When Reasoning Prompts Break Answer Decoding: Position Bias and Hidden-State Decodability in Vision-Language Models

Anonymous ACL submission

Abstract

When chain-of-thought (CoT) prompts are appended to multiple-choice VQA inputs *without* generating intermediate text, accuracy can collapse by more than 50 percentage points. We identify this as a mechanistic decoding failure: the hidden state remains answer-decodable, but the native output head is hijacked by a continuation-induced first-position prior. A linear probe (18K params) reads 83.7% from the same hidden states where the native head achieves only 50.8%—a 32.9pp decodability gap. Option-order randomization shows the bias is positional (93.5% first-position prediction, <1% A-token prediction), not label-based. Prompt-ending ablation and full-vocabulary token competition trace the origin to a next-token continuation prior: CoT endings expect reasoning tokens, compressing answer probability onto the first option as a positional shortcut. The failure is one of next-token expectation, not reasoning: an answer cue alone—with *zero* generated reasoning—recovers accuracy, and three representation-preserving interventions (input cue 90%, read-out probe 83.7%, output-head LoRA 70.9%) converge on one bottleneck, the vocabulary projection. The bias is format-dependent: on MMBench CoT-prefix *improves* accuracy by +5.6pp (93.8%→99.4%) while on AI2D it devastates accuracy by −54pp. The probe serves as a diagnostic instrument—not a deployed decoder—demonstrating that the position bias is latent in the output head and that reasoning prompts merely amplify it.

1 Introduction

Vision-language models (VLMs) have emerged as a powerful paradigm for multimodal reasoning, achieving strong performance on visual question answering (VQA) by jointly processing visual and linguistic inputs [2, 11, 12]. Chain-of-thought (CoT) prompting [26] is the dominant technique for eliciting structured reasoning in these models. In the text domain, CoT reliably improves accuracy on arithmetic, commonsense, and symbolic reasoning [7, 24]. Its adoption in vision-language settings is therefore natural—yet it produces a paradox: on multiple-choice VQA benchmarks, appending a reasoning instruction can *devastate* accuracy. On ScienceQA, Qwen2.5-VL-7B achieves 82.7% under Direct prompting, ∼75% under generation-based CoT, yet only 50.8% under CoT-prefix—a 31.9pp collapse. On AI2D, the collapse reaches 54 percentage points. What breaks?

The distinction between *generation-based CoT* (the model generates a reasoning chain and extracts the final answer) and *CoT-prefix prompting* (the reasoning instruction is appended but answer logits are extracted at the final input position without generating intermediate text) is critical. Generation-based CoT largely preserves accuracy (∼75%); CoT-prefix collapses it. This paper investigates why—and uncovers a concrete decoding vulnerability in how VLMs map hidden states to multiple-choice answer tokens. Figure 1 previews the core finding: the same hidden state can be correctly read by a small linear probe but misread by the native output head under CoT-prefix prompting.

Several recent studies have observed CoT-prefix degradation on visual reasoning benchmarks [14, 29, 28, 21], but do not identify the precise failure mode. Concurrent work on position bias [30, 22, 18, 25, 19] documents option-order sensitivity, but does not link it to CoT-prefix degradation or explain *why* reasoning prompts amplify it.

We hypothesize that CoT-prefix degradation stems from a disruption in the answer *decoding pathway*—the mapping from hidden representations to output logits—rather than from a destruction of answer information. We test this across three VLM architectures (Qwen2.5-VL 7B/3B, LLaVA-1.5-7B) and three benchmarks (ScienceQA, MMMU, AI2D) by comparing *native output* (standard logits) against *hidden-state decoding* (a lightweight linear probe), tracing the failure to a

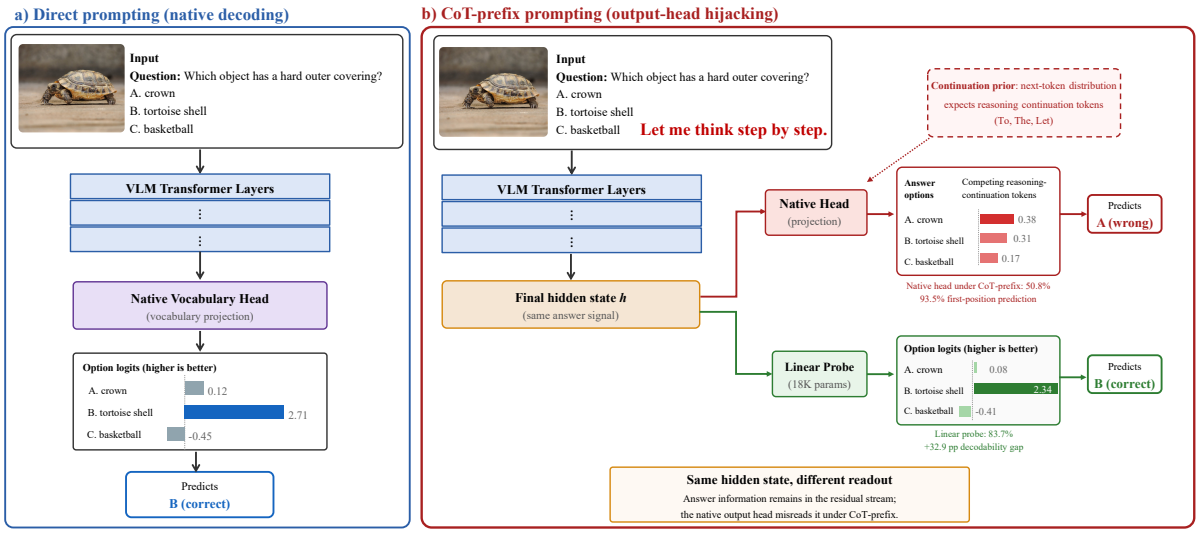

Figure 1: The CoT-prefix decoding bottleneck. (a) Under Direct prompting, the VLM's native output head reads the correct answer (B, "tortoise shell") from its final hidden state. (b) Under CoT-prefix prompting on the same image and question, the same hidden state still encodes the correct answer—an 18K-parameter linear probe recovers it (bottom path)—yet the native vocabulary projection emits the first-listed option A (top path), because its next-token distribution is dominated by reasoning-continuation tokens ("To", "The", "Let"). Answer-relevant information remains in the residual stream; only the vocabulary projection misreads it. Option-logit bars are schematic; quantitatively, under CoT-prefix the native head selects the first option in 93.5% of cases and accuracy falls to 50.8%, while the linear probe reads 83.7% (§3.1).

systematic position bias and validating it through option-order randomization, prompt-ending ablation, causal intervention, and correctability demonstration.

