POSITION, NOT CONTENT: WHAT REPAIRS QUESTION-FIRST PROMPTING IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Putting the question before the image rather than after it changes nothing about what a vision-language model is told, yet it changes what the model answers. Prior work by the present authors established this question-first paradox on three yes/no benchmarks and traced it to read-out rather than perception: the question does reach the image and visibly steers it, but the answer position cannot read it back, and repeating either the question or the image repairs the damage. What that work could not say is how far the effect reaches, what the repeated image actually contributes, and what it costs. We answer those three questions. Across 14 further benchmarks the paradox proves general, costing 7 points on average and as much as 18 on a single benchmark, yet it is not a law: on object detection, where the model returns boxes instead of a word, the effect reverses and asking first is the better choice. The repeated image turns out to carry almost no information. Shrinking it to $\frac{1}{8}$ of its resolution leaves accuracy statistically unchanged while cutting its price from thousands of extra tokens to about 100, and a reversed copy works as well as a correct one, so what matters is where the tokens sit and not what they show. Breadth also marks a limit: on video, repeating the image stops helping, while repeating the question still does. A reflective prompt optimizer given the same budget improves only the ordering our diagnosis says has room left, and never catches the simpler fix, so this is not something better wording would have found. The practical consequence is immediate: put the question after the image, and where it has to come first, repeat it afterwards. The consequence for the field is that prompt layout, which papers rarely report, moves benchmarks further than the model differences those benchmarks are built to measure.

1 INTRODUCTION

Ask a vision-language model a question about an image and you must decide, often without thinking about it, whether the question goes before the image or after it. The information content is identical either way. Under causal attention, putting the question first is the option that lets every image token see it, which is what a reader would want when told to look for something specific. It makes the model worse, and on one benchmark category it more than halves accuracy.

Prior work by the present authors (Anonymous, 2026), included in the supplementary material, established this paradox on three yes/no benchmarks and diagnosed it as a read-out failure. Section 2 states exactly what is recapped here and what is new; Sections 3 and 4 compress the diagnosis, which the new results need.

Three questions that work could not answer motivate this paper, and each turns out to have a sharp answer.

Does the paradox survive outside yes/no visual question answering? We run the full ordering ladder on 14 benchmarks not previously tested, including four video QA datasets and RF20-VL, a 24,449-item detection-style presence benchmark. Question-first hurts on 12 of 13 image and video benchmarks, costing 7.0 points on average and as much as 18.0. The same breadth exposes two boundaries the original three benchmarks could not. On video, where the echo doubles an already long frame sequence, image echoing loses on three of four datasets. On RF20-VL detection, where

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.

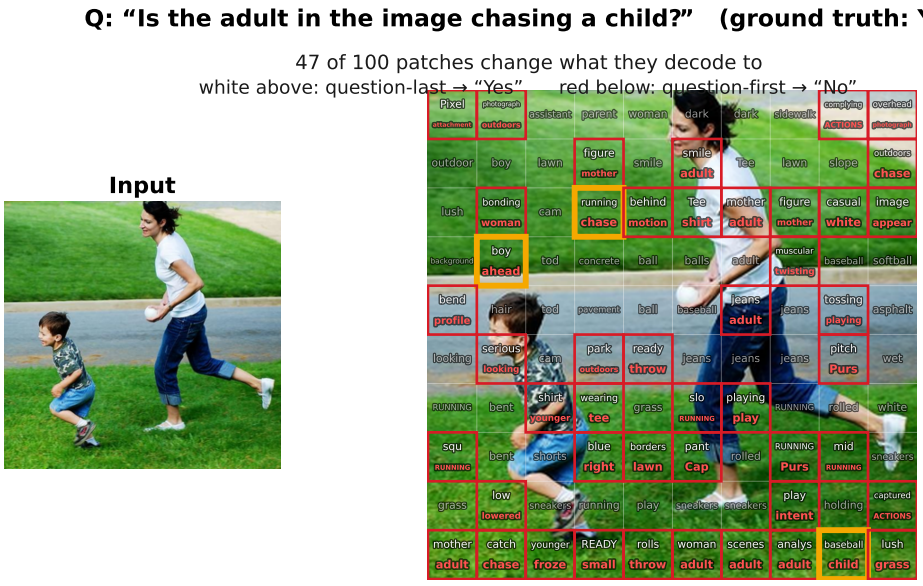


Figure 1: The ordering that looks most like it found the answer is the one that answers wrong. Every image patch of one NaturalBench example is shown once, carrying the vocabulary token it decodes to under a logit lens at layer 32. Unchanged patches carry one dimmed token; changed patches are outlined and carry both, question-last above in white and question-first below in red, with amber outlines on three of them (*running* to *chase*, *boy* to *ahead*, *baseball* to *child*). Moving the question in front of the image changes 47 of the 100 patches and takes the count decoding a word from the question itself from 1 to 12, yet the model then answers “No” where the ground truth is Yes and question-last answers “Yes”. Better perception, worse answer. Echoing the question after the image keeps all 12 steered patches and recovers the answer for 13 tokens.

