Knowing Isn't Always Saying: When Do Spatial Encodings Reach Answers in Vision-Language Models?

Zeyu Wang*
Peking University
Beijing, China
2501210409@stu.pku.edu.cn

Xinming Xu*
Tsinghua University
Beijing, China
xu-zh23@mails.tsinghua.edu.cn

Abstract

Vision-language models are known to encode spatial information in their hidden states, yet often fail to use it when answering. However, it remains unclear when and where this encoded information reaches the answer. We address this with direction patching, a class-conditioned causal intervention applied across layers, token positions, and prompt formats. Using spatial-ID directions constructed following prior encoding evidence, we find that causal influence on answer logits emerges only at mid-to-deep depths. Text chain-of-thought suppresses immediate object-word argmax-level transport in most models, while visually grounded prompts keep it open. Positive target-logit gain can remain below the argmax threshold, and transport can re-emerge at the final prefix token or at the answer step in deeper layers. Across the ten VLMs we study, these local effects form descriptive transport patterns. Complementary experiments characterize how these patterns shift across datasets, attributes, and encoding amplitudes. Together, these results reframe the encoding-grounding gap as a problem of conditional transport in VLMs.

1 Introduction

Vision-Language Models have achieved remarkable progress across a wide range of visual reasoning tasks. However, recent interpretability studies reveal a puzzling dissociation between representation and behavior: models can correctly encode visual attributes in their hidden states, yet often fail to use that information when producing answers (Nooralahzadeh et al., 2026; Liu et al., 2026; Asadi et al., 2026). Chain-of-Thought prompting is commonly adopted to mitigate reasoning failures, but on spatial tasks it can produce the opposite effect: introducing text CoT often substantially degrades performance across multiple open-source VLMs.

To understand this paradox, we move beyond static representation analysis and instead trace the causal dynamics of information flow. Prior work has established the encoding-utilization dissociation (Li et al., 2025; Kamath et al., 2023), but it remains unclear under what local conditions encoded visual information causally reaches the answer. The missing piece is a transport-level account: when the spatial signal becomes causally effective, where it is routed inside the prompt, and how reasoning prompts change that route.

We investigate these questions using *direction patching* (Kang et al., 2026), a causal intervention inspired by activation patching (Vig et al., 2020; Wang et al., 2023). The method manipulates class-conditioned hidden-state directions at controlled layers and token positions, isolating their direct causal effect on answer logits. We build on the spatial-ID directions established by Kang et al. (2026) and contribute the first systematic mapping of how these directions transport to answer logits across layer, position, and prompt format.

Applying direction patching to ten VLMs across three datasets and seven prompt formats, we analyze spatial transport at both a local and a model-level scale. Three local regularities emerge. *Late emergence*: the spatial-ID direction is constructed from early-layer representation evidence, but causal transport activates only at L12–L32. *CoT gating*: text CoT suppresses immediate object-word argmax-level transport in most models, whereas visual grounding prompts keep it open. *Position dependence*: spatial transport can reappear at the final prefix token in deeper layers, making the pathway jointly governed by prompt format and token position.

At the model level, these local patterns organize into a small number of descriptive transport groupings rather than a single uniform behavior. We further test the logit-level map with generate-mode behavioral experiments, where steering in-

*Equal contribution.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 52683–52705
October 24-29, 2026 ©2026 Association for Computational Linguistics

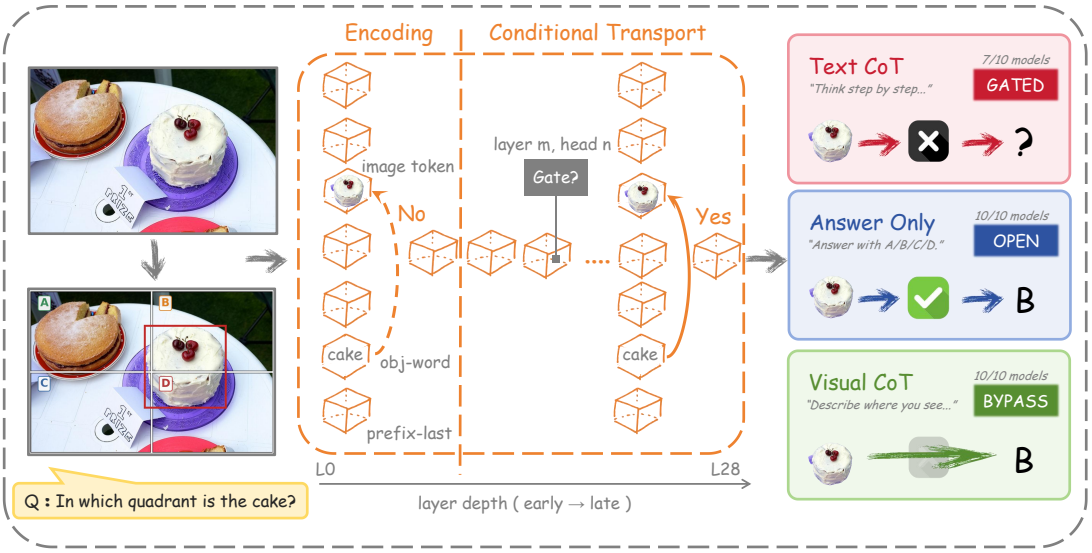

Figure 1: Direction patching reveals conditional transport. Spatial encoding is broadly present across VLMs, but whether it reaches the answer depends on prompt type, token position, and layer depth. Text CoT can block object-word transport, whereas visual grounding prompts keep this pathway open.

terventions induce substantial accuracy changes under open transport pathways, and with a head-knockout probe that provides initial, architecture-specific causal evidence for a CoT-specific gating component in one architecture. The patterns also replicate across datasets and expose attribute- and amplitude-dependent boundaries. Together, the results provide a mechanistic mapping of how text CoT can alter spatial transport (Kancheti et al., 2026) and identify visual grounding prompts (Jiang et al., 2025; Qin et al., 2025) as a prompt-only intervention that keeps the object-word pathway open.

Model	Family	Params	Layers
Qwen2.5-VL-7B	Qwen	7B	28
Qwen2.5-VL-32B	Qwen	32B	64
LLaVA-OneVision-7B	LLaVA	7B	28
InternVL2.5-8B	InternVL	8B	32
InternVL3-8B	InternVL	8B	28
InternVL3-14B	InternVL	14B	48
Gemma-3-4B-IT	Gemma	4B	34
Gemma-3-12B-IT	Gemma	12B	48
Gemma-3-27B-IT	Gemma	27B	62
Janus-Pro-7B	DeepSeek	7B	30

Table 1: Ten evaluated VLMs spanning five families, including three within-family scale pairs: Qwen 7B/32B, InternVL3 8B/14B, and Gemma-3 4B/12B/27B. Row tints group rows by family.

2 Experiment

2.1 Tasks

We evaluate on three four-way MCQ tasks that share a common pipeline but target different visual attributes. **RefCOCO spatial**: from RefCOCO/RefCOCO+ (Kazemzadeh et al., 2014; Yu et al., 2016) referring expressions, each instance becomes *"In which quadrant is the [object]? (A) top-left (B) top-right (C) bottom-left (D) bottom-right"*, with the gold label set by the bounding-box centroid. Ambiguous boundary cases are filtered. We use n=50 per quadrant for 7B/8B models and n=30 for Qwen-32B, Gemma-3, and Janus. **GQA spatial** uses analogous four-quadrant questions derived from GQA scene graphs (Hudson and Manning, 2019), providing different images and object distributions. **GQA color** replaces the spatial question with *"What color is the [object]?"* (red / blue / white / black); the centroid computation, cross-fit, and α settings are identical, only the patched direction changes from spatial-ID to color-ID.

2.2 Models

We evaluate ten open-weight VLMs from five families (Table 1): Qwen2.5-VL-7B and -32B (Bai et al., 2025), LLaVA-OneVision-7B (Li et al., 2024), InternVL2.5-8B (Chen et al., 2024), InternVL3-8B and -14B (Zhu et al., 2025), Gemma-3-4B, -12B, and -27B (Gemma Team, Google

DeepMind, 2025), and DeepSeek-Janus-Pro-7B (Chen et al., 2025b). All models expose residual-stream activations required for direction patching. The set includes three within-family scale pairs (Qwen 7B/32B, InternVL3 8B/14B, Gemma-3 4B/12B/27B), enabling controlled comparisons of how model scale affects transport.

2.3 Method: Direction Patching

We build on the spatial-ID encoding reported by Kang et al. (2026). At each layer ℓ, we compute class-conditioned centroids $C_q^{(\ell)}$ (one per quadrant or color) from hidden states at the object-word token position, using a five-fold crossfit split to avoid circularity. The intervention shifts the residual-stream activation at position p, layer ℓ:

$$h_p^{(\ell)} \leftarrow h_p^{(\ell)} + \alpha\big(C_{\text{target}}^{(\ell)} - C_{\text{source}}^{(\ell)}\big),$$

where α=5 for obj_word and α=10 for prefix_last unless stated otherwise. These canonical values were chosen from an $\alpha \in \{1, 2, 5, 10\}$ sweep as the midpoint of the monotonically increasing dose-response range (App. E). The model then completes a forward pass and we read next-token logits over {A, B, C, D}.

Metric. Δargmax = target flip rate under intervention − target flip rate under a same-norm random-direction baseline. A positive value indicates that the patched direction causally shifts the model's answer toward the target class beyond what a generic perturbation achieves. We also report the continuous target-logit gain, defined as the mean target-letter logit under the direction patch minus the corresponding mean under the random-direction baseline. Thus, target-logit gain measures directional influence even when the four-way argmax does not change. We use generate-mode experiments (128 tokens) to measure the change in parsed-answer accuracy (Δacc) as a separate behavioral validation. For these runs, we additionally report Δ_{parse}, the intervention-minus-noise contrast in the final parsed target rate; it is distinct from both Δacc and the fixed-state next-token Δargmax.

Operational definition of transport. For the coarse answer-option diagnostic, we use *argmax-level transport* to mean a non-zero Δargmax under direction patching at a given layer ℓ, token position p, and prompt format. A zero Δargmax does not imply zero target-logit gain. The measurement is causal evidence that the spatial direction influences the answer competition along the linear subspace identified by Kang et al. (2026); it is silent on transport that may occur through non-linear pathways or alternative directions outside this subspace (see Limitations).

Intervention positions. We patch at two positions: **obj_word**, the token span matched to the mentioned object expression (falling back to content-word anchor tokens when an exact span match is unavailable), and **prefix_last**, the final token of the input sequence before generation begins. The prefix_last experiments intentionally reuse the object-word-derived spatial direction rather than recomputing a position-specific centroid; they therefore test whether the same direction can be read out at a later position.

Prompt formats. We use seven instruction suffixes appended to a shared question stem, spanning a reasoning-length spectrum from *cot* ("Let's think step by step") through *brief_reason*, *answer_first*, *final_tag*, and *answer_only* ("Answer with only the letter"), plus two visual-grounding prompts (*visual_cot*, *visual_direct*) that direct reasoning toward visual observation rather than abstract spatial logic. Verbatim templates are in App. A.

3 Local Dynamics of Conditional Transport

We first examine conditional transport at the local scale, across controlled layers and token positions. Spatial-ID directions can be available before they causally affect the answer (§3.1). Transport from encoding to answer is then conditional along two axes: prompt type (§3.2, §3.3) and token position (§3.4). Layer depth tells us *when* the pathway activates; token position tells us *where* it is routed.