Our contributions are threefold:

- **Decodability gap**: CoT-prefix causes up to −54pp accuracy collapse, yet an 18K-parameter linear probe reads 83.7% from the same hidden states where the native head achieves only 50.8%—a 32.9pp gap showing answer information remains intact (§3.1).
- **Next-token continuation prior**: Option-order randomization proves the bias is positional (93.5% first-position, <1% A-token). Prompt-ending ablation and full-vocabulary token competition trace the origin to a continuation prior that shifts probability from answer tokens to reasoning tokens (§3.2).
- **Causal localization and correctability**: LoRA on only the output head (5M params) recovers 72% of degradation, confirming the vocabulary projection as the bottleneck. Probe fusion (ITAH) demonstrates the bottleneck is correctable at inference time (§5–§6).

Our goal is not to exhaustively benchmark all VLM families, but to isolate a concrete decoding mechanism under controlled conditions. §2–§8 present setup, the decodability gap and mechanism, cross-model validation, causal localization, correctability, related work, and conclusions.

2 Preliminaries

Models. We evaluate three VLM architectures: **Qwen2.5-VL-7B-Instruct** [2, 3] (28 layers, hidden dim 3584; primary model), **Qwen2.5-VL-3B-Instruct** (same family, 2048-dim), and **LLaVA-1.5-7B-hf** [11, 12] (Vicuna-7B + CLIP [17], 4096-dim). All models are loaded in bfloat16 via Hugging Face Transformers and evaluated frozen. We deliberately use instruction-tuned VLMs that predate large-scale reasoning-oriented RL post-training (e.g., long-CoT "thinking" variants such as Qwen3-VL-Thinking [4]): such training directly optimizes chain-of-thought behavior and would confound our isolation of the *latent* position-bias mechanism, since in a reasoning-RL model an absent bias cannot be distinguished from one the training has suppressed. Qwen2.5-VL and LLaVA-1.5 thus provide a clean, unconfounded testbed, and the non-thinking variants that current models still ship operate in exactly this regime.

Datasets. We use three multiple-choice VQA benchmarks spanning grade-school science to

college-level multidisciplinary material. **ScienceQA** [14]: 4,241 test questions, 2–5 choices (primary benchmark). Unless otherwise noted, results use the image-containing subset (N=2,017) for consistent visual input across all models. **MMMU** [27]: 286 college-level questions, used to test generalization to university-level multidisciplinary material. **AI2D** [6]: 3,088 science diagrams, used as a low-Direct-bias stress test.

Prompting conditions. We compare two protocols on identical forward passes. **Direct**: "{question}\nChoices:\n{A...}". **CoT-prefix**: Direct followed by "\nLet me think step by step." We extract answer-choice logits at the final input position—*before any generated tokens*—apply softmax, and select the highest-probability choice. This isolates the effect of the CoT instruction prefix on the output head, independent of generation artifacts. Unless explicitly noted (e.g., CoT-SC, generation-based CoT), all reported "CoT" results use this CoT-prefix protocol.

Hidden-state decoders. We extract the last-layer hidden state at the final input token and train decoders on frozen states from the training split (6,218 examples). The **linear probe** (18K params) is a single $h \rightarrow 5$-way projection, trained with cross-entropy and AdamW for 20 epochs, with separate probes for Direct and CoT-prefix conditions.

LoRA output-head intervention. We attach LoRA [5] to `lm_head` only (rank=32, ∼5M params), train on cached CoT-prefix hidden states, and merge back before evaluation. All transformer layers remain frozen.

Evaluation. Accuracy values report 95% Wilson score confidence intervals; paired comparisons use McNemar's test. Our primary analysis uses the ScienceQA image-containing subset (N=2,017); full-set evaluations (N=4,241) are noted where applicable. The prompt-ending ablation uses a different answer-token extraction from the main tables; the main degradation numbers always come from the bare-letter extraction.

Table 1: Native output vs. hidden-state decoding on ScienceQA (N=2,017 **image-containing examples**; 95% Wilson CIs). The decodability gap emerges specifically under CoT-prefix: a linear probe with only 18K parameters achieves 83.7% on CoT-prefix hidden states, far above the native head (50.8%). On the full N=4,241 test set the same probe scores 78.9% (Table 3); the difference is the test split, not the probe.

Method	Direct	CoT-prefix	Gap
Native Output	82.7% [81.0, 84.3]	50.8% [48.6, 53.0]	–
Linear Probe (18K)	84.8% [83.2, 86.3]	83.7% [82.0, 85.2]	32.9pp
vs. Native	+2.1pp	**+32.9pp**	–

3 The Decodability Gap

3.1 Main Result: Native vs. Probe under Direct and CoT-Prefix

Table 1 shows the key phenomenon. Under Direct prompting, native output (82.7%)[1] and hidden-state decoding (84.8%) achieve comparable accuracy. Under CoT-prefix, native output drops 31.9pp to 50.8%, but hidden-state decoding maintains 83.7%—a 32.9pp decodability gap. The 18K-parameter linear probe even slightly exceeds native Direct accuracy (82.7%), indicating that answer information is linearly separable and that CoT-prefix representations may concentrate answer-relevant features rather than disperse them.

To understand the mechanism, we analyze 200 examples where Direct and CoT-prefix predictions differ. Hidden-state similarity between Direct and CoT-prefix representations is high (91%, cosine similarity), suggesting CoT-prefix does not fundamentally alter the representation structure. Yet CoT-prefix reduces the correct-answer probability by 30.8% on average (Figure 2a), and 42.5% of predictions flip from correct (Direct) to incorrect (CoT-prefix). This pattern indicates that CoT-prefix specifically disrupts answer decoding at the output head while leaving hidden representations intact. CoT-prefix alters how the output head maps representations to answer tokens, without destroying the underlying answer signal in hidden states. The output head appears to distribute probability mass across reasoning-related tokens ("Let", "me",

[1] We validated the native baseline by cross-checking three evaluation methods on Qwen2.5-VL-7B—answer-token logits (primary, used throughout), forced generation (scoring each choice by its generation probability), and option-string logprob (averaging log-probability of the full answer text). Answer-token logits and forced generation produce identical results (80.8% Direct, 45.5% CoT-prefix on N=4,241), confirming the degradation is not an evaluation artifact.

