FROM WHAT IS SEEN TO WHAT IS MEANT: *How VLMs Infer Implications from Images*

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models can describe what an image depicts yet misinterpret what it means. This paper asks where that inference fails and examines four components: visual-evidence grounding, source-relation selection, target-domain mapping, and answer binding. We first conduct a four-stage audit of 975 incorrect II-Bench responses and assign 76.1% to source-relation selection or target-domain mapping, compared with 16.9% to explicit visual-evidence errors. Because response labels alone do not establish the computations behind these errors, we conduct controlled tests. Removing image content reduces accuracy, while three visual-attention summaries do not reliably distinguish correct from incorrect interpretations. When five true relations from the same image compete, the question-relevant relation ranks first in only 46.3% and 34.9% of full-context cases across the two models. Transferring relation-conditioned states changes both relation and answer preferences. With the correct relation fixed, implication-specific states steer preferences for held-out implications, and supplying an aligned implication improves final answer selection. These results show that the key bottleneck in image-meaning inference is not whether a model sees an image, but whether it can form and use the implication supported by the right relation among visible elements.

1 INTRODUCTION

Vision-language models (VLMs) are increasingly expected to reason about what an image means, not only what it depicts. Comics, posters, and related images often communicate indirectly, using a concrete scene to convey an intended message without stating it explicitly. Image implication understanding is the task of recovering such intended readings from depicted scenes by combining visual perception with conceptual inference (Zhang et al., 2021; Kundu et al., 2025; Liu et al., 2024). Yet current VLMs still lag behind humans. This gap leaves a central question: **Do models fail because they lack the necessary visual evidence, or because they form the wrong abstraction after acquiring it?**

Figure 1 makes this gap concrete. Across four panels, a positive individual approaches a group and gradually changes its collective mood and behavior. Qwen3-VL-2B and InternVL3.5-2B both identify the individual, the group, and the progression across panels. Nevertheless, both interpret the illustration as social proof and conformity. The intended reading instead emphasizes how individual positivity can influence a group's collective mood and behavior. The models recover the central entities and progression, but fail to compose them into the meaning conveyed by the image. The failure lies not in visual recognition but in how recognized content is organized and interpreted.

To establish that this failure is not isolated, this paper evaluates Qwen3-VL-2B and InternVL3.5-2B on all 1,399 test examples in II-Bench. The models achieve accuracies of 62.4% and 67.9%, respectively. Among the resulting 975 incorrect responses, 86.3% explicitly mention the key element required by the correct interpretation. GPT-5.5-assisted human failure-stage annotation further indicates that visual-evidence grounding is not the primary bottleneck in 83.1% of these errors. Together, the case study and aggregate evidence shift the focus from whether a model sees relevant content to how it uses that content. Motivated by these observations, this paper asks: once a VLM has accessed the relevant visual evidence, what happens between what is seen and what is meant?

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

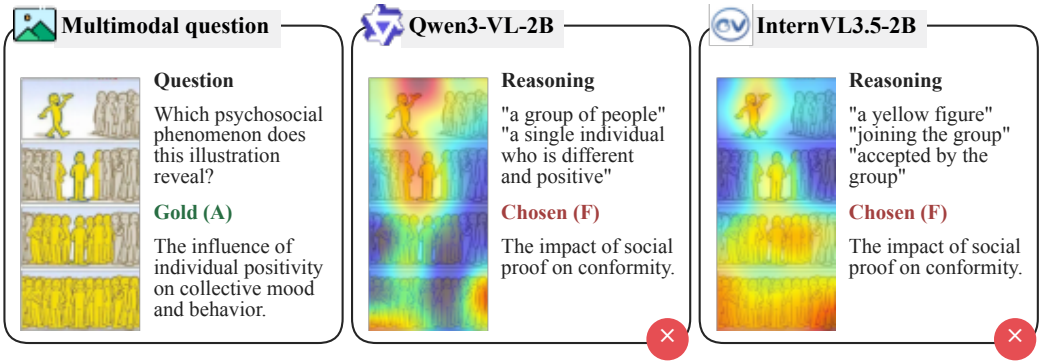

Figure 1: Both models describe salient visual content but infer an incorrect implication. Quoted phrases are excerpts from their reasoning; attention maps are independently normalized views.

To study this transition, this paper examines four components: *visual-evidence grounding*, which concerns the objects and cues needed by the question; *source-relation selection*, which determines which true visible relation organizes the interpretation; *target-domain mapping*, which connects that relation to an implication; and *answer binding*, which connects the implication to an option.

The experiments give a direct answer for each component. Image content is necessary, but the measured visual-attention patterns do not reliably predict whether its implication is recovered. The clearest behavioral bottleneck appears when several true relations compete for interpretive priority. Once a source relation is fixed, different target implications form distinguishable states that influence implication preference. Once an implication is supplied, the model shifts toward the corresponding answer. Layer-restricted adaptation further shows that late relation rewriting disrupts established behavior, whereas late target-mapping updates provide a more compatible calibration target.

The main contributions are as follows:

- To the best of our knowledge, this paper provides the first diagnostic decomposition of VLM image-implication inference into visual-evidence grounding, source-relation selection, target-domain mapping, and answer binding.
- It identifies the central behavioral bottleneck. Visual-access measures do not reliably separate correct from incorrect responses, while competition among true same-image relations prevents models from consistently forming the implication required by the question.
- It reveals that relation selection and target mapping require different forms of intervention. Relation-directed updates, especially in later layers, disrupt established relation states, whereas late target-mapping updates improve implication inference more consistently.

2 Related Work

From visual reasoning to image implication. Visual commonsense reasoning and knowledge-intensive VQA require models to go beyond object recognition by relating a scene to intentions or world knowledge (Zellers et al., 2019; Schwenk et al., 2022; Hudson & Manning, 2019). Broad MLLM benchmarks further evaluate perception and reasoning across diverse image types and question formats (Yue et al., 2024; Li et al., 2023). A complementary line of work tests compositional sensitivity to attributes, relations, order, and hard negative descriptions (Thrush et al., 2022; Yuksekgonul et al., 2023; Ma et al., 2023; Hsieh et al., 2023). Non-literal visual communication introduces a distinct challenge through multimodal metaphor understanding (Zhang et al., 2021) and visual-metaphor evaluation for VLMs (Kundu et al., 2025). II-Bench focuses this challenge on implications in abstract artworks, comics, and posters (Liu et al., 2024). These benchmarks evaluate the resulting inference; this paper instead localizes where the path from visible content to implication fails.

Intervention-based analysis of internal representations. Mechanistic analyses use controlled interventions to distinguish information that is merely present in a representation from information

that changes model behavior. Causal mediation identifies model components that mediate a specified output (Vig et al., 2020); causal abstractions formalize matched interchange interventions (Geiger et al., 2021); and causal tracing and editing locate computations that influence factual predictions (Meng et al., 2022; Hanna et al., 2023). Representation-geometry analyses further characterize how contextual states vary with input context (Ethayarajh, 2019). This paper applies these tools to image implication inference, linking source-relation states, target mappings, and final answer-option preference.

3 Experimental Setup and Diagnostic Framework

3.1 Task and Models

This paper studies image-meaning inference on II-Bench using Qwen3-VL (Bai et al., 2025) and InternVL3.5 (Wang et al., 2025). Let $x_s = (I_s, q_s, C_s, y_s^*)$ denote sample s, where I_s is the image, q_s is the question, C_s is the answer-option set, and y_s^* is the gold answer. Overall accuracy uses all 1,399 test examples. Scale and LoRA comparisons use the same 1,371 examples completed by all models; the remaining 28 Qwen-2B responses were excluded because the audit could not confidently assign a primary failure stage. The 35-example development set supports controlled diagnostics requiring manual annotations, repeated contexts, or internal-state interventions.

3.2 Four-Stage Diagnostic Framework

This paper uses four components to analyze image-meaning inference. *Visual-evidence grounding* asks whether the model accesses the objects, actions, text, and spatial cues required by the question. *Source-relation selection* asks which visible relation serves as the basis for interpretation. *Target-domain mapping* examines how a source relation is transformed into an implication, and *answer binding* asks whether that implication supports an answer option.

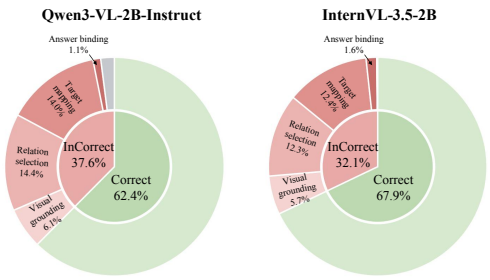

Figure 2: II-Bench accuracy and primary failure stages (% of all examples). Inner rings show correctness; outer rings show failure stages.

The audit identifies the earliest observable deviation in each incorrect response and estimates its prevalence. We then test specific operations suggested by this distribution: whether image content affects answers, whether models can prioritize the question-relevant relation among true alternatives, whether implication-specific states influence preference when the relation is fixed, and whether a supplied implication affects final option choice. These controlled tests establish decision relevance at each tested link; the audit supplies the distribution of observable failures.

GPT-5.5-assisted human annotation categorizes the 975 incorrect responses as visual-evidence grounding (165, 16.9%), source-relation selection (373, 38.3%), target domain mapping (369, 37.8%), or answer binding (38, 3.9%); 30 (3.1%) remain unresolved (see the Appendix). Relation selection and target mapping together account for 76.1% of errors, with similar distributions across models (Figure 2). The audit localizes dominant failures at scale; subsequent controlled analyses test their causes.

