WHEN DOES CHAIN-OF-THOUGHT HURT SPATIAL ANSWERS? LENGTH, RETRIEVAL, AND CONTENT CONTROLS IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Chain-of-thought (CoT) prompting often lowers the spatial accuracy of vision-language models (VLMs). In a decoder-only VLM, reasoning text cannot change the image or question states, so it can act only through how the answer position retrieves them or through what the reasoning says. We separate these routes on 2850 spatial items (CV-Bench and two synthetic tasks) in 5 open VLMs of at most 8B parameters with length, turn-structure, re-injection, regeneration, and content controls, reading every answer as one forced-choice token. CoT changes accuracy by -5.7 points on average, whereas length-matched neutral filler in a separate turn changes it by -0.6 and 1024 filler tokens by -2.4; restating the question or re-supplying the image after the reasoning recovers at most +0.8 and +1.0 points. The answer agrees with the option the reasoning states, and in four models pooled CoT accuracy is within 1.7 points of $\bar{p}$, the Direct distribution's mean probability of the correct option: CoT scores like a sample from that distribution rather than its argmax. Turn-matched filler keeps Direct's argmax step at a correct-option probability of 0.5 while CoT flattens it, and the same probabilities predict four-sample majority-vote accuracy within 1.4 points. The CoT effect thus decomposes into the argmax–expectation gap $G = \text{Direct} - \bar{p}$, which CoT gives up, and the gain $E = \text{CoT} - \bar{p}$ beyond the expectation. In nine larger or closed VLMs queried through an API (reasoning disabled where the API allows it; GPT-5 models still reason in part of their direct calls), G is at most +2.9 points, E reaches +11.2, and CoT changes accuracy by -0.7 to +11.7 points.

1 INTRODUCTION

Chain-of-thought (CoT) prompting (Wei et al., 2022) often lowers the accuracy of vision-language models (VLMs) on spatial questions. Kancheti et al. (2026) report that 7 of 8 reasoning-tuned models fail to beat their backbones across 13 spatial benchmarks, and Qwen2.5-VL-7B loses accuracy under CoT on 10 of the 12 benchmarks they list (Appendix Table 14); CoT prompting also fails to improve VSI-Bench (Yang et al., 2025). Two explanations are common. The *semantic* explanation says that the generated text states something wrong and the answer follows it. The *visual-forgetting* explanation says that attention to image tokens decays as the reasoning grows; methods built on it re-inject the image during reasoning (Sun et al., 2025) or train with rewards for visual grounding (Tian et al., 2026; Chen et al., 2026). The two explanations call for different fixes, but they are rarely tested against each other with controls that hold the image, the question, and the answer format fixed.

We start from a constraint that every explanation must satisfy. In a decoder-only VLM, every token generated after the question is invisible to the image and question tokens: their hidden states, and hence their key–value cache, are identical whether the model answers directly or after 500 tokens of reasoning. Intermediate text therefore cannot erase spatial information stored at those positions, including the spatial IDs that VLMs bind to object-word tokens (Kang et al., 2026). It can only change (a) how the answer position retrieves stored information and (b) the content that the answer position reads from the text itself. Visual forgetting is one form of (a), with the image tokens as the source. A second form of (a), which we call *retrieval distance*, puts the source at the question and object-word tokens: the answer position has to read the bound relation back across the intermediate text, and more text makes that read-out harder regardless of what the text says. The semantic explanation is (b).

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.

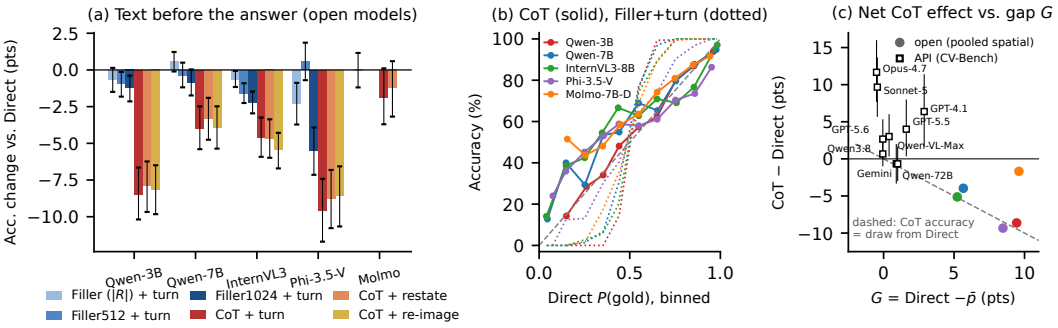


Figure 1: (a) Change in pooled spatial accuracy relative to Direct for turn-matched conditions in the five open models (2850 items per model; paired cluster-bootstrap 95% CIs; Molmo has no Filler512/1024 or re-image run). Neutral filler of the CoT's length, 512, or 1024 tokens changes accuracy by a few points at most; CoT lowers it, and restating the question or re-supplying the image after the CoT does not undo the loss. (b) Accuracy after CoT (solid) and after turn-matched filler (dotted) against Direct's probability of the gold option, binned (bins with $\geq$20 items; dashed: identity). Filler keeps Direct's argmax step at 0.5; CoT flattens it. (c) Net CoT effect against $G = \text{Direct} - \bar{p}$ in 14 models (open: pooled spatial set; API: 298–300 CV-Bench items, $\bar{p}$ from six samples, 95% CIs). On the dashed line CoT scores like a draw from the Direct distribution; height above it is $E = \text{CoT} - \bar{p}$.

The three accounts make opposing predictions for a small factorial (Table 1), which we run on 5 open VLMs over 2850 spatial items per model from CV-Bench (Tong et al., 2024) and two synthetic tasks, with CV-Bench Count and MMStar (Chen et al., 2024) as controls, reading every answer as one forced-choice token after the same answer prefix (Section 3). CoT changes pooled spatial accuracy by -5.7 points on average (range -9.3 to -1.7); turn-matched neutral filler of the same length changes it by -0.6 and 1024 filler tokens by -2.4 (Figure 1a); restating the question or re-supplying the image after the CoT recovers +0.5 and +0.1 points on average, and regenerating R after a second copy of the image does not make it more often right. The answer instead agrees with the option that R states, and in four models pooled CoT accuracy is within 1.7 points of $\bar{p}$, the accuracy expected from sampling the Direct distribution: CoT flattens Direct's argmax step (Figure 1b), so the loss is close to the argmax–expectation gap $G = \text{Direct} - \bar{p}$. Writing the net effect as $\text{CoT} - \text{Direct} = E - G$ with $E = \text{CoT} - \bar{p}$ extends the account to nine larger or closed models, in which G is small and E is near zero or positive, so CoT is neutral or helps (Figure 1c).

Inside Qwen2.5-VL (Section 5), all tested inserted texts re-encode the answer at mid-depth and draw attention away from the question and the image, but for neutral filler neither change alters the answer; patching localizes where the answer-position state becomes causally decisive (Figure 2). In conclusion edits, Qwen-3B and Phi-3.5-V follow a swapped conclusion of R, whereas Qwen-7B and InternVL3-8B follow it less often (Section 6). Our setting is short reasoning ($|R|$ of 51–131 tokens on average) in open models of at most 8B parameters, with a reduced protocol on nine larger or closed models; long reasoning is not studied.

Contributions. (i) A population-level regularity: on spatial items in four of five open VLMs, pooled CoT accuracy matches $\bar{p}$, CoT flattens the argmax step that turn-matched filler preserves, and $\text{Bin}(4, p)$ predicts majority-vote accuracy within 1.4 points, whereas a constant-$\bar{p}$ model overpredicts it by 3.7–6.6 (Section 4.4); (ii) the decomposition $\text{CoT} - \text{Direct} = E - G$ across 14 VLMs, in which G falls from 12.6 points (Qwen2.5-VL-3B) to at most +2.9 in the API models and E is positive in six of nine API models (Section 4.6); (iii) a causal-mask argument and a factorial with length-matched and 512/1024-token filler, turn-structure controls, re-injection, and regeneration that disfavor content-independent length and retrieval failure as sufficient explanations of the short-CoT loss (Sections 2–4); (iv) conclusion edits, probes, attention, and patching in Qwen2.5-VL (Sections 5–6).

Table 1: Predictions at the answer position and observations. R is the model's own CoT text, held fixed across conditions. "Turn-only" closes the assistant turn and adds a user turn with the final instruction alone, which controls for the turn structure of restate and re-image. Bold: predictions consistent with the observation; the accounts are not exclusive, and the rows do not identify a unique route. The H_{vis} and H_{ret} predictions assume that retrieval fails at the answer position; retrieval failures while R is generated are not tested by these rows.

Measurement	H_{sem} semantic	H_{vis} vis. forgetting	H_{ret} retr. distance	Observed (mean over models)
Filler (length of R, own turn) vs. Direct	**no loss**	loss	loss $\approx$ CoT	-0.6 pts (CoT: -5.7)
Filler 512 / 1024 tok., own turn, vs. Direct	**small**	loss, may grow	loss, may grow	-0.6 / -2.4 pts (four models)
Restate vs. turn-only, after CoT	**no gain**	small gain	large gain	+0.5 pts (max +0.8)
Re-image vs. turn-only, after CoT	**no gain**	large gain	gain $\leq$ restate	+0.1 pts (max +1.0)
CoT accuracy where R states a wrong option	**follows R (near 0)**	–	–	1.6% (Direct: 58.8%)
Answers following a swapped conclusion of R	follows	–	–	91.3 / 99.8% (Qwen-3B, Phi); 54.0 / 28.5% (Qwen-7B, InternVL)

2 What intermediate text can and cannot change

The causal-mask constraint. Write the input as (v, q, c, s, a): image tokens v, question tokens q, intermediate text c (empty, CoT, or filler), an optional extra user turn s, and the answer position a. With a causal mask, the hidden state $h_\ell(t)$ of token t at layer ℓ depends only on tokens at or before t. Hence $h_\ell(v)$ and $h_\ell(q)$, and the keys and values computed from them, are the same for every choice of c and s at every layer. Whatever spatial information the model has written into image or question positions, including spatial IDs on object-word tokens (Kang et al., 2026), is still there when the answer is produced. The answer logits are a function of $h_L(a)$, which differs across conditions only through what a attends to. Intermediate text can therefore act through two routes: (a) *retrieval*, by changing how attention at a (and at tokens of c and s that a reads) weights the unchanged states of v and q; and (b) *content*, by writing information into the states of c that a then reads. Route (b) requires c to carry task-relevant content, which neutral filler does not. This is a functional decomposition, not a causal identification: content can also change how later positions retrieve v and q. The controls test whether content-independent length, distance, or re-supplied inputs suffice to explain the loss; the content edits test whether changing the stated conclusion changes the answer.

Three hypotheses. H_{sem} (semantic interference) places the loss in route (b): the reasoning text states or implies a wrong relation or a wrong option, and the answer follows it. H_{vis} (visual forgetting; Sun et al., 2025; Tian et al., 2026; Chen et al., 2026) places it in route (a) with the image tokens as the source: as c grows, the answer position attends less to v. H_{ret} (retrieval distance) also places it in route (a), with the question and object-word tokens as the source: the relation is bound to those tokens, and text between them and the answer position disrupts its read-out, whatever the text says. H_{vis} and H_{ret} are not exclusive but name different sources, so they respond differently to a fresh copy of the image and a fresh copy of the question placed next to a. Both could also act while R is generated (tested in Section 4.5). Table 1 lists the predictions and the observed values from Sections 4 and 6.