"think", "step") at the expense of answer-choice tokens, while the hidden state itself retains sufficient information for accurate decoding.

A critical question is whether the observed degradation reflects a genuine failure of CoT reasoning, or merely an artifact of the CoT-prefix evaluation protocol. Generation-based CoT—where the model generates a reasoning chain before extracting the final answer—recovers accuracy to 75.2% (full set) and 80.6% (image subset), far above CoT-prefix (45.5%/45.3%) and close to Direct (80.8%/83.5%). This confirms the degradation is specific to CoT-prefix: the hidden states contain sufficient information when the model is allowed to generate.

3.2 Position Bias: The Core Mechanism

The foregoing analysis establishes that the output head fails under CoT-prefix, but not how. We now show the failure mode is a systematic position bias toward option A.

We test whether the degradation is specific to "Let me think step by step" or generalizes to alternative reasoning instructions. All four verbose triggers degrade accuracy to 45–53%, while the minimal trigger "Reasoning:" degrades far less (77.2%)—a monotonic relationship between continuation-cue strength and degradation that supports the continuation-prior hypothesis (Appendix A.2).

All three models exhibit position bias, but the strength varies dramatically: Qwen2.5-VL-7B shows extreme bias (90% A-prediction under CoT-prefix vs. 48% under Direct) while LLaVA-1.5-7B shows mild bias (47%). Crucially, CoT-prefix degradation magnitude is predicted by bias strength ($r \approx 0.99$ across models; Figure 3).

Breaking down CoT-prefix accuracy by the position of the correct answer reveals the mechanism starkly: when the correct answer is A, CoT-prefix excels (98.7% for Qwen 7B); when the correct answer is B, accuracy collapses to 7.1% (Appendix B.3). This pattern—high accuracy when A is correct, catastrophic failure when B is correct—is the signature of a position bias mechanism.

Position bias also varies with the number of answer choices: the decodability gap persists across 2–5 choices (Appendix A.3).

We hypothesize that the bias arises because reasoning-prefix endings shift the output head's next-token distribution toward continuation tokens ("Let", "me", "think", "step"), and among the remaining answer-choice tokens A wins by positional primacy. The following analysis tests this hypothesis directly at the prompt and full-vocabulary levels.

To definitively distinguish position bias from label bias, we randomize the answer-choice order across 5 seeds. Under randomization, accuracy remains unchanged ($45.8 \pm 0.7\%$) while the model predicts the first-listed option in 93.5% of cases—but predicts token "A" in fewer than 1% of cases when the correct answer is not first. The bias is positional, not label-based.

We hypothesize that the bias originates from a next-token continuation prior: when the prompt ends with "Let me think step by step.", the output head expects the next token to continue the reasoning chain rather than select an answer, and the first-listed option receives the highest residual probability as a positional shortcut. To test this, we systematically vary the prompt ending across 11 conditions. Continuation-inducing endings ("Let me think step by step.") produce the worst accuracy (70.3%) and highest first-position bias (48.5%); answer-eliciting endings ("The correct option is") recover to 83.1%, matching Direct; neutral endings maintain near-Direct accuracy (Appendix A.1). The logit margin closely tracks accuracy ($r = 0.96$), confirming the continuation prior compresses rather than eliminates the answer signal. A truncated-reasoning probe sharpens this: appending the cue "The answer is" to the CoT-prefix prompt with *zero* generated reasoning tokens already restores accuracy from 46.8% to 90.0%, and generating reasoning before answering does not improve further but slightly degrades it, flipping more initially-correct answers to wrong than the reverse (an overthinking effect; Appendix D.3). The failure is thus localized to the next-token expectation, not to a lack of reasoning. Full-vocabulary token competition analysis confirms the mechanism: under CoT-prefix the answer subspace probability falls by $\sim 14\times$ while the reasoning subspace probability rises by $\sim 4\times$ (Appendix C.2). CoT-prefix does not merely reorder a few answer tokens; it fundamentally shifts the output distribution away from the answer subspace.

4 Cross-Model and Cross-Dataset Validation

Having established position bias as the mechanism, we now show that its *severity* is strongly correlated with bias strength across both models and datasets

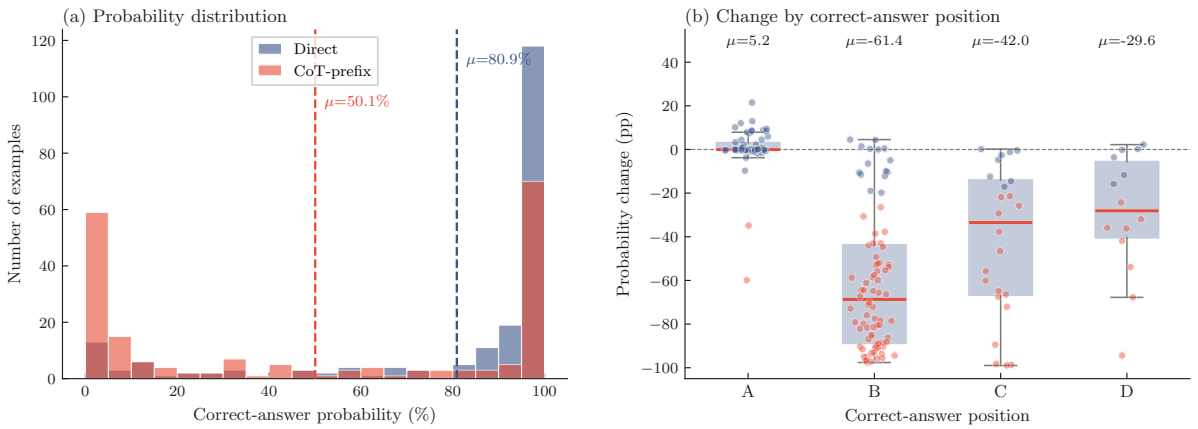

Figure 2: Correct-answer probability under Direct vs. CoT-prefix on the 200-example diagnostic subsample (Qwen2.5-VL-7B). (a) Distribution of correct-answer probabilities: Direct clusters near 100% (μ=80.9%); CoT-prefix shifts toward near-zero confidence (μ=50.1%). (b) Distribution of correct-answer probability *change* (CoT-prefix − Direct, in percentage points) split by the position of the correct option. Probabilities stay roughly flat when A is correct (the biased position), but drop by tens of points when the correct option is B, C, or D—the signature of position bias.