3.3 Analysis Protocol

Natural-trajectory analyses use pre-collected, fixed trajectories generated by greedy decoding. Both models have 28 zero-indexed decoder layers: early (0–8), middle (9–18), and late (19–27). Because fine-grained analyses use the 35-example development set, controlled variants are averaged within each sample over cyclic label rotations, reducing sensitivity to wording and label position. Group-wise comparisons use two-sided Mann–Whitney U tests (Mann & Whitney, 1947) with Benjamini–Hochberg correction (Benjamini & Hochberg, 1995); q denotes the adjusted value. State transfers

report paired sample-level effects with nonparametric bootstrap confidence intervals (Efron, 1979). Details are provided in the supplementary material.

Across sections, s, ℓ, and $\mathbf{h}$ denote samples, decoder layers, and residual-stream states. In attention analyses, p, t, and a index visual tokens, response tokens, and heads; other symbols are defined at first use.

4 SEEING IS NOT ENOUGH

Figure 1 illustrates an error despite recognition of the principal visual content, and the failure-stage audit assigns 83.1% of errors to other primary stages. This section asks two linked questions: whether image content affects implication inference, and whether successful and unsuccessful interpretations exhibit different visual-access patterns.

4.1 DO MODELS USE IMAGE CONTENT?

Let $\Delta\mathrm{Acc} = \mathrm{Acc}_{\text{mean-fill}} - \mathrm{Acc}_{\text{original}}$. This paper replaces each image with its RGB channel-wise mean while holding the question, answer options, and answer-order protocol fixed.

Table 1: Mean-fill accuracy change, $\Delta\mathrm{Acc}$ (all $q < .001$).

Setting	Qwen3-VL Δ	InternVL3.5 Δ
II-Bench 2B	-17.07 pp	-13.27 pp
II-Bench 4B	-19.26 pp	-11.89 pp
CII-Bench 2B	-12.42 pp	-11.76 pp

Table 1 shows that accuracy falls by 11.76–19.26 percentage points in every tested condition. The effect holds across both model families, at the 2B and 4B scales on II-Bench, and on the independent CII-Bench. Image content therefore makes a substantial contribution to implication inference; full statistics appear in Appendix D.

Having established this dependence, the remaining analyses ask why image use succeeds on some examples but fails on others. They compare correct and incorrect responses along three visual-access dimensions: attention to question-relevant regions, the amount of attention allocated to image tokens, and the spatial dispersion of that attention.

4.2 DO INCORRECT RESPONSES MISS KEY VISUAL REGIONS?

For each development example, annotators mark the visible objects and local cues needed for the intended interpretation. ROI lift compares the response's image-conditional attention density inside those regions with the image-wide average. A value above one indicates above-average density in the annotated regions; area normalization makes samples with differently sized regions comparable. Appendix B gives the token-level definitions of all three diagnostics.

The top row of Figure 3 shows that both correct and incorrect responses attend to the marked regions. Mean ROI lift is slightly higher for correct samples in some model–layer groups, but the shifts are small relative to within-group variation and the distributions overlap in every layer group. None of the six comparisons is significant after correction (all $q \geq .667$). Attention to key regions therefore does not stably distinguish whether a response derives the intended implication from them.

4.3 DO INCORRECT RESPONSES ATTEND LESS TO THE IMAGE?

Previous work has improved visual question answering and hallucination control by strengthening question-guided visual attention or correcting attention biases over visual tokens (Yu et al., 2025; Woo et al., 2025). This paper therefore tests whether incorrect responses allocate less attention to the image as a whole.

Image-attention amount measures the average fraction of each response token's attention allocated to image tokens, independent of where it falls within the image. It tests a different explanation from ROI lift: an incorrect response might use less of the image overall even if it visits the relevant regions.

The middle row shows that image-attention amount is organized primarily by decoder depth, with a middle-layer peak in both outcome groups. Correct–incorrect shifts differ by model: correct

samples are slightly higher for Qwen, whereas InternVL reverses this direction in the middle and late groups. The distributions overlap, and none of the six comparisons is significant after correction (all $q \geq .929$). Incorrect responses therefore do not consistently allocate less attention to the image as a whole.

4.4 Do Incorrect Responses Show More Diffuse Attention?

Comparable attention totals may still hide different spatial organization. Correct responses may concentrate attention on a few informative locations, whereas incorrect responses may distribute the same mass over background or irrelevant objects. Recent reliability analysis likewise shows that spatial attention and answer reliability need not coincide (Mann et al., 2026). This paper therefore tests whether incorrect responses have higher spatial attention entropy.

Spatial entropy measures how diffusely each response token distributes its image attention and then averages across response tokens (Shannon, 1948). Higher values indicate more diffuse attention. Tokenwise normalization prevents response length from determining the measure.

The bottom row shows that spatial entropy varies more with model family and decoder depth than with correctness. Correct samples are more concentrated in some conditions, but these shifts are small relative to within-group variation and the distributions overlap. None of the six comparisons is significant after correction (all $q \geq .929$). Incorrect responses therefore do not consistently exhibit more diffuse visual attention.

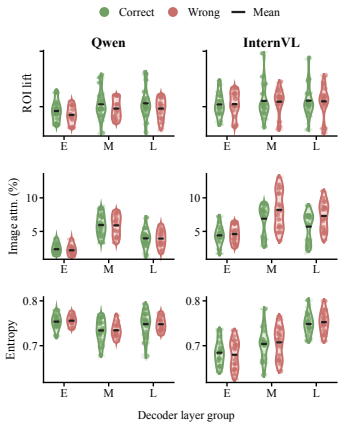

Figure 3: Visual-access diagnostics. Rows show ROI lift, image-attention amount, and spatial entropy. Points show rotations; dark points and bars show sample and group means. E/M/L denote decoder-depth groups.

Answer to visual-evidence grounding. Image-content interventions establish visual evidence as an important input to implication inference. Correctness, however, is not stably separated by attention to key regions, overall image-attention amount, or spatial dispersion, and the same result holds across model scale and benchmarks (Appendix D). Many incorrect interpretations therefore cannot be explained by the three measured forms of visual attention. This motivates the next question: how does the model choose which visible relation should support the intended implication?

5 Source-Relation Selection

Once visual evidence enters model computation, the next question is how that evidence is organized into an interpretation. A single image can contain several visible and true relations, but only one may support the implication required by the question. This section builds a three-level account of source-relation selection: it first identifies a bottleneck under explicit relation competition, then tests whether the same selection is reflected in natural inference, and finally intervenes on relation-conditioned states to measure their downstream effects.

5.1 Relation Selection Is a Capability Bottleneck

This paper turns image-meaning inference into an explicit relation diagnostic. Rather than generating an implication or answer, the model selects the source-domain relation most important for the question. Each example contains one *gold relation* and four visually true same-image competitors that are local, secondary, or insufficient. Let $R_s = \{g_s, c_{s,1}, \ldots, c_{s,4}\}$ denote these five *relation candidates*; *answer options* remain the original multiple-choice answers C_s. Candidates are length-balanced and blind-audited; each candidate occupies every A–E label position once.

This paper evaluates five contexts over the same candidates: *Rel. only*, *Q + Rel.*, *Q + I + Rel.*, *Q + Ans. + Rel.*, and *Full context*. They isolate the constraints supplied by the question, image, and answer options. For candidate relation r, let $\sigma_\rho(r; u) = \log p_\theta(\pi_\rho(r) \mid u)$ denote its score after label

rotation ρ. For sample s and available probe context u, the gold-relation margin is

$$m_s^{\mathrm{rel}}(u) = \frac{1}{5}\sum_{\rho=1}^{5}\left[\sigma_\rho(g_s;u) - \max_{1\le j\le 4}\sigma_\rho(c_{s,j};u)\right]. \tag{1}$$

Here, g_s is the gold relation, $c_{s,j}$ is competitor j, π_ρ is a label rotation, and u is the available context. A positive margin means that the gold relation beats every competitor.

Figure 4 shows that neither *Rel. only* nor *Q + Rel.* makes the gold relation reliably win. Adding the image lowers top-1 for both models as multiple true relations enter the choice space. Answer options constrain that space and improve top-1, yet *Full context* remains below 50% (46.3% and 34.9%). Even with all available context, both models frequently rank a true same-image competitor above the question-relevant relation. Relation selection is therefore a cross-model capability bottleneck after visual access.

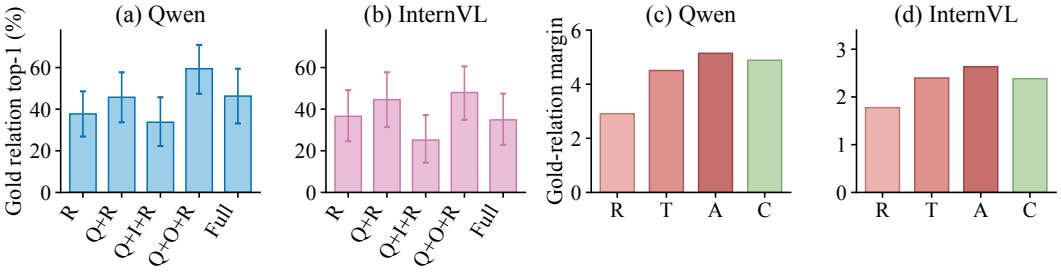

Figure 4: Relation selection in controlled choices and natural trajectories. (a,b) Gold-relation top-1 after five A–E rotations, with 95% confidence intervals. R denotes relation candidates, Q the question, I the image, and O the original answer options. (c,d) Gold-relation margins of audit-passed natural trajectories; R, T, A, and C denote relation, target-mapping, answer-binding, and correct groups.