3 Experimental setup

Models and data. Qwen2.5-VL-3B and -7B (Bai et al., 2025), InternVL3-8B (Zhu et al., 2025), Phi-3.5-Vision (Abdin et al., 2024), and Molmo-7B-D (Deitke et al., 2024), four model families, run with vLLM (Kwon et al., 2023) in bfloat16 with greedy decoding. The spatial pool has 2850 items per model: CV-Bench (Tong et al., 2024) 2D Relation (650), 3D Depth (600), and 3D Distance (600), and two balanced synthetic yes/no sets: relative position (250 images of two primitives, each asked with its true relation among left/right/above/below and with the opposite one; object centres are at

least 72 px apart along the queried axis, median 201 px) and depth ordering (500 images in which the front object is drawn on top of, and larger than, the partly occluded back object; 259 yes, 241 no). Controls are CV-Bench 2D Count (788; visual but not relational) and 500 random MMStar (Chen et al., 2024) items (general multimodal questions). Exclusions (InternVL2.5-8B, a first Molmo run with a wrong prompt wrapper, and Qwen2.5-VL-32B) are listed in Appendix G.

Conditions. Every condition shares the same first user turn (image, then question with lettered options). The text before the answer is one of: *Direct* (none); *CoT*, the model's own greedy continuation of "Let's think step by step." (at most 512 new tokens), denoted R; *Filler*, neutral prose on the history of paper making with no relational or positional words, truncated to the token length of R under the model's tokenizer; and *Filler512* and *Filler1024*, the same prose at fixed length. R and the filler occupy the assistant turn. The assistant turn either continues directly with the answer prefix, or is closed and followed by a second user turn: *turn* ("Now give your final answer."), *restate* ("Question again: ⟨question⟩" and the instruction), *re-image* (the image again, "Here is the image again." and the instruction), or *re-image+restate*. *Direct+restate* states the question twice in the first turn. Molmo accepts one image and a 4096-token context, so it has no re-image or Filler1024 conditions. All prompts and the twelve filler sentences are given verbatim in Appendix B.

Answer readout. In every condition the final assistant message ends with "Answer: (" and we read one next token. The prediction is the option letter with the highest log-probability among the top 20 tokens, and $P(\text{gold})$ is the gold letter's probability renormalized over the option letters. Items with no option letter among the top 20 are rare (Appendix B). We also parse the option that R itself states (the free-text answer, *CoT (free)*); Section 4.4 uses it to split items.

Statistics. Δ denotes a condition minus its reference in accuracy points. Confidence intervals use a paired cluster bootstrap over image groups (2000 resamples); p-values come from a paired cluster sign-flip randomization test over image groups (5000 permutations); tasks are treated as fixed. On CV-Bench each image appears once, so clustering matters only for the synthetic sets. Holm correction covers 15 tests: three pre-declared contrasts (CoT vs. Direct, Filler512 vs. Direct, CoT+restate vs. CoT+turn) for five models. Exact McNemar tests are reported in the released outputs. Cross-model statements give the mean (and where relevant the range) over the models that have the condition.

4 Behaviour: what the text before the answer does

Table 2 gives pooled spatial accuracy for every condition with the paired change relative to the reference in the last column; per-task accuracies are in Appendix Table 6.

4.1 CoT lowers spatial accuracy

Relative to Direct, the model's own CoT changes pooled spatial accuracy by -8.6 points in Qwen-3B (95% CI [-10.4, -6.9], $p < 0.001$), -4.0 in Qwen-7B, -5.1 in InternVL3-8B, -9.3 in Phi-3.5-V, and -1.7 in Molmo-7B-D; the mean is -5.7, and the loss is significant in four of the five models after Holm correction (Holm-adjusted $p = 0.003$ in each of the four, the floor of the permutation test; Molmo-7B-D $p = 0.65$). In Molmo-7B-D the gain on items where R is right nearly offsets the loss where it is wrong (Section 4.4). Closing the assistant turn after R does not change this (CoT+turn vs. CoT: +0.0 points on average). Where R states a parsable option (57.4–100.0% of items; Phi-3.5-V often ends without naming a letter), the forced answer after R agrees with it on 97.6–100.0% of items, so on these items the forced read-out measures the same decision as free-form answering.

4.2 Neutral text costs little once the turn structure is controlled

H_{vis} and H_{ret} predict that neutral filler of the same length as R costs accuracy of the order of the CoT loss. Placed in the same assistant turn directly before "Answer: (", length-matched filler changes accuracy by -4.5 points on average (range -9.8 to -1.9). This loss disappears when the assistant turn is closed and an instruction-only user turn follows: Filler+turn vs. Filler is +3.9 points on average (Qwen-7B: +10.4, $p < 0.001$), and Filler+turn vs. Direct is -0.6 (range -2.3 to +0.6). Part of this same-turn loss is a shift of the prior over option letters (Appendix Table 10).

Table 2: Forced-choice accuracy (%) on the pooled spatial set (2850 items per model). In parentheses: change in points relative to the condition in the "ref." column; *: $p < 0.001$ (paired cluster sign-flip test, uncorrected). "Turn": the assistant turn is closed and a user turn with the final instruction follows. Direct+restate repeats the question inside the first user turn. Molmo has one image slot and a 4096-token context (–).

Condition	Qwen2.5-VL-3B	Qwen2.5-VL-7B	InternVL3-8B	Phi-3.5-V	Molmo-7B-D	ref.
Direct	80.2	86.0	89.7	76.7	74.5	
CoT (own text R)	71.6 (-8.6*)	82.1 (-4.0*)	84.6 (-5.1*)	67.4 (-9.3*)	72.8 (-1.7)	direct
Filler, length of R	78.3 (-1.9)	76.2 (-9.8*)	87.3 (-2.4*)	70.4 (-6.4*)	72.6 (-1.9)	direct
Filler, 512 tok.	80.0 (-0.2)	85.4 (-0.6)	88.0 (-1.6*)	76.7 (-0.0)	74.9 (+0.4)	direct
Filler, 1024 tok.	78.3 (-1.9*)	85.2 (-0.8)	88.0 (-1.7*)	75.3 (-1.5)	–	direct
Direct + restate (in turn 1)	76.2 (-4.0*)	85.3 (-0.7)	89.2 (-0.5)	74.8 (-1.9)	77.1 (+2.6)	direct
CoT + turn	71.8 (+0.2)	82.1 (+0.0)	85.1 (+0.5)	67.2 (-0.2)	72.6 (-0.2)	cot
CoT + restate	72.4 (+0.6)	82.7 (+0.6)	85.0 (-0.1)	68.0 (+0.8)	73.3 (+0.6)	cot + turn
CoT + re-image	72.1 (+0.3)	82.1 (+0.0)	84.2 (-0.8)	68.2 (+1.0*)	–	cot + turn
CoT + re-image + restate	73.2 (+1.5*)	83.1 (+1.0*)	85.8 (+1.3)	68.1 (+0.7)	–	cot
Filler + turn	79.6 (+1.3)	86.6 (+10.4*)	89.1 (+1.8*)	74.4 (+4.1*)	74.4 (+1.9*)	filler
Filler + restate	74.9 (-4.7*)	85.7 (-0.9)	89.3 (+0.3)	75.0 (+0.6)	72.6 (-1.8)	filler + turn
Filler + re-image	79.0 (+0.7)	86.1 (+9.9*)	86.7 (-0.6)	72.4 (+2.0)	–	filler
Filler512 + restate	75.2 (-4.8*)	85.8 (+0.4)	89.3 (+1.3)	73.1 (-3.6*)	67.9 (-7.0*)	filler512
Filler512 + re-image	77.8 (-2.2*)	85.1 (-0.3)	86.8 (-1.3)	72.7 (-4.0*)	–	filler512

Table 3: Pooled spatial accuracy (%) split by whether the CoT text R states the correct or a wrong option (items without a parsable option omitted; full split with the edited-R conditions: Appendix Table 5). Molmo has no re-image condition.

	n items		CoT acc.		acc. when R is wrong	
Model	R right	R wrong	R right	R wrong	Direct	CoT+re-image
Qwen2.5-VL-3B	1979	753	99.4	0.7	56.0	1.2
Qwen2.5-VL-7B	2339	509	100.0	0.0	55.4	1.0
InternVL3-8B	2402	448	99.8	2.7	62.1	26.1
Phi-3.5-V	1149	486	98.2	3.7	66.0	2.7
Molmo-7B-D	2063	767	99.7	1.0	54.6	–

4.3 Length effects are small

Mean $|R|$ is 51–131 tokens, so 512 and 1024 filler tokens are several times longer than R. With the turn structure matched, they change accuracy by -1.6 to +0.6 and -5.5 to -0.8 points (four models; Figure 1a); 1024 tokens cost more than 512 in all four, so length does cost some accuracy, but the largest loss (Phi-3.5-V at 1024 tokens, 20 times its mean $|R|$) reaches 59% of its CoT loss, and no filler condition approaches the CoT loss at the CoT's own length (same-turn values: Table 2; quartiles of $|R|$: Appendix Table 8).

4.4 The answer agrees with the option the CoT states

Table 3 splits the pooled items by the option that R states. In every model the answer after CoT is almost always correct when R states the correct option and almost never when it states a wrong one, although Direct answers 54.6–66.0% of the latter correctly (chance is 50%); the same holds on CV-Bench Count and MMStar (Appendix Table 9). The net CoT loss is the balance of a gain where R is right and a larger loss where R is wrong, and these items are not spoiled by inserted text as such: under Filler+turn, Qwen-3B answers 56.8% of them correctly. The split is descriptive (R's correctness is a post-treatment variable); the discriminating evidence is the filler, re-injection, and regeneration results. It also does not show why R is wrong on items that Direct answers correctly.

CoT departs from Direct's argmax in a probability-dependent way. Let p be Direct's probability of the gold option and $\bar{p}$ its mean, the accuracy expected from sampling the Direct distribution. In four of five models, pooled accuracy after CoT lies within 1.7 points of $\bar{p}$ (Qwen-3B 71.6% vs. 70.8%, difference CI [-0.7, +2.2]; the CI includes zero in three, Qwen-7B [+0.5, +3.0]); equivalently, the CoT loss is close to the argmax–expectation gap (5.2–9.4 points). Molmo-7B-D exceeds $\bar{p}$ by

+7.9 (cause not identified). The contrast with filler separates this from generic noise: turn-matched filler keeps Direct's argmax step at p=0.5 (Qwen-3B: 14% for $p \in [0.4, 0.5)$ vs. 76% for $[0.5, 0.6)$), whereas CoT flattens it (48% vs. 57%); across models the jump across p=0.5 is -4 to +14 points under CoT and +24 to +62 under filler (Figure 1b; per-model curves in Appendix Figure 3). Item-level agreement with p holds in Qwen-3B (logistic slope of CoT correctness on logit p 0.96, intercept 0.08); in Qwen-7B, InternVL3-8B, and Phi-3.5-V the curve is flatter (slope 0.47–0.68), so CoT beats p where Direct is wrong and falls below it where Direct is confident, and the deviations offset in the mean. CV-Bench Count agrees with $\bar{p}$ in three of five models; on MMStar, CoT exceeds $\bar{p}$ by +1.3 to +8.2 points (CI above zero in 3 of five). The account also predicts self-consistency: if each sampled R states the gold option with probability p, a majority vote over four samples (ties split) reaches $\mathbb{E}_p[P(\mathrm{Bin}(4,p) \geq 3) + \frac{1}{2}P(\mathrm{Bin}(4,p) = 2)]$. Computed from Direct's probabilities alone, this matches the observed majority-vote accuracy on the within-item samples within 1.4 points in all four sampled models (Qwen-3B 74.5% vs. 75.5%; Qwen-7B 83.1% vs. 84.5%, CI [+0.4, +2.4]); a constant-$\bar{p}$ model overpredicts it by 3.7–6.6 points, so the prediction uses item-level information; voting gains +2.8 to +4.0 points over a single sample and, as predicted for four votes, stays below Direct (Qwen-3B 80.2%).