3.1 Spatial Encoding Does Not Immediately Become Transport

Following Kang et al. (2026), we extract object-word hidden states and compute per-quadrant centroids via five-fold crossfit. We take the early-layer class-contrastability used to construct this direction as prior encoding evidence; this subsection does not introduce an independent separability or probing curve. Under the answer-only prompt, direction patching at the object-word position yields positive Δargmax in all ten models, with peak transport clustering in L8–L16 for 7B/8B models and shifting to L20–L32 for larger models. The same

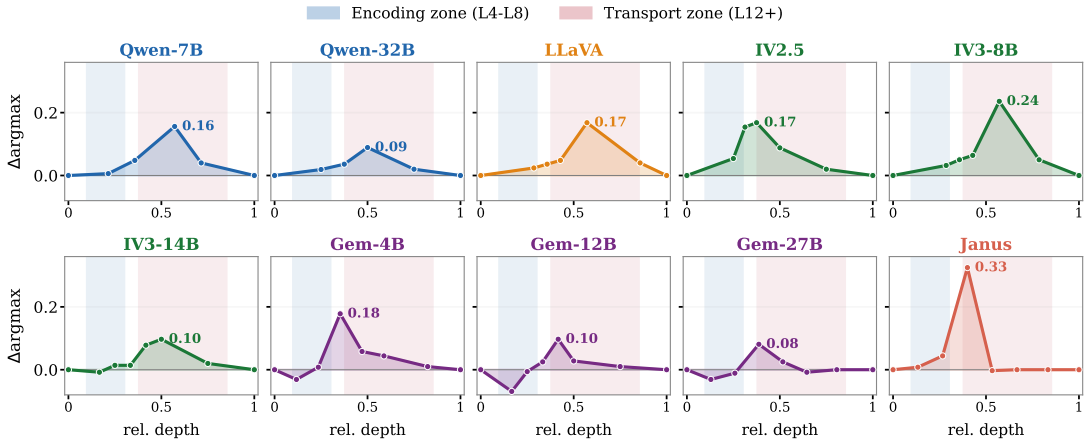

Figure 2: Layer-wise Δargmax for ten models (obj_word, answer_only, α=5). Argmax-level transport activates only at mid-to-deep layers, although the Kang-style spatial-ID direction is constructed from early-layer representation evidence.

object-word centroid construction is used across prompt formats, so the prompt effects reported below are not attributable to changing the patched spatial direction. The question is whether these available spatial directions influence the answer as soon as they are detectable, or only after later-layer processing.

Figure 2 plots Δargmax across layers for all ten models under the answer-only prompt at the object-word position. Shallow layers (L0–L8) produce near-zero Δargmax in every model, even though the spatial-ID direction is available from the early-layer representation used for its construction. Transport activates in mid-to-deep layers (Qwen-7B L16 $+0.156$; Gemma-3-4B L12 $+0.178$; Janus L12 $+0.325$, the largest peak we observe). For larger models the peak appears at a higher absolute layer index (Qwen-32B L32 of 64, $+0.089$), although the depth relative to total layer count is comparable to the 7B/8B models.

The gap between early-layer direction availability (L4–L8) and transport depth (L12–L32) shows that representation and causal use are dissociable within a single forward pass. Object-word tokens carry spatial directions from early layers, but those directions do not causally shift the answer distribution until many layers later. Thus, detecting a spatial code is not sufficient to show that the code is being transmitted to the answer; transport has its own layer condition.

Controls. Wrong-axis and non-object-position patching both stay within an order of magnitude of zero ($[-0.018, +0.031]$ and $[-0.017, +0.006]$, respectively) across all ten models, confirming the effect is direction- and position-specific (App. C).

3.2 CoT Suppresses Immediate Object-Word Transport

Having established an answer-only transport pathway at obj_word, we next change only the prompt: from answer_only to standard chain-of-thought ("Let's think step by step").

At the object-word position, this single change suppresses the immediate argmax-level transport established in the previous subsection. Eight of ten models show CoT peak Δargmax ≤ 0.006 at obj_word (we use < 0.01 as a descriptive near-zero threshold; six of these are exactly zero and LLaVA and Qwen-32B are borderline near-zero cases), including the model with the strongest answer-only peak: Janus drops from $+0.325$ to below this resolution. The three Gemma-3 variants are uniformly near-zero regardless of scale, and Qwen-32B sits within the noise floor at $+0.003$. The two exceptions are Qwen-7B ($+0.120$) and InternVL2.5 ($+0.042$, 75% reduction from answer-only); both are analysed in §4. Per-model peak values and bootstrap CIs are in App. B.

Gating is argmax-level, not signal erasure. If CoT merely weakened the signal, increasing α should systematically lift CoT Δargmax above the near-zero range. Sweeping $\alpha \in \{1, 2, 5, 10\}$ (App. E), answer-only grows across the dose range, whereas the suppressed models remain at or near zero; the two borderline cases do not show a systematic rise into the open range. The block is at the argmax-flip level, not at the underlying logit: under CoT, IV3-8B's Δargmax is 0 while the target-

Model	CoT	Vis-CoT	Ans-only
Qwen-7B	+.081	+.136	+.169
Qwen-32B	+.006	**+.117***	+.089
LLaVA	.000	+.103	+.158
InternVL2.5	.000	**+.153***	+.114
InternVL3-8B	.000	+.164	+.236
InternVL3-14B	.000	**+.117***	+.097
Gemma-3-4B	.000	+.092	+.178
Gemma-3-12B	.000	+.064	+.097
Gemma-3-27B	.000	+.022	+.081
Janus	.000	**+.403***	+.069

Table 2: Δargmax at obj_word at each model's visual_cot peak layer (RefCOCO, α=5). *: visual_cot exceeds answer_only. All ten models recover positive transport under visual_cot, including eight suppressed under text CoT.

letter logit gain remains +0.421 (App. D), a sub-threshold directional signal the head knockout in §4.2 exploits. This connects to the accuracy-side observation of Kancheti et al. (2026), who report CoT-induced spatial accuracy degradation without identifying a mechanism: in our data, CoT suppresses the obj_word argmax-level route in the majority, so the patched direction does not reliably change the answer option competition at that position.

Answer-step timing control. The canonical readout fixes the prompt state, layer, and token position so that it measures a comparable immediate answer-option effect. To test whether this choice misses a later answer readout, we generated a clean CoT prefix, located the model's actual answer-letter step, reconstructed the prefix immediately before that letter, and repeated the same obj_word direction patch. In LLaVA-OV-7B, Qwen2.5-VL-7B, and InternVL3-8B, the target-logit gain is positive across L8–L20 and fades at L24; at L16 the values are +0.430, +0.309, and +0.538, respectively (App. L). This timing control supports the interpretation that the first-step result is not caused only by a mismatch between the fixed readout and the eventual answer step. It does not imply that the effect is identical at every generated token.

3.3 Visual Grounding Bypasses the Gate

The previous subsection established that text CoT suppresses immediate spatial transport at obj_word. Is the gating triggered by having intermediate reasoning tokens, or specifically by text-abstract reasoning? We keep the intervention fixed at obj_word and vary the prompt attribute.

We compare text CoT to two visual-grounding prompts that also generate intermediate tokens but redirect the chain toward visual evidence (*visual_cot*: "describe where you see the object"; *visual_direct*: "note its position relative to the center"; full templates in App. A).

Table 2 compares peak Δargmax at obj_word for text CoT, visual_cot, and answer-only. All ten models recover positive obj_word Δargmax under visual_cot, including the eight that are suppressed under text CoT. Four models produce visual_cot effects that even exceed answer-only: Qwen-32B (+.117 vs. +.089), IV2.5 (+.153 vs. +.114), IV3-14B (+.117 vs. +.097), and most strikingly Janus (+.403 vs. +.069, a 5.8× amplification). The Janus jump shows visual grounding does not just "unblock" but actively elicits a spatial direction that answer-only under-extracts. Because visual_cot generates intermediate tokens like text CoT, a token-count explanation for obj_word gating fails: the conditional variable is the reasoning format (text-abstract "think step by step" closes the pathway; visual grounding "describe what you see" leaves it open).

3.4 Transport Pathways Are Position-Dependent

Patching at prefix_last (the final token before generation) tests whether closing the obj_word pathway removes spatial transport entirely or leaves a later-position route open. Here we reuse the object-word-derived spatial direction and change only the injection position, so the comparison tests relocation/readout compatibility rather than a separately optimized prefix_last basis. We use GQA Spatial for its full position × prompt grid at α=5; the RefCOCO prefix_last ladder (App. F) gives converging evidence at α=10.

Most models show a relocation pattern (Figure 3): obj_word Δargmax peaks in early-to-mid layers and declines, while prefix_last Δargmax emerges deeper. InternVL3-8B is the clearest case (obj_word L12 +0.112, prefix_last L20 +0.455), and CoT is near-zero at obj_word but *not* at prefix_last. The suppression is therefore local to one transport route, and the conditional pathway is jointly determined by prompt type and token position.

We call this two-step shape (obj_word peak in early-to-mid layers, prefix_last re-emergence in deeper layers) the *relocation pattern* and use this term throughout §4. Not all models follow it: Janus concentrates transport at obj_word (+0.315 at L12)

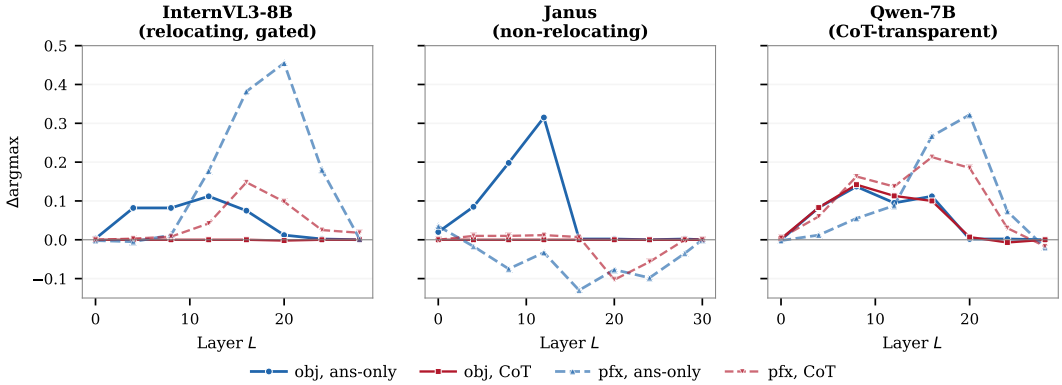

Figure 3: Position-dependent transport (GQA Spatial, three models). Left: InternVL3-8B shows the relocation pattern (obj_word L12 → prefix_last L20; CoT gated only at obj_word). Center: Janus concentrates at obj_word. Right: Qwen-7B transports at both positions under both prompts.

Pattern	Models	Obj. route	Reloc.
Open	**Qwen-7B**	**Open**	**Yes**
Reduced supp.,reloc.	**LLaVA**, **IV2.5**, **IV3-8B**, **IV3-14B**	**Supp./red.**	**Yes**
Suppressed, non-reloc.	**Gemma-3** (4/12/27B), **Janus**	**Supp.**	**No**
Prompt-selective	**Qwen-32B**	**Near-zero**	**Sel.**

Table 3: Four descriptive transport patterns across ten models. "Reloc.": whether transport reappears at prefix_last (Yes / Sel. = prompt-selective / No). Colors match Figure 4.

with prefix_last Δargmax going negative (patching pushes the answer *away* from target, indicating that prefix_last is dominated by competing features rather than spatial content), and Gemma-3 shows only weak, dataset-dependent prefix_last effects. These differences motivate §4.1.