5.2 Relation Selection Appears in Natural Inference

The explicit diagnostic establishes that models struggle when true relations compete, but it does not by itself show that the same selection occurs during ordinary answering. Treating relation selection as part of natural inference yields a direct prediction: trajectories that fail at relation selection should contain weaker evidence for the gold relation than trajectories that proceed to target mapping, answer binding, or a correct answer.

This paper tests the prediction using the models' pre-collected natural trajectories. Audit-passed responses are grouped as relation-selection errors (R), target-mapping errors (T), answer-binding errors (A), and correct responses (C). For a trajectory z_s, the same $m_s^{\mathrm{rel}}(u)$ is evaluated with $u = u_{s,z}^{\mathrm{rel}} = (I_s, q_s, R_s, z_s)$. A larger value indicates stronger probability evidence for the gold relation.

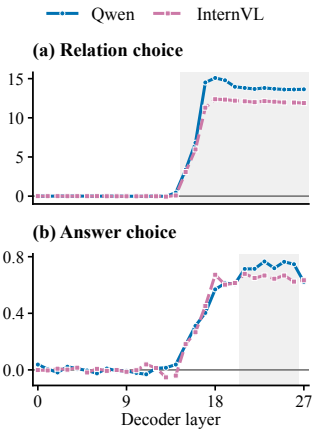

Figure 5: Relation-state effects; gray marks shared significance.

Figure 4(c,d) shows that R has the lowest gold-relation margin in both models. Compared with correct trajectories, relation-selection-error trajectories carry weaker gold-relation evidence in Qwen ($q = .0002 < .05$) and InternVL ($q = .0563 < .10$). Relation selection therefore leaves a measurable signature in natural inference: trajectories that fail to establish the question-relevant relation remain distinguishable from successful trajectories.

5.3 Relation States Influence Downstream Choices

The explicit diagnostic and natural trajectories show that relation selection is associated with answer correctness, but they do not establish whether relation-conditioned states affect later decisions. This

experiment therefore holds the surrounding context fixed, replaces only the relation-conditioned state, and measures whether relation and answer preferences change (Geiger et al., 2021; Meng et al., 2022).

The first probe inserts either a gold-relation state or a same-image competitor state into the same answer-blind relation-choice context. Figure 5(a) shows that the gold-relation state produces a consistently larger gold-relation margin from L15 onward ($q < .05$). Relation information at these layers can therefore change explicit relation choice.

The second probe performs the same comparison within the model's natural answer trajectory. Figure 5(b) shows that gold-relation states increase the gold-answer margin over L21–L26 ($q < .05$). Relation-conditioned states therefore affect relation choice from L15 onward and later contribute to answer selection. Exact state mixing and aggregation are provided in Appendix B.

Answer to source-relation selection. Source-relation selection is a capability bottleneck between visual evidence and image implication. When several true relations are available, both models frequently fail to give the question-relevant relation highest priority, even with the full context. This difficulty appears both in explicit competition and in the weaker gold-relation evidence of natural error trajectories; state interventions further show that relation-conditioned states change relation and answer preferences. The central challenge is therefore not representing which relations are visible, but determining which relation should govern the current interpretation.

6 From Relation to Implication

Section 5 identifies which visible relation should guide an interpretation. That relation is still a concrete description; the model must turn it into the implication required by the question. This section asks two questions: whether different implications produce different internal states from the same selected relation, and whether changing those states changes implication preference (Geiger et al., 2021; Meng et al., 2022). All analyses start from the same image, question, and selected source relation.

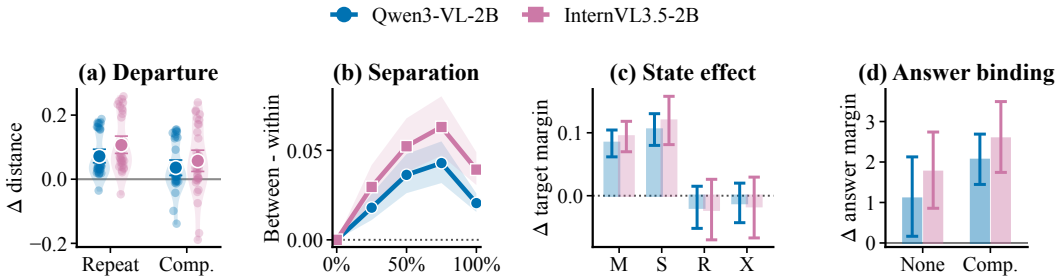

Figure 6: From source relation to implication and answer preference. (a,b) Starting from the same relation anchor, implication continuations produce an additional state change and separate by implication meaning. (c) M is the effect on the corresponding implication; S is its advantage over directions constructed for the other two implications; R and X are norm-matched random and cyclic cross-sample controls. (d) An aligned implication raises gold-answer margin relative to no implication (None) and a same-image competitor (Comp.). Points and bars show sample means; bands and whiskers show 95% bootstrap intervals.

6.1 Different Implications Produce Different States

All branches are identical up to a shared relation anchor: they contain the same image, question, and explicit gold source relation. After that point, a branch expresses a target implication, repeats the selected relation, or describes another true same-image relation. This design isolates what changes when the model moves from a concrete relation to an implication.

Figure 6(a) measures the residual-state displacement of each branch from the shared relation anchor in L19–L27 (Ethayarajh, 2019). Target continuations produce a larger displacement than either

relation control in both models (all $q < .006$). Expressing an implication therefore produces an internal update beyond merely repeating or replacing the selected relation.

Extra movement alone does not show that the update represents a particular meaning. Figure 6(b) therefore compares different implications with different paraphrases of the same implication. All branches begin at the same state; as the continuation unfolds, states for different implications become more separated than paraphrases of one implication in both models ($q < .001$). The update is therefore organized by implication meaning rather than by wording alone. Appendix B gives the exact distance measures.

6.2 Changing the State Changes the Preferred Implication

The separation above shows that implication states differ, but not whether those differences affect model behavior. If a state direction represents implication meaning rather than surface wording, it should generalize to an unseen paraphrase of the same implication. This paper therefore estimates the direction associated with one implication, injects it at the shared relation anchor, and evaluates a held-out paraphrase. Each direction is constructed from other paraphrases and contrasted with competing implications, so the test cannot be solved by reusing the wording that defines the direction.

The intervention changes only the anchor-token state; the image, input tokens, and all other residual positions remain unchanged. Figure 6(c) shows that moving the state in the matched direction selectively increases the probability of the corresponding held-out implication in both models (all $q < .001$). The positive S contrast shows that a direction has a larger effect on its corresponding implication than directions constructed for the other two implications; random and cross-sample controls do not reproduce the effect. The exact intervention and target-margin definitions appear in Appendix B.

Answer to target-domain mapping. Starting from the same source-relation anchor, implication continuations induce target-specific residual states: paraphrases of the same implication remain closer, different implications diverge, and steering the anchor along a matched direction increases the probability of a held-out implication. The controlled mapping experiment therefore identifies a semantic state transition that can causally steer the model from a selected relation toward a specific implication.

7 Answer Binding

Section 6 shows that target-specific states can change which implication the model prefers. The fourth stage asks whether the resulting implication enters the original answer-option competition. This paper compares no implication, an image-aligned implication, and a same-image competitor while holding the original task fixed. The outcome is the gold-answer log-probability margin over the strongest wrong option, averaged within sample across implication paraphrases and all six A–F label orders. Appendix B gives the exact aggregation.

Figure 6(d) shows that an image-aligned implication significantly increases gold-answer margin relative to both controls in both models (all $q \leq .0495$). Because the same-image competitor supplies comparable explanatory text without the aligned meaning, the contrast attributes the shift to implication content rather than text presence alone. The same effect reaches discrete answer choice: gold-answer Top-1 rises from 65.2% to 77.0% in Qwen and from 61.0% to 77.0% in InternVL. Sample-averaged $W \to C$ and $C \to W$ rates are 14.0% and 2.2% in Qwen, and 22.1% and 6.0% in InternVL, respectively.

Answer to answer binding. An aligned implication raises both the gold-answer margin and discrete Top-1 accuracy, with corrections substantially outnumbering harms in both models. Implication content is therefore behaviorally available to the final option competition, although answer binding remains imperfect. Together, the mapping and answer-binding experiments connect a selected relation to an implication-specific state and connect the resulting implication to answer preference.

8 From Diagnosis to Training

The preceding analyses separate two sources of failure after visual access: selecting the relation that should govern interpretation and mapping that relation to an implication. This section asks what scaling improves, whether the two stages should receive the same supervision, and where adaptation should act. Table 2 compares model scale, training target, and layer placement. Acc. is identity-order accuracy from the six-order evaluation, S6 requires correctness in all six orders, and Net is $W{\to}C - C{\to}W$ versus the matching 2B base.

What does scale improve? Scaling raises identity-order accuracy from 2B to 8B in both families. Qwen also improves monotonically under the six-order criteria, while InternVL reaches its highest S6 at 4B. Greater capacity therefore improves implication inference and order robustness, although these gains alone do not specify which computation training should target.