Within-item test. We sample 4 CoT texts per spatial item at temperature 0.7 and read the forced answer after each. On items whose samples include both a correct and a wrong stated option (889–1427 items per model), accuracy after a sample that states the correct option is 96.3–99.8% and after one that states a wrong option 0.3–5.5% (within-item difference +91.0 to +99.5 points, four models; Qwen-3B 95% CI [+95.2, +96.9]), while Direct answers 72.2–77.5% of the same items correctly.

Task effects are heterogeneous: CoT is below turn-matched filler in 18 of 25 model–task cells (CI below zero in 15) and reliably above it in 2 (Appendix Table 7); on the controls CoT changes CV-Bench Count by -1.1 and MMStar by -0.0 points on average.

4.5 Distance and visual access while R is generated

Visual forgetting or retrieval distance could act while R is generated and make R wrong, which the answer then copies. We therefore regenerate R with greedy decoding under two changes and read the forced answer after it (four models). Supplying the image a second time directly after the question, so that fresh image tokens precede the reasoning, changes final accuracy by -0.7 to +0.6 points and the share of items on which R states a wrong option by -1.8 to +0.4 points relative to the original R. Putting 512 filler tokens and a re-supplied image between the question and the reasoning changes accuracy by -3.0 to -1.0 points. A second copy of the image before the reasoning thus does not make R more often right, and 512 tokens between the question and the reasoning (with the image re-supplied) cost less than the CoT loss in each model (Appendix Table 11).

4.6 Larger and closed models

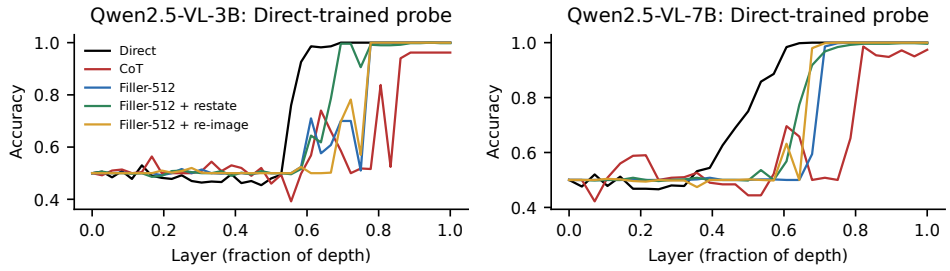
Through an API we run a reduced protocol on the same CV-Bench items for nine larger or closed VLMs (Table 4; details and filler results in Appendix I, Table 13), estimating $\bar{p}$ from six Direct answers at temperature 1, so for these models G is a greedy-minus-sampled gap. Reasoning is disabled in Direct calls where the API allows it; logged Direct calls reason in at most 1% of cases except in GPT-5.5 and GPT-5.6-Sol, where `reasoning_effort=none` is not enforced. In all nine models, CoT fails on every item whose R states a wrong option (6–29 per model). Within Qwen2.5-VL, G falls from 3B to 72B together with the CoT loss; across the API models G stays at or below +2.9 points, and E is within noise in three models and positive in six, including both Claude models, whose Direct calls never reason, and GPT-5.5 on Depth, where only 1% of its Direct calls reason ($E = +8.2$). Filler in the final-answer turn is not neutral for the API models (GPT-4.1, GPT-5.5, GPT-5.6-Sol: -7.3 to -5.7 points), so the filler controls hold for the open models only.

5 Inside Qwen2.5-VL: representation shifts and where the answer becomes fixed

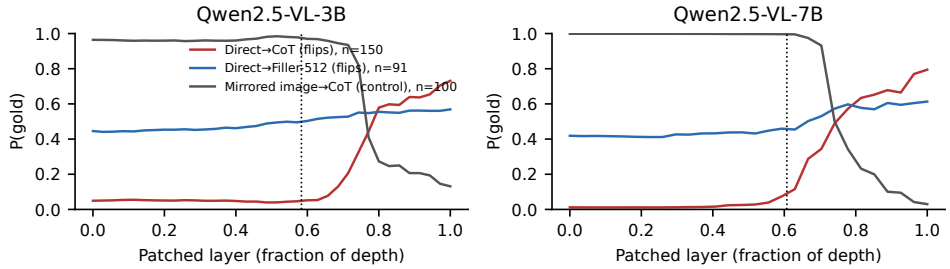
We analyse Qwen2.5-VL-3B (36 blocks) and -7B (28 blocks) with HuggingFace Transformers, using the behavioural prompts, image resolution, and forced-answer readout. Items are the 500 synthetic

Table 4: Decomposition of the CoT effect on 298–300 CV-Bench spatial items (accuracy in %, differences in points; brackets: 95% bootstrap CIs over items). G = Direct $-\bar{p}$, E = CoT $-\bar{p}$, and CoT−Direct $= E - G$; CoT is read after a final-answer turn. R_D: share of logged Direct calls that contain reasoning (–: not logged). †Main-run outputs on the same items with $\bar{p}$ from exact probabilities; the API models estimate $\bar{p}$ from six samples.

Model	R_D (%)	Direct	CoT	$\bar{p}$	G	$E = \text{CoT} - \bar{p}$	CoT−Direct
Qwen2.5-VL-3B†	0	77.3	66.0	64.7	+12.6	+1.3 [-3.6, +6.1]	−11.3 [-16.7, -5.7]
Qwen2.5-VL-7B†	0	87.3	84.3	78.5	+8.8	+5.8 [+2.0, +9.4]	−3.0 [-7.3, +1.3]
Qwen2.5-VL-72B	–	95.3	94.7	94.4	+0.9	+0.2 [-2.2, +2.8]	−0.7 [-3.3, +2.0]
Qwen3.8-Max	1	95.7	96.3	95.7	−0.1	+0.6 [-0.8, +2.3]	+0.7 [-1.0, +2.7]
Qwen-VL-Max	0	95.0	98.0	94.6	+0.4	+3.4 [+0.9, +6.1]	+3.0 [+0.3, +6.0]
Gemini-3.8-Flash	0	94.3	93.7	93.3	+1.0	+0.3 [-1.6, +2.3]	−0.7 [-3.0, +1.7]
GPT-4.1	–	83.9	90.3	81.0	+2.9	+9.3 [+5.6, +12.8]	+6.4 [+1.7, +11.4]
GPT-5.5	56	89.0	93.0	87.4	+1.6	+5.6 [+2.6, +8.9]	+4.0 [+0.3, +8.0]
GPT-5.6-Sol	72	90.7	93.3	90.7	−0.1	+2.6 [+0.2, +5.0]	+2.7 [+0.0, +5.3]
Claude-Sonnet-5	0	82.0	91.7	82.4	−0.4	+9.2 [+5.7, +12.8]	+9.7 [+5.7, +13.7]
Claude-Opus-4.7	0	82.0	93.7	82.5	−0.5	+11.2 [+7.5, +15.0]	+11.7 [+7.7, +16.0]



(a) Probe trained on Direct states, tested on held-out images of each condition (synthetic relative position)



(b) $P(\text{gold})$ after patching the Direct residual into the answer position of the target run

Figure 2: Inside Qwen2.5-VL (x-axis: fraction of depth). (a) Every inserted text moves the answer representation off the Direct direction at mid-depth, including filler that leaves answers unchanged; later layers bring it back. (b) Patching flips CoT errors (red) only in the second half of the network; Filler512 flips (blue) are near-threshold items; grey: mirrored-image control on correct CoT runs; dotted: first readable layer.

relative-position questions (250 images, each asked with its true relation and the opposite one) and the 650 CV-Bench Relation questions. Conditions are Direct, CoT, Filler, Filler512, and their restate or re-image variants (Appendix D). For probes and attention, layer ℓ is the residual stream after ℓ blocks (`hidden_states[`ℓ`]`) at the answer position, the token after "Answer: (". Patching uses 0-based block indices b; block b outputs layer $\ell{=}b{+}1$.

5.1 PROBES AND ATTENTION: SHIFTS THAT DO NOT CHANGE ANSWERS

At every layer we fit logistic-regression probes with 5-fold cross-validation grouped by image (Appendix D). As in a preliminary fixed-prefix study (Appendix Table 12, Figure 5), a *fresh* probe is trained and tested within a condition; a *transfer* probe is trained on Direct states and tested on held-out images of another condition. The first readable layer ℓ^* is the first layer at which the Direct fresh probe reaches 0.9 (layer 21 of 36 on synthetic relative position in Qwen-3B). At ℓ^* the fresh probe reads the label under every condition, but the Direct probe transfers at chance under every inserted text (0.50 under CoT, 0.50 under Filler, 0.52 under Filler512) and returns near behavioural accuracy by the last layer (Figure 2a; this set is near ceiling); on synthetic depth ordering, where CoT does lower accuracy, the last-layer read-out separates CoT from filler while the mid-depth shift does not (Appendix D). Attention from the answer position to the image falls from 0.030 under Direct to 0.015 under CoT and 0.016 under Filler, and attention to the object words falls under every inserted text; R receives about twice the attention that filler does (0.241 vs. 0.115). Inserted text of any kind thus re-encodes the answer at mid-depth and draws attention away from the image and the question, but for filler neither change alters the answer, which is inconsistent, at the answer position, with the attention predictions of H_{vis} and H_{ret} (attention is not causal; Appendix Figure 4).

5.2 ACTIVATION PATCHING

We select items that Direct answers correctly and a target condition answers wrongly in the behavioural (vLLM) runs, over the five spatial sets (at most 150 per target, shuffled), run the target condition, and replace the residual stream at the answer position after block b with the Direct run's residual at the same block. By the causal mask this changes nothing at earlier positions; later blocks at the answer position can still read the target context. As a control for the patch site, on correct CoT runs of left/right questions we patch in the Direct residual computed on the horizontally mirrored image, whose correct answer is the opposite letter. Flip confirmation and final-norm details are in Appendix E.

For CoT targets in Qwen-3B, the unpatched $P(\text{gold})$ is 0.05; patching leaves it unchanged up to the middle of the network, it starts to rise at block 24 of 0–35, and reaches 0.70 at block $L-2$; the mirrored-image control flips correct CoT answers in the same blocks (majority at block 27), and Qwen-7B restores from block 16 of 0–27 (Figure 2b). Filler512 flips are near-threshold items (unpatched $P(\text{gold})$ 0.44), whereas CoT flips are confident.

Patching thus localizes where the answer-position state becomes causally decisive, from 69% (Qwen-3B) and 59% (Qwen-7B) of depth (Figure 2b); it does not show what later blocks read, and its onset does not separate the hypotheses.