4 Model-Level Patterns and Boundary Conditions

§3 isolated the local conditions under which spatial information reaches the answer: layer depth, prompt type, and token position. We now move to a model-level view: whether these local patterns organize into structured cross-model variation (§4.1), whether the logit-level map is reflected beyond next-token logits (§4.2), and where the map transfers or changes across datasets, attributes, and encoding amplitudes (§4.3).

4.1 Transport Groupings Across Models

Figure 4 summarizes the per-model gating signatures, and Table 3 organizes the ten models into four descriptive patterns. The *reduced/suppressed, relocating* pattern is the modal pattern (4/10 models): LLaVA and the three InternVL variants suppress or sharply reduce transport at obj_word but show positive prefix_last transport in deeper layers. The *suppressed, non-relocating* pattern (4/10) comprises the three Gemma-3 variants and Janus: spatial information either stays at the object-word token (Janus, $+0.325$ at L12) or does not recover at prefix_last (Gemma-3). Qwen-7B is the sole *open* model: CoT does not block transport at either position; §4.3 treats this as an amplitude-sensitive boundary case. Qwen-32B is *prompt-selective*: near-zero CoT transport at obj_word but a prompt-format-dependent prefix_last pathway (final_tag $+0.431$ vs. answer-only $+0.019$ at L55; App. F). With $N{=}10$ models, two of the four patterns have a single member, so these labels summarize the observed combinations rather than define a statistically validated classification. Grouping is consistent within family in two of three multi-scale families (InternVL3-8B and -14B share identical gating; Gemma-3 4B/12B/27B all fall in the suppressed, non-relocating pattern), but the Qwen family shifts from open (7B) to prompt-selective (32B), so family is not a sufficient predictor of grouping in our data.

Janus: dual role. Janus is the strongest visual-bypass case (§3.3): visual_cot lifts obj_word Δargmax from $+0.069$ to $+0.403$, a $5.8\times$ amplification no other model approaches. The same model is also the cleanest counter-example to a strong reading of "linear-subspace transport ⇒ behaviour": despite the $+0.403$ logit effect, generate-mode steering changes accuracy by 0pp (baseline 50–58%), suggesting that Janus's unified archi-

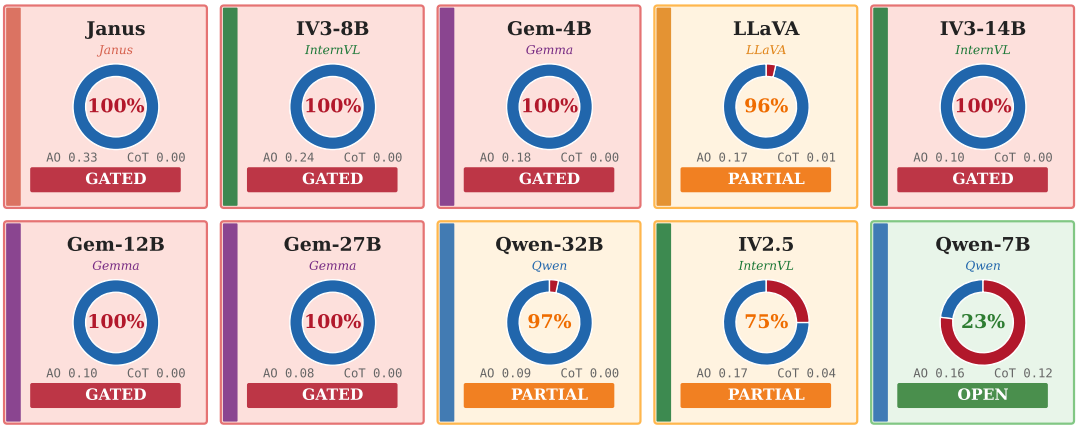

Figure 4: Per-model CoT gating signature (obj_word, RefCOCO, α=5). Eight of ten models fall below the descriptive 0.01 threshold: six are exact-zero and two are near-zero/borderline; Qwen-7B retains substantial CoT transport. Numerical values in App. B.

tecture routes generation through a pathway misaligned with our linear subspace. This is a concrete boundary for the operational definition in §2.3 (see Limitations).

4.2 Behavioral Consequences and Head-Level Probe

We next test whether the logit-level map has support beyond next-token logits.

Behavioral validation. Full 128-token generation with the same intervention provides a behavioral check on the logit-level map. Under answer-only, steering toward incorrect quadrants drops accuracy by 8–21pp in nine of ten models (Qwen-7B 74.2% $\rightarrow$ 53.6%; IV3-8B 77.5% $\rightarrow$ 56.9%; App. G). Janus is the boundary case, with no answer-only accuracy drop in this generation protocol. Under CoT, models with intact CoT baselines (LLaVA 55.8%, IV3-14B 60.8%, Gem-12B 45.8%) still show 13–16pp accuracy drops under the same object-word intervention, even though their fixed-state obj_word Δargmax is near zero for CoT. Because the intervention is applied before autoregressive generation, the perturbation can affect later hidden states and the final parsed answer without changing the immediate answer-option argmax. The independent prefix_last sweep identifies a compatible later readout route in relocating models; the generate-mode table itself does not patch prefix_last. Under visual_cot, steering produces sizable drops for many models, while IV2.5, Gemma-3, and Janus show weaker changes. This pattern is consistent with visual grounding preserving an active spatial pathway, but also highlights the model dependence of generate-mode validation.

For Qwen-32B and IV2.5, CoT collapses the baseline well below 4-way chance (9.2%, 4.2%); the $\leq$ 2.2pp post-steering drop is consistent with logit-level gating but cannot be cleanly separated from generic CoT derailing, so behavioral evidence rests on the intact-baseline group above.

Head-level evidence: proof of concept. We sweep attention head knockouts downstream of the steer layer: for each of 32 heads at a candidate layer, we zero its output projection and measure CoT Δargmax recovery at obj_word. In InternVL3-8B, a single head L27.H21 produces +0.063 recovery under CoT (n=100; CoT baseline +0.095), while all 31 other heads stay within $\pm$0.02 of baseline (Figure 5a). H21 has no effect under answer-only (Figure 5b, +0.007), consistent with a CoT-specific gating role. A parallel scan in LLaVA-OV-7B finds no comparable concentration: the best candidate (L22.H25) recovers only +0.013 at n=100 (App. Q). We do not test Qwen-7B because it is open at obj_word and has no suppressed route to localize. Head-level concentration is thus evident in at least one architecture but architecture-dependent; a full circuit account (upstream triggers, downstream completion, cross-architecture comparison) remains open.

4.3 Generality and Boundary Conditions

We test the generality of the transport map across datasets, visual attributes, and encoding amplitudes.

Cross-dataset generalization. We derive four-quadrant spatial MCQs from GQA scene graphs

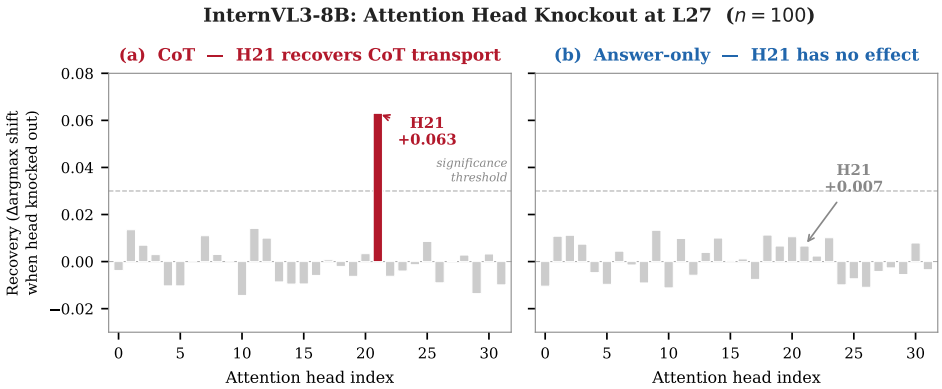

Figure 5: Attention head knockout at L27 of InternVL3-8B (n=100, RefCOCO, α=5). (a) Under CoT, knocking out H21 recovers +0.063 Δargmax, the only head above the significance threshold among all 32 heads. (b) Under answer-only, H21 has no effect (+0.007). H21 is a CoT-specific gating candidate in this scan, rather than a general transport component.

Prediction	Match	Note
CoT suppression at obj_word	9/10	Qwen-7B: remains open
Peak layer ±8L	6/6	original 6 models
pfx_last deeper than obj_word	8/10	Janus, Gem-27B anomalous

Table 4: Structural transfer from RefCOCO to GQA Spatial. Three transport-map predictions tested with independent images, scene graphs, and object distributions.

Model	Spatial peak	Color peak	Shift
Qwen-7B	L8	L4	−4
Qwen-32B	L24	L16	−8
LLaVA	L8	L4	−4
InternVL2.5	L8	L4	−4
InternVL3-8B	L12	L4	−8
InternVL3-14B	L12	L12	0
Gemma-3-4B	L8–12	L8	0 to −4
Gemma-3-12B	L18	L12	−6
Gemma-3-27B	L24	L16	−8
Janus	L12	L8	−4

Table 5: Obj_word peak layer on GQA Spatial vs. GQA Color (answer_only), all ten models. Color peaks ∼ 4–8 layers earlier than spatial in 9/10 models; InternVL3-14B is the lone exception (same peak layer).

(Hudson and Manning, 2019) and test three structural predictions from the RefCOCO map (Table 4): CoT suppression status transfers in 9/10 models, peak layers stay within ±8 layers in the 6 models with full coverage, and 8/10 reproduce the relocation pattern. Janus stays non-relocating; Gemma-3-27B shows a weak prefix_last effect (+0.187 at L40) not captured by the cross-dataset ladder.

Cross-attribute generalization. We extend the framework to a non-spatial attribute via GQA color MCQs on all ten models. CoT gates color transport in 10/10, including Qwen-7B (open on spatial), whose color answer-only peak (+0.082 at L4) collapses to 0.000 under CoT. Color peaks lie ∼ 4–8 layers earlier than spatial peaks (Table 5), consistent with color being a shallower visual attribute.

Amplitude as a boundary condition. Why does Qwen-7B flip between open (spatial) and fully suppressed (color)? Across the ten models, gating strength is not monotone in encoding amplitude: Janus encodes at +0.325 yet is fully suppressed, while Qwen-7B at +0.156 is not, so there is no universal amplitude threshold across architectures. Within Qwen-7B alone, however, lowering the encoding amplitude flips the outcome: on spatial tasks (GQA +0.137; RefCOCO +0.156), CoT does not gate; on color (+0.082), gating is complete. The model, pipeline, and intervention are identical; only the encoding amplitude differs. This is consistent with a model-specific threshold: amplitude can flip a single architecture's gating behaviour even when no universal threshold separates architectures.

5 Related Work

VLM spatial probing and grounding failures. VLMs are known to encode rich visual and spatial properties internally while failing to use them at output (Nooralahzadeh et al., 2026; Liu et al., 2026; Asadi et al., 2026; Li et al., 2025; Kamath et al., 2023; Liu et al., 2023). Closest to our setup, Kang et al. (2026) isolate linear spatial-

ID directions, with concurrent probes on fine-tuning origins (Naghashyar et al., 2026) and the encoder/backbone split (Cui et al., 2026). Broader probes cover grounding, hidden representations, corruption, and storage–transfer (Yu and Lee, 2025; Liu et al., 2025; Golovanevsky et al., 2025; Basu et al., 2024); Wu et al. (2026) decompose visual-to-language flow via PID and attention knockouts. These are largely static probes of what is encoded; we take encoding as given and ask about transport.