Should relation selection and mapping use the same objective? The results indicate that they should not. Full-layer relation LoRA reduces performance in both model families, whereas mapping LoRA is less disruptive and improves Qwen. A likely explanation follows from the preceding diagnostics. Relation learning must preserve several true same-image relations while calibrating which one should dominate for the question. Mapping supervision directly trains implication selection from the image, question, and six answer candidates, making it better aligned with the downstream decision.

Table 2: Scale and representative structured-LoRA results. Results use the shared 1,371-example evaluation set. Full results appear in Appendix E.

	Qwen3-VL			InternVL3.5		
Condition	Acc.	S6	Net	Acc.	S6	Net
Scale						
2B	63.3	38.0	–	68.2	52.2	–
4B	75.9	61.7	+173	75.2	60.2	+96
8B	79.4	64.5	+220	76.1	59.0	+108
Adaptation from 2B						
Relation r2	50.1	23.2	−181	57.0	34.5	−154
Mapping r2	**68.9**	51.0	**+76**	65.8	49.5	−33
Mapping r4	68.1	**51.1**	+65	62.9	45.4	−72
Rel. L0–15	54.0	31.6	−128	61.9	38.7	−86
Rel. L15–27	43.1	20.1	−277	44.3	17.6	−328
Map. L21–27	65.3	47.8	+27	**68.9**	**54.6**	**+10**

Where should adaptation act? Layer placement makes the contrast sharper. Relative to L0–L15, L15–L27 relation LoRA reduces S6 by 11.5 points in Qwen and 21.1 in InternVL, and both relation windows remain below their bases. L21–L27 mapping instead improves S6 over base in both families. Later layers are therefore poorly suited to rewriting established relation states but provide a compatible location for calibrating relation-to-implication mapping. Three-seed replications and direct transfer of II-Bench-trained LoRA to CII-Bench preserve this contrast (Appendix E).

Training implication. Relation learning should preserve relational diversity and calibrate question-conditioned priority. Mapping objectives can target the later transformation that connects a selected relation to implication and answer competition. The diagnostic framework therefore implies distinct training designs for the two principal reasoning failures.

9 Conclusion

We examined where VLM image-implication errors appear and which operations affect subsequent decisions. A four-stage response audit assigns most errors to relation selection or target mapping. Controlled tests show that image content matters, but the measured attention patterns do not reliably distinguish correct interpretations. Models often fail to prioritize the question-relevant relation among several true alternatives, and relation-conditioned states affect answer preference. With the correct relation fixed, implication-specific states steer implication preference; supplying an aligned implication improves final answer choice. Layer-restricted adaptation further favors late mapping updates over late relation rewriting. These findings motivate distinct training objectives for relation priority and implication mapping.

AI USE STATEMENT

Generative AI tools, including GPT-5.5, were used to produce candidate synthetic annotations and relation–implication descriptions for the controlled experiments, and to assist with language polishing and copyediting. All AI-generated candidates were reviewed by the authors against the source images, benchmark annotations, and model outputs before inclusion. The authors verified the resulting data and revised all AI-edited text, and take responsibility for the final content of the paper.

ETHICS STATEMENT

This study analyzes publicly released benchmark images, model outputs, and internal states, and does not collect private personal data or recruit human subjects. Human involvement is limited to research annotation and adjudication of benchmark examples. Any released artifacts will follow the licenses and redistribution terms of the original datasets and models. The stage labels are task-specific research judgments; unresolved cases are retained rather than assigned a forced label. The authors have read and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

The main paper specifies the evaluated models, datasets, layer indexing, answer-order controls, statistical tests, and principal interventions. The Appendix provides the prompt templates, annotation protocol, diagnostic definitions, aggregation procedures, training configurations, random seeds, cross-benchmark evaluations, and representative cases needed to reproduce and audit the reported results.

REFERENCES

Mohammad Asadi, Jack W. O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage: The illusion of visual understanding, 2026. URL https://arxiv.org/abs/2603.21687.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 01 1995. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1): 1 – 26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL https://aclanthology.org/D19-1006/.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan

(eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 76033–76060. Curran Associates, Inc., 2023. doi: 10.52202/075280-3322. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/efbba7719cc5172d175240f24be11280-Paper-Conference.pdf.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 31096–31116. Curran Associates, Inc., 2023. doi: 10.52202/075280-1355. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/63461de0b4cb760fc498e85b18a7fe81-Paper-Datasets_and_Benchmarks.pdf.

Drew A. Hudson and Christopher D. Manning. 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)*, June 2019.

Manishit Kundu, Sumit Shekhar, and Pushpak Bhattacharyya. Looking beyond the pixels: Evaluating visual metaphor understanding in VLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23137–23158, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1257. URL https://aclanthology.org/2025.findings-emnlp.1257/.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023. URL https://arxiv.org/abs/2307.16125.

Ziqiang Liu, Feiteng Fang, Xi Feng, Xinrun Du, Chenhao Zhang, Zekun Wang, Yuelin Bai, Qixuan Zhao, Liyang Fan, Chengguang Gan, Hongquan Lin, Jiaming Li, Yuansheng Ni, Haihong Wu, Yaswanth Narsupalli, Zhigang Zheng, Chengming Li, Xiping Hu, Ruifeng Xu, Xiaojun Chen, Min Yang, Jiaheng Liu, Ruibo Liu, Wenhao Huang, Ge Zhang, and Shiwen Ni. Ii-bench: An image implication understanding benchmark for multimodal large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 46378–46480. Curran Associates, Inc., 2024. doi: 10.52202/079017-1474. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/52764eb83bf0a0bd32766ce5c01612e5-Paper-Datasets_and_Benchmarks_Track.pdf.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10910–10921, June 2023.

H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60, 1947. doi: 10.1214/aoms/1177730491. URL https://doi.org/10.1214/aoms/1177730491.

Logan Mann, Yi Xia, Ajit Saravanan, Ishan Dave, Saadullah Ismail, Shikhar Shiromani, Emily Huang, Ruizhe Li, and Kevin Zhu. Visuals lie, consistency speaks: Disentangling spatial attention from reliability in vision-language models, 2026. URL https://arxiv.org/abs/2606.17389.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 146–162, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20074-8.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27 (3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5238–5248, June 2022.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1927–1951, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.99. URL https://aclanthology.org/2025.findings-acl.99/.

Runpeng Yu, Weihao Yu, and Xinchao Wang. Attention prompting on image for large vision-language models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 251–268, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73404-5.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, June 2024.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh*

International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Chenhao Zhang, Xi Feng, Yuelin Bai, Xeron Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, Xingwei Qu, Yifei Zhang, Qixuan Zhao, Yiming Liang, Ziqiang Liu, Feiteng Fang, Min Yang, Wenhao Huang, Chenghua Lin, Ge Zhang, and Shiwen Ni. Can MLLMs understand the deep implication behind Chinese images? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14369–14402, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.700. URL https://aclanthology.org/2025.acl-long.700/.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. MultiMET: A multimodal dataset for metaphor understanding. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3214–3225, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.249. URL https://aclanthology.org/2021.acl-long.249/.

A Prompts, Procedures, and Experimental Map

This supplement documents the fixed inputs, controlled variations, and sample-level aggregation underlying the main paper.

A.1 Shared Conventions

For example s, I_s, q_s, and C_s denote image, question, and displayed answer options; z_s is the response. Controlled experiments vary only the field in Table S1; cyclic rotations are averaged within sample before groupwise tests.

A.2 Prompt Templates

Table S2 gives schematic prompt templates. Bracketed fields are instantiated from the corresponding item. Full prompts, parsers, and model configurations are retained with the experimental records.

A.3 Analysis Procedure

Each analysis first obtains the original response, completes stage annotation before measuring downstream quantities, varies one controlled factor, and aggregates repeated variants within sample:

```
response[s] ← MODEL_RESPONSE(s)
stage[s] ← STAGE_ANNOTATION(s, response[s])
score[s,u] ← MODEL_SCORE(s,u)   for each control u
aggregate score[s,*] within s; report the prespecified
statistic
```

Audit independence. Stage review precedes every attention, relation-margin, residual-state, and adaptation analysis. Reviewers receive the image, task context, gold interpretation, and raw response, but never model identity or downstream measurements.

Table S1: Experimental map. Each row identifies the analysis unit, the controlled factor, and the recorded outcome.

Main-paper result	Data and unit	Controlled variation	Aggregation and inference	Recorded output
Figure 2: failure stages	975 incorrect test responses	None; original task responses	Per-response primary bottleneck	Stage label, evidence quote, and review record
Figure 3: visual access	35 development examples	Six A–F answer-label rotations	Six-rotation mean per sample; Mann–Whitney U with within-family BH correction	ROI lift, image-attention amount, spatial entropy
Figure 4(a,b): relation competition	35 development examples	Context condition and five A–E relation-label rotations	Mean across rotations per sample	Gold-relation margin and win rate
Figure 4(c,d): natural relation evidence	Original test responses	Natural trajectory retained as context	Response-level group comparison	Gold-relation margin by R, T, A, C group
Figure 5: relation-state transfer	35 development examples	Donor relation state, recipient context, and layer	Paired sample effects with bootstrap intervals	Relation-choice and answer-margin effects
Figure 6(a–c): target mapping	35 development examples	Target continuation, paraphrase, and transfer direction	Sample-level branch aggregates with BH correction	Anchor departure, target separation, implication margin
Figure 6(d): answer binding	35 development examples	No, aligned, or competitor implication; six A–F rotations	Sample-level averages over rotations and held-out paraphrases	Gold-answer margin
Table 2: adaptation	Shared 1,371-example evaluation set	Scale, LoRA target, rank, layer window, and seed	Six-order behavior; seed summaries where reported	Accuracy, S6, and net $W{\to}C - C{\to}W$

B Diagnostic Measures and Intervention Details

This section gives the exact definitions summarized in the main text. Let $\mathcal{V}_s$ and $\mathcal{O}_s$ denote the visual-token and response-token sets for sample s, $\mathcal{G}$ a decoder-layer group, and N_{h} the number of attention heads. For layer ℓ, head a, response query t, and visual token p, let $\alpha_{s,t,p}^{\ell,a}$ denote the attention weight.