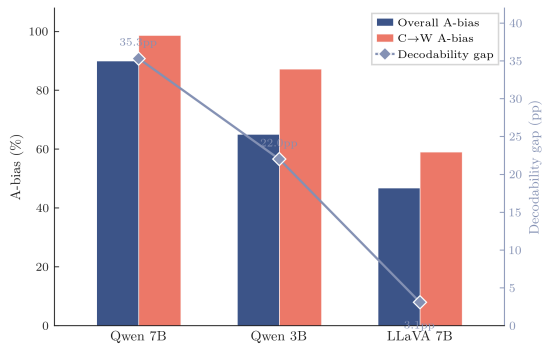

Figure 3: Position bias strength (bars, left axis) tracks degradation severity (green line, right axis). Qwen 7B: 90% A-bias → 35.3pp gap. LLaVA 7B: 47% A-bias → 3.1pp gap.

in our controlled settings. We do not claim a universal law; rather, we use this relationship as a *diagnostic signature* of the proposed mechanism.

Table 2 shows cross-model results: the decodability gap generalizes but varies substantially in magnitude (31.9pp for Qwen7B, 22.0pp for Qwen3B, 3.1pp for LLaVA).

*Probe accuracy on held-out test split (517 examples for 3B/LLaVA; train/val/test = 1200/300/517). ITAH accuracy evaluated on full 2,017 examples.

The cross-architecture variation is explained by position bias strength (Section 3.2). Qwen2.5-VL exhibits strong bias (90% for 7B, 65% for 3B), producing large gaps. LLaVA-1.5-7B exhibits mild bias (47%), producing a small gap. This correlation ($r \approx 0.99$) suggests the decodability gap is not an arbitrary architecture artifact but a systematic consequence of how strongly the output head defaults to option A under CoT-prefix. Despite vastly different gap magnitudes, linear probes achieve strong accuracy on CoT-prefix hidden states across all three models (83.7%, 69.8%, 65.6%), with probe accuracy correlating with native Direct accuracy ($r \approx 0.99$).

We also validate across datasets. On MMMU (N=286, 10 college-level subjects), CoT-prefix degrades 8 of 10 subjects with an aggregate gap of −16.1pp and 97.6% A-bias, confirming the mechanism generalizes beyond grade-school science (Appendix E.1). On AI2D [6] (N=3,088), a low-Direct-bias stress test, Qwen2.5-VL-7B collapses from 82.48% to 28.47% (−54.02pp—the largest gap in our study), with A-bias driven from 24% to 95% (Appendix E.2). The severity of degradation depends on the *margin* between Direct-mode capability and Direct-mode bias: benchmarks with already-high A-bias (ScienceQA, MMMU) have limited room for additional damage, while AI2D's low Direct bias leaves maximal room for CoT-prefix to devastate. Figure 4 summarizes this relationship. As a negative control, MMBench [13] uses a different option format where CoT-prefix is *not merely neutral but reverses direction*: CoT improves accuracy by +5.6pp (99.4% CoT vs. 93.8% Direct). The same prompt that destroys 54pp of accuracy on AI2D adds 5.6pp on MMBench, confirming the bias is format-dependent rather than CoT-dependent. Generation-based CoT recovers

Table 2: Cross-model validation on ScienceQA test set (N=2,017 image-containing examples). The decodability gap generalizes across architectures but varies substantially in magnitude. ITAH-Add (best α per model) provides recovery proportional to the gap size.

Method	**Qwen2.5-VL-7B (3584-dim)**	**Qwen2.5-VL-3B (2048-dim)**	**LLaVA-1.5-7B (4096-dim)**
Native Direct	82.7% [81.0,84.3]	80.1% [78.3,81.7]	66.1% [64.0,68.2]
Native CoT-prefix	50.8% [48.6,53.0]	58.1% [55.9,60.3]	63.1% [60.9,65.1]
Linear Probe CoT-prefix	83.7% [82.0,85.2]	69.8%*	65.6%*
ITAH-Add CoT-prefix (best α)	83.3% (α=0.1)	76.7% (α=0.1)	72.0% (α=0.1)
Decodability Gap	32.9pp	22.0pp	3.1pp
ITAH Recovery	+32.5pp	+18.5pp	+8.9pp

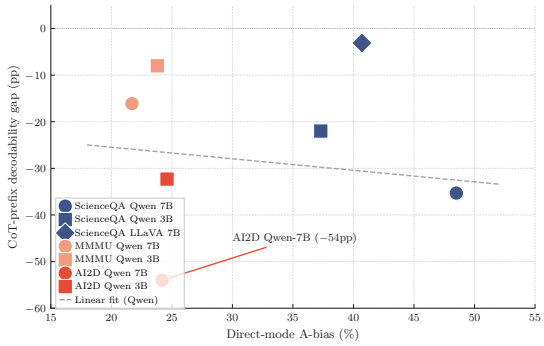

Figure 4: Bias–degradation diagnostic trend across evaluated settings. Each point represents one model–benchmark pair. Bias strength strongly tracks degradation severity ($r \approx 0.99$). Qwen models (circles/squares) cluster in the high-bias, high-degradation quadrant. LLaVA (diamond) remains in the low-bias, low-degradation quadrant.

accuracy across all 11 tested prompt endings (76–87%), confirming the degradation is specific to CoT-prefix (Appendix D.1).

5 Causal Localization

Individual case inspection shows that when the Direct prediction is correct and the CoT-prefix prediction is wrong, generation-based CoT still identifies the correct answer and the linear probe still recovers it—only the CoT-prefix output head fails. The decodability gap persists across probe architectures: even a linear probe (18K parameters) achieves 78.8% on CoT-prefix hidden states. Answer information concentrates in final layers (21–28), where probe accuracies rise sharply and the Direct–CoT-prefix gap widens monotonically. Cosine similarity between Direct and CoT-prefix hidden states remains high (∼0.91) across all layers—the hidden state does not change, only its readout does (Appendix C.1).

To establish causation, we fine-tune only the output head (all 28 transformer layers frozen). We add LoRA adapters [5] (rank=32, ∼5M parameters) to the `lm_head`, train on CoT-prefix hidden states from ScienceQA with answer-token supervision, then merge the adapter back. Whereas mechanistic-interpretability work [15] localizes factual associations to specific MLP layers, we constrain edits to the vocabulary projection—a clean causal test.