Causal diagnostics and intervention. Selective heads and grounding circuits have been mapped via attention inspection (Ma et al., 2026; Chen et al., 2025a; Bi et al., 2025; Kang et al., 2025; Basile et al., 2025), but attention overlap is correlational. Activation patching, causal mediation, and representation engineering trace or steer LM computation via direction shifts (Meng et al., 2022; Geva et al., 2023; Wang et al., 2023; Vig et al., 2020; Zou et al., 2023; Turner et al., 2023; Li et al., 2023), and causal abstraction (Geiger et al., 2024) formalizes such groupings. Our direction patching shares this primitive but localizes it across layer, position, and prompt to map transport rather than steer behavior.

CoT and visual grounding prompts. CoT (Wei et al., 2022; Kojima et al., 2022) is the default for complex reasoning, but Kancheti et al. (2026) report that text CoT *degrades* VLM spatial accuracy and Mehrafarin et al. (2026) find correct solutions still recoverable from hidden states; neither localizes where the pathway is blocked. Visually grounded prompts (Jiang et al., 2025; Zhang et al., 2025; Qin et al., 2025) are typically framed as better recipes. Recent decoding-time work also uses the temporal evolution of LVLM output logits to preserve visual grounding: Residual Decoding aggregates logits from semantically stable historical steps to provide history-aware guidance against language-prior hallucinations (Chen et al., 2026). Our analysis complements this line by localizing the upstream transport of spatial evidence across residual-stream layers and token positions under different prompt interfaces. We provide complementary transport evidence: text CoT suppresses immediate obj_word argmax-level transport while visual grounding keeps that route open, and a prefix-token route remains available in many suppressed models.

6 Conclusion

This work traces when encoded spatial information becomes causally available for VLM answers. Direction patching shows that spatial transport is conditional on layer depth, prompt type, and token position: spatial-ID directions can be available before they produce answer-logit effects, text CoT often suppresses immediate object-word argmax-level transport, and visually grounded prompts keep this pathway open. Across models, these local effects form descriptive transport groupings, with behavioral steering and a proof-of-concept head knockout providing additional, architecture-dependent evidence. Together, the results refine the encoding-grounding dissociation: VLMs may encode spatial information, but whether they use it depends on how the prompt changes the layer, position, and route through which that information reaches the answer logits. The answer-step and stepwise controls further show that an immediate argmax-level suppression does not by itself establish trajectory-wide erasure.

Limitations

Direction patching probes the linear subspace identified by Kang et al. (2026): it manipulates a class-conditioned direction in the residual stream and reads the resulting change in answer logits. All claims of "transport", "gating", and "closed pathway" in this paper therefore refer to causal influence *along this linear subspace* under the tested prompt state, layer, and token position. In particular, "gated" denotes suppression of the four-way answer-option argmax effect below the descriptive threshold; it does not denote complete erasure of the target-logit signal or of every later decoding route. The crossfit protocol and noise baselines control for spurious effects within the subspace, but direction patching is silent on transport that may occur through non-linear pathways or through alternative directions outside this subspace. The Janus case (§4.1: $+0.403$ visual_cot Δargmax but 0pp generate-mode accuracy change) is a concrete example of this gap: the linear subspace is highly responsive to patching, yet the architecture's actual generation appears to rely on a distinct pathway that our method cannot reach. A fuller account of gating in VLMs will likely require complementary non-linear probes (e.g. sparse autoencoders, dictionary learning) in addition to direction patching.

We test two visual attributes (spatial position and

color) in a four-way forced-choice format. Free-form spatial descriptions, fine-grained localization, and other attribute types (size, shape, material) remain untested.

All ten models are open-weight, ranging from 4B to 32B parameters. Closed-source models and architectures above 32B may exhibit different transport patterns.

Generate-mode behavioral validation uses 128-token generation and a single object-word intervention at the specified layer. The resulting parsed-answer changes can reflect effects that propagate through later autoregressive states, so they should not be read as a direct estimate of the fixed-state next-token Δargmax. Longer generation or multi-turn interaction could produce different accuracy patterns that our single-pass evaluation does not capture.

The visual grounding prompts we test are manually designed. Systematic optimization of prompt wording for maximal bypass could yield stronger effects but is beyond the scope of this study.

Our head-level localization is proof-of-concept, not a complete circuit. The InternVL3-8B knockout identifies a single CoT-specific gating head (L27.H21) whose removal reliably reopens transport, and a parallel LLaVA-OV scan rules out a comparable single head there. Tracing the upstream features that activate H21, the downstream computation that completes the gate, and the architectural reasons different families concentrate versus distribute the same function are open questions.

References

Mohammad Asadi, Jack W. O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. 2026. MIRAGE: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*.

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

Lorenzo Basile, Valentino Maiorca, Diego Doimo, Francesco Locatello, and Alberto Cazzaniga. 2025. Head pursuit: Probing attention specialization in multimodal transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. 2024. Understanding information storage and transfer in multi-modal large language models. In *Advances in Neural Information Processing Systems*.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. 2025. Unveiling visual perception in language models: An attention head analysis approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. 2025a. Why is spatial reasoning hard for VLMs? an attention mechanism perspective on focus areas. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025b. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Xinrong Chen, Xu Chu, Yingmin Qiu, Hengyuan Zhang, Jing Xiong, Shiyu Tang, Shuai Liu, Shaokang Yang, Cheng Yang, Hayden Kwok-Hay So, and Ngai Wong. 2026. Residual decoding: Mitigating hallucinations in large vision-language models via history-aware residual guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25281–25292.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, and Zhaoyang Liu. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*. InternVL2.5 technical report.

Kelly Cui, Nikhil Prakash, Ayush Raina, David Bau, Antonio Torralba, and Tamar Rott Shaham. 2026. The dual mechanisms of spatial reasoning in vision–language models. *arXiv preprint arXiv:2603.22278*.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning (CLeaR)*, volume 236 of *PMLR*, pages 160–187.

Gemma Team, Google DeepMind. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*. Covers 4B, 12B, and 27B variants.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. 2025. What do VLMs NOTICE? A mechanistic interpretability pipeline for Gaussian-noise-free text-image corruption and evaluation. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Albuquerque, New Mexico. Association for Computational Linguistics.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chaoya Jiang, Yongrui Heng, Wei Ye, Han Yang, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. 2025. VLM-R^3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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

Raphi Kang, Hongqiao Chen, Georgia Gkioxari, and Pietro Perona. 2026. Linear mechanisms for spatiotemporal reasoning in vision language models. In *International Conference on Learning Representations (ICLR)*.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798. Association for Computational Linguistics.

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*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research (TMLR)*.

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*.

Rang Li, Lei Li, Shuhuai Ren, Hao Tian, Shuhao Gu, Shicheng Li, Zihao Yue, Yudong Wang, Wenhan Ma, Zhe Yang, Jingyuan Ma, Zhifang Sui, and Fuli Luo. 2025. GroundingME: Exposing the visual grounding gap in MLLMs through multi-dimensional evaluation. *arXiv preprint arXiv:2512.17495*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*.

Benlin Liu, Amita Kamath, Madeleine Grunde-McLaughlin, Winson Han, and Ranjay Krishna. 2025. Visual representations inside the language model. In *Conference on Language Modeling (COLM)*.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Hanqing Lu, Benoit Dumoulin, and Hanghang Tong. 2026. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in VLMs. In *International Conference on Learning Representations (ICLR)*.

Xueqi Ma, Shuo Yang, Yanbei Jiang, Shu Liu, Zhenzhen Liu, Jiayang Ao, Xingjun Ma, Sarah Monazam Erfani, and James Bailey. 2026. Attention in space: Functional roles of VLM heads for spatial reasoning. *arXiv preprint arXiv:2603.20662*.

Houman Mehrafarin, Amit Parekh, and Ioannis Konstas. 2026. When chain-of-thought fails, the solution hides in the hidden states. *arXiv preprint arXiv:2604.23351*.

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

Lachin Naghashyar, Hunar Batra, Ashkan Khakzar, Philip Torr, Ronald Clark, Christian Schroeder de Witt, and Constantin Venhoff. 2026. Towards understanding multimodal fine-tuning: A case study into spatial features. *arXiv preprint arXiv:2602.08713*.

Farhad Nooralahzadeh, Omid Rohanian, Yi Zhang, Jonathan Fürst, and Kurt Stockinger. 2026. Arbitration failure, not perceptual blindness: How vision-language models resolve visual-linguistic conflicts. *arXiv preprint arXiv:2604.09364*.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and Xudong Wang. 2025. Chain-of-visual-thought: Teaching VLMs to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*.

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*.

Hongxuan Wu, Yukun Zhang, and Xueqing Zhou. 2026. How vision becomes language: A layer-wise information-theoretic analysis of multimodal reasoning. *arXiv preprint arXiv:2602.15580*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision (ECCV)*.

Zhuoran Yu and Yong Jae Lee. 2025. How multimodal LLMs solve image tasks: A lens on visual grounding, task reasoning, and answer decoding. In *Conference on Language Modeling (COLM)*.

Xinyu Zhang, Yuxuan Dong, Lingling Zhang, Chengyou Jia, Zhuohang Dang, Basura Fernando, Jun Liu, and Mike Zheng Shou. 2025. CoFFT: Chain of foresight-focus thought for visual language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, and Jie Shao. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*. Covers 8B and 14B variants.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, and Ann-Kathrin Dombrowski. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A Prompt Templates

All experiments in the paper hold the question stem fixed and vary only the instruction suffix appended to it. This isolates the prompt as the manipulated variable so that any change in transport (e.g. CoT gating in §3.2, visual bypass in §3.3) can be attributed to the suffix rather than to differences in the upstream visual or linguistic input.

Table 6 lists the verbatim suffix for each of the seven prompt formats. The first five (`cot`, `brief_reason`, `answer_first`, `final_tag`, `answer_only`) span a reasoning-length spectrum from full text CoT down to direct letter output, and are used in the prompt ladder of §4.1. The last two (`visual_cot`, `visual_direct`) are visual- grounding prompts that introduce intermediate tokens but redirect the reasoning chain to visual evidence, isolating reasoning *format* from token *count*.

B Per-Model Peak Δargmax Values

Figure 4 summarises CoT gating as donut cards; this section provides the underlying numerical values. Table 7 reports per-model peak Δargmax under CoT and answer-only at the object-word position, together with the layer at which each peak is attained.

Eight of ten models fall below the descriptive 0.01 threshold: six have CoT peak Δargmax exactly 0.000, while LLaVA (+0.006) and Qwen-32B (+0.003) are near-zero/borderline cases. These values support the descriptive transport patterns in §4.1, not a validated classification system. InternVL2.5 is *partially* gated: CoT collapses from +0.168 to +0.042, a 75% reduction, but the residual is non-zero. Qwen-7B is the only *open* model: its CoT peak (+0.120) remains close to the answer-only peak (+0.156).

The peak layer is itself informative. Most 7B–8B models peak in the L8–L16 band, but Qwen-32B and Gemma-3-27B shift to L24–L32; this late-layer shift in larger models also appears in §3.1 and Appendix O. The peak values in this table are the canonical numbers used in the abstract, the descriptive grouping, and the threshold analysis of §4.3.