the model emits boxes rather than an answer token, the paradox itself reverses: question-first is 2.80 mAP *better* and wins on 17 of 20 datasets, which places the effect in read-out rather than in prompt order as such.

Which patches change, and can we find them without cherry-picking? The qualitative evidence for steering has so far been chosen by hand, one persuasive image at a time, which is the kind of evidence a reader should distrust. We replace it with a procedure: decode every image patch under both orderings with a logit lens, drop the patches whose top-1 token is unchanged, and rank what survives by whether the new token belongs to the question’s own vocabulary. The figure reports every patch that changed, and the ranking, not the author, decides which are worth naming. Figure 1 is that output.

What is the echo actually made of? If the echo re-anchors read-out rather than supplying a second look, it should survive being destroyed as an image. It does. Downscaling the echoed copy to $\frac{1}{8}$ resolution, $\frac{1}{64}$ of the pixels, leaves NaturalBench statistically indistinguishable from a full-resolution echo ($p = 1.00$) while still beating question-last ($p = 0.005$), and cuts the price from 3,352 extra tokens per query to 107. RF20 is the lone exception, and an informative one: its targets are small, so the second look there does real perceptual work.

Would careful prompt engineering do the same job? We give GEPA (Agrawal et al., 2026), a reflective prompt optimizer, the same budget under both orderings. It improves question-first on all three benchmarks it is run on and finds nothing under question-last, which is what the read-out diagnosis predicts, since only the broken ordering has headroom. It closes at most 83% of the gap and never reaches the plain question-last baseline.

Together these say something the original result did not: the repair is not about giving the model a better look at the image but about where the tokens sit, it costs about a hundred tokens rather than a doubled prompt, and rewording cannot replace it. We also report where it fails, which is video and any task whose evidence is small enough that a degraded second look loses it.

2 Related Work

Prompt order and position effects. That the position of information in a prompt changes what a model does is established in text. Few-shot example order alone swings accuracy across a wide range (Lu et al., 2022), and long-context models recover facts placed at the ends of a context better than facts placed in the middle (Liu et al., 2024b), and both concern where evidence sits relative to the query. The question-first paradox is the multimodal counterpart with a sharper edge: our two layouts hold the same tokens in the same quantity, so context length, truncation and retrieval difficulty cannot distinguish them. Re-reading the question also helps text reasoning (Xu et al., 2024), the same intervention as our question echo reached from a different direction; that work motivates it as deliberation, whereas our diagnosis says it restores an attention path.

Prompt optimization. An alternative to reasoning about layout is to search over prompt text. DSPy compiles declarative pipelines into tuned prompts (Khattab et al., 2024), and GEPA evolves prompts by reflecting on failures (Agrawal et al., 2026). We use GEPA not as a competitor but as a test of the diagnosis, because it can rewrite the system message and cannot move the image: Section 7.3 shows it recovers part of the loss under the broken ordering and nothing under the working one, which is what a read-out account predicts and a wording account does not.

Interpretability of vision-language models. We read intermediate states with a logit lens (nostalgebraist, 2020; Belrose et al., 2023) and establish causality by severing attention edges, the method used to localize factual recall in language models (Geva et al., 2023; Meng et al., 2022). Our use is diagnostic rather than editing: we do not intervene to change behaviour at deployment, only to identify which path an answer depends on. The patch-difference procedure of Section 4.2 adds a small piece of method here, turning per-patch logit-lens decodes into an exhaustive, rankable list of changes rather than a chosen illustration.