Table 3 shows: fine-tuning only the output head recovers CoT-prefix accuracy from 45.5% to 70.9% (+25.4pp, 72% recovery) on the 7B model. On the 3B model, the same intervention recovers from 55.7% to 73.0% (+17.3pp, 56% recovery). Direct accuracy is unchanged in both models, confirming that the intervention targets CoT-prefix-specific disruption without affecting Direct-mode decoding. This provides causal evidence that the output head is the primary bottleneck across model scales.

The remaining 8pp gap relative to the linear probe (78.9%) reflects LoRA's constraint of projecting to all 152K vocabulary tokens (vs. 5 answer choices for the probe). Even this minimal in-situ modification substantially recovers accuracy, confirming that most of the degradation originates in the output projection rather than in hidden representations.

6 Inference-Time Correction (ITAH)

The preceding analysis establishes that CoT-prefix degradation originates in the output head. We now demonstrate that this bottleneck is correctable without modifying model architecture or retraining. ITAH fuses native answer probabilities with linear probe predictions at inference time:

$$P_{\text{ITAH}} = \alpha P_{\text{native}} + (1-\alpha) P_{\text{probe}}, \qquad (1)$$

where $P_{\text{probe}} = \text{softmax}(W_{\text{probe}} h)$ is computed via a frozen linear probe pre-trained on CoT-prefix hidden states, and $\alpha \in [0,1]$ controls the

Table 3: Causal intervention via LoRA on lm_head. Qwen2.5-VL-7B evaluated on ScienceQA full test set (N=4,241); Qwen2.5-VL-3B evaluated on image subset (N=300 held-out). Fine-tuning only the output head recovers a substantial fraction of CoT-prefix degradation in both models.

	Qwen2.5-VL-7B (N=4,241)		**Qwen2.5-VL-3B** (N=300)	
Method	**Direct**	**CoT-prefix**	**Direct**	**CoT-prefix**
Native VLM	80.8%	45.5%	86.3%	55.7%
LoRA lm_head (5M)	80.6%	**70.9%**	86.3%	**73.0%**
Recovery	–	+25.4pp (72%)	–	+17.3pp (56%)

weight on native logits. We also evaluate ITAH-Product ($P \propto P_{\text{native}}^{\alpha} \cdot P_{\text{probe}}^{(1-\alpha)}$) and ITAH-Max ($P \propto \max(P_{\text{native}}, P_{\text{probe}})$). At α=0.1, ITAH-Add achieves 83.3% CoT-prefix accuracy, matching the linear probe (83.7%). The probe-dominated optimal α indicates that under CoT-prefix the native output head primarily contributes noise. Inference-time cost is $1\times$ (one extra $d\times 5$ projection per forward pass), plus a one-off offline probe training pass (<30 minutes on a single GPU).

ITAH transfers to Qwen2.5-VL-3B and LLaVA-1.5-7B using the same α=0.1 (Table 2), suggesting the relative reliability of native vs. probe signals is architecture-independent at this scale. Table 4 compares against CoT Self-Consistency (CoT-SC) [24] and the trained LoRA intervention.

*CoT-SC evaluated on 500 randomly sampled examples due to computational cost (5× forward passes per example). Mean inter-trace agreement: 84.7%.

†LoRA lm_head result from 4,241-sample evaluation (including non-image examples); comparable Direct accuracy (80.6%) confirms the intervention is non-destructive.

‡ITAH inference cost is $1\times$; offline cost is one linear-probe training pass (<30 min on a single GPU, 6,218 training examples).

Additive and product fusion with low α perform nearly identically (83.3% vs. 83.0%), both matching the pure probe (83.7%). The near-complete recovery with probe fusion confirms that the information required for correct answering remains in the hidden state; under CoT-prefix the native output head contributes primarily as a uniform smoothing term. Max fusion underperforms (72.7%) because taking the maximum across distributions can amplify false confidence.

7 Related Work

CoT in vision–language models. Chain-of-thought prompting [26] elicits step-by-step reasoning in LLMs and reliably improves multi-step reasoning [7], but its effect on visual question answering is mixed [14, 29]. Recent concurrent work has independently observed CoT-prefix degradation on visual benchmarks. [28] reported that CoT-prefix hurts performance on 8 of 11 VQA benchmarks and proposed training-based solutions (data distillation + DPO). [21] described a "dual nature of multimodal reasoning," showing that prolonged reasoning chains impair perceptual grounding. These works document the phenomenon but do not identify position bias as the mechanistic cause, establish a predictive relationship between bias strength and degradation severity, or propose lightweight inference-time correction.

Position bias in multiple-choice models. [30] showed that LLMs are not robust multiple-choice selectors, with option reordering causing dramatic accuracy fluctuations. [22] identified position bias in multi-image VLMs, [25, 19] studied position bias in LLM-as-a-judge and text-based settings, and [18] evaluated prompt-format sensitivity across 7 MLLMs. These works focus on (i) pure-text models without CoT, (ii) multi-image spatial position rather than answer-option position, or (iii) prompt-format artifacts rather than the CoT–position-bias interaction. To our knowledge, this is the first work to demonstrate that reasoning prompts systematically amplify position bias in VLM answer decoding, to trace this amplification to a next-token continuation prior, and to causally localize the bottleneck to the vocabulary projection layer.

CoT faithfulness and decoding. [8] examined whether CoT reasoning is faithful to the model's actual computation, and [23] showed that CoT explanations can be unfaithful when models are influenced by biased features. We extend this line by showing that CoT-prefix disrupts output mechanisms while reasoning signals remain intact in hidden states, identifying the precise mechanism (next-token continuation prior) and establishing a quantitative diagnostic.

Table 4: Recovery control comparison on ScienceQA (N=2,017). The near-complete recovery with probe fusion (ITAH-Add: 83.3%) confirms that the information required for correct answering remains in the hidden state. LoRA on lm_head provides causal evidence of output-head localization.

Method	**CoT-prefix Acc**	**Recovery**	**Inference Cost**
Native Direct	82.7%	–	1×
Native CoT-prefix	50.8%	–	1×
Recovery Methods			
CoT-SC (N=5, T=0.7)	82.4%*	+31.6pp	5×
Linear Probe (separate)	83.7%	+32.9pp	1× (offline)
LoRA lm_head (in-situ)	70.9%†	+25.4pp	1× (trained)
ITAH Fusion Variants			
ITAH-Add (α=0.1)	83.3%	+32.5pp	1×‡
ITAH-Product (α=0.1)	83.0%	+32.2pp	1×‡
ITAH-Max	72.7%	+21.9pp	1×‡
ITAH-Min	74.3%	+23.5pp	1×‡

Probing and inference-time intervention. Probing classifiers on frozen representations have been used to extract linguistic properties [20] and factual knowledge [16, 15]; in the multimodal domain, [9, 12, 1] show that visual information is encoded in early layers while semantic information concentrates in later layers. Inference-time intervention methods [10] steer LLM behavior toward desired outputs. Our probe-fusion approach extends this paradigm to the VLM decoding bottleneck, using the probe as a diagnostic instrument to demonstrate that the decoding bottleneck is correctable without retraining.