C Controls

A positive Δargmax at the object-word position could in principle arise from two trivial sources rather than from direction-specific spatial transport: any sufficiently large perturbation of the residual

Format	Instruction suffix
`cot`	"Let's think step by step."
`brief_reason`	"Give a brief reason, then answer with the letter."
`answer_first`	"First, answer with only the letter (A–D). Then explain briefly."
`final_tag`	"Think about it, then give your final answer as Final Answer: X"
`answer_only`	"Answer with only the letter (A–D)."
`visual_cot`	"First, carefully look at the image and describe where you see the object mentioned above. Then answer with only the letter."
`visual_direct`	"Look at the object in the image. Note its position relative to the center. Answer with only the letter (A–D)."

Table 6: Verbatim instruction suffixes appended to a shared question stem. The first five suffixes (`cot` through `answer_only`) span a reasoning-length spectrum and are used in the prompt ladder of §4.1; the last two (`visual_cot`, `visual_direct`) are visual-grounding prompts evaluated in §3.3.

Model	CoT	Ans-only	Gated?
Qwen-7B	+0.120 L16	+0.156 L16	No
InternVL2.5	+0.042 L12	+0.168 L12	Weak
LLaVA	+0.006 L10	+0.168 L16	Yes
InternVL3-8B	0.000 all	+0.236 L16	Yes
InternVL3-14B	0.000 all	+0.097 L24	Yes
Gemma-3-4B	0.000 all	+0.178 L12	Yes
Gemma-3-12B	0.000 all	+0.097 L20	Yes
Gemma-3-27B	0.000 all	+0.081 L24	Yes
Janus	0.000 all	+0.325 L12	Yes
Qwen-32B	+0.003 L32	+0.089 L32	Yes

Table 7: Peak obj_word Δargmax under CoT and answer-only (RefCOCO, α=5), with the layer at which each peak is attained. The sample count is n=50/quadrant for 7B/8B models and n=30/quadrant for Qwen-32B, Gemma-3, and Janus; each source sample is paired with the three non-source target quadrants. Top two rows (above the midrule) are open or partially suppressed; the bottom eight rows fall at or below 0.006.

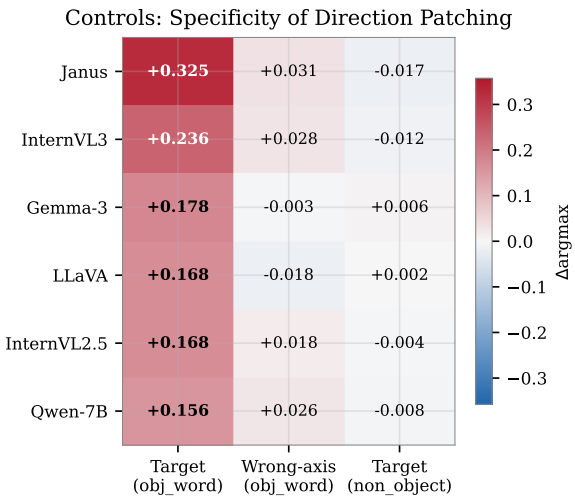

Figure 6: Wrong-axis and non-object patching produce near-zero Δargmax, confirming direction and position specificity.

Model	Lyr	Target	Wrong	Non-obj
Qwen-7B	16	+.156	+.026	−.008
LLaVA	16	+.168	−.018	+.002
IV2.5	12	+.168	+.018	−.004
IV3-8B	16	+.236	+.028	−.012
IV3-14B	24	+.097	+.006	+.003
Gem-4B	12	+.178	−.003	+.006
Gem-12B	20	+.097	−.006	−.017
Gem-27B	24	+.081	−.011	+.000
Janus	12	+.325	+.031	−.017

Table 8: Direction and position controls (RefCOCO, ans-only, α=5). Wrong: perpendicular direction. Non-obj: correct direction at a non-object token.

stream might bias the answer logits, or the model might be sensitive to any change at the object-word token regardless of its semantic content. We address both possibilities here.

Figure 6 and Table 8 report two control conditions matched in L_2 norm to the target patch. *Wrong-axis* patches inject the perpendicular spatial direction at the correct token, holding norm constant. *Non-object* patches inject the correct direction at a non-object token (a content word that is not the referent). Both controls produce Δargmax within ±0.03 across all ten models, an order of magnitude below the matched target direction (e.g. Janus target +0.325 vs. wrong-axis +0.031 vs. non-obj −0.017).

Together, these controls rule out the two trivial explanations: the target effect requires both the correct spatial direction and the correct object-word position. The same axis-specific behaviour also holds on COCO-Spatial (Appendix N), where horizontal centroids do not shift vertical answers and vice versa.

D Bootstrap Confidence Intervals

The main text reports point estimates for Δargmax and target logit gain. We additionally report

Model	Pmt	Lyr	Δ_{argmax}	95% CI
Qwen-7B	ans	16	+.156	$[+.122, +.192]$
Qwen-7B	cot	16	+.120	$[+.090, +.150]$
LLaVA	ans	16	+.168	$[+.134, +.202]$
LLaVA	cot	16	+.002	$[+.000, +.006]$
IV2.5	ans	12	+.168	$[+.134, +.204]$
IV2.5	cot	12	+.042	$[+.026, +.060]$
IV3-8B	ans	16	+.236	$[+.192, +.279]$
IV3-8B	cot	16	+.000	$[+.000, +.000]$

Table 9: 95% sample-cluster bootstrap CIs (2000 resamples, n=50/quadrant) for obj_word Δargmax under each prompt at each model's canonical layer (RefCOCO, α=5). IV3-8B CoT is exactly zero with zero-width CI; LLaVA CoT lower bound touches zero.

Model	Pmt	Lyr	gain	95% CI
Qwen-7B	ans	16	+1.252	$[+1.130, +1.386]$
Qwen-7B	cot	16	+0.500	$[+0.417, +0.589]$
LLaVA	ans	16	+0.855	$[+0.753, +0.964]$
LLaVA	cot	16	+0.146	$[+0.075, +0.217]$
IV2.5	ans	12	+1.527	$[+1.367, +1.679]$
IV2.5	cot	12	+0.311	$[+0.263, +0.362]$
IV3-8B	ans	16	+1.850	$[+1.676, +2.028]$
IV3-8B	cot	16	+0.421	$[+0.341, +0.501]$

Table 10: 95% bootstrap CIs for the raw shift in the target-letter logit (n=50/quadrant, 2000 resamples, RefCOCO, α=5). Target logit gain is positive even when Δargmax is zero (IV3-8B CoT $+0.421$), showing that CoT gating suppresses the argmax flip rather than erasing the directional signal.

sample-cluster bootstrap CIs (2000 resamples; resampling unit is the sample id rather than the sample–target pair) for the original four models at their canonical layers (RefCOCO, α=5, n=50/quadrant). CIs use heldout centroids within each bootstrap fold, so no in-fold leakage inflates significance.

Table 9 reports CIs for obj_word Δargmax. For ans-only the CIs are tight and exclude zero by a wide margin (e.g. Qwen-7B $[+.122, +.192]$). For CoT, LLaVA's lower bound touches zero ($[+.000, +.006]$), and IV3-8B's CI is exactly $[+.000, +.000]$. These intervals support the distinction between exact zero and borderline near-zero cases; they do not turn the descriptive threshold into a universal significance test.

Table 10 reports CIs for target logit gain (the raw shift in the target-letter logit, before argmax). The relevant contrast is between argmax and logit gain. Even when Δargmax is exactly zero (IV3-8B CoT), the logit gain is positive ($+0.421$, $[+.341, +.501]$). The direction is still pushing the target logit upward; it is simply not enough to flip the argmax. The tested CoT interface therefore shows a *threshold* phenomenon (directional influence on the logit distribution survives while argmax-level transport is suppressed), not complete signal erasure. This distinction matters for the head-knockout analysis (§4.2): recovering even small amounts of Δargmax is meaningful because the underlying logit signal is not zeroed out.

E Alpha Sweep

If CoT gating were merely a small effect size, scaling the patch strength α upward should eventually push the CoT Δargmax above zero. We test this with an α sweep at both intervened positions and compare the dose–response curve of answer-only to that of CoT for the same model and layer.

We sweep $\alpha \in \{1, 2, 5, 10\}$ at obj_word (Table 11) and at prefix_last (Table 12). The qualitative pattern is the same at both positions and is summarised in Figure 7 for obj_word: answer-only is monotonic and grows with α (e.g. Qwen-7B $+.011 \rightarrow +.144$, IV3-8B $+.017 \rightarrow +.167$), while the suppressed models remain at or near zero across the sweep. LLaVA and Qwen-32B stay in the near-zero range, while the larger non/partially suppressed effects occur for Qwen-7B and InternVL2.5. The sweep therefore supports the same descriptive grouping without turning the 0.01 threshold into a claim of exact equality.

The prefix_last sweep adds two specific observations. Qwen-7B CoT Δargmax is non-monotonic in α, peaking at α=5 ($+.356$) and collapsing at α=10 ($+.069$), suggesting that the prefix_last pathway saturates earlier than the obj_word pathway. InternVL3-14B is the only model with an exact-zero pattern at prefix_last under both prompts at every α; this is consistent with its non-relocating transport pattern in §4.1. The α=5 value used throughout the main text sits comfortably in the linear regime for both positions, justifying our choice of a single canonical α.

F Prompt Ladder at Prefix_last

§4.1 distinguishes the *prompt-selective* pattern (Qwen-32B) from the suppressed–relocating pattern (LLaVA, IV2.5, IV3-8B/14B) based on whether the prefix_last pathway responds uniformly across prompt formats. This appendix supplies the underlying data. Holding the patch position fixed at prefix_last and the layer at each

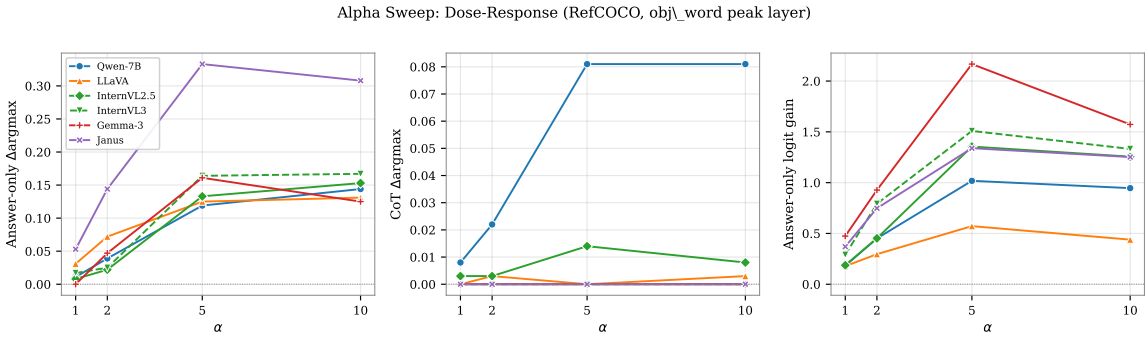

Figure 7: Alpha sweep at obj_word (RefCOCO, ten models). Answer-only: increasing dose-response across the tested range. CoT: near-zero for the suppressed majority, with LLaVA and Qwen-32B as borderline cases.