Relation to our own prior work. This paper extends a workshop paper by the present authors (Anonymous, 2026), included in the supplementary material for reviewers. We state the overlap plainly. That paper established the question-first paradox on NaturalBench, POPE and Winoground, diagnosed it as a read-out failure using the perception probe, the read-out probe and the attention knockout, proposed question and image echoing as the repair, and reported the reversed-echo control. Sections 3 and 4 recap that material in compressed form, with all numbers reproduced on our own runs, because the new results are not interpretable without it. Everything in Sections 6 to 7 is new: fourteen further benchmarks including video and object detection, the reversal of the effect under a detection objective, the echo-resolution sweep and its exact token accounting, the automated patch-difference procedure, and the prompt-optimizer comparison. The earlier paper named the second image's token cost as an open limitation and did not measure it; measuring it, and showing how little of the echo is needed, is the core of this one.

3 The Question-First Paradox

This section and the next recap the phenomenon and diagnosis this paper extends (Anonymous, 2026), compressed to what the new results need. All numbers are reproductions on our own runs.

3.1 Notation and setup

A prompt has three sections: the system message (S), the image (I), which expands to many visual tokens, and the task or question (T). An ordering is written as its sequence, so `SIT` is system-image-question and `STI` is system-question-image. An ordering may repeat a section: `STIT` repeats the question after the image, `SITIT` repeats the image after the question, and `SITIT_rev` is `SITIT` with the second image copy spatially reversed. The two baselines are `SIT`, which is what most inference code emits by default, and `STI`, which is what a reader is told to do when looking for something specific.

Qwen3-VL-8B (Bai et al., 2025a) is the primary model and carries every analysis; for breadth we also run Gemma-3-27B (Gemma Team, 2025), Gemma-4-31B (Gemma Team, 2026), Qwen2.5-VL-7B (Bai et al., 2025b), InternVL3-8B (Zhu et al., 2025) and LLaVA-1.5 (Liu et al., 2024a). All

Table 1: Moving the question in front of the image is the worst choice on all four benchmarks, and echoing it back is the best on three. Ordering ladder for Qwen3-VL-8B, reproducing Anonymous (2026) and adding RF20. Bold marks the best ordering in each row. POPE is the exception to the repair: plain question-last already wins there. N is 1,900 groups for NaturalBench, 9,000 items for POPE, 400 examples for Winoground and 24,449 for RF20.

Benchmark	Metric	IST	SIT	STI	STIT	SITIT	SITIT_rev
NaturalBench	pair acc.	78.93	79.21	75.29	78.95	**80.36**	80.26
NaturalBench	group acc.	33.95	35.05	26.95	35.00	**37.37**	**37.37**
POPE	F1	88.22	**88.35**	86.13	87.61	88.12	88.25
Winoground	group acc.	28.00	31.75	22.25	37.50	40.25	**41.00**
RF20	F1	78.92	78.57	65.59	78.61	**80.56**	79.78

results use greedy decoding with a fixed answer-token budget, and every ordering of a benchmark shares one system message, one question string per item, one image preprocessing path and one scoring function, so ordering is the only difference between columns.

Three benchmarks carry the diagnosis: NaturalBench (Li et al., 2024a), POPE (Li et al., 2023) and Winoground (Thrush et al., 2022). Section 6 adds fourteen not tested in prior work: RF20-VL, twenty detection datasets drawn from RF100-VL (Robicheaux et al., 2025), used both for detection, scored by mAP, and as a $24{,}449$-question yes/no presence task over the same images; nine image benchmarks (BLINK (Fu et al., 2024), CV-Bench (Tong et al., 2024a), HR-Bench (Wang et al., 2025), MMStar (Chen et al., 2024), MMVP (Tong et al., 2024b), RealWorldQA (xAI, 2024), WorldMedQA-V (Matos et al., 2024), TallyQA (Acharya et al., 2019), VQAv2 (Goyal et al., 2017)); and four video QA benchmarks (MVBench (Li et al., 2024b), NExT-QA (Xiao et al., 2021), MSVD-QA (Xu et al., 2017), TGIF-QA (Jang et al., 2017)), where each `I` expands to six uniformly sampled frames. Sizes are in the tables. For NaturalBench and Winoground we pair on the group or example index and report an exact two-sided McNemar test with a paired bootstrap 95% interval over $5{,}000$ resamples.

3.2 The phenomenon: moving the question earlier makes the model worse

Table 1 runs six orderings on four benchmarks with Qwen3-VL-8B. Every column uses the same system message, question string, image, decoding and scoring code; `STI` and `SIT` contain the same tokens in the same quantity. Only the position of the question moves. Question-first is the worst ordering on all four: NaturalBench group accuracy falls from 35.05 to 26.95 ($p = 1.3 \times 10^{-12}$), Winoground from 31.75 to 22.25 ($p = 7.7 \times 10^{-5}$), RF20 F1 from 78.57 to 65.59.