6 WHAT DOES NOT REMOVE THE LOSS, AND WHAT THAT IMPLIES

Restating the question and re-supplying the image. Both interventions follow from route (a): re-supplying the image targets H_{vis}, as in methods that re-inject visual tokens during reasoning (Sun et al., 2025), and restating the question puts the object words next to the answer position, as H_{ret} requires. Relative to CoT+turn, which has the same turn structure, CoT+restate changes pooled spatial accuracy by +0.5 points on average (range -0.1 to +0.8; Holm-adjusted $p = 0.042$ in Phi-3.5-V and larger in the other models), and CoT+re-image by +0.1 (range -0.8 to +1.0); both together change it by +1.1 relative to CoT (Table 2). On items where R states a wrong option, the two interventions leave accuracy at 5.3% and 7.8% on average (Table 5). The one sizeable partial recovery is InternVL3-8B with a re-supplied image on items where R states a wrong option (26.1%, vs. 2.7% after CoT and 62.1% for Direct); its pooled effect is -0.8 points because re-imaging also flips items on which R was right. This is what H_{sem} predicts: a second copy of the image or the question does not change what R concludes, and the answer position keeps reading that conclusion. The same re-prompts are not neutral without CoT and lower accuracy in some models (Appendix C.1).

Editing the content of R. The split in Section 4.4 leaves open whether the answer depends on the lettered conclusion of R or on the text before it. Cutting R before its last lettered sentence changes accuracy by at most +0.7 points, but the remaining text usually still names an option word (Appendix F). We therefore append "Therefore, the answer is (Y) ⟨option⟩." with R's own option or

a different one to the cut text, so both have the same position and template. The original conclusion is followed on 99.8–100.0% of items and the swapped one on 91.3% (Qwen-3B) and 99.8% (Phi-3.5-V), but on 54.0% (Qwen-7B) and 28.5% (InternVL3-8B), which follow a swap toward the gold option more often than one away from it (in-place swaps and details: Appendix Table 11). The two smaller models thus take the answer from the final statement of R, while the answer position of the two larger ones integrates evidence beyond it.

7 Related work

CoT and spatial reasoning. Kancheti et al. (2026) show that CoT and reasoning tuning lower accuracy on most of 13 spatial benchmarks, and CoT, self-consistency, and tree-of-thoughts do not improve VSI-Bench (Yang et al., 2025). Liu et al. (2025) link CoT failures to tasks where thinking hurts humans.

Visual forgetting and re-injection. TVC (Sun et al., 2025) attributes errors in long multimodal reasoning to declining attention to image tokens and re-injects the image at intermediate steps. Tian et al. (2026) show that replaying the image during reasoning improves accuracy and propose VAPO, an RL objective that rewards visually anchored steps; Remember-R1 (Chen et al., 2026) adds process-level rewards for sustained visual dependence. Unlike these methods, which act during generation and target long reasoning, we re-supply the image before or after a short R; neither recovers the CoT loss.

Faithfulness of CoT. Prior faithfulness studies causally perturb rationales, including truncating the reasoning, adding mistakes, and replacing it with filler, to test whether answers depend on the stated chain of thought (Lanham et al., 2023); stated reasoning can also omit the features that drive the answer (Turpin et al., 2023), and filler tokens can carry computation when a model is trained to use them (Pfau et al., 2024). Our content edits adapt this intervention family (conclusion removal itself is not new) and combine it with token-length, dialogue-turn, question-restatement, and image controls under one forced-choice readout in VLMs.

Spatial binding and read-out. Kang et al. (2026) show that VLMs bind linear spatial IDs to object-word tokens and that swapping them changes the model's spatial answer in a band of intermediate layers. Wang & Xu (2026) find with direction patching that text CoT suppresses the transport of spatial directions to the answer in most VLMs they study. Fu et al. (2025) show that VLMs underuse information their vision encoders contain. We add the causal-mask decomposition, a factorial on length, distance, and re-supplied inputs, and answer-position probing (Belinkov, 2022; Hewitt & Liang, 2019) and patching (Meng et al., 2022; Geiger et al., 2025) under length-matched text.

8 Conclusion

We asked when chain-of-thought hurts spatial answers in vision-language models. Starting from the causal-mask constraint, we separated retrieval and content accounts of the loss with length, turn-structure, re-injection, regeneration, and content controls. In five open VLMs, neither the length of the intermediate text nor its distance from the image and question explains the loss: the answer agrees with the option the reasoning states, and CoT scores like a sample from the direct answer distribution, giving up the gap between its argmax and its expectation. Our contributions are a causal-mask decomposition with a factorial of controls that separates what the reasoning says from how much text it adds, a sampling regularity that predicts majority-vote accuracy, and a two-term account, CoT $-$ Direct $= E - G$, that places fourteen open and closed VLMs on one axis. Beyond these benchmarks, remedies for CoT failures should target how often the reasoning commits to the right option rather than access to the image, and because G needs only direct answers and E one reasoning trace per item, whether a CoT effect carries over from one model to another becomes a measurement.

AI USE STATEMENT

Generative AI (Claude) was used to draft and edit prose and to help write experiment and figure scripts. It was not used to design the scientific claims, to choose which results to report, or to produce any measurement. Every number from the main experiments is written into the paper by a script that reads the saved per-item outputs, and all figures are regenerated directly from those outputs.

ETHICS STATEMENT

The experiments use public open-weight VLMs, public benchmarks, and procedurally generated images. There are no human subjects and no personal data.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix B give the models, prompts (verbatim), filler text, decoding settings, readout, statistics, probe and patching settings. The anonymous code, which writes every table, figure, and in-text number from saved per-item outputs, is in the supplementary material.

REFERENCES

Marah Abdin et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 2022.

Jianmin Chen, Jiaqi Tang, Wei Wei, Xiaogang Xu, Jiafei Wu, Zhe Liu, Qianzhou Wang, Yingying Yan, Botong Geng, Yuyang Xia, Lei Zhang, and Qifeng Chen. Remember-R1: Mitigating long-context visual forgetting through reinforcement learning. In *ACM Multimedia*, 2026. arXiv:2608.01314.

Lin Chen, Jinsong Li, Xiaoyi Dong, et al. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024.

Matt Deitke, Christopher Clark, Sangho Lee, et al. Molmo and PixMo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: VLMs overlook their visual representations. *arXiv preprint arXiv:2506.08008*, 2025.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26, 2025.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *EMNLP*, 2019.

Sai Srinivas Kancheti, Aditya Sanjiv Kanade, Vineeth N. Balasubramanian, and Tanuja Ganu. Chain-of-thought degrades visual spatial reasoning capabilities of multimodal LLMs. *arXiv preprint arXiv:2604.16060*, 2026.

Raphi Kang, Hongqiao Chen, Georgia Gkioxari, and Pietro Perona. Linear mechanisms for spatiotemporal reasoning in vision language models. In *ICLR*, 2026. arXiv:2601.12626.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, et al. Efficient memory management for large language model serving with PagedAttention. In *SOSP*, 2023.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. In *ICML*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, 2022.

Jacob Pfau, William Merrill, and Samuel R. Bowman. Let's think dot by dot: Hidden computation in transformer language models. In *Conference on Language Modeling (COLM)*, 2024.

Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. Mitigating visual forgetting via take-along visual conditioning for multi-modal long CoT reasoning. In *ACL*, 2025. arXiv:2503.13360.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter Henry Tu, and Jing Zhang. More thought, less accuracy? On the dual nature of reasoning in vision-language models. In *ICLR*, 2026. arXiv:2509.25848.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *NeurIPS*, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *NeurIPS*, 2023.

Zeyu Wang and Xinming Xu. Knowing isn't always saying: When do spatial encodings reach answers in vision-language models? In *EMNLP*, 2026. arXiv:2608.22916.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A LIMITATIONS

Answers are read as one forced-choice token (which agrees with free-form answers where R states an option; Section 4.1), after one greedy R per item except in the within-item test. Reasoning is short (mean $|R|$ of 51–131 tokens) and the open models have at most 8B parameters; larger and closed models get a reduced API protocol (300 items, no log-probabilities, reasoning not disabled in GPT-5.5 and GPT-5.6-Sol); mechanism analyses cover Qwen2.5-VL only; long reasoning is not studied. Filler is one neutral text matched to R in tokens only; content edits change only R's conclusion; patching transplants a residual from another position (calibrated by a mirrored image). Filler controls hold for the open models only: in the API models a filler turn is not neutral. The decomposition CoT − Direct $= E - G$ is an accounting identity; E is a residual associated with the reasoning, not an identified causal component. Trained methods that reward visually grounded reasoning steps (Tian et al., 2026; Chen et al., 2026) may change the content of R and are not covered by our prompt-level interventions.

B IMPLEMENTATION DETAILS OF THE MAIN EXPERIMENTS

In the released code, the behavioural factorial is E1, the probes and attention analyses E2, activation patching E3, and the content edits of R E5.

Models and inference. Behavioural experiments (E1, E5) use vLLM 0.8.4 in bfloat16 with temperature 0, prefix caching, seed 0, and a maximum context of 12288 tokens (Molmo: 4096, one image per prompt; other models: up to two images). Image settings: Qwen2.5-VL at most $1280\times28\times28$ pixels per image; InternVL3-8B at most 6 dynamic tiles (`max_dynamic_patch=6`); Phi-3.5-Vision 4 crops (`num_crops=4`); Molmo-7B-D default processor with a single image. The CoT text R is generated once per item and model (greedy, at most 512 new tokens) and reused by every condition that contains it. Mean $|R|$ is 124, 75, 131, 51, and 99 tokens (Qwen-3B, Qwen-7B, InternVL3-8B, Phi-3.5-V, Molmo-7B-D), and the share of R truncated at 512 tokens is $<$0.1, 0.0, 0.0, 0.0, and 0.0%.

Prompts. All prompts use each model's chat template; for Molmo the prompt is rendered with a plain "User: … Assistant: …" template and vLLM's own prompt wrapper is disabled (Appendix G). The first user turn is always ⟨image⟩ followed by the dataset question Q with its lettered options; synthetic items append "(A) yes" and "(B) no" on separate lines. The constants in the code are:

```
COT_PREFIX   = "Let's think step by step."
FORCE_PREFIX = "Answer: ("
FINAL_INSTR  = "Now give your final answer."
```

The CoT generation prompt is the first user turn followed by a partial assistant message `COT_PREFIX`, continued greedily. Answer prompts (`\n` is a newline; C is R or filler text):

- Direct: user [⟨image⟩, Q]; assistant "`Answer: (`".
- Direct+restate: user [⟨image⟩, "Q`\n\nQuestion again:` Q`\nNow give your final answer.`"]; assistant "`Answer: (`".
- Text in the same turn: user [⟨image⟩, Q]; assistant "C`\nAnswer: (`".
- Text with an extra turn: user [⟨image⟩, Q]; assistant "C" (turn closed); user turn; assistant "`Answer: (`". The user turn is "`Now give your final answer.`" (turn), "`Question again:` Q`\nNow give your final answer.`" (restate), [⟨image⟩, "`Here is the image again.\nNow give your final answer.`"] (re-image), or [⟨image⟩, "`Here is the image again.\nQuestion again:` Q`\nNow give your final answer.`"] (re-image+restate).

The final assistant message is left open, so the model continues directly after "Answer: (".