Model	Lyr	α=1	α=2	α=5	α=10
Answer-only					
Qwen-7B	16	+.011	+.039	+.119	+.144
LLaVA	16	+.031	+.072	+.125	+.131
IV2.5	12	+.008	+.022	+.133	+.153
IV3-8B	16	+.017	+.025	+.164	+.167
Gem-4B	12	.000	+.047	+.161	+.125
Janus	12	+.053	+.144	+.333	+.308
Qwen-32B	32	+.008	+.008	+.083	+.294
IV3-14B	16	−.006	.000	+.014	+.075
Gem-12B	16	+.003	+.003	+.036	+.019
Gem-27B	24	+.022	+.028	+.089	+.189
CoT					
Qwen-7B	16	+.008	+.022	+.081	+.081
LLaVA	16	.000	+.003	.000	+.003
IV2.5	12	+.003	+.003	+.014	+.008
IV3-8B	16	.000	.000	.000	.000
Gem-4B	12	.000	.000	.000	.000
Janus	12	.000	.000	.000	.000
Qwen-32B	32	.000	.000	+.006	.000
IV3-14B	16	.000	.000	.000	.000
Gem-12B	16	.000	.000	.000	.000
Gem-27B	24	.000	.000	.000	.000

Table 11: Obj_word Δargmax across $\alpha \in \{1, 2, 5, 10\}$ at each model's canonical obj_word layer (RefCOCO). Answer-only grows across the dose range; the suppressed models remain at or near zero. LLaVA and Qwen-32B are borderline near-zero cases, while Qwen-7B and InternVL2.5 retain larger CoT effects.

model's prefix_last peak, we vary the prompt suffix across the reasoning-length ladder of Appendix A.

Figure 8 and Table 13 reveal three prefix_last behaviours. The *prompt-stable* pattern is exemplified by Qwen-7B at L28: Δargmax stays near 0.80 across all five formats, so the prefix_last route is open regardless of prompt. The *prompt-selective* pattern is exemplified by Qwen-32B at L55: Δargmax reaches +0.431 under `final_tag` but only +0.019 under `answer_only`, a 23× swing that tracks whether the prompt creates an explicit answer-anchor token. The *near-zero* pattern covers IV3-14B, Gem-12B and Gem-27B, which remain at zero across every layer and every prompt at prefix_last, consistent with their non-relocating transport pattern. LLaVA and IV2.5 sit in between: the prefix_last route opens under reasoning-style prompts (CoT, Brief, Ans-1st) but does not match Qwen-7B's universal-open behaviour. This prompt ladder is the data underlying the four descriptive transport patterns in §4.1.

G Generate-Mode Accuracy

The fixed-state transport results use next-token logit Δargmax at t=0. To test whether an intervention at that state can influence the actual decoded answer, we run 128-token greedy generation under the same direction patches and compare parsed-answer accuracy before and after the intervention. These are related measurements with different denominators and readout times.

Model	Lyr	α=1	α=2	α=5	α=10
Answer-only					
Qwen-7B	20	+.014	+.047	+.150	+.275
LLaVA	20	+.014	+.031	+.075	+.214
IV2.5	24	+.003	+.008	+.017	+.039
IV3-8B	20	+.008	+.025	+.097	+.461
IV3-14B	36	+.028	+.044	+.086	+.344
Gem-4B	24	.000	+.031	+.050	−.025
Gem-12B	30	+.008	+.014	+.033	+.006
Gem-27B	40	+.011	+.017	+.056	+.069
Janus	20	+.039	+.094	+.178	+.228
Qwen-32B	55	.000	.000	+.008	+.003
CoT					
Qwen-7B	20	+.072	+.211	+.356	+.069
LLaVA	20	.000	.000	+.008	+.028
IV2.5	24	+.008	+.006	+.022	+.100
IV3-8B	20	.000	+.006	+.033	+.097
IV3-14B	36	.000	.000	.000	.000
Gem-4B	24	.000	.000	+.006	−.003
Gem-12B	30	.000	.000	−.056	+.008
Gem-27B	40	.000	.000	−.003	+.008
Janus	20	.000	.000	.000	+.064
Qwen-32B	55	+.017	+.019	−.014	−.022

Table 12: Prefix_last alpha sweep (RefCOCO). Qwen-7B CoT non-monotonic (peak α=5: +.356). IV3-14B: hard-zero at prefix_last.

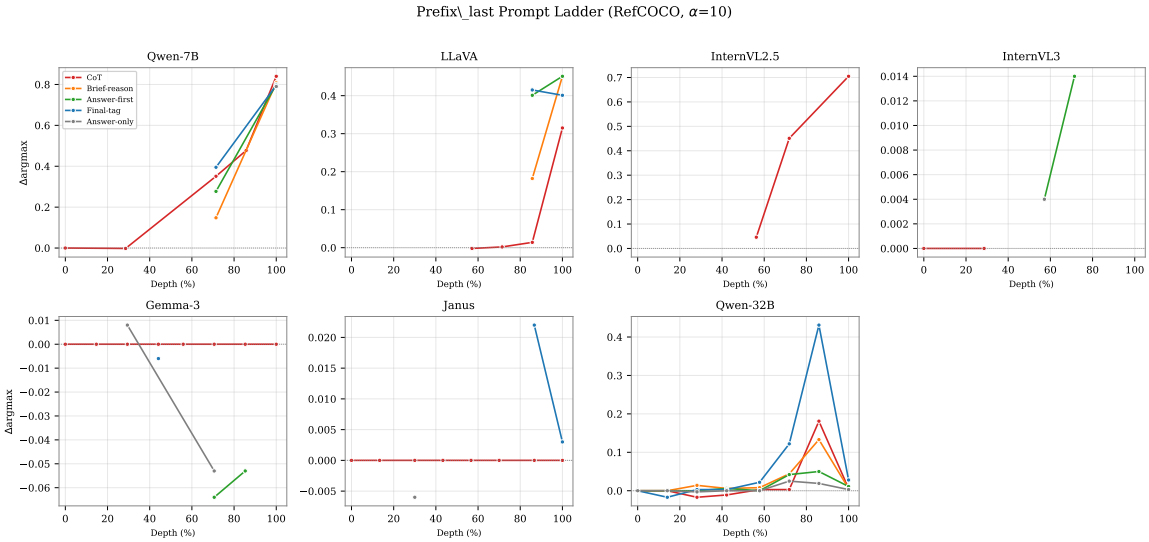

Figure 8: Prompt ladder at prefix_last (RefCOCO, α=10). Qwen-7B prompt-stable; Qwen-32B prompt-selective.

Table 14 shows that steering toward incorrect quadrants under answer-only drops accuracy by 8 to 21pp in nine of ten models (Qwen-7B 74.2% → 53.6%; IV3-8B 77.5% → 56.9%; Gem-12B 50.8% → 32.8%). Under text CoT the picture splits: models whose CoT baselines have collapsed to near-chance (Qwen-32B 9.2%, IV2.5 4.2%) show ≤ 2.2pp further drop, whereas models with intact CoT baselines (LLaVA 55.8%, IV3-14B 60.8%, Gem-12B 45.8%) still show 13 to 16pp drops despite near-zero immediate obj_word Δargmax. An obj_word perturbation can affect later autoregressive states and the final parsed answer even when it does not change the first answer-option argmax; the independent prefix_last experiments provide the position-level route analysis.

Two cases deserve note. Visual_cot steering produces sizable accuracy drops for many models, while IV2.5, Gemma-3, and Janus show weaker or absent changes. This is consistent with visual grounding preserving an active behavioral pathway in many, but not all, generation tests. Janus is the principal exception: its visual_cot Δargmax is the strongest in the study (+0.403), yet generate-

Model	Lyr	CoT	Brief	Ans-1st	Final	Ans
Qwen-7B	20	+.351	+.148	+.277	+.395	—
Qwen-7B	28	+.840	+.808	+.800	+.796	+.790
Qwen-32B	46	+.003	+.044	+.042	+.122	+.025
Qwen-32B	55	+.181	+.133	+.050	+.431	+.019
LLaVA	24	+.014	+.182	+.401	+.415	—
LLaVA	28	+.315	+.451	+.451	+.401	—
IV2.5	23	+.451	—	—	—	—
IV2.5	32	+.705	—	—	—	—
IV3-8B	16–20	+.004	+.004	+.004	—	+.004
Gem-4B	0–34	0	−.064	−.053	−.006	0
Janus	0–30	0	+.003	+.003	+.022	0
All zero across layers and prompts: IV3-14B, Gem-12B, Gem-27B						

Table 13: Prefix_last Δargmax under five prompt suffixes at each model's prefix_last layer (RefCOCO, α=10). Qwen-7B is prompt-stable; Qwen-32B is prompt-selective (final_tag $+0.431$ vs. ans-only $+0.019$); IV3-14B, Gem-12B, Gem-27B are zero at every layer and every prompt. Dashes: condition not run.

mode accuracy moves by $\leq$ 3pp under any prompt. The unified vision–language architecture appears to route generation through a pathway decoupled from the linear spatial subspace we patch; this is the sharpest encoding-without-behavior boundary case in our data.

H GQA Spatial Layer Sweep

The RefCOCO-derived MCQs in the main text use referring expressions and COCO images. GQA Spatial uses different images, scene graphs, and an independent object distribution, so reproducing the transport map on GQA tests whether the patterns are dataset-specific or properties of the underlying models. Table 15 reports the complete layer-by-layer sweep for four representative models that span the four descriptive transport patterns.

Three structural predictions from RefCOCO are preserved on GQA. First, the suppression status is stable: LLaVA, IV3-8B and Janus remain near-zero at obj_word under CoT (peak ≤ 0.003), and Qwen-7B remains open ($+.142$ under CoT). Second, peak layers are stable within ± 8 layers: IV3-8B's obj_word peak is at L12 on GQA and L16 on RefCOCO, well within the 8-layer tolerance. Third, the relocation pattern holds for the suppressed–relocating models: IV3-8B's prefix_last peaks at L20 with $+0.455$ (its largest cell), while obj_word saturates at L12 ($+0.112$). Janus stays non-relocating with negative prefix_last throughout.

The takeaway is that the transport map is a model property, not a quirk of RefCOCO scene composition. The two minor deviations are consistent with the boundary analysis in §4.3: Gemma-3-27B shows a weak prefix_last effect on GQA ($+0.187$ at L40) not visible at the RefCOCO ladder resolution, and Janus reproduces its sharp non-relocating shape across both datasets, suggesting the architecture is the controlling factor rather than the input distribution.

I GQA Color Sweep

We also test whether the transport apparatus applies to a non-spatial visual attribute. GQA Color MCQs are constructed by the same procedure as GQA Spatial, replacing the spatial question with a "what color is the X" question grounded in scene-graph color annotations. The framework is held constant; only the attribute changes.

All ten models were evaluated on GQA Color, and CoT gates color transport in 10/10 (compared with 8/10 on spatial; per-model peak layers in Table 5). Table 16 gives the full obj_word layer sweep for the six models with $\leq$ 30-layer LMs; peak layer summaries for the larger variants (Qwen-32B, InternVL3-14B, Gemma-3-12B, Gemma-3-27B) are in Table 5 of the main text. Color peaks are systematically four to eight layers earlier than the corresponding spatial peaks, consistent with color being a shallower visual feature. Peak amplitudes are also smaller on color (typical maxima $+0.08$ to $+0.14$ for color vs. $+0.10$ to $+0.32$ for spatial), reflecting a weaker color direction in our centroid construction.