8 Discussion and Conclusion

The key findings form a closed causal chain. CoT-prefix causes severe multiple-choice accuracy collapse (up to −54pp), yet generation-based CoT largely recovers across all 11 tested prompt endings (76–87%); a linear probe reads 83.7% from the same hidden states where the native output head achieves only 50.8%—a 32.9pp decodability gap. The underlying bias is positional (93.5% first-position prediction, <1% A-token prediction). Prompt-ending ablation traces the origin to a next-token continuation prior: continuation-inducing endings degrade accuracy while answer-eliciting endings recover it, and full-vocabulary token competition confirms CoT-prefix shifts probability away from answer tokens toward reasoning-continuation tokens (4.0× increase). LoRA on only the output head recovers 72% (7B) and 56% (3B) of the degradation, causally confirming the vocabulary projection layer as the bottleneck. The bias generalizes across datasets (ScienceQA, AI2D, MMMU) but reverses on MMBench, where CoT-prefix improves accuracy by +5.6pp—a stark format-dependence that further confirms the mechanism is not about CoT semantics but about how CoT endings interact with the option-token distribution.

Three corrections, one bottleneck. Three independent corrections converge on the same diagnosis: an input-side answer cue (zero reasoning) recovers 90%, a read-out linear probe 83.7%, and an output-head LoRA 70.9%—none altering the transformer's hidden representations, all localizing the bottleneck to the output head's next-token expectation rather than to missing answer information. The cue recovers most because it intervenes furthest upstream, reframing the expected next token *before* the head reads the hidden state; generation-based CoT is thus neither necessary nor optimal, as the cue exceeds full reasoning, which slightly overthinks (Appendix D.3). Our contribution is accordingly a *diagnosis*, not a decoder.

The central claims here are mechanistic rather than leaderboard-oriented: they rely on controlled prompt-ending interventions, option-order randomization, hidden-state probes, and output-head-only causal intervention. More broadly, reasoning prompts can expose latent decoding vulnerabilities in VLMs—position bias is not introduced by CoT but amplified by it—so evaluation protocols should routinely include position-bias diagnostics, and training procedures should incorporate position-invariant objectives.

9 Limitations

Our study has several limitations that should be considered when interpreting the results.

Model scale and reasoning post-training. All experiments use models with at most 7B parameters (Qwen2.5-VL-7B/3B, LLaVA-1.5-7B). Whether the position-bias mechanism persists, weakens, or changes form in larger VLMs (e.g., 72B or 100B+ scale) remains an open question; the LoRA recovery rate (72% for 7B, 56% for 3B) hints at scale-dependent correctability, but we do not extrapolate beyond the tested range. We also exclude reasoning-RL "thinking" VLMs (e.g., Qwen3-VL-Thinking [4]) by design: these models are post-trained to generate and exploit long CoT, which both confounds the mechanism we isolate and likely suppresses its surface symptom. Our generation-based-CoT control—which recovers 75–80% accuracy—already shows that *actually reasoning* mitigates the failure; the contribution of this paper is to expose and localize the latent output-head vulnerability in the zero/short-generation regime that non-thinking models continue to operate in. Whether the bias re-emerges under adversarial prompt formats even in reasoning-RL models is left to future work.

Task scope. Our analysis is restricted to multiple-choice VQA benchmarks. Open-ended generation, fill-in-the-blank, and ranking-based evaluation formats may exhibit different CoT-prefix dynamics. The MMBench negative control already shows format-dependence, suggesting the mechanism is sensitive to how answer options are tokenized and presented.

Mechanistic granularity. While we identify the output head as the primary bottleneck via causal intervention, the precise token-level dynamics of why the first-listed option becomes the positional default are not fully characterized. The continuation-prior hypothesis explains the direction of the effect but not the specific mapping from next-token expectation to first-position primacy.

Correction scope. The proposed diagnostic (linear probe) and inference-time correction (ITAH) demonstrate that the bias is correctable, but they do not eliminate it at source. Training-time solutions—such as position-invariant objectives or debiased CoT fine-tuning—could provide more fundamental remedies and represent important future work.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

[3] 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, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

[4] Shuai Bai and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

[6] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pages 235–251.

[7] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

[8] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

[9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742.

[10] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time

intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*.

[11] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*.

[13] 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. 2024. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233.

[14] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*.

[15] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*.

[16] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763.

[18] Fabio Rosenthal, Sebastian Schmidt, Thorsten Graf, Thorsten Bagodonat, Stephan Günnemann, and Leo Schwinn. 2025. Unexplored flaws in multiple-choice vqa evaluations. *arXiv preprint arXiv:2511.22341*.

[19] Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in llm-as-a-judge. In *Proceedings of the 2025 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (IJCNLP-AACL)*.

[20] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

[21] Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter Tu, and Jing Zhang. 2025. More thought, less accuracy? on the dual nature of reasoning in vision-language models. *arXiv preprint arXiv:2509.25848*. Under review at ICLR 2026.

[22] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. 2025. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10599–10609.

[23] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*.

[24] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

[25] Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. 2025. Eliminating position bias of language models: A mechanistic approach. In *International Conference on Learning Representations*.

[26] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

[27] 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, and 3 others. 2024. 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*, pages 9556–9567.

[28] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2025. Improve vision language model chain-of-thought reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

[29] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alexander Smola. 2024. Multimodal chain-of-thought reasoning in language mod-

els. *Transactions on Machine Learning Research*. ArXiv:2302.00923.

[30] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*.

Appendix

A Ablation Studies

A.1 Prompt-Ending Ablation Details

Table 5 shows the full prompt-ending ablation across 11 conditions.

A.2 Alternative CoT-Prefix Triggers

Table 6 shows that all four verbose triggers degrade accuracy to 45–53%, with A-bias rising to 86–94%. The minimal trigger "Reasoning:" degrades far less (77.2%, 56% A-bias).

Table 6: Alternative CoT-prefix trigger phrases (ScienceQA full test set N=4,241, Qwen2.5-VL-7B). All verbose triggers induce severe degradation; the minimal trigger "Reasoning:" is comparatively mild.