B.1 Visual-Access Measures

The attention assigned to image token p, averaged over response queries, layers, and heads, is

$$\bar{\alpha}_{s,\mathcal{G}}(p) = \frac{1}{|\mathcal{O}_s||\mathcal{G}|N_{\mathrm{h}}} \sum_{t\in\mathcal{O}_s} \sum_{\ell\in\mathcal{G}} \sum_{a=1}^{N_{\mathrm{h}}} \alpha_{s,t,p}^{\ell,a}. \tag{2}$$

Normalizing within the image gives

$$\beta_{s,\mathcal{G}}(p) = \frac{\bar{\alpha}_{s,\mathcal{G}}(p)}{\sum_{v\in\mathcal{V}_s} \bar{\alpha}_{s,\mathcal{G}}(v)}. \tag{3}$$

If $\omega_s(p) \in [0,1]$ is the fraction of token p covered by the annotated region, area-normalized ROI lift is

$$\mathrm{Lift}_{s,\mathcal{G}} = \frac{\sum_{p\in\mathcal{V}_s} \omega_s(p)\beta_{s,\mathcal{G}}(p)}{|\mathcal{V}_s|^{-1}\sum_{p\in\mathcal{V}_s} \omega_s(p)}. \tag{4}$$

Thus, lift equals one under attention proportional to region area and exceeds one when the annotated region receives above-average density.

Table S2: Prompt templates for the analyses in the main paper. Bracketed fields are instantiated from each item.

Procedure	Schematic input	Output used by the analysis
Task response	`[image] [question] [A--F options]; brief reasoning followed by a fixed-format answer.`	Parsed answer and natural response z_s
Failure-stage audit	`[image] [question] [options] [gold interpretation] [raw response]; identify the earliest blocking deviation.`	Primary stage, exact evidence quote, confidence, review flag
Relation competition	`[optional image] [question] [optional answer options]; choose the relation from [A--E candidates].`	Gold-relation probability and maximum competitor probability
Relation-state transfer	`Visible relation: [gold or competitor relation]. Continuation:`	Residual state at a declared layer and downstream relation/answer scores
Target mapping	`Visible relation: [gold source relation]. Continuation: [target or source-domain text]`	Branch state trajectory and implication scores
Answer binding	`[image] [question] [A--F options]; implication: [none/aligned/competitor phrase].`	Gold-answer margin under each implication condition

Image-attention amount is the average fraction of each response query's attention assigned to visual tokens:

$$\mathrm{Amount}_{s,\mathcal{G}} = \frac{1}{|\mathcal{O}_s||\mathcal{G}|N_\mathrm{h}} \sum_{t\in\mathcal{O}_s}\sum_{\ell\in\mathcal{G}}\sum_{a=1}^{N_\mathrm{h}}\sum_{p\in\mathcal{V}_s} \alpha_{s,t,p}^{\ell,a}. \tag{5}$$

For spatial entropy, each response query is first normalized within the image:

$$\beta_{s,\mathcal{G},t}(p) = \frac{\sum_{\ell\in\mathcal{G}}\sum_{a=1}^{N_\mathrm{h}} \alpha_{s,t,p}^{\ell,a}}{\sum_{v\in\mathcal{V}_s}\sum_{\ell\in\mathcal{G}}\sum_{a=1}^{N_\mathrm{h}} \alpha_{s,t,v}^{\ell,a}}. \tag{6}$$

The normalized sample-level entropy is

$$\mathrm{Ent}_{s,\mathcal{G}} = -\frac{1}{|\mathcal{O}_s|\log|\mathcal{V}_s|} \sum_{t\in\mathcal{O}_s}\sum_{p\in\mathcal{V}_s} \beta_{s,\mathcal{G},t}(p)\log\beta_{s,\mathcal{G},t}(p). \tag{7}$$

B.2 Relation-State Transfer

For the answer-blind relation-choice recipient ξ_s, let $\mathbf{h}_{s,\ell}^{g}$ and $\mathbf{h}_{s,\ell}^{c}$ be the layer-ℓ states from the gold- and competitor-relation donors. The relation-state advantage is

$$\Delta_\ell^{\mathrm{rel}} = \frac{1}{|\mathcal{D}|}\sum_{s\in\mathcal{D}}\left[m_s^{\mathrm{rel}}\left(\mathbf{h}_{s,\ell}^{g}\to\xi_s\right) - m_s^{\mathrm{rel}}\left(\mathbf{h}_{s,\ell}^{c}\to\xi_s\right)\right]. \tag{8}$$

A positive $\Delta_\ell^{\mathrm{rel}}$ means that the gold-relation state more strongly favors the gold relation in the shared recipient context.

For the natural-trajectory recipient, the intervention mixes its state with a norm-matched donor at $\eta = 0.75$:

$$\widetilde{\mathbf{h}}_{s,\ell}^{\mathrm{don}} = (1-\eta)\mathbf{h}_{s,\ell}^{\mathrm{nat}} + \eta\frac{\|\mathbf{h}_{s,\ell}^{\mathrm{nat}}\|_2}{\|\mathbf{h}_{s,\ell}^{\mathrm{don}}\|_2}\mathbf{h}_{s,\ell}^{\mathrm{don}}. \tag{9}$$

Let $m_s^{\mathrm{ans}}(\cdot)$ be the gold-answer log-probability margin over the highest-scoring wrong answer. The layerwise advantage of a gold-relation donor over the four same-image competitors is

$$\Delta_\ell^{\mathrm{ans}} = \frac{1}{|\mathcal{D}|}\sum_{s\in\mathcal{D}}\left[m_s^{\mathrm{ans}}(\widetilde{\mathbf{h}}_{s,\ell}^{g}) - \frac{1}{4}\sum_{j=1}^{4} m_s^{\mathrm{ans}}(\widetilde{\mathbf{h}}_{s,\ell}^{c_j})\right]. \tag{10}$$

B.3 TARGET-STATE INTERVENTION

Let $d_{\cos}(\mathbf{h}, \mathbf{h}') = 1 - \cos(\mathbf{h}, \mathbf{h}')$. For sample s, let $\mathbf{h}^{(\ell)}_{s,\mathrm{rel}}$ be the residual state at the source-relation anchor in layer ℓ, and let $\mathbf{h}^{(\ell)}_{s,b,\mathrm{end}}$ be the state after branch b ends. The branch endpoint displacement used in Figure 6(a) is

$$\delta^{(\ell)}_{s,b} = d_{\cos}\left(\mathbf{h}^{(\ell)}_{s,\mathrm{rel}}, \mathbf{h}^{(\ell)}_{s,b,\mathrm{end}}\right). \tag{11}$$

Let $\mathbf{h}^{(\ell)}_{s,k,i}(\tau)$ denote the state of paraphrase i for implication identity k at normalized continuation progress τ, where $\tau = 0$ is the shared relation anchor. The separation statistic used in Figure 6(b) is

$$\Gamma_s(\tau) = \overline{d}_{s,\mathrm{between}}(\tau) - \overline{d}_{s,\mathrm{within}}(\tau). \tag{12}$$

Here, $\overline{d}_{s,\mathrm{between}}$ averages $d_{\cos}$ over $\mathcal{L}_{\mathrm{late}}$, distinct implication identities, and paraphrase pairs; $\overline{d}_{s,\mathrm{within}}$ uses the same layer average over distinct paraphrases of one implication.

For target k, a leave-one-paraphrase-out direction $\hat{\mathbf{v}}^{(-f,\ell)}_{s,k}$ is estimated from the other paraphrases' displacement at 75% progress relative to the source-relation anchor after subtracting the mean displacement of competing targets. The intervention changes only the anchor-token residual state:

$$\widetilde{\mathbf{h}}^{(\ell,k,f)}_{s,\mathrm{rel}} = \mathbf{h}^{(\ell)}_{s,\mathrm{rel}} + \hat{\mathbf{v}}^{(-f,\ell)}_{s,k}. \tag{13}$$

Let $\lambda_{s,\chi,k,f}$ be the average token log-likelihood of held-out target text k under intervention condition χ. The target margin is

$$m^{\mathrm{tgt}}_{s,\chi,k,f} = \lambda_{s,\chi,k,f} - \max_{j \neq k} \lambda_{s,\chi,j,f}. \tag{14}$$

Figure 6(c) aggregates these interventions as follows. **M** is the within-sample change in target margin relative to the no-op condition when each direction is evaluated on the implication identity from which it was constructed, averaged over the three implication identities and three held-out-paraphrase folds. **S** is this corresponding-implication change minus the mean change produced by directions constructed from the other two implication identities. **R** replaces the direction constructed for the evaluated implication with a deterministic random vector of the same norm. **X** uses the corresponding direction from the next sample in a fixed cyclic order and rescales it to the norm of the current sample's direction. Both controls are evaluated on a held-out paraphrase of the evaluated implication. All identities and folds are averaged within sample before statistical inference.

