal{V}_1, \mathcal{T}_1, \text{"Answer:"} \rangle$ with unrestricted attention. All blocks match $\mathcal{V}$ in token count except Copy $\mathcal{T}$; Lorem is fixed Lorem ipsum text. Boldface marks the best block, and underlining marks the best content-free block in each row.

Model	Full	$\mathcal{R}$: image block				$\mathcal{R}$: text block			
		Copy $\mathcal{V}$	White	Yellow	Noise	Copy $\mathcal{T}$	Dots	Spaces	Lorem
Qwen2.5-VL-3B	+35.3	**+28.7**	+1.6	+0.0	+28.0	+13.2	+0.8	+11.3	-0.2
Qwen2.5-VL-7B	+39.4	**+34.0**	+2.2	-1.3	+33.4	+33.1	-0.9	+32.2	+8.3
Qwen2.5-VL-32B	+44.4	+38.2	-0.9	-0.8	+21.1	+38.8	+31.7	**+40.9**	-0.9
MiniCPM-V 4.5	+37.2	+30.1	**+32.0**	+31.7	+30.5	+23.8	+0.6	+22.5	-1.3
MiniCPM-o 4.5	+39.4	+34.5	**+35.8**	+33.6	+23.9	+20.0	-0.8	+0.2	+0.5
Gemma-3-12B-IT	+33.8	+34.9	+34.8	**+35.0**	+33.2	+31.0	+31.1	+29.2	+30.8

Image blocks generally outperform text blocks. The best image block outperforms the best text block for five of six models, with an average margin of 6.9 points across all six (Table 3); among content-free blocks alone, the best one is also an image block for the same five models. Qwen2.5-VL-32B is the only exception, for which spaces work best. The block's modality thus matters: the visual block is the more effective scratchpad.

Even a white image can improve reasoning. The white image yields substantial probe gains on MiniCPM-o 4.5 and Gemma-3-12B-IT but not on Qwen2.5-VL-32B. We test whether this pattern carries over to standard generation by inserting a white image between the two question copies in $\langle \mathcal{V}, \mathcal{T}, \mathcal{T} \rangle$, giving $\langle \mathcal{V}, \mathcal{T}, \square, \mathcal{T} \rangle$. Across the seven reasoning benchmarks, the white image improves scores over $\langle \mathcal{V}, \mathcal{T}, \mathcal{T} \rangle$ by +1.6 points on MiniCPM-o 4.5 and +0.9 on Gemma-3-12B-IT, while Qwen2.5-VL-32B gains only +0.1 (Table 4). This matching pattern supports the scratchpad interpretation: like filler tokens in language models (Pfau et al., 2024), a blank image block retains answer-relevant information without adding task-relevant pixels.

Breaking exact spatial correspondence further improves SLP. The image copy is not always the best block: a content-free block beats it in the scratchpad probe for four of six models (Table 3), and a white image beats it in reasoning with MiniCPM-o 4.5 under standard generation (Table 4). We hypothesize that exact correspondence draws $\mathcal{V}_2$'s attention toward $\mathcal{V}_1$ at the expense of attention to the question and computation within $\mathcal{V}_2$. We therefore cyclically shift $\mathcal{V}_2$ by 7–28 pixels. Averaged across shifts and directions, reasoning scores increase over intact SLP by +0.3, +0.5, and +1.3

Table 4: White-image gains. Probe gains are from Table 3. The Δ columns compare $\langle \mathcal{V}, \mathcal{T}, \square, \mathcal{T} \rangle$ with each prompt under standard generation. Underlining marks Qwen2.5-VL-32B's near-zero probe gain and small gain over $\langle \mathcal{V}, \mathcal{T}, \mathcal{T} \rangle$.

Model	Probe gain	Δ vs. $\langle \mathcal{V}, \mathcal{T}, \mathcal{T} \rangle$		Δ vs. $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$		Δ vs. $\langle \mathcal{V}, \mathcal{T} \rangle$	
		Reas.	Perc.	Reas.	Perc.	Reas.	Perc.
Qwen2.5-VL-32B	<u>−0.9</u>	<u>+0.1</u>	<u>+0.2</u>	−1.9	+0.1	+0.2	+0.5
MiniCPM-o 4.5	+35.8	+1.6	+0.1	+1.1	−0.2	+2.0	+0.4
Gemma-3-12B-IT	+34.8	+0.9	−0.4	−0.3	0.0	+1.1	+1.3

points for Qwen2.5-VL-32B, Qwen2.5-VL-3B, and MiniCPM-o 4.5. On 100 LogicVista prompts, an attention probe on Qwen2.5-VL-32B shows that rightward shifts move $\mathcal{V}_2$'s attention away from $\mathcal{V}_1$ toward the question and $\mathcal{V}_2$ itself (Appendix F).

Findings 1 and 2 identify two complementary functions of the added image block: its tokens attend to $\mathcal{V}_1$ under guidance from the question, and its hidden states retain answer-relevant input information that the model reads when answering. The two functions are linked: the scratchpad states are written by reading the image and question. In SLP, $\mathcal{T}_2$ is the first block that can read this scratchpad during prefill, processing the question again with access to the question-guided $\mathcal{V}_2$ and completing unified multimodal encoding. $\langle \mathcal{V}, \mathcal{T}, \mathcal{V} \rangle$ lacks this step: appending $\mathcal{T}_2$ raises the average gain from $+0.5$ to $+0.8$, and SLP yields the largest gain in each model family (Table 2). The replayed segments thus support contextual computation as well as another view of the same pixels.

3.3.3 Why Does Reasoning Not Resolve Partial Observation?

FINDING 3 **Reasoning does not resolve cross-modal partial observation.** *Generated text cannot reorganize visual states formed during prefill; SLP remains beneficial alongside reasoning.*

Findings 1 and 2 show that the second look reorganizes visual information under question guidance and provides a prefill scratchpad. We next ask whether generated reasoning can replace these functions. In text-only models, reasoning may repeat the question and reduce the benefit of prompt repetition, as suggested by prior work (Leviathan et al., 2025). We compare direct-answer instructions with default benchmark prompts, which permit reasoning without requiring it. The text-only experiment compares $\langle \mathcal{T} \rangle$ and $\langle \mathcal{T}, \mathcal{T} \rangle$ on Qwen2.5-32B over five text benchmarks; the VLM experiment compares $\langle \mathcal{V}, \mathcal{T} \rangle$ and $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$ on Qwen2.5-VL-32B over the 12 main benchmarks (per-benchmark results in Appendix G).

Table 5: Repetition gains (points) on 32B models. Δ = Default − Direct is the difference between gains.

Input	Model	Direct	Default	Δ
$\langle \mathcal{T}, \mathcal{T} \rangle$	Qwen2.5-32B	+3.1	+0.7	-2.4
$\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$	Qwen2.5-VL-32B	+1.3	+1.4	**+0.0**

The text-only repetition gain is smaller with default prompts. Repeating the question improves the average score by 3.1 points with direct-answer instructions and 0.7 with default prompts, a reduction of 2.4 points (Table 5).

Generated reasoning does not replace prefill replay. SLP improves the VLM average by 1.3 points with direct-answer instructions and 1.4 with default prompts (Table 5). The similar gains indicate that generated reasoning does not substitute for the unified encoding of visual and textual information established by replaying the pair during prefill. From the modality perspective, reasoning produces text, which may be a less effective scratchpad than visual tokens. This is consistent with the generally stronger image-block results in Finding 2, though that probe compares fixed input blocks rather than generated rationales (Section 3.3.2). From the functional perspective, the next-token prediction objective alone may not be sufficient to train generated rationales to integrate multimodal input in the same way as prefill replay. Together, these differences suggest that input-side replay can complement generation-time reasoning.