Trigger phrase	Accuracy
Direct (no trigger)	80.8%
"Let me think step by step."	45.5%
"Let's think step by step."	49.2%
"Let's reason step by step."	52.8%
"Let's solve this step by step."	52.5%
"Reasoning:"	77.2%

A.3 Accuracy by Number of Answer Choices

Table 7 reports accuracy across 2–5 choices on Qwen2.5-VL-7B: native CoT-prefix stays in the 37–48% band while the linear probe maintains 66–82%, showing the decodability gap is consistent across the choice axis.

Table 7: Accuracy by number of answer choices on ScienceQA (N=4,241, Qwen2.5-VL-7B). The decodability gap is large and roughly constant across 2–5 choices.

Method	2-choice	3-choice	4-choice	5-choice
Native Direct	86.2%	77.5%	75.4%	71.1%
Native CoT-prefix	48.2%	42.5%	40.1%	36.8%
Linear CoT-prefix	82.3%	76.4%	75.0%	65.8%

B Position Bias Analysis

B.1 Option-Order Randomization

Table 8 shows the option-order randomization results.

Table 8: Option-order randomization on ScienceQA (N=4,241, Qwen2.5-VL-7B). Mean and std. over 5 independent shuffle seeds.

Condition	CoT Acc.	1st-pos.	A-pred.
Original order	45.6%	94.1%	94.1%
Shuffled (mean)	$45.8 \pm 0.7\%$	$93.5 \pm 0.4\%$	$< 1\%$

B.2 Position Bias Quantification

Table 9 quantifies position bias across models.

Table 9: Position bias quantification on ScienceQA (N=4,241). A-bias = fraction of CoT-prefix predictions that are A.

Model	Direct A%	CoT-prefix A%	C→W A%	Gap
Qwen2.5-VL-7B	48.5%	**90.0%**	**98.7%**	35.3pp
Qwen2.5-VL-3B	37.3%	65.1%	87.2%	22.0pp
LLaVA-1.5-7B	40.7%	46.8%	59.0%	3.1pp

B.3 CoT-Prefix Accuracy Breakdown by Correct-Answer Position

Table 10 breaks down CoT-prefix accuracy by the position of the correct answer. When the correct answer is A, CoT-prefix excels (98.7% for Qwen 7B); when it is B, accuracy collapses to 7.1%.

C Mechanistic Analysis Details

C.1 Layer-Wise Decodability

C.2 Token Competition Details

Table 11 provides the full-vocabulary token competition analysis. Under CoT-prefix the answer subspace probability falls by $\sim 14\times$ while the reasoning subspace probability rises by $\sim 4\times$.

Table 11: Full-vocabulary token competition analysis (ScienceQA, Qwen2.5-VL-7B). Answer subspace $= \sum P(\text{A..E})$. Reasoning subspace $= \sum P(\text{reasoning tokens})$.

Metric	Direct	CoT-prefix
Answer subspace P	2.15×10^{-8}	1.51×10^{-9}
Reasoning subspace P	3.73×10^{-2}	1.49×10^{-1}
Top-1 is answer token (%)	0.0	0.0
Top-1 is reasoning token (%)	4.3	9.9
Correct answer rank (mean)	5,251	26,562
Correct answer in top-10 (%)	0.0	0.0

D Generation-Based CoT Recovery

This section shows that generation-based CoT—where the model produces a reasoning chain before extracting the final answer—recovers accuracy across all tested prompt endings. We compare Direct, CoT-prefix, and generation-based CoT (§D.1),

Table 5: Prompt-ending ablation organized by expected next-token type (ScienceQA, Qwen2.5-VL-7B, N=2,017).

Category	Ending	%Acc	%1st-pos	Margin
Continuation-inducing				
	"Let me think step by step."	70.3	48.5	2.24
	"step by step."	76.8	37.2	3.02
	"Let me think"	81.7	32.2	4.33
Answer-eliciting				
	"The answer is"	85.1	26.2	6.12
	"Answer:"	85.3	27.8	6.14
	"...step by step. The correct option is"	**83.1**	34.2	3.50
	"...step by step. Answer:"	74.0	47.1	2.49
Neutral				
	Direct (no suffix)	82.3	31.3	4.98
	"Look at the image carefully."	83.0	30.0	4.74
	"Let's analyze this."	80.2	32.0	4.17

Table 10: CoT-prefix accuracy breakdown by correct-answer position (ScienceQA, Qwen2.5-VL-7B/3B).

	Qwen 7B				Qwen 3B			
Correct	N	Acc	Pred A	Pred B	N	Acc	Pred A	Pred B
A	78	98.7%	98.7%	1.3%	713	88.5%	88.5%	5.0%
B	84	7.1%	92.9%	7.1%	726	30.3%	63.4%	30.3%
C	24	25.0%	75.0%	0.0%	375	47.5%	45.9%	2.9%
D	14	50.0%	50.0%	0.0%	195	73.3%	22.1%	1.5%

and report per-ending results across 11 prompt variants (§D.2).

D.1 Generation-Based CoT Comparison

Table 12 compares Direct, CoT-prefix, and generation-based CoT; Table 13 shows generation-based CoT across all 11 endings.

Table 12: Direct vs. CoT-prefix vs. generation-based CoT (Qwen2.5-VL-7B). Full test set N=4,241; image subset N=2,017.

	Full (N=4,241)		Image (N=2,017)	
Method	Acc.	A-bias	Acc.	A-bias
Direct	80.8%	48.5%	83.5%	44.8%
CoT-prefix	45.5%	94.1%	45.3%	89.6%
Generation-based CoT	75.2%	42.1%	80.6%	35.9%

D.2 Generation-Based CoT Across Prompt Endings

Table 13 shows generation-based CoT results across all 11 prompt endings.

Table 13: Generation-based CoT across 11 prompt endings (ScienceQA N=2,017, Qwen2.5-VL-7B). All endings recover to near-Direct accuracy under generation.