These two effects together explain the Qwen-7B gating flip discussed in §4.3. Qwen-7B is open on spatial tasks, but its color answer-only peak is only

Model	Pmt	Lyr	Pos	Base	Steer	Δacc	Δ_{parse}
Qwen-7B	ans	16	obj	74.2	53.6	−20.6	+.125
	vis	16	obj	72.5	57.8	−14.7	+.114
	cot	16	obj	20.0	18.6	−1.4	−.006
Qwen-32B	ans	32	obj	77.5	61.1	−16.4	+.072
	vis	32	obj	70.8	54.7	−16.1	+.044
	cot	32	obj	9.2	6.9	−2.2	−.019
LLaVA	ans	16	obj	55.8	36.9	−18.9	+.106
	vis	16	obj	57.5	36.9	−20.6	+.131
	cot	16	obj	55.8	40.6	−15.3	+.117
IV2.5	ans	16	obj	69.2	60.3	−8.9	+.053
	vis	16	obj	67.5	60.6	−6.9	+.053
	cot	16	obj	4.2	4.4	+0.3	+.019
IV3-8B	ans	16	obj	77.5	56.9	−20.6	+.164
	vis	16	obj	78.3	57.2	−21.1	+.122
	cot	16	obj	33.3	27.5	−5.8	+.064
IV3-14B	ans	20	obj	81.7	65.6	−16.1	+.072
	vis	20	obj	82.5	64.4	−18.1	+.094
	cot	20	obj	60.8	47.8	−13.1	+.067
Gem-4B	ans	12	obj	40.0	26.7	−13.3	+.125
	vis	12	obj	44.2	31.9	−12.2	+.078
	cot	12	obj	30.0	26.7	−3.3	+.028
Gem-12B	ans	18	obj	50.8	32.8	−18.1	+.139
	vis	18	obj	50.8	38.9	−11.9	+.036
	cot	18	obj	45.8	30.3	−15.6	+.061
Gem-27B	ans	24	obj	55.8	46.4	−9.4	+.100
	vis	24	obj	60.0	53.3	−6.7	+.067
	cot	24	obj	55.8	49.4	−6.4	+.042
Janus	ans	16	obj	50.0	50.0	+0.0	−.006
	vis	16	obj	57.5	60.3	+2.8	−.014
	cot	16	obj	41.7	41.1	−0.6	+.017

Table 14: Full 128-token greedy-generation accuracy (%, RefCOCO, n=30/quadrant, 120 source samples and 360 source–target intervention records per cell) before steering (Base) and after steering toward an incorrect quadrant (Steer). Pos: obj = obj_word. Δacc is the change in parsed-answer accuracy in percentage points; Δ_{parse} is the intervention-minus-noise contrast in the final parsed target rate. It is not the fixed-state next-token Δargmax. Pmt: ans = answer-only, vis = visual_cot, cot = text CoT.

Model	Cond	L0	L4	L8	L12	L16	L20	L24	L28
Qwen-7B	o,a	0	+.083	**+.137**	+.095	+.112	+.002	+.002	0
	o,c	+.003	+.083	**+.142**	+.113	+.100	+.007	−.007	0
	p,a	−.002	+.012	+.055	+.088	+.267	**+.322**	+.073	−.020
	p,c	+.005	+.060	+.163	+.137	**+.213**	+.185	+.030	−.017
LLaVA	o,a	+.002	+.085	**+.117**	+.090	+.050	−.003	0	0
	o,c	0	0	+.002	**+.003**	+.002	0	0	0
	p,a	0	+.025	+.037	+.085	**+.347**	+.245	+.055	+.068
	p,c	+.002	+.040	+.020	+.042	**+.182**	+.142	+.010	+.048
IV3-8B	o,a	+.003	+.082	+.082	**+.112**	+.075	+.012	+.002	0
	o,c	0	0	0	0	0	−.002	0	0
	p,a	−.002	−.005	+.012	+.177	+.382	**+.455**	+.180	+.002
	p,c	0	+.003	+.008	+.042	**+.148**	+.098	+.025	+.018
Janus	o,a	+.020	+.085	+.198	**+.315**	+.002	+.002	0	+.002
	o,c	0	0	0	0	0	0	0	0
	p,a	+.035	−.017	−.075	−.033	−.130	−.077	−.098	−.035
	p,c	0	+.010	+.010	+.012	+.007	−.103	−.057	0

Table 15: GQA Spatial layer sweep (4 models). o = obj_word, p = prefix_last, a = ans-only, c = CoT. Bold: peak. IV3-8B: relocation (obj L12, pfx L20) with position-specific gating. Janus: obj peaks at L12; pfx negative.

$+0.082$, below the threshold needed to sustain CoT transport, and CoT color Δargmax collapses to 0.000. The architecture, pipeline, and intervention are identical; only the encoding amplitude differs.

J Visual Prompt Comparison

§3.3 contrasts text CoT (`cot`) with visual grounding (`visual_cot`) and reports that the latter bypasses obj_word gating. We compare both visual prompts (`visual_cot` and the briefer `visual_direct`) against text CoT and answer_only at both intervened positions, testing whether the bypass is a property of visual grounding in general or a property of the specific prompt wording.

Table 17 shows that both visual prompts bypass CoT gating at obj_word in every gated model: Vis-C and Vis-D both produce positive Δargmax even when standard CoT is exactly zero (e.g. IV3-8B Vis-C $+.164$, Vis-D $+.144$, CoT $.000$). The bypass is not specific to one wording. A second observation is that visual prompts sometimes open routes that even answer-only cannot: Janus's Vis-C/Vis-D values ($+.403$/$+.397$) exceed its answer-only peak ($+.069$) by a factor of ~ 6, suggesting that visual grounding prompts can amplify spatial transport beyond the canonical answer-only baseline in models that route through visual-thinking pathways.

At prefix_last, the bypass is also visible: Gem-4B reaches $+.206$ under Vis-C despite a null prefix_last under both CoT and answer-only, indicating that visual grounding can reopen a position-level route that standard prompts leave this route near zero. Qwen-32B shows a similar pattern, with prefix_last Vis-C $+.217$ and Vis-D $+.200$, consistent with its prompt-selective grouping in which the prefix_last route depends on prompt format.

Across families, the two visual prompts generally agree in direction and magnitude: the Pearson correlation between Vis-C and Vis-D at obj_word is $r > 0.99$ across the ten models, and the largest Vis-C / Vis-D discrepancy is Gem-12B ($+.064$ vs. $+.103$). This confirms that visual grounding bypass is a property of the prompt class, not a quirk of one template. Janus remains the most amplified case: its Vis-C and Vis-D values at obj_word ($+.403$ / $+.397$) also extend to prefix_last ($+.244$ / $+.203$), making it the only non-relocating model with strong positive prefix_last under visual prompts.

As an additional axis-specific control, horizontal-axis centroids do not shift vertical-axis answers, and vice versa (8 axis-transfer experiment files on RefCOCO), so the spatial transport documented here is direction-specific rather than a generic sensitivity to any centroid difference.

K Stepwise CoT Generation Probe

The obj_word gating result in §3.2 measures Δargmax only at the first generation step. If CoT merely delays rather than blocks the spatial pathway, the patched direction should surface at some later step in the reasoning trace and shift the parsed final answer. We test this with a stepwise probe.

We record per-step A/B/C/D logits at every greedy decoding step for Qwen-7B and LLaVA at L16, α=5, and report both the best single-step target-logit gain over the trajectory and the final answer hit rate (Table 18). Object-word patching does move later-step logits in absolute terms (best gain $+1.32$ at Qwen-7B step 65; $+2.88$ at LLaVA step 109). These are non-trivial logit shifts, but they do not by themselves establish a stable answer readout at those steps.

What does not change, however, is the parsed answer. Across 32, 64 and 128-step traces, the final-answer hit rate moves by at most $+0.10$ over the noise control. The intervention pushes the target logit upward, but rarely enough to flip the argmax against the rest of the sequence's accumulated evidence. The probe therefore does not support a simple "first token versus later token" explanation: later logit shifts can occur without systematic parsed-answer recovery. It also does not prove that every generated step has the same transport status. The empty-answer fractions in the rightmost column (intervention/noise) are matched, ruling out the explanation that the intervention simply destabilises generation into unparseable output.

L Answer-Step Timing Control

The stepwise probe above follows every generated token, but it does not force the readout to occur at the model's eventual answer step. We therefore run a targeted timing control on RefCOCO. For each clean text-CoT generation, we locate the step at which the model produces its answer letter, reconstruct the prefix immediately before that letter, and read the A/B/C/D logits there. We then apply the same object-word direction patch and same-norm random-direction control at that answer step. The direction, crossfit procedure, α=5, and 128-token budget are otherwise unchanged.

Table 19 reports the target-logit gain on all sam-

Model	L0	L4	L8	L12	L16	L20	L24	L28
Qwen-7B	0	**+.082**	+.040	+.028	+.010	0	−.002	0
LLaVA	+.002	**+.103**	+.062	+.018	+.007	−.002	+.002	0
IV2.5	+.003	**+.102**	+.098	+.023	+.018	+.005	0	−.002
IV3-8B	+.002	**+.107**	+.073	+.033	−.002	0	0	0
Gem-4B	+.047	+.017	**+.078**	+.018	+.005	0	−.002	+.002
Janus	+.015	+.077	**+.082**	+.052	+.002	0	0	0

Table 16: Layer-wise Δargmax at obj_word on GQA Color (answer-only, α=5, n=50/quadrant). Bold: per-model peak layer. Color peaks at L4–L8, approximately four layers earlier than the spatial-attribute peaks at L8–L16.

Obj_word peak					*Prefix_last peak*				
Model	CoT	Vis-C	Vis-D	Ans	Model	CoT	Vis-C	Vis-D	Ans
Qwen-7B	+.081	+.136	+.111	+.169	**Qwen-7B**	—	+.122	+.169	—
Qwen-32B	+.006	+.117	+.097	+.089	**Qwen-32B**	—	+.217	+.200	—
LLaVA	.000	+.103	+.117	+.158	**LLaVA**	—	+.108	+.097	—
IV2.5	.000	+.153	+.128	+.114	**IV2.5**	—	+.058	+.047	—
IV3-8B	.000	+.164	+.144	+.236	**IV3-8B**	—	+.144	+.156	—
IV3-14B	.000	+.117	+.108	+.097	**IV3-14B**	—	+.108	+.097	—
Gem-4B	.000	+.092	+.100	+.178	**Gem-4B**	—	+.206	+.142	—
Gem-12B	.000	+.064	+.103	+.097	**Gem-12B**	—	+.131	+.022	—
Gem-27B	.000	+.022	+.061	+.081	**Gem-27B**	—	+.022	+.072	—
Janus	.000	+.403	+.397	+.069	**Janus**	—	+.244	+.203	—

Table 17: Peak Δargmax under four prompt types at obj_word (left) and prefix_last (right), RefCOCO, α=5. Vis-C = visual_cot, Vis-D = visual_direct, Ans = answer_only. Both visual prompts produce positive Δargmax in every gated model. Dashes: condition not run.

Model	steps	n/q	best step	best Δargmax	final Δ	empty
Qwen-7B	32	20	19	+.008	+.017	.82/.80
Qwen-7B	64	10	13	+.008	−.025	.63/.48
Qwen-7B	128	20	65	+.031	+.025	.41/.35
LLaVA	32	20	3	+.013	+.042	.25/.37
LLaVA	64	10	60	+.030	+.025	.07/.10
LLaVA	128	20	109	+.200	+.104	.02/.01

Table 18: Stepwise CoT probe. "best step": generation index with maximal target-logit gain; "best Δargmax": target-option argmax-rate contrast at that step. "final Δ": change in parsed-answer hit rate (intervention − noise). "empty": fraction of unparseable outputs (intervention/noise).

ples for which an answer step was found. Each source sample is paired with the three non-source target quadrants, giving 240, 177, and 216 records for LLaVA, Qwen, and InternVL3, respectively. The target-logit gain is positive from L8 through L20 in all three models and fades at L24. The same sign pattern holds on the clean-correct subset, although the secondary Δargmax is smaller and model-dependent. This control addresses the possibility that the first fixed readout is mistimed without claiming that all decoding steps share one transport profile.