3.4 Learning to Use the Second Look

SLP leverages an existing input capability without adaptation. We next test whether continued training amplifies this latent capability. Starting from LLaVA-OneVision-7B-SI (Li et al., 2025), we conduct two training runs on the same single-turn, single-image instruction subset, one using $\langle \mathcal{V}, \mathcal{T} \rangle$ (Standard) and the other $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$ (SLP), with optimizer steps and hyperparameters matched (Appendix H).

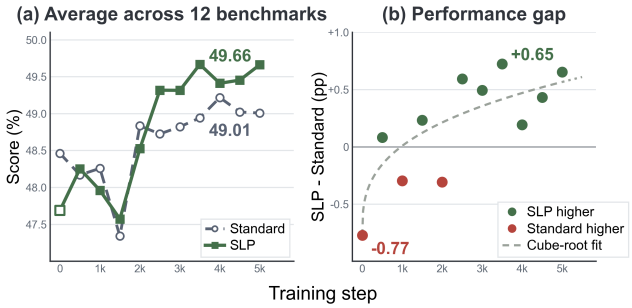


Figure 5: Continued training on $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$ vs. a matched $\langle \mathcal{V}, \mathcal{T} \rangle$ control. (a) Overall average over 12 benchmarks, with each run evaluated using its training prompt; step 0 uses the shared LLaVA-OneVision-7B-SI checkpoint. (b) Per-checkpoint SLP−Standard gap with a descriptive cube-root fit; the dotted segment is extrapolated beyond the observed checkpoints.

Table 6: Scores for each training and inference prompt combination, averaged over the same 12 benchmarks.

	Inference	
Training	Standard	SLP
Standard	49.01	48.41
SLP	49.00	**49.66**

Training unlocks the second-look gain. At initialization, the single-image checkpoint has never seen interleaved inputs, scoring 47.69 under SLP versus 48.46 under Standard. Under continued training, it learns to utilize the repeated presentation: SLP gains $+1.97$ points versus $+0.55$ for Standard, establishing a $+0.65$-point lead overall (Figure 5). The gap widens on both suites, from -0.03 to $+0.94$ points on reasoning and from -1.81 to $+0.25$ on perception.

SLP training preserves the standard-prompt average. At step 5,000, the SLP-trained and Standard-trained checkpoints achieve nearly identical macro averages under $\langle \mathcal{V}, \mathcal{T} \rangle$ inference (49.00 vs. 49.01 across 12 benchmarks; Table 6). Individual benchmark scores change in both directions (Appendix H). Under $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$ inference, the SLP-trained checkpoint outperforms the Standard-trained checkpoint by 1.25 points on the same 12-benchmark macro average.

3.5 Inference Efficiency

We measure end-to-end (E2E) latency on all nine open-source models using default prompts and a shared 10% subsample of each benchmark (Appendix I.1). Family-average changes are $+0.16$ seconds for Qwen2.5-VL, -0.16 seconds for Gemma-3-IT, and $+0.03$ seconds for MiniCPM-V/o 4.5 (Table 7). Shorter generations may offset the extra prefill cost: mean output length in characters decreases for seven of nine models (Appendix I.2), also indicating that SLP's gains do not rely on longer generations.

Table 7: Mean end-to-end latency (s), averaged within families. Δ is SLP minus Standard.

Family	$\langle \mathcal{V}, \mathcal{T} \rangle$	$\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$	Δ
Qwen2.5-VL	3.12	3.28	+0.16
Gemma-3-IT	2.32	2.16	-0.16
MiniCPM-V/o 4.5	4.81	4.84	+0.03

4 Related Work

Prompt repetition and reasoning strategies. Reasoning strategies include CoT (Wei et al., 2022; Kojima et al., 2022), self-consistency (Wang et al., 2023), and structured search (Yao et al., 2023; Besta et al., 2024). Other work studies sensitivity to input order (Liu et al., 2024a; Chen et al., 2024b) and improves reasoning, embeddings, or sequence labeling by repeating the input (Xu et al., 2024; Springer et al., 2025; Kukić et al., 2026). Repetition gains vary by task (Shaier et al., 2024) and can diminish when reasoning is permitted (Leviathan et al., 2025). We show that repeating an

image–question pair establishes unified multimodal encoding, with gains that persist when reasoning is permitted.

Input order and cross-modal conditioning. Input order affects VLMs (Tan et al., 2024; Wardle & Sušnjak, 2025; Wei et al., 2026). Architectural designs enable question-conditioned visual features (Dai et al., 2023; Ganz et al., 2024). Concurrent work studies $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$ for visual steering and question access in natural-image VQA and compositional matching (Abhinandan et al., 2026). However, transfer to broader reasoning tasks and frontier proprietary VLMs remains untested. We evaluate SLP across 12 reasoning and perception benchmarks and 12 open-source and proprietary VLMs, and investigate how visual replay supports computation beyond re-exposure to the image.

Reasoning with visual intermediates. Thinking-with-images methods interactively extract, crop, mark, or sketch images during reasoning (Su et al., 2025; Gao et al., 2025; Zhou et al., 2024; Hu et al., 2024; Liu et al., 2024d; Cao et al., 2024; Zheng et al., 2026); multimodal CoT produces rationales conditioned on both modalities (Zhang et al., 2024; Mitra et al., 2024; Zhao et al., 2024); and VGR replays detected image regions during reasoning (Wang et al., 2026). Unlike these generation-time methods, SLP operates on the input side.

Computation with additional tokens. Pause tokens (Goyal et al., 2024), filler tokens (Pfau et al., 2024), and vision-transformer registers (Darcet et al., 2024) show how additional token positions can support computation, typically with training. SLP instead repeats the image–question pair to enable question-conditioned visual processing during prefill. Our analyses further show that the added visual tokens can serve as a scratchpad without additional training (Section 3.3.2).

5 Limitations

SLP is a simple input-side intervention, and its scope and costs have three main limitations. First, duplicate images may interfere with position-sensitive tasks: in stress tests, counting and spatial relations hold (-0.12 on CountQA; $+0.54$ on SpatialMQA), but grounding drops by 1.11 points on RefCOCOg (Appendix J). Second, training on SLP takes $2.25\times$ the wall-clock time of standard training due to the duplicated prefix (32.0 vs. 14.2 hours for 5,000 steps on 16 A100 GPUs; Appendix H). Third, SLP does not yet offer a unified treatment of multi-image or interleaved image–text inputs: our experiments are limited to prompts whose original form is $\langle \mathcal{V}, \mathcal{T} \rangle$, and how to replay other input organizations is left for future work.

6 Conclusion

We introduce Second-Look Prompting (SLP), a training-free input presentation that addresses the cross-modal partial-observation problem by establishing unified multimodal encoding under causal attention. SLP reshapes prefill computation: the repeated image revisits the earlier visual content under the question's guidance, and the repeated question then conditions on the resulting visual states. These added visual tokens also act as a prefill scratchpad, retaining answer-relevant input information, so that even a blank second look can aid reasoning. SLP's gains persist when reasoning is allowed, unlike those of text-only repetition. Continued training improves performance under SLP while retaining a similar average under standard prompting. These results position input-side prompting as an accessible control surface complementary to generation-time reasoning. They also suggest that the potential of input presentation remains underexplored and raise a broader question: what properties distinguish visual from textual tokens during prefill, and how do these differences shape the integration of the image and question?