B.4 ANSWER-BINDING AGGREGATION

For answer binding, let $\mathcal{Z}_s(u)$ be the anonymous implication phrases supplied under condition u, and let $\pi_\rho(C_s)$ denote answer-label rotation ρ. The within-sample margin is

$$\bar{m}^{\mathrm{ans}}_s(u) = \frac{1}{6|\mathcal{Z}_s(u)|} \sum_{\rho=1}^{6} \sum_{z \in \mathcal{Z}_s(u)} m^{\mathrm{ans}}_s(\pi_\rho(C_s), z). \tag{15}$$

A larger value indicates a stronger preference for the gold answer under implication condition u.

C FOUR-STAGE ERROR AUDIT

The audit records the earliest response deviation that blocks the gold interpretation. It assigns one primary bottleneck to each incorrect response for Figure 2.

C.1 INDEPENDENT ANNOTATION AND ADJUDICATION

The final audit uses two independent GPT-5.5 reviewers under the same prespecified annotation guidelines. Each reviewer is blind to the other label. Disagreements, unresolved labels, low-confidence records, and incomplete evidence are sent to expert review; a human expert adjudicates these records and completes a final consistency pass. The audit provides the behavioral stratification used to relate large-scale failure patterns to the controlled analyses.

Stage labels are finalized before relation margins, attention summaries, residual-state probes, and LoRA outcomes are analyzed. The audit record retains both proposals, the adjudication decision, evidence quote, final label, coverage, adjudication rate, and annotation-guideline version.

Table S3: Annotation rules for the primary-bottleneck audit.

Stage	Positive operational definition	Required evidence
Visual-evidence grounding	The response makes an affirmative incorrect claim about a visually necessary object, action, text cue, or spatial configuration, and correcting that claim removes the earliest obstacle to the gold interpretation.	Exact erroneous response span plus the conflicting visual evidence. Non-mention alone is insufficient.
Source-domain relation selection	The response relies on a true but non-diagnostic same-image relation rather than the visible relation required to organize the gold interpretation.	The competing relation selected by the response and the gold relation it displaces.
Target-domain mapping	The response has reached the relevant source relation but maps it to an implication incompatible with the gold interpretation.	The relation-compatible response span and the incompatible implication.
Answer binding	The response reaches an implication compatible with the gold interpretation but selects or binds it to an incompatible answer option.	The implication span, selected option, and incompatible binding.
Unresolved	The earliest blocking deviation cannot be determined reliably from the available image, task context, and response.	A concise uncertainty note.

Retained audit record. Each record stores the reviewer identifier, annotation-guideline version, proposed stage, evidence quote, confidence, review flag, stage-pair agreement indicator, and adjudication decision. Final validation checks response-span coverage, data validity, unique sample coverage, and consistency of the primary stage. These records are stored independently of Figures 3–6, model identity, historical predictions, and adaptation results.

C.2 An Unresolved Primary-Stage Example

In II-Bench test-325, the image text answers "Why are you wearing that again?" with "Because I paid for it and I have a washing machine." Qwen3-VL-2B instead selects the uniform-related option and explains only, "She claims it's part of a uniform required for her job." This response does not establish whether the model missed the image text, selected a different visible relation, or mapped the relevant relation to the wrong implication. The audit therefore cannot confidently assign a primary failure stage and marks this response as unresolved.

D Additional Visual-Evidence Analyses

Figure 3 asks whether visual-attention statistics distinguish correct from incorrect responses. This section first tests whether image content causally affects correctness, then reports the complete II-Bench results, a larger-model check, and an independent CII-Bench replication.

D.1 Causal Visual-Input Checks Across Model Scale and Benchmarks

Recent work reports that VLMs can produce detailed image descriptions and retain high benchmark scores even without image input (Asadi et al., 2026). To verify that the models use image content and that image content affects their inference of image implications, this paper complements Figure 3 with an image-content intervention. Each image is replaced by its RGB channel-wise mean while the question, answer options, and answer-order protocol remain fixed. Let $\Delta_{\mathrm{acc}} = \mathrm{Acc}_{\text{mean-fill}} - \mathrm{Acc}_{\mathrm{original}}$ denote the paired change in discrete semantic accuracy. A negative value indicates that removing image content lowers answer accuracy.

Table S4 shows that mean fill significantly lowers answer accuracy in every tested condition (all $q < .001$), with drops of 11.76–19.26 percentage points. The effect appears in both model families, at the 2B and 4B scales on II-Bench, and on the independent CII-Bench. Image content is therefore

Table S4: Causal visual-input checks across model scale and benchmarks. Mean fill replaces only image content; all text and answer-order inputs remain fixed. Δ_{acc} is paired within item. All six comparisons form one BH family.

Benchmark	Model	Examples	Δ_{acc} (mean-fill − original)	BH q
II-Bench	Qwen3-VL-2B	1,371	-17.07 pp	$< .001$
II-Bench	InternVL3.5-2B	1,371	-13.27 pp	$< .001$
II-Bench	Qwen3-VL-4B	1,371	-19.26 pp	$< .001$
II-Bench	InternVL3.5-4B	1,371	-11.89 pp	$< .001$
CII-Bench	Qwen3-VL-2B	765	-12.42 pp	$< .001$
CII-Bench	InternVL3.5-2B	765	-11.76 pp	$< .001$

Table S5: ROI lift after six answer-label rotations. Qwen and InternVL denote Qwen3-VL-2B and InternVL3.5-2B. The unit of inference is the per-sample rotation mean.

Model	Layers	Correct	Incorrect	Δ	Rot. SD	p	q
Qwen-2B	Early	0.9227	0.8515	$+0.0712$	0.0024	.2320	.6668
Qwen-2B	Middle	1.0422	0.9668	$+0.0754$	0.0073	.3372	.6668
Qwen-2B	Late	1.0608	0.9665	$+0.0943$	0.0115	.5557	.6668
InternVL-2B	Early	1.0414	1.0461	-0.0047	0.0041	.6799	.6799
InternVL-2B	Middle	1.1040	1.0897	$+0.0143$	0.0096	.4577	.6668
InternVL-2B	Late	1.1087	1.0966	$+0.0121$	0.0196	.4187	.6668

important for obtaining the correct implication. The models use image content; the three diagnostics tested in Figure 3, however, do not stably distinguish correct from incorrect responses.

D.2 II-Bench Attention Diagnostics

This subsection provides the complete results underlying Figure 3. Question-relevant regions are independently annotated as the core source objects and local cues needed by the gold interpretation, then projected onto the 8×8 visual-token grid. Attention from response tokens to image tokens is collected under six cyclic answer-label rotations and averaged within each sample before comparing correct and incorrect groups. Each model contributes 35 sample-level observations, obtained by averaging the six rotations within each sample.

Table S5 gives the complete ROI-lift results underlying Figure 3. None of the six comparisons is significant after BH correction (all $q \geq .667$).

Table S6 reports the complete global attention diagnostics underlying the middle and right panels of Figure 3. None of the twelve image-attention-amount and conditional-spatial-entropy comparisons is significant after BH correction (all $q \geq .929$). Together, the three diagnostics show that, in both 2B models, a substantial fraction of failures cannot be primarily attributed to insufficient attention to the regions required for the gold interpretation, less attention to the image as a whole, or more diffuse image attention.

D.3 Qwen3-VL-4B Attention Diagnostics

This subsection repeats the three Figure 3 diagnostics for Qwen3-VL-4B on the II-Bench development set. Attention is measured for each natural response under six cyclic answer orders and averaged within sample before comparing the 24 correct and 11 incorrect responses. Tables S7 and S8 report the early, middle, and late decoder groups.

ROI lift, image-attention amount, and conditional spatial entropy do not stably distinguish correct from incorrect responses in any layer group (ROI: all $q \geq .367$; global diagnostics: all $q \geq .733$). Thus, in a larger Qwen checkpoint, a substantial fraction of failures likewise cannot be primarily attributed to insufficient attention to the regions required for the gold interpretation, less attention to the image as a whole, or more diffuse image attention.

Table S6: II-Bench global attention diagnostics. Qwen and InternVL denote Qwen3-VL-2B and InternVL3.5-2B. Differences are correct minus incorrect after averaging six answer orders within each sample. BH correction is applied over the twelve amount and entropy comparisons.

Model	Layers	Δ entropy	p	q	Δ amount	p	q
Qwen-2B	Early	−.00210	.8531	.9597	+.00160	.3905	.9597
Qwen-2B	Middle	−.00033	.6986	.9597	+.00050	.9597	.9597
Qwen-2B	Late	+.00014	.7491	.9597	+.00037	.9329	.9597
InternVL-2B	Early	+.00460	.8045	.9597	−.00193	.7791	.9597
InternVL-2B	Middle	−.00355	.6799	.9597	−.01319	.2548	.9597
InternVL-2B	Late	−.00429	.5415	.9597	−.01593	.0774	.9292

Table S7: Qwen3-VL-4B ROI lift on II-Bench after six answer-label rotations. The unit of inference is the per-sample rotation mean.

Layers	Correct	Incorrect	Δ	Rot. SD	p	q
Early	1.0368	.9191	+.1178	.0026	.1222	.3665
Middle	1.0627	.9457	+.1170	.0135	.4449	.6673
Late	.9872	.9239	+.0632	.0203	.8451	.8451

D.4 CII-Bench 2B Attention Diagnostics

To test whether the Figure 3 pattern transfers across datasets, this paper repeats all three diagnostics on the independent 35-example CII-Bench development split. As on II-Bench, attention is measured for each natural response under six cyclic answer orders and averaged within sample before comparing correct and incorrect groups. Tables S9 and S10 report the complete early, middle, and late results.