Filler. The filler cycles through the following twelve sentences, joined by spaces, and is truncated to exactly n tokens under the model's tokenizer (n=$|R|$ for length-matched Filler, 512 or 1024 otherwise):

1. The history of paper making spans many centuries and several cultures.
2. Early writers recorded ideas on bark, cloth, clay and animal skin.
3. Later, workers learned to soak plant fibers and press them into thin sheets.
4. The method spread slowly through trade and the movement of skilled craftsmen.
5. Mills powered by water made production faster and cheaper.
6. Printed books became common, and literacy rates rose over time.
7. Today most paper is produced from wood pulp in large industrial plants.
8. Recycling programs recover a large share of used paper every year.
9. Some artists still prefer handmade sheets for their texture and strength.
10. Libraries preserve old documents by controlling humidity and temperature.
11. Scholars study watermarks to estimate the age of historical manuscripts.
12. The chemistry of acid-free paper helps documents last for generations.

Readout. We generate one token with the top-20 log-probabilities. Each option letter receives the maximum log-probability over decoded tokens equal to that letter; letters absent from the top 20 receive probability 0 after renormalization over the options. The largest share of items with no option letter in the top 20, over all conditions and item groups, is 0.0% (Qwen-3B), 0.0% (Qwen-7B), 0.0% (InternVL3-8B), 0.0% (Phi-3.5-V), and 0.6% (Molmo-7B-D). The free-text answer of R is parsed with three patterns tried in order ("answer is (X)", "(X)", and a bare letter followed by ".", ")" or ":"), taking the last match; R with no match is labelled "no letter".

Statistics. Confidence intervals use a paired cluster bootstrap over image groups (2000 resamples; percentile 95% intervals); p-values come from a paired cluster sign-flip randomization test over image groups (5000 permutations, two-sided); tasks are treated as fixed. A group is one image: the two synthetic relative-position questions about an image share a group, and on CV-Bench each image appears once, so clustering matters only for the synthetic sets. Holm correction covers 15 tests: three pre-declared contrasts (CoT vs. Direct, Filler512 vs. Direct, CoT+restate vs. CoT+turn) for five models. Holm-adjusted values are, for CoT vs. Direct, $p = 0.003$, $p = 0.003$, $p = 0.003$, $p = 0.003$, and $p = 0.65$; for Filler512 vs. Direct, $p = 1.00$, $p = 0.98$, $p = 0.003$, $p = 1.00$, and $p = 1.00$; for CoT+restate vs. CoT+turn, $p = 0.65$, $p = 0.13$, $p = 1.00$, $p = 0.042$, and $p = 0.059$ (Qwen-3B, Qwen-7B, InternVL3-8B, Phi-3.5-V, Molmo-7B-D). Other p-values in the paper are uncorrected. Exact McNemar tests on the discordant pairs are reported in the released outputs; per-task and control-set p-values in the released outputs are McNemar tests. Cross-model means, minima and maxima are over the models that have the condition.

Content edits (E5). *CoT without its lettered conclusion* ("CoT w/o answer" in the code and in Table 5) cuts R at the start of the sentence that contains the last mention of an answer (the word "answer", "(X)", "X)", or "option X"), repeating up to four times while the remaining tail still contains one and at least 30% of R remains. *CoT swap* appends "Therefore, the answer is (Y) ⟨option text⟩." with Y the other option, on two-option items where R states a parsable option. Both are scored with the same readout; the restate variant adds the restate turn. The audit reconstructs every edited text and checks whether a parenthesised option letter remains anywhere and whether the last remaining sentence contains the text of an option as printed in the question (e.g. "left", "yes", or an object name); the second case occurs in 60.7, 63.5, 58.8, and 62.8% of edited texts (Qwen-3B, Qwen-7B, InternVL3-8B, Phi-3.5-V).

C SUPPLEMENTARY BEHAVIOURAL RESULTS

Table 5: Pooled spatial accuracy (%) split by the option that the CoT text R states ("no letter": no parsable option). Direct, CoT+restate and CoT+re-image are evaluated on the same items. Last two columns (Section 6): accuracy after R without its lettered conclusion ("R w/o answer"), and the share of answers that follow a swapped conclusion. Controls: Appendix Table 9.

Model	R states	n	Direct	CoT	CoT+restate	CoT+re-image	R w/o answer	swapped answer followed
Qwen2.5-VL-3B	correct	1979	89.7	99.4	98.5	99.8	99.0	89.0
	wrong	753	56.0	0.7	5.0	1.2	3.6	97.1
	no letter	118	75.4	58.5	63.6	59.3	58.5	–
Qwen2.5-VL-7B	correct	2339	92.7	100.0	99.2	99.7	99.6	49.6
	wrong	509	55.4	0.0	6.9	1.0	1.8	74.7
	no letter	2	100.0	50.0	50.0	100.0	100.0	–
InternVL3-8B	correct	2402	94.8	99.8	99.0	95.1	99.8	21.1
	wrong	448	62.1	2.7	10.0	26.1	2.9	69.4
Phi-3.5-V	correct	1149	85.1	98.2	99.9	100.0	98.9	99.7
	wrong	486	66.0	3.7	1.0	2.7	6.2	99.8
	no letter	1215	73.1	63.8	64.6	64.3	63.9	–
Molmo-7B-D	correct	2063	82.1	99.7	99.4	–	–	–
	wrong	767	54.6	1.0	3.4	–	–	–
	no letter	20	55.0	55.0	55.0	–	–	–

A constant-probability control shows that the self-consistency prediction in Section 4.4 uses item-level information: predicting majority-vote accuracy from the pooled mean $\bar{p}$ alone gives 79.4, 90.0, 93.5, and 76.2% (Qwen-3B, Qwen-7B, InternVL3-8B, Phi-3.5-V), 4–7 points above the observed 75.5, 84.5, 86.9, and 72.5%, and an argmax account predicts Direct accuracy; the item-level prediction is within 1.4 points. Molmo-7B-D has no sampled texts.

New CoT errors lie mostly at $0.5 \leq p < 0.8$ in Qwen-3B, Phi-3.5-V, and Molmo-7B-D (78, 67, 80%), but 42–45% lie at $p \geq 0.8$ in Qwen-7B and InternVL3-8B.

Table 6 gives accuracy per task and condition for every model, including the two controls. Table 8 splits the spatial pool by quartile of $|R|$ (items are the same in every row of a model). Contrasts for

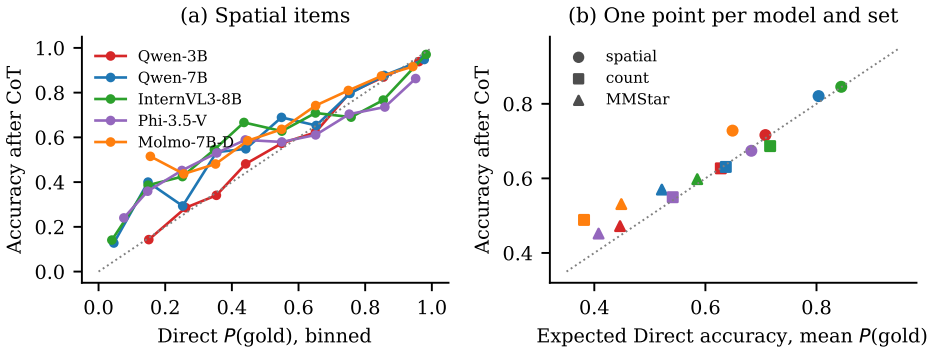


Figure 3: (a) Accuracy after CoT on spatial items against Direct's probability of the gold option, binned (bins with $\geq$20 items); dotted: identity. (b) Accuracy after CoT against expected Direct accuracy (mean $P(\text{gold})$), one point per model and item set. Most pooled points lie near the identity; InternVL3-8B Count and all Molmo-7B-D sets are clear exceptions, and MMStar lies above it. Computed by `calibration_v2.py` from the saved E1 and E6 outputs.

every condition pair, with confidence intervals, McNemar p-values and discordant counts, are written by the released analysis script.

Table 6: Forced-choice accuracy (%) per task, the task-stratified version of Table 2. CoT (free): the option stated in R, counted as wrong when R states none. F512/F1024: 512/1024 filler tokens in the same turn; rest.: restate; reim.: re-image.

Model	Task	Direct	CoT	CoT (free)	Filler	F512	F1024	CoT+rest.	CoT+reim.	F512+rest.	F512+reim.
Qwen2.5-VL-3B	CVB Rel. (n=650)	74.5	75.5	74.3	72.5	72.9	69.8	76.5	76.3	72.6	72.2
	CVB Depth (n=600)	78.2	70.8	70.5	75.7	77.3	73.7	71.0	71.7	77.8	76.2
	CVB Dist. (n=600)	70.2	59.5	50.8	65.5	69.5	70.5	62.2	60.0	64.8	71.5
	Syn. pos. (n=500)	100.0	96.2	96.2	100.0	100.0	100.0	96.2	96.2	100.0	100.0
	Syn. depth (n=500)	82.6	57.4	57.4	82.8	85.2	82.6	57.0	57.4	63.0	72.6
	CVB Count (n=788)	68.3	62.7	53.8	66.2	65.7	66.5	62.2	63.6	61.0	67.4
	MMStar (n=500)	50.4	47.2	29.2	44.6	48.0	49.0	47.8	47.2	50.4	49.2
Qwen2.5-VL-7B	CVB Rel. (n=650)	89.2	86.3	86.3	85.1	87.2	86.2	87.7	86.9	86.2	87.7
	CVB Depth (n=600)	84.8	82.7	82.7	72.7	80.8	79.5	83.0	82.7	79.8	80.0
	CVB Dist. (n=600)	79.0	76.5	76.5	64.7	77.2	77.5	77.2	76.3	80.2	78.7
	Syn. pos. (n=500)	100.0	94.6	94.6	99.2	100.0	100.0	94.6	94.6	100.0	100.0
	Syn. depth (n=500)	77.8	70.0	70.0	59.8	83.8	85.4	70.6	69.6	85.0	80.8
	CVB Count (n=788)	66.4	63.1	63.1	61.4	63.6	62.9	62.8	63.6	63.1	64.8
	MMStar (n=500)	56.2	57.0	51.8	45.0	51.4	52.0	56.8	56.2	55.8	56.8
InternVL3-8B	CVB Rel. (n=650)	90.5	90.3	90.2	90.0	89.8	89.7	90.2	90.8	90.5	87.7
	CVB Depth (n=600)	86.2	86.3	86.0	85.7	86.3	85.0	85.8	78.3	83.7	81.2
	CVB Dist. (n=600)	83.5	81.0	80.8	84.8	84.3	82.5	82.0	79.8	85.5	80.5
	Syn. pos. (n=500)	99.8	97.8	97.4	98.8	99.0	99.0	99.2	99.4	99.6	99.6
	Syn. depth (n=500)	90.2	66.0	65.6	77.2	81.2	84.8	66.6	73.0	88.8	87.0
	CVB Count (n=788)	75.0	68.7	68.5	73.7	73.2	72.2	69.4	73.2	72.8	74.9
	MMStar (n=500)	62.4	59.8	34.4	61.4	61.8	61.6	59.4	60.4	62.2	60.8
Phi-3.5-V	CVB Rel. (n=650)	79.1	76.6	42.8	66.3	78.6	78.6	76.5	76.6	76.8	78.6
	CVB Depth (n=600)	76.2	53.3	14.0	63.5	71.7	69.0	53.3	53.8	67.0	66.8
	CVB Dist. (n=600)	67.8	59.2	32.3	58.5	68.5	67.7	61.3	61.5	69.0	70.5
	Syn. pos. (n=500)	97.6	90.2	76.8	97.4	96.8	96.6	90.4	91.6	94.8	93.0
	Syn. depth (n=500)	64.2	59.4	41.8	71.0	70.0	66.2	60.2	59.0	58.8	54.4
	CVB Count (n=788)	56.0	54.9	38.2	47.7	46.2	51.9	54.3	52.4	48.9	53.4
	MMStar (n=500)	43.8	45.2	30.0	34.8	39.4	44.4	47.8	47.2	42.0	42.6
Molmo-7B-D	CVB Rel. (n=650)	73.7	71.2	71.1	65.1	64.6	–	71.4	–	67.5	–
	CVB Depth (n=600)	68.0	71.0	70.0	66.0	74.0	–	71.2	–	68.0	–
	CVB Dist. (n=600)	50.7	62.7	61.5	50.3	54.3	–	63.2	–	56.7	–
	Syn. pos. (n=500)	96.8	86.4	87.0	98.2	97.0	–	87.4	–	84.6	–
	Syn. depth (n=500)	89.6	75.6	75.4	91.2	91.8	–	76.2	–	65.0	–
	CVB Count (n=788)	38.3	48.9	48.7	41.5	40.1	–	49.4	–	46.7	–
	MMStar (n=500)	49.8	53.2	50.0	47.6	50.2	–	52.6	–	46.8	–