AI USE STATEMENT

We used generative AI tools to aid and polish writing (language editing, manuscript organization, terminology and repetition checks, and LaTeX formatting) and for retrieval and discovery (finding related work and verifying bibliographic entries). We have reviewed all AI-assisted content and take full responsibility for the final content of this paper.

ETHICS STATEMENT

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

REPRODUCIBILITY STATEMENT

Prompt construction is described in Section 2. Prompt templates, decoding settings, and benchmark sampling are listed in Appendix A. Protocols for the attention analyses and scratchpad probe are given in Appendices D and E; the continued-training settings and results appear in Appendix H. We will release our implementation and evaluation code.

REFERENCES

Rakshanda Hassan Abhinandan, John Galeotti, Deva Ramanan, and Gautam Rajendrakumar Gare. Ask Twice, Look Twice: Prompt Echoing Resolves the Question-First Paradox in Vision-Language Models. In *ECCV 2026 Workshop on Explainable Computer Vision (eXCV)*, 2026. URL https://arxiv.org/abs/2607.15565.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923, 2025. URL https://arxiv.org/abs/2502.13923.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, Mar 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i16.29720. URL https://doi.org/10.1609/aaai.v38i16.29720.

Yuhang Cao, Pan Zhang, Xiaoyi Dong, Dahua Lin, and Jiaqi Wang. DualFocus: Integrating Macro and Micro Perspectives in Multi-modal Large Language Models. arXiv preprint arXiv:2402.14767, 2024. URL https://arxiv.org/abs/2402.14767.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are We on the Right Way for Evaluating Large Vision-Language Models? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087. Curran Associates, Inc., 2024a. doi: 10.52202/079017-0850. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2f8ee6a3d766b426d2618e555b5aeb39-Paper-Conference.pdf.

Xinyun Chen, Ryan Andrew Chi, Xuezhi Wang, and Denny Zhou. Premise Order Matters in Reasoning with Large Language Models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6596–6620. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/chen24i.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John

Schulman. Training Verifiers to Solve Math Word Problems. arXiv preprint arXiv:2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. MiniCPM-o 4.5: Towards Real-Time Full-Duplex Omni-Modal Interaction. arXiv preprint arXiv:2604.27393, 2026. URL https://arxiv.org/abs/2604.27393.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 49250–49267. Curran Associates, Inc., 2023. doi: 10.52202/075280-2142. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9a6a435e75419a836fe47ab6793623e6-Paper-Conference.pdf.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 2632–2652, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0b408293619f725fd30162af057e531a-Paper-Conference.pdf.

Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. Question Aware Vision Transformer for Multimodal Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13861–13871, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ganz_Question_Aware_Vision_Transformer_for_Multimodal_Reasoning_CVPR_2024_paper.html.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-Modal Chain-of-Thought. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19520–19529, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Gao_Interleaved-Modal_Chain-of-Thought_CVPR_2025_paper.html.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training Language Models With Pause Tokens. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 27896–27923, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/76917808731dae9e6d62c2a7a6afb542-Paper-Conference.pdf.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information*

Processing Systems Track on Datasets and Benchmarks, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah Smith, and Ranjay Krishna. Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 139348–139379. Curran Associates, Inc., 2024. doi: 10.52202/079017-4423. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/fb82011040977c7712409fbdb5456647-Paper-Conference.pdf.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A Diagram is Worth a Dozen Images. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 235–251, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0. doi: 10.1007/978-3-319-46493-0_15. URL https://link.springer.com/chapter/10.1007/978-3-319-46493-0_15.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213. Curran Associates, Inc., 2022. doi: 10.52202/068431-1613. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

Matija Luka Kukić, Marko Čuljak, David Dukić, Martin Tutek, and Jan Šnajder. Sequence Repetition Enhances Token Embeddings and Improves Sequence Labeling with Decoder-only Language Models. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 6442–6456, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.339. URL https://aclanthology.org/2026.findings-eacl.339/.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Prompt Repetition Improves Non-Reasoning LLMs. arXiv preprint arXiv:2512.14982, 2025. URL https://arxiv.org/abs/2512.14982.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=zKv8qULV6n.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. SEED-Bench-2-Plus: Benchmarking Multimodal Large Language Models with Text-Rich Visual Comprehension. arXiv preprint arXiv:2404.16790, 2024. URL https://arxiv.org/abs/2404.16790.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.

Jingping Liu, Ziyan Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, and Tong Ruan. Can Multimodal Large Language Models Understand Spatial Relations? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 620–632, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.31. URL https://aclanthology.org/2025.acl-long.31/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the*

Association for Computational Linguistics, 12:157–173, 2024a. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is Your Multi-modal Model an All-Around Player? In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, volume 15064 of *Lecture Notes in Computer Science*, pp. 216–233, Cham, 2024b. Springer Nature Switzerland. ISBN 978-3-031-72658-3. doi: 10.1007/978-3-031-72658-3_13. URL https://link.springer.com/chapter/10.1007/978-3-031-72658-3_13.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: on the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 67(12):220102, Dec 2024c. ISSN 1869-1919. doi: 10.1007/s11432-024-4235-6. URL https://doi.org/10.1007/s11432-024-4235-6.

Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-Spot: Interactive Reasoning Improves Large Vision-Language Models. arXiv preprint arXiv:2403.12966, 2024d. URL https://arxiv.org/abs/2403.12966.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/663bce02a0050c4a11f1eb8a7f1429d3-Paper-Conference.pdf.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and Comprehension of Unambiguous Object Descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11–20, June 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Mao_Generation_and_Comprehension_CVPR_2016_paper.html.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional Chain-of-Thought Prompting for Large Multimodal Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14420–14431, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Mitra_Compositional_Chain-of-Thought_Prompting_for_Large_Multimodal_Models_CVPR_2024_paper.html.

Jacob Pfau, William Merrill, and Samuel R. Bowman. Let's Think Dot by Dot: Hidden Computation in Transformer Language Models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=NikbrdtYvG.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma GongQue, Shanglin Lei, YiFan Zhang, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Xiao Zong, Yida Xu, Peiqing Yang, Zhimin Bao, Muxi Diao, Chen Li, and Honggang Zhang. We-Math: Does Your Large Multimodal Model Achieve Human-like Mathematical Reasoning? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20023–20070, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.983. URL https://aclanthology.org/2025.acl-long.983/.

Sagi Shaier, Mario Sanz-Guerrero, and Katharina von der Wense. Asking Again and Again: Exploring LLM Robustness to Repeated Questions. arXiv preprint arXiv:2412.07923, 2024. URL https://arxiv.org/abs/2412.07923.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. VisualPuzzles: Decoupling Multimodal Reasoning Evaluation from Domain Knowledge. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/63662.

Jacob Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. Repetition Improves Language Model Embeddings. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 93543–93579, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e9b0c9480b03f9a6bca24ed8c08ee058-Paper-Conference.pdf.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, Linjie Li, Yu Cheng, Heng Ji, Junxian He, and Yi R. Fung. Thinking with Images for Multimodal Reasoning: Foundations, Methods, and Future Frontiers. arXiv preprint arXiv:2506.23918, 2025. URL https://arxiv.org/abs/2506.23918.

Jayant Sravan Tamarapalli, Rynaa Grover, Nilay Pande, and Sahiti Yerramilli. CountQA: How Well Do MLLMs Count in the Wild? arXiv preprint arXiv:2508.06585, 2025. URL https://arxiv.org/abs/2508.06585.