ROI lift, image-attention amount, and conditional spatial entropy again do not stably distinguish correct from incorrect responses (ROI: all $q \geq .318$; global diagnostics: all $q \geq .257$). Thus, on an independent image-implication development set, a substantial fraction of failures likewise cannot be primarily attributed to insufficient attention to the regions required for the gold interpretation, less attention to the image as a whole, or more diffuse image attention.

E Training Analyses and External Transfer

This section supplies the complete structured-LoRA results, training configuration, three-seed replication, and external CII-Bench behavior underlying Table 2.

E.1 Structured Supervision

All adapters start from the 2B base and are trained on the same 35 development examples. Every semantic target occupies each displayed A–F position exactly once per epoch, yielding 210 position-balanced tasks per epoch. **Relation** supervision presents the image, question, five audited same-image relation candidates, and an abstain alternative, and trains the displayed position of the intended relation. **Mapping** supervision presents the image, question, and six answer candidates, each expressing a candidate implication, and trains the displayed position of the intended implication. **Joint** supervision is the union of these two task sets. **Merge** averages the independently trained relation and mapping updates in float32, $\frac{1}{2}\Delta_{\mathrm{rel}} + \frac{1}{2}\Delta_{\mathrm{map}}$, and refactorizes the result at the original rank.

LoRA is attached only to the language decoder's self-attention q, k, v, and output projections; the vision encoder remains frozen. The primary sweep trains all decoder layers at ranks $r \in \{2, 4, 8\}$ with $\alpha = 2r$. The layer-window study fixes $r = 2$ and compares relation updates in L0–L15 and L15–L27 with mapping updates in L21–L27. Layer indices are zero-based and endpoints are inclusive; both relation windows contain L15. Table S11 gives the shared optimization configuration. The three-seed replication fixes $r = 2$ and uses seeds 20260722, 20260723, and 20260724.

Table S8: Qwen3-VL-4B global attention diagnostics on II-Bench. Differences are correct minus incorrect after averaging six answer orders within each sample. BH correction is applied over the six amount and entropy comparisons.

Layers	Δ entropy	p	q	Δ amount	p	q
Early	$-.00322$	.7357	.7357	$-.00203$	.2202	.7326
Middle	$-.02298$	.4884	.7326	$-.00721$	.4884	.7326
Late	$-.02976$	.4449	.7326	$+.00254$	.6828	.7357

Table S9: CII-Bench ROI lift after six answer-label rotations. Qwen and InternVL denote Qwen3-VL-2B and InternVL3.5-2B. The unit of inference is the per-sample rotation mean.

Model	Layers	Correct	Incorrect	Δ	Rot. SD	p	q
Qwen-2B	Early	.9529	.9226	$+.0304$	.0022	.4779	.7043
Qwen-2B	Middle	1.0520	1.0112	$+.0408$	.0117	.7043	.7043
Qwen-2B	Late	1.0661	1.0241	$+.0420$	.0196	.7043	.7043
InternVL-2B	Early	1.1099	1.1395	$-.0297$	.0020	.1060	.3179
InternVL-2B	Middle	1.1530	1.2115	$-.0585$	.0078	.0860	.3179
InternVL-2B	Late	1.1986	1.2374	$-.0388$	.0139	.2113	.4226

Table S10: CII-Bench global attention diagnostics. Qwen and InternVL denote Qwen3-VL-2B and InternVL3.5-2B. Differences are correct minus incorrect after averaging six answer orders within each sample. BH correction is applied over the twelve amount and entropy comparisons.

Model	Layers	Δ entropy	p	q	Δ amount	p	q
Qwen-2B	Early	$-.00021$	.6799	.9066	$-.00144$	.4577	.6866
Qwen-2B	Middle	$+.00481$	.2985	.6605	$-.00651$	.3141	.6605
Qwen-2B	Late	$+.00177$	.3302	.6605	$+.00054$	.9080	.9601
InternVL-2B	Early	$+.02416$	.0512	.2573	$+.00132$	.8545	.9601
InternVL-2B	Middle	$+.03105$	.0404	.2573	$+.00079$	.9601	.9601
InternVL-2B	Late	$+.01502$	.0643	.2573	$+.00542$	.4533	.6866

Table S11: Structured-LoRA training configuration shared by Qwen3-VL-2B and InternVL3.5-2B.

Component	Configuration
Training data	35 development examples; each relation or answer target occurs once at each A–F position per epoch (210 tasks per epoch)
Adapter placement	Language-decoder self-attention q, k, v, and output projections only; vision encoder frozen
Primary structured sweep	Relation, mapping, joint, and merge objectives; ranks $2/4/8$ with $\alpha = 4/8/16$ ($\alpha/r = 2$)
Layer-window sweep	Relation L0–L15, relation L15–L27, and mapping L21–L27; rank 2, $\alpha = 4$
Optimization	Two epochs; AdamW with $\beta_1 = .9$, $\beta_2 = .999$, $\epsilon = 10^{-8}$, and zero weight decay; constant learning rate 2×10^{-4} without warmup; per-device batch size 1; gradient accumulation 4; gradient clipping at 1.0; LoRA dropout .05
Precision and seeds	CUDA bfloat16; fixed seeds 20260722, 20260723, and 20260724 for the replication

E.2 Evaluation and Metrics

Adapters are evaluated on the shared 1,371-example test frame under all six answer orders. The frame was established during the initial Qwen-2B audit, when the primary failure stage could not be assigned confidently for 28 responses. Each evaluation receives only the image, question, displayed candidates, sample ID, and answer order. Gold answers, error-stage labels, historical predictions, source relations, and measurements from Figures 3–6 are joined only after scores are computed. Ev-

ery model–condition–seed evaluation has complete coverage, six finite normalized candidate probabilities per item, no evaluation errors, and matching model configurations.

Identity is fixed-order answer accuracy. *S6* is the fraction of items answered correctly under all six displayed orders, and *Stable-6* is the fraction for which the same semantic candidate is selected under every order. $W{\to}C$ and $C{\to}W$ count changes with respect to the corresponding 2B base; their difference is the Net column in Table 2. These metrics distinguish a raw accuracy change from a change that persists under answer-order variation.

E.3 Complete Structured-LoRA Sweep

Table S12 reports every rank and objective from the single-run sweep, including joint and merged adapters omitted from the compact main-paper table. Relation-only updates lower S6 at every rank in both families; mapping-only updates retain more of the base's six-order accuracy.

Table S12: Complete scale and single-run structured-LoRA sweep. Acc. is identity-order accuracy from the six-order evaluation, S6 requires correctness under all six orders, and Net is $W{\to}C - C{\to}W$ relative to the matching 2B base. Rel. and Map. denote relation- and mapping-directed updates; the best adaptation results are bold.

	Qwen3-VL			InternVL3.5		
Condition	Acc.	S6	Net	Acc.	S6	Net
Scale						
2B	63.3	38.0	–	68.2	52.2	–
4B	75.9	61.7	+173	75.2	60.2	+96
8B	79.4	64.5	+220	76.1	59.0	+108
Adaptation from 2B						
Base	63.3	38.0	–	68.2	52.2	–
Relation-only LoRA						
r2	50.1	23.2	−181	57.0	34.5	−154
r4	52.1	32.5	−154	55.3	34.1	−177
r8	44.8	19.0	−254	38.3	4.3	−410
Mapping-only LoRA						
r2	**68.9**	51.0	**+76**	65.8	49.5	−33
r4	68.1	**51.1**	+65	62.9	45.4	−72
r8	65.1	47.5	+25	65.7	50.5	−34
Joint LoRA						
r2	61.9	46.2	−20	62.7	46.7	−76
r4	66.6	50.0	+45	63.0	47.9	−71
r8	66.3	47.6	+41	63.9	47.3	−59
Merged LoRA						
r2	66.0	48.6	+37	65.4	49.2	−38
r4	66.4	50.9	+42	62.9	44.9	−73
r8	65.1	48.4	+24	59.2	35.2	−124
Layer-restricted r2						
Rel. L0–15	54.0	31.6	−128	61.9	38.7	−86
Rel. L15–27	43.1	20.1	−277	44.3	17.6	−328
Map. L21–27	65.3	47.8	+27	**68.9**	**54.6**	**+10**

E.4 Three-Seed Layer-Window Replication

Table S13 reports the complete rank-2 layer-window replication. Relation-only adaptation reduces identity, S6, and Stable-6 in both families; moving the relation update from L15–L27 to L0–L15 reduces, but does not remove, this damage. The late relation window has the most negative Net in both families. Mapping is comparatively stable: full-layer mapping improves Qwen relative to its 2B base, while L21–L27 mapping improves all three order-robust metrics for InternVL. These outcomes motivate the main paper's training insight that robust relation learning should preserve

competing same-image relations and calibrate their question-dependent priority, whereas mapping supervision offers a more localized target near candidate competition.

Table S13: Three-seed rank-2 structured-LoRA replication on the shared evaluation set. Values, including Net, are mean ± sample SD across seeds, except for the unadapted base.