Table 7: CoT minus Filler+turn accuracy (points) per model and task, with paired cluster-bootstrap 95% CIs over image groups (2000 resamples). CoT is below the turn-matched filler in 18 of 25 cells (CI below zero in 15); it is above it with a CI above zero in 2 cells (Molmo-7B-D on CV-Bench Relation and Distance).

Model	CVB Rel.	CVB Depth	CVB Dist.	Syn. RelPos	Syn. Depth
Qwen2.5-VL-3B	$+1.8$ [-1.5, +5.2]	-8.0 [-11.7, -4.3]	-9.0 [-14.0, -4.2]	-3.8 [-5.6, -2.2]	-23.6 [-28.2, -19.0]
Qwen2.5-VL-7B	-2.6 [-5.4, +0.0]	-1.7 [-5.0, +1.7]	-4.2 [-8.0, -0.3]	-5.4 [-7.4, -3.6]	-10.0 [-14.0, -5.8]
InternVL3-8B	$+1.2$ [-1.2, +3.5]	$+1.2$ [-1.5, +3.7]	-3.5 [-6.2, -0.8]	-1.8 [-3.2, -0.6]	-22.6 [-27.4, -17.6]
Phi-3.5-V	$+1.2$ [-2.9, +5.2]	-14.5 [-19.7, -9.2]	-11.5 [-16.2, -6.8]	-7.2 [-10.2, -4.4]	-3.2 [-9.0, +2.8]
Molmo-7B-D	$+5.5$ [+0.9, +9.8]	$+1.0$ [-3.3, +5.2]	$+8.2$ [+3.2, +13.2]	-11.4 [-14.4, -8.6]	-16.0 [-20.4, -11.8]

Table 8: Pooled spatial accuracy (%) by quartile of $|R|$ (in tokens). Longer R goes with lower Direct accuracy in several models, so this split does not isolate the effect of length.

Model	Condition	Q1	Q2	Q3	Q4
Qwen2.5-VL-3B	$\|R\|$ range	39–95	95–120	120–143	143–512
	direct	80.2	84.7	79.8	75.9
	cot	77.6	76.6	68.2	63.3
	filler	78.4	84.3	79.8	70.4
	filler + restate	77.2	79.3	74.5	68.3
	cot + restate	78.5	76.3	68.3	65.4
Qwen2.5-VL-7B	$\|R\|$ range	23–49	49–64	64–91	91–279
	direct	92.9	88.7	84.2	79.8
	cot	85.4	85.0	83.0	76.6
	filler	88.3	78.6	71.5	67.4
	filler + restate	94.1	88.7	81.6	79.6
	cot + restate	85.5	85.9	84.1	77.2
InternVL3-8B	$\|R\|$ range	23–78	78–114	114–181	181–339
	direct	97.6	91.3	88.3	82.0
	cot	85.4	86.4	88.1	79.2
	filler	93.7	86.1	87.9	82.1
	filler + restate	96.9	91.3	88.0	81.7
	cot + restate	86.9	85.8	88.4	79.6
Phi-3.5-V	$\|R\|$ range	2–39	39–50	50–63	63–161
	direct	77.0	78.3	74.7	76.0
	cot	68.7	65.1	69.5	66.7
	filler	73.6	70.9	70.0	66.1
	filler + restate	74.1	76.4	75.2	73.3
	cot + restate	69.0	66.0	69.9	67.8
Molmo-7B-D	$\|R\|$ range	32–73	73–96	96–121	121–278
	direct	87.9	82.4	69.2	58.5
	cot	82.2	77.5	68.5	64.0
	filler	86.0	79.9	67.1	57.4
	filler + restate	82.1	79.9	68.8	58.7
	cot + restate	83.2	78.1	69.0	64.0

Task effects are heterogeneous (Table 6). Synthetic depth ordering falls under CoT in every model (e.g. InternVL3-8B 90.2% to 66.0%), whereas CV-Bench Relation changes little (e.g. InternVL3-8B 90.5% to 90.3%) and CV-Bench Distance rises in Molmo-7B-D (50.7% to 62.7%). Pooled estimates weight tasks by their item counts.

Table 9 gives the split by the option that R states on the two control sets, and Table 10 the letter prior and prior-matched accuracy of Section 4.2.

C.1 Re-prompting without CoT

The same interventions are not neutral without CoT. Repeating the question in the first turn (Direct+restate) changes accuracy by -0.9 points on average and by -4.0 in Qwen-3B ($p < 0.001$). After Filler512, restating changes accuracy by -4.8 (Qwen-3B), +0.4 (Qwen-7B), and -3.6 (Phi-3.5-V), and re-supplying the image by -2.2, -0.3, and -4.0.

Table 9: Accuracy (%) on the two control sets split by the option that the CoT text R states (items where R states no parsable option are omitted). As on the spatial items (Table 5), the answer after CoT copies the stated option.

		CV-Bench Count			MMStar		
Model	R states	n	Direct	CoT	n	Direct	CoT
Qwen2.5-VL-3B	correct	424	93.4	99.8	146	76.7	97.9
	wrong	237	27.4	1.3	180	32.8	9.4
Qwen2.5-VL-7B	correct	497	92.6	100.0	259	87.3	99.6
	wrong	291	21.6	0.0	168	24.4	4.2
InternVL3-8B	correct	540	94.1	100.0	172	86.0	97.1
	wrong	247	33.2	0.4	128	45.3	23.4
Phi-3.5-V	correct	301	88.0	100.0	150	71.3	94.0
	wrong	190	22.1	0.0	141	19.9	3.5
Molmo-7B-D	correct	384	64.8	100.0	250	77.6	99.6
	wrong	404	13.1	0.2	195	20.5	1.0

Table 10: Letter prior and prior-matched accuracy on the pooled spatial set (two-option items). Top: share of items on which the forced answer is "A". Bottom: raw accuracy and prior-matched accuracy (p.-m.), for which the threshold on the log-odds of "A" versus "B" is set so that the condition predicts "A" as often as Direct does; the correction uses no labels. Direct's prior-matched value differs slightly from its raw accuracy because the threshold is an empirical quantile of the log-odds and is not exactly zero.

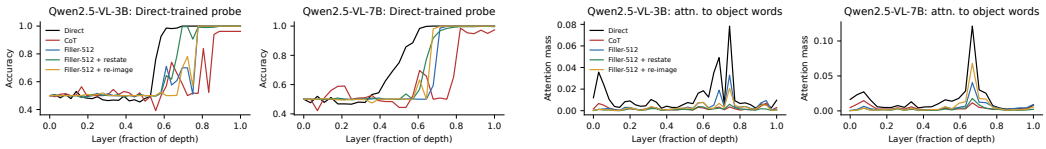
	Share of "A" predictions (%)					
Model	Gold	Direct	CoT	Filler	Filler+turn	Filler512
Qwen2.5-VL-3B	50.5	55.6	62.6	48.6	51.9	54.3
Qwen2.5-VL-7B	50.5	48.6	49.8	29.9	47.5	46.4
InternVL3-8B	50.5	50.0	52.7	46.6	49.7	47.6
Phi-3.5-V	50.5	46.3	58.2	57.6	46.1	48.5
Molmo-7B-D	50.5	60.8	53.8	54.2	52.0	49.5

	Direct		CoT		Filler		Filler+turn		Filler512	
Model	raw	p.-m.	raw	p.-m.	raw	p.-m.	raw	p.-m.	raw	p.-m.
Qwen2.5-VL-3B	80.2	80.8	71.6	68.9	78.3	78.3	79.6	79.6	80.0	79.6
Qwen2.5-VL-7B	86.0	86.1	82.1	82.4	76.2	81.3	86.6	86.7	85.4	85.4
InternVL3-8B	89.7	89.8	84.6	84.2	87.3	87.5	89.1	88.8	88.0	88.0
Phi-3.5-V	76.7	76.6	67.4	67.4	70.4	72.0	74.4	73.9	76.7	77.1
Molmo-7B-D	74.5	74.5	72.8	70.3	72.6	71.7	74.4	74.1	74.9	73.5

D PROBE AND ATTENTION DETAILS

E2 runs HuggingFace Transformers in bfloat16 with eager attention (to return attention weights), with the E1 prompts and image resolution, and stores `hidden_states[`ℓ`]` for ℓ=0, . . . , L at the answer position (0 is the embedding output; L=36 for Qwen-3B and 28 for Qwen-7B). Probes are scikit-learn logistic regressions (C=0.1, at most 3000 iterations) on standardized features, evaluated with `GroupKFold(5)`; the group is the image file (synthetic) or the source image (CV-Bench), so the two questions about a synthetic image are always in the same fold. Folds whose training set contains one class are dropped. The fresh probe is trained and tested within a condition; the transfer probe is trained on Direct states of the training images and tested on the held-out images of the target condition. The first readable layer is the first ℓ at which the Direct fresh probe reaches 0.9. The relation probe uses as label the relation that holds in the image (whether object A is left of object B), which on the synthetic set is uncorrelated with the answer letter because each image is asked with both wordings; on CV-Bench Relation the two labels coincide up to sign (correlation -1.00), so we report the relation probe for the synthetic set only. Attention spans: image tokens of the first image; the question text; object words (the two noun phrases of the question, a subset of the question span); the intermediate text C; the text of the extra user turn; and the tokens of the second image when present. Attention is averaged over heads, summed over the tokens of a span, and averaged over layers $\ell \geq \ell^*$ and over items. On CV-Bench Relation (Qwen-3B) the corresponding values are:

Direct question 0.230; CoT R 0.238 and question 0.110; Filler text 0.106 and question 0.179; second image under re-image 0.038.