Zhijie Tan, Xu Chu, Weiping Li, and Tong Mo. Order Matters: Exploring Order Sensitivity in Multimodal Large Language Models. arXiv preprint arXiv:2410.16983, 2024. URL https://arxiv.org/abs/2410.16983.

Jiacong Wang, Zijian Kang, Haochen Wang, Xiao Liang, Ya Wang, Jiawen Li, Bohong Wu, Jiao Ran, Haiyong Jiang, Chao Feng, and Jun Xiao. VGR: Visual Grounded Reasoning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 89506–89528, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/90aeee6dfe75ab6b5d6958bce40d3e16-Paper-Conference.pdf.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95095–95169. Curran Associates, Inc., 2024a. doi: 10.52202/079017-3014. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad0edc7d5fa1a783f063646968b7315b-Paper-Datasets_and_Benchmarks_Track.pdf.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290. Curran Associates, Inc., 2024b. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad236edc564f3e3156e1b2feafb99a24-Paper-Datasets_and_Benchmarks_Track.pdf.

Grant Wardle and Teo Sušnjak. Image First or Text First? Optimising the Sequencing of Modalities in Large Language Model Prompting and Reasoning Tasks. *Big Data and Cognitive Computing*, 9 (6):149, June 2025. ISSN 2504-2289. doi: 10.3390/bdcc9060149. URL https://doi.org/10.3390/bdcc9060149.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Sheng-Lun Wei, Yu-Ling Liao, Hen-Hsen Huang, and Hsin-Hsi Chen. Do Multimodal LLMs Understand Order? Measuring the Fragility of Multimodal Reasoning under Input Order Perturbations. In Stelios Piperidis, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral (eds.), *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 9118–9128, Palma de Mallorca, Spain, May 2026. ELRA Language Resource Association. doi: 10.63317/4jtpgzks8pbr. URL https://aclanthology.org/2026.lrec-1.716/.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. LogicVista: Multimodal LLM Logical Reasoning Benchmark in Visual Contexts. arXiv preprint arXiv:2407.04973, 2024. URL https://arxiv.org/abs/2407.04973.

Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, Jian-Guang Lou, and Shuai Ma. Re-Reading Improves Reasoning in Large Language Models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15549–15575, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.871. URL https://aclanthology.org/2024.emnlp-main.871/.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822. Curran Associates, Inc., 2023. doi: 10.52202/075280-0517. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/271db9922b8d1f4dd7aaef84ed5ac703-Paper-Conference.pdf.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Chi Chen, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Ronghua Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Efficient GPT-4V level multimodal large language model for deployment on edge devices. *Nature Communications*, 16(1):5509, Jul 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-61040-5. URL https://doi.org/10.1038/s41467-025-61040-5.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. MiniCPM-V 4.5: Cooking Efficient MLLMs via Architecture, Data, and Training Recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11704–11715, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yu_MiniCPM-V_4.5_Cooking_Efficient_MLLMs_via_Architecture_Data_and_Training_CVPR_2026_paper.html.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal Chain-of-Thought Reasoning in Language Models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=y1pPWFVfvR.

Xueliang Zhao, Xinting Huang, Tingchen Fu, Qintong Li, Shansan Gong, Lemao Liu, Wei Bi, and Lingpeng Kong. BBA: Bi-Modal Behavioral Alignment for Reasoning with Large Vision-Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7255–7279, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.433. URL https://aclanthology.org/2024.findings-acl.433/.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "Thinking with Images" via Reinforcement Learning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/cdb347d7516d52a9280ea9d6708911f6-Paper-Conference.pdf.

Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. Image-of-Thought Prompting for Visual Reasoning Refinement in Multimodal Large Language Models. arXiv preprint arXiv:2405.13872, 2024. URL https://arxiv.org/abs/2405.13872.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. DynaMath: A Dynamic Visual Benchmark for Evaluating Mathematical Reasoning Robustness of Vision Language Models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 48337–48383, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/78b248ea6f627431bba5029d92be8a3d-Paper-Conference.pdf.

A Prompt and Decoding Details

A.1 Prompt Templates

Default prompts. All main experiments use each benchmark's official prompt verbatim, including the question, any benchmark-provided hint, the option list, and the official answer-format instruction. A schematic multiple-choice prompt is

```
Question: {question}
Options:
A. {A}
B. {B}
...
Please select the correct answer from the options above.
```

Free-form benchmarks pose the question with their official answer-format hint (e.g., DynaMath's JSON output with `solution` and `short answer` keys). These prompts permit but do not require intermediate reasoning. The same prompt is used for all models and input presentations: a presentation such as $\langle \mathcal{V}, \mathcal{T}, \mathcal{V} \rangle$ or $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$ only reorders or duplicates the image and question segments verbatim. Each model uses its official chat template and native image tokens.

Direct-answer prompts (Section 3.3.3). For the direct-answer ablation, we append an instruction that asks for the final answer without intermediate reasoning (Table 8); for DynaMath, it replaces the official JSON-format instruction.

Table 8: Direct-answer instructions, appended to the default prompt. A few benchmarks keep their official phrasing (e.g., VisualPuzzles keeps "the option letter from the given choices"; We-Math embeds the instruction in its official hint).

Answer type	Benchmarks	Instruction
Option letter	LogicVista, VisualPuzzles, MMMU, We-Math, AI2D, SEED-Bench-2-Plus, MMBench, MMStar	`Return only the final option letter. Do not write 'Answer:'. Do not include explanation, reasoning, derivation, punctuation, or extra text. Output only one option letter, such as A, B, C, D, E, or F.`
Free-form	MATH-Vision, MathVista	`Return only the final answer. Do not write 'Answer:'. Do not include explanation, reasoning, derivation, units unless required by the question, or extra text.`
Recognized text	OCRBench	`Return only the recognized text. Do not write 'Answer:'. Do not include explanation, reasoning, punctuation not present in the image, or extra text.`
Number	DynaMath	`Return only the final numeric answer as a three-digit floating-point number. Do not output JSON. Do not include explanation, reasoning, derivation, units, or extra text.`

A.2 Decoding Configurations

Table 9 lists the decoding configurations. Within each comparison, all input presentations share the same configuration. Open-source models are served with vLLM, and proprietary models are queried through OpenAI-compatible chat-completion endpoints.

Table 9: Decoding configurations. Unlisted sampling parameters keep engine or provider defaults. Qwen2.5-VL on LogicVista uses near-greedy sampling (temperature 0.01, top-p 1.0, repetition penalty 1.05). On the perception suite, MiniCPM-V uses regular CoT throughout; MiniCPM-o follows the vendor defaults, using no thinking on AI2D, OCRBench, and SEED-Bench-2-Plus and regular CoT on MMBench and MMStar.

Model	Decoding	Temp.	Max tokens	Reasoning control
Qwen2.5-VL (3B–72B)	greedy	0	2,048	none
Gemma-3-IT (4B–27B)	greedy	0	4,096	none
MiniCPM-V/o 4.5, reasoning	top-p 1.0	0.7	16,384	long-CoT thinking
MiniCPM-V/o 4.5, perception	greedy, rep. pen. 1.2	0	16,384	per benchmark
GPT-5.4 mini	API	0	32,768	reasoning effort: minimal
Gemini 3 Flash	API	0	32,768	non-thinking
Claude Haiku 4.5	API	0	32,768	provider default

A.3 Benchmark Splits and Subsampling

Open-source models are evaluated on the full splits. Due to API cost, proprietary models are evaluated on a deterministic subset of at most 1,000 examples per benchmark, drawn with the benchmark-specific seed `int(sha256("{seed}:{dataset}")[:16], 16)` and global seed 20260627. LogicVista, MathVista `testmini`, and OCRBench have at most 1,000 examples and are used in full.