Qwen3-VL-2B					
Condition	Window	Identity	S6	Stable-6	Net
2B base	–	63.31	38.00	42.01	–
Relation	All	50.13±2.15	25.04±3.85	29.61±6.08	−181±30
Relation	L0–L15	53.63±1.70	30.10±1.30	35.23±2.32	−133±23
Relation	L15–L27	43.69±2.60	18.70±2.50	26.21±2.94	−269±36
Mapping	All	67.47±1.42	49.21±1.91	56.16±2.37	+57±20
Mapping	L21–L27	65.28±.22	47.85±1.13	55.17±1.69	+27±3
InternVL3.5-2B					
Condition	Window	Identity	S6	Stable-6	Net
2B base	–	68.20	52.15	58.42	–
Relation	All	58.13±3.09	36.91±6.16	43.84±8.02	−138±42
Relation	L0–L15	62.44±.60	41.94±3.48	49.21±4.12	−79±8
Relation	L15–L27	44.08±.82	19.50±2.50	25.09±3.68	−331±11
Mapping	All	65.06±1.20	49.11±1.82	58.04±1.13	−43±17
Mapping	L21–L27	69.12±.18	54.85±.26	62.65±.48	+13±3

E.5 CII-Bench External Transfer

CII-Bench is an ACL 2025 benchmark of Chinese image implications, with six-option questions spanning life, art, society, politics, environment, and traditional culture (Zhang et al., 2025). This analysis uses its 765-question test split as an external behavioral transfer check. It evaluates 2B, 4B, and 8B base models together with three seeds of rank-2 adapters trained on the II-Bench development set and evaluated directly on CII-Bench. Every condition is evaluated under six cyclic answer orders. The displayed gold position has a detectable effect for all six bare base models on CII-Bench (all BH $q \leq .0011$), so Tables S14 and S15 report fixed-order identity together with S6 and Stable-6.

Scaling improves Qwen from 2B to 8B on identity, S6, and Stable-6, whereas InternVL-4B exceeds InternVL-8B on S6 and Stable-6. The transferred adapters then reproduce a stage–layer asymmetry. In both families, L15–L27 relation adaptation is the most damaging placement, while moving relation updates to L0–L15 reduces this damage. In Qwen, L21–L27 mapping improves identity, S6, and Stable-6 across all three seeds; in InternVL, mapping remains closer to the base behavior than either relation window. These external results are consistent with the main paper's account: relation formation is distributed across earlier layers and fragile to late rewriting, whereas target mapping admits more localized late calibration.

Table S14: Qwen3-VL external CII-Bench transfer on 765 test questions with six cyclic answer orders. Adapter rows report mean ± sample SD across three II-Bench training seeds. Identity is fixed-order accuracy, S6 is all-order correctness, Stable-6 is semantic answer stability, and Net is $W{\to}C - C{\to}W$ relative to the matching CII 2B base.

Condition	Identity	S6	Stable-6	Net
Scale				
2B base	38.69	17.91	27.97	–
4B base	52.68	36.47	51.37	+107
8B base	55.56	41.57	56.60	+129
II-Bench-trained rank-2 adapters (three seeds)				
Rel. L0–L15	39.43±1.76	18.00±2.21	25.62±4.78	+6±13
Rel. L15–L27	35.99±.87	12.11±2.13	21.35±4.92	−21±7
Map. L21–L27	42.40±.15	23.88±.27	39.83±.67	+28±1

Table S15: InternVL3.5 external CII-Bench transfer under the same protocol as Table S14. Adapter rows report mean ± sample SD across three II-Bench training seeds.

Condition	Identity	S6	Stable-6	Net
Scale				
2B base	46.54	29.54	44.58	–
4B base	52.03	36.34	51.63	+42
8B base	48.89	26.01	33.20	+18
II-Bench-trained rank-2 adapters (three seeds)				
Rel. L0–L15	42.88±1.47	26.62±1.26	43.83±1.53	−28±11
Rel. L15–L27	40.26±1.81	16.86±2.15	25.32±2.20	−48±14
Map. L21–L27	47.06±.13	30.59±.47	47.36±.72	+4±1

F Case Studies

This section grounds the diagnostic framework in sample-level case studies. The first group traces the same image from visual evidence through source-relation and target-implication competition. The second group shows the earliest divergence in one audited error for each primary bottleneck.

F.1 Shared Cases for Visual Evidence and Semantic Competition

Figure S1 displays five development examples used across the visual-evidence, relation-competition, and target-mapping protocols. Blue boxes mark independently annotated source objects and local text cues; orange boxes mark relation cues. The same image then supplies one intended source relation, four visually true competing relations, one intended implication, and two competing implications. Thus, the examples make explicit that a candidate can be visually supported without being the relation or implication required by the question.



Figure S1: Five shared development examples with independently annotated visual evidence. Blue boxes are source objects or local text cues; orange boxes are relation cues. From left to right: British spelling, Lego plastic surgery, a Shiba Inu tank driver, dessert/desert wordplay, and a circular comparison chain.

Table S16: Source-relation candidates for the shared examples. The intended relation is the gold relation; all four alternatives are grounded in the same image but do not organize the gold interpretation.

Case	Intended source relation	Visually true competing relations
dev-2	British spelling and slang align the accused soldier with the enemy in the others' eyes.	A soldier questions the written report; a soldier points to the extra *u*; the accused protests in British slang while being seized; another soldier calls him an English spy and orders his capture.
dev-3	Lego figures treat plastic surgery as swapping detachable plastic body parts during an operation.	Lego figures wait outside plastic-surgery doors; two surgeons attend a patient; the doors separate operating and waiting areas; a doctor holds a separate head piece beside the patient.
dev-4	A cute tank driver makes a soldier unable to fire at an otherwise dangerous military target.	Commanders discuss changed tactics; a soldier orders a shot; the assigned soldier develops heart-shaped pupils; a Shiba Inu rests on the tank turret.
dev-7	A waiter confuses *dessert* with *desert*, serving sand and cactus as food.	The waiter serves the cactus plate; the diner calls it dry; the offer contains the misspelling; the diner accepts food before delivery.
dev-12	Each woman wishes she resembled the next, forming a circular chain of mutual envy.	Arrows run clockwise between women; the same caption appears four times; the women have different outfits; each woman wears sunglasses.

Table S17: Target-implication candidates for the same shared examples. All three candidates are compatible with the visible source relation; the intended implication is benchmark-aligned and the alternatives isolate nearby, but distinct, semantic emphases.

Case	Intended implication	Competing implications
dev-2	Linguistic details reveal political allegiance.	Minor spelling clues invite wartime suspicion; cultural speech differences trigger public punishment.
dev-3	Cosmetic change becomes superficial part replacement.	Identity appears modular and easily reassembled; medical rituals become toylike consumer choices.
dev-4	Affection can override violent duty.	Charm turns warfare into emotional hesitation; softness can disarm organized aggression.
dev-7	Misspelling turns dessert into literal desert.	Language errors create absurd service outcomes; wordplay makes misunderstanding visually concrete.
dev-12	People covet others' perceived advantages.	Comparison traps everyone in dissatisfaction; mutual envy prevents self-recognition.

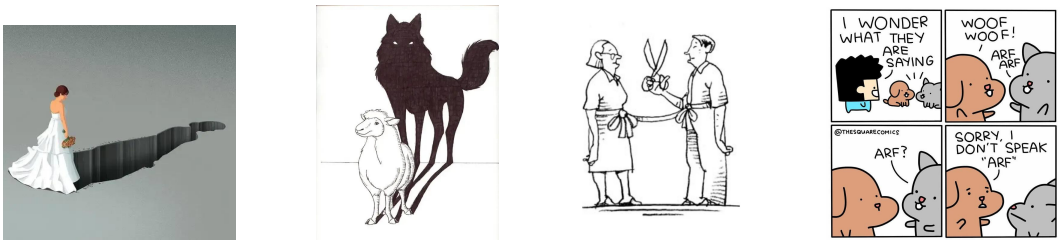

Figure S2: One audited II-Bench error for each primary bottleneck. From left to right: visual-evidence grounding, source-domain relation selection, target-domain mapping, and answer binding.

Table S18: Illustrative high-confidence primary-bottleneck records. Evidence is quoted from the frozen original responses; no wording is reconstructed for presentation.

Stage and case	Gold interpretation	Original response evidence	Why this is the earliest blocking deviation
Visual grounding Qwen3-VL-2B, test-1141	The abyss has the shape of a woman in a dress.	"The abyss is shaped like a jagged chasm."	The visible dress-shaped contour is asserted to be an ordinary chasm, blocking the required relation at the visual-evidence stage.
Relation selection InternVL3.5-2B, test-1061	Beneath a gentle exterior may lie a cunning heart.	"A contrast between the sheep's innocence and the wolf's cunning nature."	The response sees the sheep and wolf-shaped shadow, but treats them as independent opposing entities rather than the sheep's own shadow.
Target mapping InternVL3.5-2B, test-1053	The act calls for personal space and individual independence within a partnership.	"Break free from societal or familial expectations and define the relationship on their own terms."	The response recovers a man cutting the binding between two people but maps that relation to external expectations rather than independence within the partnership.
Answer binding InternVL3.5-2B, test-873	People with different linguistic and cultural backgrounds may experience barriers to communication.	"People with different linguistic and cultural backgrounds experiencing barriers to communication." The selected option is F, not gold option E.	The response explicitly expresses the gold implication but binds it to a non-equivalent option.

F.2 Four Primary-Bottleneck Examples

Figure S2 shows one high-confidence audited error for each stage. The accompanying evidence in Table S18 is quoted from the original model response. These cases illustrate the ordered annotation rule: the selected label is the earliest deviation that blocks the gold interpretation.