(a) Transfer probe (trained on Direct) (b) Attention to object words

Figure 4: Answer-position analyses on synthetic relative position (x-axis: fraction of depth). (a) Accuracy of a probe trained on Direct states and tested on held-out images of each condition; it drops to chance under every inserted text at mid-depth and returns near the behavioural accuracy by the last layer. (b) Attention mass from the answer position to the object words of the question.

D.1 Full probe results

On synthetic depth ordering, where CoT lowers Qwen-3B accuracy from 82.6% to 57.4%, the same shift appears at ℓ^* (transfer 0.52–0.51), but at the last layer transfer returns to 0.93 under Filler+turn and 0.91 under Filler512 and stays at 0.59 under CoT (Qwen-7B: 0.99, 0.88, 0.60): the last-layer read-out separates CoT from filler where the mid-depth shift does not.

At every layer we fit logistic-regression probes with 5-fold cross-validation grouped by image. A *fresh* probe is trained and tested within a condition and asks whether the answer label is linearly readable. A *transfer* probe is trained on Direct states and tested on held-out images of another condition and asks whether the label is still encoded along the Direct direction. The first readable layer ℓ^* is the first layer at which the Direct fresh probe reaches 0.9: layer 21 of 36 (58% of depth) on synthetic relative position and 28 on CV-Bench Relation in Qwen-3B, and 17 of 28 on synthetic relative position in Qwen-7B.

At ℓ^* in Qwen-3B the label is readable under every condition (fresh accuracy from 0.84 under Filler to 0.97 under CoT), but the Direct probe transfers at chance under every inserted text (Figure 4a): 0.50 under CoT, 0.50 under Filler, 0.52 under Filler512, and 0.50–0.57 with a restated question or re-supplied image. This reproduces, with image-grouped folds and text in the assistant turn, the prefix-induced read-out shift that a preliminary probing study found with fixed prefixes (Appendix H). The shift does not explain errors: under Filler and Filler512 the model answers this set as accurately as under Direct (behavioural accuracy 100.0% and 100.0% vs. 100.0%), and by the last layer the Direct probe transfers at 1.00 (Filler), 1.00 (Filler512), and 0.96 (CoT), which matches CoT accuracy on this set (96.2%). Inserted text of any kind thus re-encodes the answer at mid-depth, later layers bring it back to the Direct direction, and the remaining CoT errors are items on which the output carries the other label. On CV-Bench Relation the shift at ℓ^* is smaller (transfer 0.60 under CoT, 0.66 under Filler), and last-layer transfer stays at 0.56–0.77 under inserted text (Direct: 0.89); the cause of this difference from the synthetic set is not identified.

A second probe targets the relation that holds in the image ("A is left of B"), which is independent of the question wording on the synthetic set (correlation with the answer letter 0.00). It is readable from layer 21 under Direct and under every condition (fresh 0.86–0.96), so the image relation reaches the answer position whatever text precedes it.

D.2 Full attention results

We average attention weights over heads at the answer position, sum them over token spans, and average over layers from ℓ^* on. Under Direct, the question receives 0.269 of the mass, its object words 0.017, and the image 0.030 (synthetic relative position, Qwen-3B). With CoT, R receives 0.241 and the question drops to 0.078; with Filler, the filler receives 0.115 and the question keeps 0.166 (Filler512: 0.119 and 0.192). Attention to the image falls to 0.015 under CoT and 0.016 under Filler, and attention to object words falls under every inserted text (Figure 4b). A restate turn takes 0.228 of the mass, and a re-supplied image 0.061. Filler thus lowers attention to the image as much as CoT does and attention to object words by more than half, while leaving answers unchanged,

which is inconsistent, at the answer position, with the attention predictions of H_{vis} and H_{ret}; CoT differs from filler in that its text receives about twice the attention and, by Section 4.4, its content determines the answer. Attention weights are not a causal measure, and head and layer averaging can hide specialised heads.

E PATCHING DETAILS

E3 runs HuggingFace Transformers with SDPA attention. Block b (0-based) is the b-th decoder block; patching replaces its output at the answer position, which equals `hidden_states[`b+1`]`. An E3 block index b therefore corresponds to E2 layer $\ell{=}b{+}1$; Figure 2b plots $b/(L{-}1)$ and Figure 4a plots ℓ/L. Flip sets: over the five spatial sets, items with Direct correct and target wrong (target CoT or Filler512), shuffled with seed 0, at most 150 per target (Qwen-3B: 150 of 458 CoT flips and 91 of 91 Filler512 flips). Mirror control: synthetic items with a left/right relation and CV-Bench Relation items whose question mentions both "left" and "right", with Direct and CoT both correct, at most 100 items (Qwen-3B: 100 of 520); the source run is Direct on the horizontally mirrored image, whose correct answer is the other option. We report the mean $P(\text{gold})$ over items per block; the onset is the first block at which the mean differs from the unpatched value by more than 0.05. The crossover block is the first block at which the majority of items have $P(\text{gold}){>}0.5$ (mirror control: <0.5). The patching script stores `hidden_states[1:]`; the last entry is taken after the final norm, so its value is not directly comparable with the source run's $P(\text{gold})$, and the main text reports block $L{-}2$ (Qwen-3B CoT flips: 0.70 at block $L{-}2$ and 0.73 after the final norm; source 0.70). Flip sets are selected from the vLLM outputs and patched under Transformers; restricting to flips confirmed under Transformers keeps 139 of 150 (Qwen-3B CoT), 59 of 91 (Qwen-3B Filler512), 144 of 150 (Qwen-7B CoT), and 72 of 98 (Qwen-7B Filler512) items, and the mean curves are unchanged by this restriction.

F WITHIN-ITEM, GENERATION-TIME, AND MATCHED-EDIT CONTROLS

All three controls use the 2850 pooled spatial items of Section 3 and the forced readout; InternVL3-8B, Phi-3.5-V and both Qwen models are run (Molmo-7B-D is excluded because its prompt wrapper had to be patched in a model copy that was not kept).

Sampled reasoning (Section 4.4). For each item we sample 4 CoT texts at temperature 0.7 (top-p 1, at most 512 tokens) and parse the stated option of each. The share of samples with a parsable option is 93.5, 99.9, 99.9, and 57.9% (Qwen-3B, Qwen-7B, InternVL3-8B, Phi-3.5-V), and the forced answer agrees with the stated option on 98.9, 99.9, 99.2, and 96.0% of those samples. The within-item difference is the mean over mixed items of (accuracy after samples stating the correct option) − (accuracy after samples stating a wrong option); its 95% interval resamples items.

Conclusion edits (Section 6). *CoT without its lettered conclusion* cuts R before the last sentence that states a lettered answer, keeping 88.5–90.2% of R on average. An audit of the edited texts shows that this removes the lettered answer but often not the verbal conclusion: parenthesised letters remain in 0.0% of Qwen-3B texts, and in 58.8–63.5% of edited texts the last remaining sentence names an option word (e.g. "left"). In Qwen-3B, the edit yields 72.1% (Direct 80.2%, CoT 71.6%), and 3.6% on the items where R stated a wrong option. *CoT swap* appends "Therefore, the answer is (Y) ⟨option⟩." to the edited text, where Y differs from the option that R stated; the in-place swap replaces the letter and option word of R's own conclusion.

G EXCLUDED RUNS

- InternVL2.5-8B: under Transformers 4.51 its tokenizer returned a boolean instead of token ids in our pipeline; the cause is not identified, and the model was dropped (InternVL3-8B represents the family; Phi-3.5-Vision was added).
- Molmo-7B-D, first run: vLLM wrapped every prompt as "User: … Assistant:", so the forced answer prefix was followed by an extra "Assistant" tag and the read-out token did not continue "Answer: (". The run was discarded and repeated with the wrapper disabled and the processor's

Table 11: Controls added after the main factorial. †Share of items on which the regenerated R states a wrong option when the image is supplied again after the question; the corresponding share for the original R is 26.4, 17.9, 15.7, and 17.1%. With the filler in the same assistant turn the models stop almost immediately, so that row does not test reasoning under distance. Changes are paired over items; confidence intervals and randomization p-values are in the released outputs.

	Qwen-3B	Qwen-7B	InternVL3-8B	Phi-3.5-V
Mixed items	1427	976	889	1047
Acc. after correct R (%)	98.7	99.8	99.0	96.3
Acc. after wrong R (%)	2.6	0.3	5.5	5.4
Direct on same items (%)	74.2	74.7	77.5	72.2
Regenerated R (greedy): forced accuracy (%), change vs. original R in points				
Original R	71.6	82.1	84.6	67.4
Image again after question	70.9 (-0.7)	82.5 (+0.4)	85.1 (+0.6)	67.3 (-0.1)
Filler512, turn with image	70.4 (-1.2)	79.1 (-3.0)	83.6 (-1.0)	66.2 (-1.2)
Filler512, same turn	79.2	84.9	86.5	72.2
mean $\|R\|$ (tokens), same turn	1	1	41	9
Wrong R: original / image again (%)	25.5†	17.4†	13.9†	17.4†
Matched conclusion edits: share of answers that follow the appended / replaced conclusion (%)				
R w/o lettered concl. + original	100.0	100.0	99.8	100.0
R w/o lettered concl. + swapped	91.3	54.0	28.5	99.8
In-place swap of R's conclusion	72.8	21.7	13.9	94.4
Filler with matched turn structure: change vs. Direct (points)				
Filler512 + turn	-0.9	-0.4	-1.6	+0.6
Filler1024 + turn	-1.2	-0.8	-2.2	-5.5

default sequence length raised from 1536 to 4096 tokens; only the repeated run is reported. The content edits (E5) were not run for Molmo-7B-D.

- Qwen2.5-VL-32B: not run (the download failed).

H Additional analyses: preliminary probing study with fixed prefixes

This appendix summarises a preliminary probing study that motivated the main experiments; its numbers are from separate runs and are not generated from the main outputs. It differs from the probes of Section 5.1 in three ways: prefixes are fixed strings appended to the question in the user turn, the readout is the last prompt token, and cross-validation folds are not grouped by image.

Setup. Images are 448×448 synthetic scenes with two or three primitives. Relative position asks "Is the A to the left of the B?" or "…to the right of …", so each image appears twice with opposite labels; depth ordering asks one question per image. Content strings (spatial CoT, Python, Chinese, repeated characters, numbers) are built to an exact token length of 50 to 300 tokens. Probes are logistic regressions on standardized features ($C=1$). The *transfer* probe is fit on all no-prefix states of a layer and scored unchanged on the prefixed states of the same items; the *fresh* probe is retrained within each condition with stratified 5-fold cross-validation. Because relative-position folds were not grouped by image, a held-out item's twin with the opposite label sat in the training folds, which pushed held-out accuracy below chance at layers that do not yet combine the question word with the image; these values carry no information, and the image-grouped probes of Section 5.1 supersede them.

Readable depth. Without a prefix, the fresh probe reads relative position at 1.00 from layer 25 of 36 in Qwen2.5-VL-3B (69% of depth) and from layer 19 of 28 in Qwen2.5-VL-7B (68%), and depth ordering at $\geq$0.95 from layer 20 (3B) and layer 12 (7B); the depth runs saved `hidden_states[`ℓ`+1]` under the label ℓ.