B Confidence Intervals for the Main Results

Table 10 gives a 95% confidence interval for each model's overall gain in Table 1. Because $\langle \mathcal{V}, \mathcal{T} \rangle$ and $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$ are evaluated on the same benchmarks with identical decoding, we treat the benchmark as the paired unit: we resample the 12 per-benchmark gains with replacement 20,000 times and report the percentile interval of the mean. The interval thus reflects how consistently the gain holds across benchmarks. For reference, within-benchmark sampling contributes a standard error of about 0.5 points to each overall average under a binomial model.

Table 10: Overall gain of $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$ over $\langle \mathcal{V}, \mathcal{T} \rangle$ (12-benchmark average) with 95% paired-bootstrap confidence intervals. Improved/declined counts benchmarks with positive/negative gains; the rest are ties at the reported precision.

Model	Δ	95% CI	Improved / declined
GPT-5.4 mini	+1.68	[+0.80, +2.68]	10 / 1
Gemini 3 Flash	+1.08	[+0.21, +2.15]	8 / 4
Claude Haiku 4.5	+1.25	[+0.83, +1.72]	12 / 0
Qwen2.5-VL-72B	+1.26	[+0.59, +1.92]	10 / 1
Qwen2.5-VL-32B	+1.36	[+0.73, +1.99]	11 / 1
Qwen2.5-VL-7B	+0.10	[−0.53, +0.71]	5 / 5
Qwen2.5-VL-3B	+0.78	[+0.03, +1.62]	7 / 4
MiniCPM-o 4.5	+0.76	[+0.18, +1.37]	10 / 2
MiniCPM-V 4.5	+0.63	[+0.07, +1.17]	11 / 1
Gemma-3-27B-IT	+0.89	[−0.10, +1.87]	9 / 3
Gemma-3-12B-IT	+1.33	[+0.26, +2.62]	10 / 2
Gemma-3-4B-IT	+0.22	[−1.07, +1.47]	6 / 5

C Order Reversal: Per-Benchmark Results

Table 11 gives the per-benchmark results behind Figure 1 (left). Reversing $\langle \mathcal{V}, \mathcal{T} \rangle$ to $\langle \mathcal{T}, \mathcal{V} \rangle$ lowers the overall average of every Qwen2.5-VL and Gemma-3-IT model by 2.1–6.3 points. The drop is larger on perception (3.9–12.4 points) than on reasoning (0.2–2.9 points).

Table 11: Order reversal. Indented rows evaluate the model above with $\langle \mathcal{T}, \mathcal{V} \rangle$ instead of $\langle \mathcal{V}, \mathcal{T} \rangle$; deltas on the averages are relative to $\langle \mathcal{V}, \mathcal{T} \rangle$.

(a) Reasoning

Model	Dyna Math	MATH-Vision	Logic Vista	Visual Puzzles	MMMU	MathVista	We-Math	**Avg.**
Qwen2.5-VL-72B	66.7	38.6	57.5	41.6	67.7	73.5	48.8	56.3
$\langle \mathcal{T}, \mathcal{V} \rangle$	66.4	39.3	57.7	44.5	66.3	70.6	48.5	56.2 (-0.2)
Qwen2.5-VL-32B	67.0	37.3	54.4	38.2	65.8	72.7	46.9	54.6
$\langle \mathcal{T}, \mathcal{V} \rangle$	66.8	37.9	56.8	37.2	63.4	71.8	43.0	53.8 (-0.8)
Qwen2.5-VL-7B	54.2	25.4	46.5	32.9	53.8	66.6	35.3	45.0
$\langle \mathcal{T}, \mathcal{V} \rangle$	53.6	27.7	45.4	32.6	49.5	64.9	31.9	43.7 (-1.3)
Qwen2.5-VL-3B	43.6	21.7	36.0	28.5	48.4	62.1	21.6	37.4
$\langle \mathcal{T}, \mathcal{V} \rangle$	40.1	22.9	37.8	28.2	47.7	50.2	21.9	35.6 (-1.9)
Gemma-3-27B-IT	50.6	36.7	47.2	15.2	57.6	60.1	42.0	44.2
$\langle \mathcal{T}, \mathcal{V} \rangle$	42.8	38.2	38.3	17.2	54.5	58.5	40.2	41.4 (-2.9)
Gemma-3-12B-IT	30.8	31.2	36.9	16.4	51.6	54.1	35.3	36.6
$\langle \mathcal{T}, \mathcal{V} \rangle$	33.0	32.5	34.5	17.6	50.2	51.6	26.7	35.2 (-1.5)
Gemma-3-4B-IT	25.9	22.2	30.7	13.8	42.0	43.8	25.7	29.1
$\langle \mathcal{T}, \mathcal{V} \rangle$	23.5	22.4	26.9	13.8	44.1	42.5	27.5	28.7 (-0.5)

(b) Perception and overall

Model	AI2D	OCRBench	SEED-Bench-2-Plus	MMBench	MMStar	**Avg.**	**Overall**
Qwen2.5-VL-72B	88.6	88.6	73.7	88.2	71.0	82.0	67.0
$\langle \mathcal{T}, \mathcal{V} \rangle$	82.5	87.8	69.8	82.6	61.3	76.8 (-5.2)	64.8 (-2.3)
Qwen2.5-VL-32B	83.6	85.5	72.6	85.5	66.8	78.8	64.7
$\langle \mathcal{T}, \mathcal{V} \rangle$	80.7	81.8	70.6	79.3	62.1	74.9 (-3.9)	62.6 (-2.1)
Qwen2.5-VL-7B	85.1	88.3	70.9	83.2	63.9	78.3	58.8
$\langle \mathcal{T}, \mathcal{V} \rangle$	74.0	85.6	62.2	71.5	54.8	69.6 (-8.7)	54.5 (-4.4)
Qwen2.5-VL-3B	81.5	82.3	68.5	77.1	55.7	73.0	52.3
$\langle \mathcal{T}, \mathcal{V} \rangle$	66.6	77.9	56.1	61.5	41.0	60.6 (-12.4)	46.0 (-6.3)
Gemma-3-27B-IT	83.3	75.3	70.1	79.5	56.5	72.9	56.2
$\langle \mathcal{T}, \mathcal{V} \rangle$	76.0	73.7	64.6	72.1	47.9	66.8 (-6.1)	52.0 (-4.2)
Gemma-3-12B-IT	80.3	72.9	67.2	75.2	54.7	70.1	50.5
$\langle \mathcal{T}, \mathcal{V} \rangle$	74.8	70.1	62.4	73.5	47.8	65.7 (-4.3)	47.9 (-2.6)
Gemma-3-4B-IT	73.5	67.9	61.5	66.1	45.9	63.0	43.2
$\langle \mathcal{T}, \mathcal{V} \rangle$	64.1	63.9	55.3	63.2	41.6	57.6 (-5.4)	40.7 (-2.5)

D Question-Guided Attention: Case-Study Protocol

The three examples in Figure 3 compare $\langle \mathcal{V}, \mathcal{T}, \mathcal{V} \rangle$ with $\langle \mathcal{V}, \mathcal{V}, \mathcal{T} \rangle$ on Qwen2.5-VL-32B, using the same prompt template and identical image copies.