Ending	Gen-CoT Acc.	A-bias	Parse Fail
"therefore"	**87.5%**	34.2%	39
"in_conclusion"	86.9%	33.7%	37
"the_answer_is"	86.9%	33.9%	47
"hence"	86.7%	33.8%	36
"so"	86.4%	34.5%	39
"restate"	86.4%	35.0%	69
"analyze"	82.4%	34.8%	85
"observe_carefully"	82.2%	34.2%	67
"think_step_by_step"	80.6%	35.9%	121
"work_through"	79.5%	37.7%	156
"reason_before"	76.8%	37.5%	198

D.3 Truncated-Reasoning Answer Commitment

To separate the effect of the answer cue from the effect of reasoning content, we probe answer commitment as a function of reasoning length (Qwen2.5-VL-7B, N=500 image subset). We greedily generate the CoT-prefix reasoning chain and, at truncation lengths $t \in \{0, 4, 8, 16, 32, 64, 128\}$, append the cue "The answer is" to the first t generated tokens and read the answer-choice logits—i.e., "if the model stops reasoning after t tokens and is forced to answer, what does it commit to?" Figure 6 shows the result. The answer cue alone, with *zero* generated reasoning tokens (t=0), already restores accuracy from the no-cue CoT-prefix baseline of 46.8%

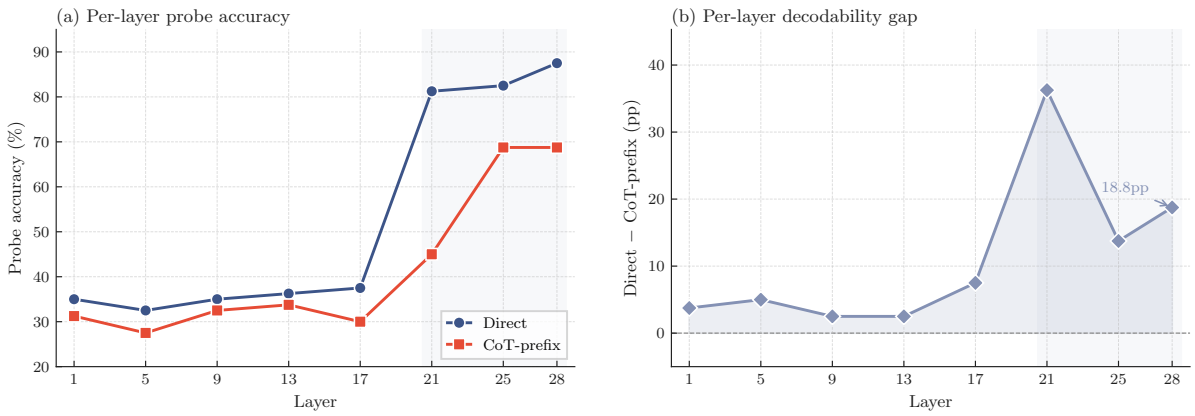

Figure 5: Layer-wise decodability on the 200-example diagnostic subsample (Qwen2.5-VL-7B, ScienceQA). (a) Per-layer probe accuracy: both Direct and CoT-prefix remain near chance (∼35%) until layer 17, then rise sharply in the late-layer band (21–28, shaded). (b) Per-layer decodability gap. The gap is essentially zero through the middle of the network and widens monotonically in the late-layer band, peaking around layer 28.

to 90.0%; generating reasoning before answering does *not* improve on this and slightly degrades it (78–87%). Comparing t=0 to full reasoning on the same items, 51 initially-correct answers flip to wrong while only 17 flip correct—a net overthinking cost consistent with [21]. This localizes the CoT-prefix failure to the next-token expectation at the output head rather than to a lack of reasoning: reframing what token the model expects to emit next is sufficient to recover accuracy.

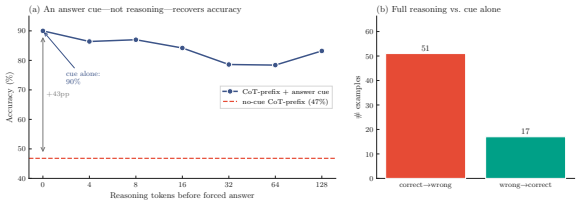

Figure 6: Truncated-reasoning answer commitment (Qwen2.5-VL-7B, ScienceQA N=500 image subset). (a) Accuracy when the model is forced to answer after t generated reasoning tokens (with the cue "The answer is"), against the no-cue CoT-prefix baseline. The cue alone (t=0, no reasoning) recovers +43pp; additional reasoning does not help. (b) Within-example answer changes from t=0 to full reasoning: reasoning flips more correct answers to wrong (51) than wrong to correct (17), an overthinking effect.

E Cross-Benchmark Results

E.1 MMMU Per-Subject Results

Table 14 provides the full per-subject breakdown on MMMU.[2]

Table 14: MMMU validation full results per subject (N=286 total, Qwen2.5-VL-7B). CoT-prefix degrades native accuracy on 8 of 10 subjects.

Subject	Direct	CoT-prefix	Gap
Computer Science	51.9%	33.3%	−18.5pp
Math	51.7%	44.8%	−6.9pp
Physics	44.8%	31.0%	−13.8pp
Chemistry	47.8%	21.7%	−26.1pp
Biology	41.4%	27.6%	−13.8pp
Accounting	33.3%	36.7%	+3.3pp
Economics	44.8%	44.8%	0.0pp
History	63.3%	33.3%	−30.0pp
Art	66.7%	26.7%	−40.0pp
Music	33.3%	16.7%	−16.7pp
Aggregate	**47.9%**	**31.8%**	**−16.1pp**

E.2 AI2D Stress Test Details

Table 15 provides the full AI2D results as a low-Direct-bias stress test.

[2]The per-subject analysis here uses the 286-question multiple-choice multimodal subset where we can extract clean native-VLM logits under both Direct and CoT-prefix. As an additional probe-stability check, we also ran the linear probe on the full MMMU validation set (N=2,118): probe accuracy is 32.9% under Direct and 30.2% under CoT-prefix (probe gap −2.7pp), confirming that on this much larger and harder benchmark the probe is essentially unperturbed by CoT-prefix—consistent with the decodability-gap story on ScienceQA. We retain the smaller N=286 subset for the main MMMU analysis because it admits a clean per-subject breakdown.

Table 15: AI2D as a low-Direct-bias stress test (N=3,088). CoT-prefix drives A-bias from $\sim$24% to 95% (7B) / 67% (3B), producing the largest gap in our study ($-$54pp).

Model	Qwen 7B	Qwen 3B
Direct accuracy	82.48%	79.02%
CoT-prefix accuracy	28.47%	46.73%
Gap	$-$54.02pp	$-$32.29pp
Direct A-bias	$\sim$24%	$\sim$24%
CoT-prefix A-bias	94.66%	66.71%
% of C$\rightarrow$W errors predicting A	99.2%	86.4%