Content sweep at the first readable layer. Figure 5 and Table 12 compare five 50-token prefixes. In the 3B model at layer 25, the fresh probe is 0.99–1.00 under every prefix, and the transfer probe falls under every prefix: 0.74 (spatial CoT), 0.87 (repeated characters), 0.95 (Python), 0.63 (Chinese), 0.93 (numbers). In the 7B model at layer 19, fresh is 1.00 everywhere; transfer is 0.98–1.00 under the four non-spatial strings and 0.52 under spatial CoT. Deeper layers restore the transfer readout: 1.00 at 3B layers 30 and 35 under four strings (Python: 0.99, 0.87), and ≥0.99 at 7B layer 24 under all strings. This is the prefix-induced read-out shift that Section 5.1 reproduces with grouped folds and text in the assistant turn.

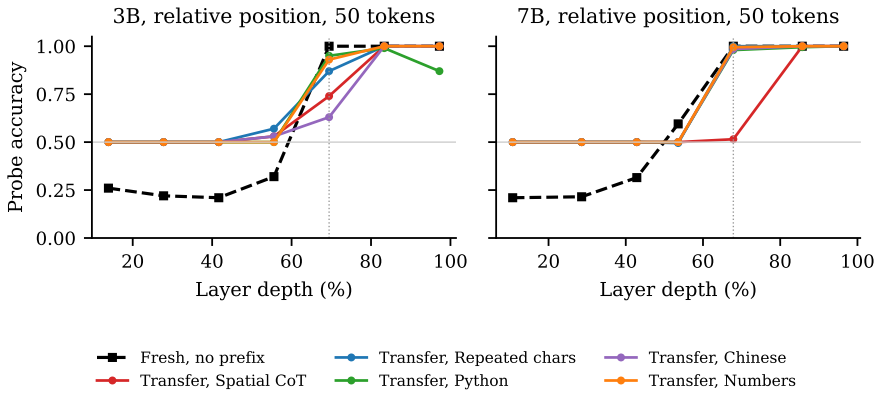


Figure 5: Relative position, 50-token prefixes. Solid: transfer probe. Dashed black: no-prefix fresh probe. Dotted vertical line: first readable layer.

Table 12: Relative position (chance 0.50). L20 and L15 are below the readable depth. *Training-set accuracy of the transfer probe.

	3B			7B		
Prefix (50 tokens)	L20 transfer	L25 transfer	L25 fresh	L15 transfer	L19 transfer	L19 fresh
None	1.00*	1.00*	1.00	1.00*	1.00*	1.00
Spatial CoT	0.53	0.74	0.99	0.50	**0.52**	1.00
Repeated characters	0.57	0.87	1.00	0.49	1.00	1.00
Python	0.50	0.95	0.99	0.50	0.98	1.00
Chinese	0.53	0.63	0.99	0.50	0.98	1.00
Numbers	0.50	0.93	1.00	0.50	0.99	1.00

Prefix length. For the strings that logged real tokens up to 300 tokens (spatial CoT did not), transfer accuracy at the first readable layer falls with length in the 3B model (0.65–0.69 at 100 tokens and 0.54–0.58 at 300 for repeated characters, Chinese, and numbers; Python non-monotone), while fresh accuracy stays at 0.97–1.00; in the 7B model the same three strings stay at 0.94–1.00 up to 300 tokens, while Python falls from 0.99 to 0.59.

Free-text pilot. On 190 mixed synthetic items per model answered in free text, CoT changed accuracy by -8.9 points in the 3B model (exact McNemar p=0.003) and by $+4.2$ in the 7B model (p=0.13); the 7B difference is not significant, rests on 190 items with a different CoT instruction and free-text parsing, and is superseded by the main experiments, in which Qwen-7B CoT changes accuracy by -4.0 points on 2850 items ($p < 0.001$). Geometry analyses (CKA, subspace overlap, intrinsic dimension) and a Procrustes check from this study are not reported; the Procrustes check applied a probe trained on standardized features to unstandardized features.

I Larger and closed models through an API

We sample 100 items per task from CV-Bench Relation, Depth, and Distance (seed 0) and query nine models through an OpenAI-compatible gateway: Qwen2.5-VL-72B-Instruct, Qwen3.8-Max, Qwen-

VL-Max, GPT-4.1, GPT-5.5, GPT-5.6-Sol, Gemini-3.8-Flash, Claude-Sonnet-5, and Claude-Opus-4.7 (the newest Claude Opus model the gateway serves). The gateway returns no log-probabilities, so each item receives a greedy Direct answer ("Answer with the option's letter only."), six Direct answers at temperature 1, whose share of gold answers estimates $P(\text{gold})$, a greedy CoT ("Let's think step by step, and end with 'Answer: (X)'."), and a final-answer turn ("Now give your final answer with the option's letter only.") after either the model's own CoT or the neutral filler of Appendix B truncated to the character length of the CoT. Items with an unparsable Direct answer or fewer than three parsable samples are dropped. Table 13 gives the results; the Qwen2.5-VL-3B/7B reference values in Section 4.6 are computed from the main runs on the same items.

Reasoning in the Direct condition. Direct answers should contain no reasoning, hidden or visible. Extended thinking is not enabled for the Claude models; Gemini-3.8-Flash, GPT-5.5, and GPT-5.6-Sol receive `reasoning_effort=none`; Qwen3.8-Max receives `enable_thinking=false` and `thinking_budget=0`; Qwen2.5-VL-72B, Qwen-VL-Max, and GPT-4.1 have no reasoning mode. We log completion and reasoning tokens of the greedy Direct call and, for all but the first 100–140 items of GPT-5.5 and GPT-5.6-Sol and the earlier runs of Qwen2.5-VL-72B and GPT-4.1, of the six samples, and count a call as reasoning if the gateway reports reasoning tokens or the reply exceeds 12 tokens (column R_D). GPT-5.5 and GPT-5.6-Sol reason in 56% and 72% of Direct calls although the request sets `reasoning_effort=none` (`minimal` and `reasoning.effort=none` behave the same in a probe), mostly on Distance (greedy calls: 72% and 95%) and rarely on Depth (1% and 13%). The rows "$R_D{=}0$ items" restrict these two models to items whose greedy Direct call has no reasoning; their samples may still reason, and the restriction selects easier items. On Depth, where GPT-5.5 reasons in 1% of greedy Direct calls, its CoT$-\bar{\hat{p}}$ is +8.2 points (CI [+1.3, +15.8]). With thinking disabled, Qwen3.8-Max ignores the letter-only instruction on about half of the Direct calls and writes visible reasoning of 50–250 tokens; we therefore append "Do not explain." to its letter-only instructions (Direct, samples, and both final-answer turns), after which R_D is 1%. Prefilling "Answer: (" as a partial assistant message also removed the reasoning but lowered Direct accuracy from 100% to 67% on items where the unmodified prompt already produced a bare letter, and is not used.

Other details. †After the filler turn, Qwen2.5-VL-72B, Qwen-VL-Max, and Gemini-3.8-Flash sometimes reply that they cannot see the image or that the conversation is inconsistent (parsable letters in 80.7%, 76.9%, and 91.7% of items); the filler column is therefore not comparable to the local filler conditions and is not used in the main text. Unparsed filler replies were re-queried once and parsed with a more tolerant pattern. Mean CoT length is 96–131 characters in the GPT-5 and Claude models and 361–1224 in the others; GPT-5.5 and GPT-5.6-Sol also produce hidden reasoning in the CoT call. With six samples, $\hat{p}$ has a resolution of 1/6 and is used only in item means; the gateway's temperature behaviour for reasoning models is not verified. The code is `scripts/v2api/` in the supplementary material; the API key is read from an environment variable.

Table 13: Accuracy (%) on 298–300 CV-Bench spatial items through an API. R_D: share (%) of logged Direct calls that contain reasoning (–: not logged); $R_D{=}0$ items: items whose greedy Direct call contains no reasoning. $\bar{\hat{p}}$: mean share of gold answers among six Direct samples at temperature 1. Brackets: bootstrap 95% CIs over items. †Filler accuracy over parsable replies (share in parentheses).

Model	n	R_D	Direct	CoT+turn	Filler+turn†	$\bar{\hat{p}}$	CoT−Direct	Filler−Direct†	CoT−$\bar{\hat{p}}$
Qwen2.5-VL-72B	300	–	95.3	94.7	65.7 (81%)	94.4	−0.7 [-3.3, +2.0]	−30.6 [-37.2, -24.0]	+0.2 [-2.2, +2.8]
Qwen3.8-Max	300	1	95.7	96.3	67.4 (99%)	95.7	+0.7 [-1.0, +2.7]	−28.2 [-33.9, -22.1]	+0.6 [-0.8, +2.3]
Qwen-VL-Max	299	0	95.0	98.0	63.5 (77%)	94.6	+3.0 [+0.3, +6.0]	−31.3 [-38.3, -24.3]	+3.4 [+0.9, +6.1]
GPT-4.1	298	–	83.9	90.3	76.8 (100%)	81.0	+6.4 [+1.7, +11.4]	−7.1 [-12.8, -1.3]	+9.3 [+5.6, +12.8]
GPT-5.5	300	56	89.0	93.0	81.7 (100%)	87.4	+4.0 [+0.3, +8.0]	−7.3 [-12.0, -3.0]	+5.6 [+2.6, +8.9]
$R_D{=}0$ items	183		87.4	97.3		90.1	+9.8 [+4.4, +15.3]		+7.2 [+3.0, +11.5]
GPT-5.6-Sol	300	72	90.7	93.3	85.0 (100%)	90.7	+2.7 [+0.0, +5.3]	−5.7 [-10.3, -1.3]	+2.6 [+0.2, +5.0]
$R_D{=}0$ items	121		89.3	96.7		92.6	+7.4 [+2.5, +13.2]		+4.1 [-0.3, +8.8]
Gemini-3.8-Flash	300	0	94.3	93.7	66.2 (92%)	93.3	−0.7 [-3.0, +1.7]	−28.4 [-34.9, -22.2]	+0.3 [-1.6, +2.3]
Claude-Sonnet-5	300	0	82.0	91.7	82.7 (100%)	82.4	+9.7 [+5.7, +13.7]	+0.7 [-4.3, +5.3]	+9.2 [+5.7, +12.8]
Claude-Opus-4.7	300	0	82.0	93.7	82.3 (100%)	82.5	+11.7 [+7.7, +16.0]	+0.3 [-4.3, +4.7]	+11.2 [+7.5, +15.0]

J PUBLISHED BENCHMARK NUMBERS

The table below transcribes published results; it is context, not a measurement from our code.

Table 14: Qwen2.5-VL-7B scores transcribed from Kancheti et al. (2026). Δ is CoT minus Direct. Under that definition CoT is lower on 10 of the 12 rows. We do not compute a weighted 3D/2D mean because the per-benchmark sample sizes are not in our notes.

Benchmark	**Type**	**Direct**	**CoT**	Δ
MindCube	3D	83.78	76.09	-7.69
RealWorldQA	2D/3D	69.02	63.05	-5.97
OmniSpatial	3D	45.23	40.40	-4.83
MMSIBench	3D	35.11	30.83	-4.28
SAT	2D/3D	63.11	59.22	-3.89
VSR	2D	85.38	81.83	-3.55
V*Bench	2D	79.06	76.27	-2.79
BLINK	2D	56.04	53.44	-2.60
CV-Bench2D	2D	77.17	75.92	-1.25
SpatialBench	2D	62.87	61.75	-1.12
MMVP	2D	26.87	27.47	$+0.60$
3DSRBench	3D	55.38	57.11	$+1.73$