We use prefill self-attention in the language decoder, with queries from the second image and keys from the first; attention inside the vision encoder is excluded. Each query row is normalized over the first image's keys, so the maps show how attention is distributed within the first image rather than its total mass. We then average over all 40 heads, the final four layers (60–63), and all second-image queries, and plot the per-patch difference between $\langle \mathcal{V}, \mathcal{T}, \mathcal{V} \rangle$ and $\langle \mathcal{V}, \mathcal{V}, \mathcal{T} \rangle$. The outlined patches form the smallest set that accounts for 50% of the total absolute difference.

E SCRATCHPAD PROBE: PROTOCOL AND ADDITIONAL BLOCKS

E.1 PROTOCOL

We use the strict single-answer multiple-choice subsets of DynaMath, We-Math, MMBench, MMStar, and AI2D, sampling 128 questions per benchmark for each model (640 per model). The probe prefills $\langle \mathcal{V}_1, \mathcal{T}_1, \mathcal{R}, \text{“Answer:”} \rangle$. Besides the image and question copies, $\mathcal{R}$ is a white or yellow image, $1/f^2$ grayscale noise, repeated periods (Dots) or spaces, a fixed Lorem ipsum sequence, or a shuffled version of the same Lorem tokens. Within a model, all blocks except the question copy match the image copy in core token count and readout position. Noise and shuffled Lorem are averaged over three fixed seeds.

During readout, the forced “Answer:” cue attends only to $\mathcal{R}$ and itself, so it cannot bypass $\mathcal{R}$ to reach $\mathcal{V}_1$ or $\mathcal{T}_1$. We predict the option letter with the highest next-token log-probability. The probe gain is the accuracy gain over a cue-only baseline in which the cue is the only available context. This empirical baseline reflects the model’s option-label preferences rather than uniform guessing.

Blocked-context control. To test whether the gains depend on the earlier context, we additionally block $\mathcal{R}$’s attention to all preceding keys during prefill. For every block in Table 3, this returns accuracy to within 4.6 points of the cue-only baseline; the largest residual is for the yellow image on MiniCPM-o 4.5. Attention masks were verified with an independent implementation, and examples with different masks were never batched together.

E.2 SHUFFLED LOREM CONTROL

Matching the token count does not make blocks interchangeable: on Qwen2.5-VL-32B, spaces yield $+40.9$ points, whereas Lorem yields -0.9 (Table 3). Token order also matters. On Gemma-3-12B-IT, Lorem yields $+30.8$ points, but shuffling the same tokens reduces the gain to -1.0 (Table 12). Qwen2.5-VL-7B shows a smaller reduction, and on the other models Lorem is already ineffective.

Table 12: Probe gains (points) for Lorem and a shuffled version with the same token multiset. Lorem values are repeated from Table 3.

Model	Lorem	Shuffled Lorem
Qwen2.5-VL-3B	–0.2	–0.1
Qwen2.5-VL-7B	+8.3	+3.1
Qwen2.5-VL-32B	–0.9	–1.3
MiniCPM-V 4.5	–1.3	–2.0
MiniCPM-o 4.5	+0.5	–1.4
Gemma-3-12B-IT	+30.8	–1.0

E.3 UNIFORM IMAGES ACROSS ARCHITECTURES

In the scratchpad probe, white and yellow images yield at most $+2.2$ points on the three Qwen2.5-VL models but $+31.7$ to $+35.8$ points on MiniCPM-V/o 4.5 and Gemma-3-12B-IT (Table 3). One possible explanation lies in where positional information enters the vision encoder. Gemma-3-12B-IT and MiniCPM-V/o 4.5 use SigLIP-based encoders that add learned absolute position embeddings to the patch embeddings before the first Transformer layer, so even a uniform image yields visual tokens that differ across positions.[1] MiniCPM’s 64 learned resampler queries alone would not produce this variation: the resampler adds position only to its keys, so identical encoder features would yield identical outputs for every query. Qwen2.5-VL’s vision encoder has no additive position embedding and injects position only through rotary embeddings on queries and keys.[2] Rotary embeddings change attention weights but not the values being averaged, so identical patches remain identical through the encoder and the patch merger, and a uniform image enters the decoder as repeated copies of a single vector. Noise images, whose visual tokens are distinct, are effective on Qwen2.5-VL (Table 3),

[1] See the SigLIP embeddings, and the MiniCPM-V 4.5 resampler and configuration.

[2] See the Qwen2.5-VL vision encoder.

consistent with this account. The analysis concerns image blocks only; text blocks follow a different embedding and position path and may behave differently. Testing whether identical visual tokens cause the near-zero gains requires measuring and intervening on visual-token states.

F SHIFTED-IMAGE CONTROL

Section 3.3.2 hypothesizes that an exact image copy draws $\mathcal{V}_2$'s attention toward $\mathcal{V}_1$ at the expense of attention to the question and computation within $\mathcal{V}_2$. We test this hypothesis by shifting $\mathcal{V}_2$, measuring its effect on generation and on prefill attention.

Generation. We cyclically shift $\mathcal{V}_2$ by 7, 14, or 28 pixels in each of the four directions, which preserves all pixels while breaking their spatial correspondence with $\mathcal{V}_1$. Table 13 reports the change relative to intact $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$ under default prompts. Reasoning scores never decrease at the reported precision, whereas perception changes are small.

Table 13: Score change (points) from shifting $\mathcal{V}_2$, relative to intact $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$ and averaged over four shift directions.

Model	Suite	7 px	14 px	28 px	Avg.
Qwen2.5-VL-32B	Reasoning	+0.4	0.0	+0.4	+0.3
	Perception	+0.2	−0.1	−0.1	0.0
Qwen2.5-VL-3B	Reasoning	+0.4	+0.4	+0.6	+0.5
	Perception	+0.4	+0.2	+0.1	+0.2
MiniCPM-o 4.5	Reasoning	+1.4	+1.3	+1.1	+1.3
	Perception	+0.6	+0.4	+0.2	+0.4

Prefill attention. We probe Qwen2.5-VL-32B on 100 LogicVista prompts in $\langle \mathcal{V}, \mathcal{T}, \mathcal{V} \rangle$, shifting only $\mathcal{V}_2$ to the right by 7, 14, or 28 processed pixels (half a ViT patch, one ViT patch, or one LLM visual token); the baseline uses an exact copy. For each visual query in $\mathcal{V}_2$, we sum its last-layer attention mass over keys in $\mathcal{V}_1$, $\mathcal{T}_1$, and $\mathcal{V}_2$ (causally accessible keys only), then average over queries and prompts. Masses are not renormalized over the three groups; other prompt tokens receive the remainder.

Table 14: Change in last-layer attention mass (percentage points) from $\mathcal{V}_2$ after rightward shifts, relative to the unshifted $\langle \mathcal{V}, \mathcal{T}, \mathcal{V} \rangle$ baseline (Qwen2.5-VL-32B, 100 LogicVista prompts).

Shift	$\mathcal{V}_2 \to \mathcal{V}_1$	$\mathcal{V}_2 \to \mathcal{T}_1$	$\mathcal{V}_2 \to \mathcal{V}_2$
7 px	−0.28	+0.12	+0.13
14 px	−0.30	+0.13	+0.13
28 px	−0.41	+0.19	+0.16

All three shifts move attention from $\mathcal{V}_1$ toward the question and $\mathcal{V}_2$ itself, and larger shifts produce larger changes (Table 14). Together with the generation gains, this supports the hypothesis, although the probe does not establish how the attention changes produce the score gains.