M Blank-Image and Full-CoT Sanity

Two further controls address residual confounds. *Blank-image*: if the patched centroids were dominated by linguistic priors rather than visual content, patching at a blank canvas should still reproduce the original effect. *Full-CoT generation*: if the gating result depended on the specific phrasing of our `cot` suffix, a longer or differently worded CoT chain might still expose transport at later steps.

Table 21 shows that the blank-image substitution collapses the original-image effect (+.156 for Qwen-7B, +.168 for LLaVA) to a residual ±0.04 at n=20/q. This residual is within the noise floor of the bootstrap CIs in Appendix D for the same models. The direction we patch is therefore image-grounded, not image-independent. The full-CoT rows are intentionally underpowered (n=10/q) and serve only as a cross-check against the more informative 128-step probe in Table 18; both point in the same direction.

Model	answer-found n	L8	L12	L16	L20	L24
LLaVA-OV-7B	80	+.100	+.280	+.430	+.080	−.019
Qwen2.5-VL-7B	59	+.221	+.204	+.309	+.111	−.000
InternVL3-8B	72	+.649	+.763	+.538	+.176	−.081

Table 19: Target-logit gain at the model's actual generated answer step under an obj_word direction patch (RefCOCO, text CoT, α=5, five-fold crossfit). Values use all answer-found samples; the answer step is the position immediately before the generated A/B/C/D letter. The corresponding Δargmax at L16 is $+.125$, $+.011$, and $-.005$ for LLaVA, Qwen, and InternVL3, respectively, so the continuous gain is the primary timing-control readout.

N COCO-Spatial Axis Transfer

The four-quadrant RefCOCO MCQs combine horizontal (left/right) and vertical (above/below) information into a single answer letter, so the quadrant centroid we patch in principle mixes both axes. To check that each axis is independently encoded and transported, we construct a parallel COCO-Spatial benchmark from 2-object captions and convert it to two separate binary MCQs (left/right; above/below).

Table 20 reports L16, n=50/class, α=5, with 5-fold heldout centroids per axis. Horizontal transport reproduces cleanly: Qwen-7B reaches Δargmax $+.220$ under CoT and $+.210$ under answer-only, mirroring its open RefCOCO behaviour, and LLaVA shows the same CoT-zero / ans-positive split as on RefCOCO (.000 vs. $+.170$). Vertical transport is systematically weaker than horizontal: both models drop by roughly an order of magnitude on the above/below task ($+.070$ for Qwen-7B; .000 / $-.010$ for LLaVA).

This horizontal–vertical asymmetry is consistent with the broader literature on VLM spatial reasoning, where left/right has stronger linguistic and visual scaffolding than above/below (Kamath et al., 2023; Liu et al., 2023). For our purposes the relevant result is that each axis is independently steerable, so the quadrant centroid we patch in RefCOCO is not an artefact of axis confounding.

Model	Ax	CoT$_\Delta$	Ans$_\Delta$	CoT$_g$	Ans$_g$
Qwen-7B	H	+.220	+.210	+1.04	+0.34
LLaVA	H	.000	+.170	+0.44	+0.40
Qwen-7B	V	+.070	+.060	+0.35	+0.17
LLaVA	V	.000	−.010	+0.16	+0.06

Table 20: Axis-specific transport on COCO-Spatial binary MCQs (n=50/class, α=5, L16, 5-fold heldout centroids). Ax: H = horizontal (left/right), V = vertical (above/below). Δ = Δargmax; g = target-letter logit gain. Horizontal axis reproduces the RefCOCO CoT-zero / ans-positive pattern; vertical axis is systematically weaker.

Model	Pmt	Eval	Img	n/q	Contrast	gain
Qwen-7B	cot	gen	orig	10	−.033	—
LLaVA	cot	gen	orig	10	+.042	—
Qwen-7B	ans	logit	blank	20	−.042	+.064
LLaVA	ans	logit	blank	20	+.042	−.094

Table 21: Blank-image and full-CoT sanity controls (α=5, L16). Eval: gen = full 128-token generation hit rate, logit = next-token Δargmax. Img: orig = true image, blank = uniform-grey canvas. Blank-image patching collapses the original-image effect to within noise; the full-CoT rows are intentionally underpowered (n=10/q) and cross-checked by Table 18.

O Qwen-32B Scaling Sanity

Qwen-32B has 64 layers, more than twice Qwen-7B's 28, so the late-emergence pattern observed at L16 in Qwen-7B is expected to shift to proportionally deeper layers. This appendix performs the sanity check.

Table 22 reports a layer sub-sweep on n=30/q and a larger n=168 rerun at L32. CoT argmax stays at 0.000 at every layer tested, while answer-only Δargmax grows monotonically from $+.019$ at L16 to $+.089$ at L32, with the n=168 rerun confirming $+.085$ $[.060, .111]$. This mirrors Qwen-7B's late-emergence shape at a deeper layer index, supporting the scaling interpretation of §3.1.

The sweep also clarifies Qwen-32B's transport pattern. Within Qwen-7B the obj_word route is open under both prompts (the open pattern), but Qwen-32B's CoT Δargmax is null even at the deepest answer-only peak. The transition from 7B to 32B therefore moves the model from the open pattern into the prompt-selective pattern identified in §4.1. The Qwen family is the only family where the transport pattern is not stable across scales, providing a within-family counter-example to treating family as a sufficient predictor explicitly flag in the main text.

Run	Pmt	Lyr	Δ_{argmax}	95% CI	gain
n=30/q layer sub-sweep					
	cot	16	−.003	—	+.057
	cot	24	+.003	—	+.095
	cot	32	+.003	—	+.108
	ans	16	+.019	—	+.165
	ans	24	+.036	—	+.360
	ans	32	+.089	—	+.676
n=168 rerun at L32					
	cot	32	+.000	[.000, .000]	+.089
	ans	32	+.085	[.060, .111]	+.721

Table 22: Qwen-32B obj_word Δargmax and target logit gain across layers (top: n=30/q sub-sweep without CIs; bottom: n=168 rerun with 95% bootstrap CIs). Answer-only Δargmax grows monotonically from L16 to L32 while CoT stays at zero, mirroring Qwen-7B's late-emergence pattern at a deeper layer index.

P Implementation Details

This section documents the implementation choices that the main text references but does not spell out: how the object-word position is located in each tokenizer, how centroids are constructed without in-fold leakage, how the bootstrap is parametrized, where the patching hook is inserted, and what the compute budget was.

Object-word position rule. For each MCQ sample, we tokenize the question with the model's own tokenizer and locate the contiguous substring matching the referring expression for the gold object. The patch is applied to all tokens in this matched span. If an exact span match is unavailable, we fall back to the matched content word anchor tokens after removing common stopwords. This is the implementation used for the obj_word cells and matches the token-span rule used by the evaluation scripts.

Centroid construction. For sample s in fold f, we compute heldout centroids $C_q^{(\ell)}$ from folds $\{1, \ldots, 5\} \setminus \{f\}$ with gold quadrant q. Patching uses $\alpha(C_{\text{target}}^{(\ell)} - C_{\text{source}}^{(\ell)})$, guaranteeing no in-fold leakage. The same object-word-derived centroids are reused for prefix_last interventions; no separate prefix_last basis is fitted. This holds the direction source fixed while changing the injection position. Random-direction baselines sample isotropic Gaussian directions, project to unit norm, and rescale to the L2 norm of the Δ-centroid.

Bootstrap. Sample-cluster bootstrap with 2000 resamples; the resampling unit is the sample id (not the (sample, target) pair), so CIs reflect across-sample variability.

Hook implementation. Direction patching is a forward-pre-hook on the residual stream at the chosen LM layer. The hook adds $\alpha(C_{\text{target}} - C_{\text{source}})$ to the residual at every matched object-word token (or at the final prefix token for prefix_last). Logits at the next-token position are read for {A, B, C, D}.

Compute. All 7B–8B models run on a single A800 80GB GPU. Qwen-32B and Gemma-3-27B use two A800 GPUs with model parallelism. Each canonical cell (one layer × one prompt × one position) takes ∼20–40 minutes per model; the source-sample count follows the per-model protocol in Section 2. The full experimental grid (10 models × 7 prompts × multi-layer sweeps × two positions) totals approximately 200 A800 GPU-hours.

Q Attention Head Knockout Details

The head-knockout experiment in §4.2 reports a single number: L27.H21 in InternVL3-8B recovers $+0.063$ Δargmax under CoT. This appendix documents the full sweep used to localise that head, so that the result is not the product of multiple comparisons or a cherry-picked layer.

InternVL3-8B multi-layer scan. We zero each of the 32 attention heads at each of four candidate downstream layers (L17, L20, L24, L27) under CoT, n=50/q. For each layer we report the best-head recovery. Table 23 shows that only L27.H21 exceeds the $+0.03$ significance threshold ($+0.062$ at n=50, $+0.063$ at n=100 confirmation). All other $31 \times 4 = 124$ head–layer cells are at or below threshold. Under answer-only the same head produces only $+0.007$, ruling out a generic "always-helpful" interpretation of H21 and supporting its identification as a CoT-specific gating candidate in this architecture.

LLaVA-OV-7B: distributed gating. A nine-layer scan over L17–L27 in LLaVA, also under CoT n=50/q, finds no head exceeding the same $+0.03$ threshold. The best candidate is L22.H25 at $+0.040$ (n=50), which shrinks to $+0.013$ at n=100 confirmation and remains $+0.003$ under answer-only. LLaVA's gating is therefore distributed rather than concentrated in a single head, consistent with the broader observation that gate

KO Layer	Baseline Δ	Best Head	Recovery
L17	+0.072	H14	+0.032
L20	+0.078	H21	+0.028
L24	+0.080	H18	+0.006
L27	+0.080	**H21**	**+0.062**
L27 confirmation (n=100)			
L27 (CoT)	+0.095	**H21**	**+0.063**
L27 (AO)	+0.236	H21	+0.007

Table 23: InternVL3-8B head knockout (steer L16, obj_word, RefCOCO, α=5). KO Layer is the layer at which a single attention head is zeroed out; Baseline Δ is the CoT Δargmax without knockout; Recovery is the change after knockout (positive value = gate is opened); AO = answer-only. Bold values mark the only head and condition above the $+0.03$ significance threshold.

localisation, not gate *existence*, is architecture-dependent.

We do not run head knockout on Qwen-7B because it is open at obj_word and therefore has no gate to localise. A full upstream–downstream circuit account, including which heads write the gating signal and which downstream paths it suppresses, remains open and is discussed in the Limitations section of the main text.

R Responsible NLP Checklist

Reproducibility. All hyperparameters are reported: 5-fold crossfit, α=5 canonical / α=10 for the prefix_last ladder, sample-cluster bootstrap with 2000 resamples. Code and scripts will be released.

Data. All evaluation uses three public research datasets: RefCOCO and RefCOCO+ (Kazemzadeh et al., 2014; Yu et al., 2016), the GQA scene-graph benchmark (Hudson and Manning, 2019), and COCO captions (Lin et al., 2014) for the axis-transfer probe. Our MCQ derivation reuses the existing bounding-box and scene-graph annotations and introduces no new human labels. No personally identifiable information is collected or released.

Compute. Total: ∼200 A800 GPU-hours for all experiments.

AI assistants. AI coding assistance was used for codebase management and LaTeX formatting; all scientific claims, experimental design, and analysis are by the authors.

Ethical considerations. This paper studies model internals on a spatial MCQ task and does not involve user data, deployed systems, or sensitive content.