G REPETITION UNDER DIRECT AND DEFAULT PROMPTS: PER-BENCHMARK RESULTS

Tables 15 and 16 give the per-benchmark results behind Table 5; averages match that table after rounding. Within each pair, only the input changes, while the prompt template and decoding configuration stay fixed.

Text-only model. We evaluate Qwen2.5-32B on OpenBookQA (Mihaylov et al., 2018), GSM8K (Cobbe et al., 2021), MATH-500 (Hendrycks et al., 2021; Lightman et al., 2024), the text-only math split of OlympiadBench (He et al., 2024), and MMLU-Pro (Wang et al., 2024b), where $\langle \mathcal{T}, \mathcal{T} \rangle$ repeats the question verbatim. Repetition helps on every benchmark under both prompt

settings, but the average gain shrinks from $+3.09$ with direct-answer instructions to $+0.67$ with default prompts.

VLM. We evaluate Qwen2.5-VL-32B on the 12 main benchmarks, comparing $\langle \mathcal{V}, \mathcal{T} \rangle$ with $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$. The overall gain is $+1.3$ with direct-answer instructions and $+1.4$ with default prompts.

Table 15: Question repetition with Qwen2.5-32B: accuracy of $\langle \mathcal{T} \rangle$ and $\langle \mathcal{T}, \mathcal{T} \rangle$ under direct-answer and default prompts.

Prompt	**Input**	OpenBookQA	GSM8K	MATH-500	OlympiadBench	MMLU-Pro	**Avg.**
Direct	$\langle \mathcal{T} \rangle$	93.80	39.80	40.00	20.46	57.67	50.35
	$\langle \mathcal{T}, \mathcal{T} \rangle$	95.40	47.31	44.40	22.18	57.88	53.43
	Δ	+1.60	+7.51	+4.40	+1.72	+0.21	**+3.09**
Default	$\langle \mathcal{T} \rangle$	95.80	95.60	80.20	48.36	68.78	77.75
	$\langle \mathcal{T}, \mathcal{T} \rangle$	96.60	95.68	81.60	49.05	69.14	78.41
	Δ	+0.80	+0.08	+1.40	+0.69	+0.36	**+0.67**

Table 16: Image–question repetition with Qwen2.5-VL-32B. Shaded rows use $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$ instead of $\langle \mathcal{V}, \mathcal{T} \rangle$ under the prompt setting above; deltas on the averages are relative to $\langle \mathcal{V}, \mathcal{T} \rangle$.

(a) Reasoning

Prompt	Dyna Math	MATH-Vision	Logic Vista	Visual Puzzles	MMMU	MathVista	We-Math	**Avg.**
Direct	44.8	31.8	44.5	37.0	60.3	69.0	40.5	46.8
+SLP	49.0	33.0	47.0	38.8	61.7	70.3	43.0	**49.0** (+2.1)
Default	67.0	37.3	54.4	38.2	65.8	72.7	46.9	54.6
+SLP	67.5	38.8	57.7	40.8	67.7	74.9	49.4	**56.7** (+2.1)

(b) Perception and overall

Prompt	AI2D	OCRBench	SEED-Bench-2-Plus	MMBench	MMStar	**Avg.**	**Overall**
Direct	86.7	85.4	73.5	86.8	68.5	80.2	60.7
+SLP	87.3	85.3	73.3	87.3	68.9	**80.4** (+0.2)	**62.1** (+1.3)
Default	83.6	85.5	72.6	85.5	66.8	78.8	64.7
+SLP	84.4	85.8	71.8	86.1	67.7	**79.2** (+0.4)	**66.1** (+1.4)

H Controlled Continued Training: Full Setting and Results

H.1 Curriculum and Data Construction

LLaVA-OneVision trains in stages (Li et al., 2025): a single-image phase on 3.2M instructions produces the released SI checkpoint, and the subsequent OneVision phase uses a 1.6M mixture of single-image, multi-image, and video data whose roughly 800K single-image component is a balanced, high-quality subset of the single-image data. We initialize from the public LLaVA-OneVision-7B-SI checkpoint and train on this single-image subset, keeping only strict single-turn, single-image examples (729,599 in total).

The two runs use the same sample-ID groups and targets. The Standard run serializes each example as $\langle \mathcal{V}, \mathcal{T} \rangle$; the SLP run serializes it as $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$ with verbatim image and question copies. Both supervise the answer once. The runs use different offline packing caps to fit the longer SLP sequences but share sample assignments, optimizer steps, data seed, and optimization hyperparameters (Table 17).

Table 17: Training and evaluation settings for the controlled continued-training study.

Component	Setting
Initialization	`llava-hf/llava-onevision-qwen2-7b-si-hf`
Training data	729,599 unique strict single-turn, single-image examples; identical sample-ID groups in both runs
Presentations	Standard: $\langle \mathcal{V}, \mathcal{T} \rangle$; SLP: $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$ with exact image and question copies; one supervised answer in both runs
Schedule	5,000 optimizer steps; checkpoints and evaluation every 500 steps; seed and data seed 42
Optimization	fused AdamW; learning rate 10^{-5} for non-vision parameters and 2×10^{-6} for the vision tower; cosine decay; 3% warmup; weight decay 0; gradient norm clipped at 1; sample-token-mean loss normalization
Batch and packing	16 GPUs; one offline-packed microbatch per GPU; no gradient accumulation; 16 packed sequences and on average 145.92 examples per optimizer step; packing cap 32,075 tokens for Standard and 62,725 for SLP
Compute and numerics	2 nodes $\times$ 8 NVIDIA A100-SXM4-80GB; FSDP2 HSDP (shard size 8, replicate degree 2); BF16 compute, TF32 enabled, FP32 master optimizer state; padding removed for variable-length FlashAttention-2
Runtime	14.2 hours for Standard and 32.0 hours for SLP
Evaluation	vLLM 0.6.3 with the OneVision post-norm parity fix; BF16, eager mode, greedy decoding; maximum context 32,768 and maximum generation 2,048 tokens
Benchmarks	The 12 main benchmarks (MathVista on `testmini`); macro averages weight benchmarks equally

H.2 Checkpoint Trajectory

Table 18 lists the values plotted in Figure 5. Step 0 evaluates the shared initialization under each presentation; later steps evaluate each run under its training presentation. Relative to step 0, Standard gains 0.55 points and SLP gains 1.97. The SLP−Standard gap moves from -0.03 to $+0.94$ on the reasoning suite and from -1.81 to $+0.25$ on the perception suite.

Table 18: Matched-run trajectory (12-benchmark macro averages). Wins and ties count benchmarks on which SLP scores higher than and equal to Standard, respectively.

Step	Standard	SLP	SLP−Standard	Wins / ties
0	48.46	47.69	–0.77	5 / 1
500	48.17	48.25	+0.08	7 / 0
1,000	48.26	47.96	–0.30	4 / 0
1,500	47.34	47.57	+0.23	6 / 0
2,000	48.83	48.52	–0.31	4 / 0
2,500	48.72	49.32	+0.59	8 / 0
3,000	48.82	49.31	+0.49	8 / 1
3,500	48.94	49.67	+0.72	7 / 0
4,000	49.22	49.41	+0.19	6 / 0
4,500	49.02	49.45	+0.43	7 / 0
5,000	49.01	49.66	+0.65	8 / 2

H.3 Cross-Presentation Evaluation

The trajectory varies the training and inference presentations together. At step 5,000, we additionally evaluate each checkpoint under both presentations (Table 19; summarized in Table 6). Under $\langle \mathcal{V}, \mathcal{T} \rangle$ inference, the two checkpoints are tied (49.00 vs. 49.01); under $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$ inference, the SLP-trained checkpoint leads by 1.25 points. On 10 of 12 benchmarks, SLP training helps more under SLP inference than under Standard inference.

Table 19: Per-benchmark results at step 5,000 for each training (rows of Table 6) and inference presentation. Bold marks the highest macro average.

	Standard inference		SLP inference	
Benchmark	Std.-trained	SLP-trained	Std.-trained	SLP-trained
MATH-Vision	18.5	18.9	18.6	18.5
We-Math	21.6	23.6	20.0	22.8
OCRBench	69.7	69.1	71.1	70.7
AI2D	82.4	81.8	79.8	82.2
DynaMath	31.2	31.3	32.5	33.2
LogicVista	33.6	33.3	34.2	34.1
VisualPuzzles	26.4	29.0	29.0	28.3
MMMU	45.0	43.9	43.9	46.1
MathVista	58.3	57.8	56.3	58.3
MMBench	78.6	78.1	77.6	78.3
MMStar	57.3	56.5	54.7	58.1
SEED-Bench-2-Plus	65.5	64.8	63.2	65.6
Macro avg.	49.01	49.00	48.41	**49.66**

I Inference Latency and Generation Length

I.1 Per-Model Latency

Latency is measured with vLLM on 4×A800 GPUs with tensor parallelism 4. For each benchmark, we sample 10% of the examples uniformly at random, giving 2,610 examples per model, and evaluate $\langle \mathcal{V}, \mathcal{T} \rangle$ and $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$ on the same subset with default prompts. Table 20 reports mean per-example end-to-end latency, expanding the family averages in Table 7. Changes range from -0.49 to $+0.46$ seconds: five models become slower, three become faster, and one is unchanged at the reported precision. These measurements reflect default serving behavior; attributing the changes to prefill and decoding would require separate stage timings.

Table 20: Mean end-to-end latency (seconds) with default prompts. Δ is $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$ minus $\langle \mathcal{V}, \mathcal{T} \rangle$.

Model	$\langle \mathcal{V}, \mathcal{T} \rangle$	$\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$	Δ
Qwen2.5-VL-72B	5.01	5.47	+0.46
Qwen2.5-VL-32B	5.73	5.87	+0.14
Qwen2.5-VL-7B	0.95	1.00	+0.05
Qwen2.5-VL-3B	0.78	0.78	+0.00
MiniCPM-o 4.5	4.42	4.32	–0.10
MiniCPM-V 4.5	5.20	5.36	+0.16
Gemma-3-27B-IT	3.96	3.47	–0.49
Gemma-3-12B-IT	1.87	1.93	+0.06
Gemma-3-4B-IT	1.13	1.08	–0.05

I.2 Generation Length

Table 21 reports mean generation length in characters with default prompts, averaged over the 12 main benchmarks. SLP shortens the output on seven of nine models; the two increases (Qwen2.5-VL-72B, $+9$ characters; MiniCPM-V 4.5, $+83$) are below 2% of the $\langle \mathcal{V}, \mathcal{T} \rangle$ lengths. SLP's gains therefore do not come from longer generations. MiniCPM-V/o 4.5 lengths carry extra variance because their reasoning-suite runs use sampling (Appendix A.2).

Table 21: Mean generation length (characters) with default prompts, averaged over the 12 main benchmarks. Δ is $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$ minus $\langle \mathcal{V}, \mathcal{T} \rangle$.

Model	$\langle \mathcal{V}, \mathcal{T} \rangle$	$\langle \mathcal{V}, \mathcal{V}, \mathcal{T} \rangle$	$\langle \mathcal{V}, \mathcal{T}, \mathcal{T} \rangle$	$\langle \mathcal{V}, \mathcal{T}, \mathcal{V} \rangle$	$\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$	Δ
Qwen2.5-VL-72B	562	550	592	573	571	+9
Qwen2.5-VL-32B	1176	972	1232	1101	1166	−10
Qwen2.5-VL-7B	642	568	617	590	570	−72
Qwen2.5-VL-3B	493	440	468	482	466	−27
MiniCPM-o 4.5	5494	5432	5518	5501	5465	−29
MiniCPM-V 4.5	8013	8021	7698	8034	8096	+83
Gemma-3-27B-IT	379	378	375	331	353	−26
Gemma-3-12B-IT	389	383	362	358	320	−69
Gemma-3-4B-IT	432	417	340	358	354	−78

J Duplicate-Image Stress Tests

SLP presents the image as two separate visual-token blocks, which could lead a model to count instances once per copy, confuse spatial relations across copies, or emit less reliable box coordinates. We test these risks on CountQA (Tamarapalli et al., 2025), SpatialMQA (Liu et al., 2025), and RefCOCOg (Mao et al., 2016), reporting accuracy for counting and spatial relations and Acc@0.5 for grounding (Table 22).

Counting and spatial relations. Counting changes are mixed: five models improve, four decline, and the mean change is -0.12 points, so repetition shows no consistent degradation. SLP improves SpatialMQA accuracy on eight of nine models, by $+0.54$ points on average.

Grounding. RefCOCOg is the only task with a consistent decline. Five of the six Qwen2.5-VL and MiniCPM-V/o 4.5 models decline, with changes from -2.70 to $+1.00$ points and a mean of -1.11. Gemma-3-IT rarely emits valid boxes and is excluded from this average. We list this decline as a limitation in Section 5.

Table 22: Duplicate-image stress tests. Std. is $\langle \mathcal{V}, \mathcal{T} \rangle$, SLP is $\langle \mathcal{V}, \mathcal{T}, \mathcal{V}, \mathcal{T} \rangle$, and Δ is their difference in points. Gemma-3-IT rarely emits valid boxes, so its RefCOCOg results (gray) are excluded from that average.

	CountQA			**SpatialMQA**			**RefCOCOg**		
Model	Std.	SLP	Δ	Std.	SLP	Δ	Std.	SLP	Δ
Qwen2.5-VL-72B	31.48	32.00	+0.52	43.87	45.72	+1.86	83.02	84.02	+1.00
Qwen2.5-VL-32B	24.54	23.69	−0.85	43.68	44.61	+0.93	87.34	86.80	−0.53
Qwen2.5-VL-7B	21.20	20.55	−0.65	37.36	38.20	+0.84	86.26	85.07	−1.20
Qwen2.5-VL-3B	15.31	14.20	−1.11	34.67	35.97	+1.30	83.21	81.50	−1.71
MiniCPM-o 4.5	17.28	17.80	+0.52	38.48	40.06	+1.58	35.12	33.60	−1.52
MiniCPM-V 4.5	28.14	25.65	−2.49	40.33	40.52	+0.19	78.42	75.72	−2.70
Gemma-3-27B-IT	18.46	19.24	+0.79	38.85	38.94	+0.09	0.06	0.49	+0.43
Gemma-3-12B-IT	17.28	18.59	+1.31	37.45	38.66	+1.21	0.00	0.02	+0.02
Gemma-3-4B-IT	14.86	15.71	+0.85	31.23	28.07	−3.16	0.00	0.00	+0.00
Avg.	20.95	20.83	−0.12	38.44	38.97	+0.54	75.56	74.45	−1.11