Intuition predicts the opposite. Placing the question first is what a reader would do when told to look for something specific, and under causal attention it is the only layout in which every image token can see the question.

4 Anatomy of the Paradox

Two explanations come to mind, and both are wrong. A length or budget effect is ruled out immediately: `STI` and `SIT` hold the same tokens in a different order, so their prompts are identical in length (Table 11). The other, that question-first degrades perception, takes the rest of this section to reject.

4.1 Perception probe: the question reaches the image, and rewrites it

Reading image patches with a logit lens, question-first makes 6.975 patches per image decode a question content word against 2.800 for question-last, a factor of 2.49 over 160 NaturalBench candidates. Figure 1 shows this patch by patch: asking "Is the adult in the image chasing a child?" before the image turns *running* into *chase*, *boy* into *ahead* and *baseball* into *child*, taking the count of patches decoding a question word from 1 to 12, and the model then answers "No" to a question whose ground truth is Yes.

The rewrite shows in the representations too: against a forward pass on the image alone, STI's layer-18 image tokens sit at mean cosine 0.8627 where SIT's sit at 0.9110 and IST's are untouched (Appendix F). Causal masking supplies a control that matters later: a question copy placed *after* the image provably cannot alter the image tokens, and the measurement confirms it, $\cos(\texttt{STI}, \texttt{STIT}) = 0.99999$.

4.2 Finding the rewrite without cherry-picking

The evidence above comes from a procedure, not from inspection, because the alternative is choosing the persuasive image and hoping it is typical. We run the model under both orderings, project every image patch through the unembedding, take the top-1 token per patch, and drop the patches whose token is unchanged. Of those that remain we mark the ones whose new token belongs to the question's vocabulary, matched by stem, so a patch turning *throws* into *chase* counts and one turning *lush* into *grass* does not. The output is a count with a denominator rather than a gallery, it is exhaustive so a reader can check the changes we did not name, and it costs two forward passes. We ship it as an interactive viewer in the supplementary material, which steps through the changes one at a time; the paper figures are frames of it.

4.3 Read-out probe: the question is under-attended, and late

If perception is intact, the failure is downstream of it. Probing the answer position on the 150 NaturalBench pairs where question-first is wrong and question-last is right, question-first sends $2.19\times$ less attention to the question at the peak read-out layer (0.0676 vs 0.1479) and over twice as much to the image. The timing matters as much as the size: the two orderings are indistinguishable at chance through layer 25 of 36, then fork at layer 26 and reach $P(\text{correct})$ of 0.0986 and 0.8747 (Figure 4). A failure that appears only in the last quarter of the network is not an encoding failure.

4.4 Causal intervention: each ordering depends on a different path

The attention numbers are correlational; severing the edges makes the dependence causal and doubly dissociated (Appendix B). Cutting answer-to-question costs question-last 5.6 points (95% CI $[+1.99, +9.60]$) and question-first nothing ($+0.4$); cutting answer-to-image costs question-first 9.6 points and question-last nothing; an equal-length random span moves both by at most 0.8. Question-last answers by reading the question. Question-first answers by reading the image, into which it has already folded the question, and that path does not carry the answer.

5 From Diagnosis to Prompt: Echoing

The diagnosis prescribes the prompt with no design freedom left. If the answer position cannot read the question, put something after the image that it can read. There are exactly two things available.

Question echoing (`STIT`). Repeat the question after the image. This restores question-answer adjacency while provably leaving perception untouched, since causal masking means a later copy cannot alter earlier patches ($\cos(\texttt{STI}, \texttt{STIT}) = 0.99999$, Section 4.1). It costs 13 tokens, and it is the controlled intervention the diagnosis calls for: matched on content, matched on perception, differing only in whether the answer position has a question to read.

Image echoing (`SITIT`). Repeat the image after the question, so the answer position has both a nearby question and image tokens encoded with the question already in context. Prior work also reversed the echoed copy (`SITIT_rev`) and found it changed little, which we reproduce: NaturalBench group accuracy moves from 37.37 to 37.37 ($p = 1.00$) and Winoground from 40.25 to 41.00 ($p = 0.68$).

Two results in Table 1 keep us honest. Question echoing alone is not reliably better than asking the question last (NaturalBench 35.00 against 35.05, $p = 1.00$), so its role is to repair question-first prompting, not to improve on question-last. Image echoing is the real gain and the claim we defend ($+0.085$ on Winoground, $p = 2.4 \times 10^{-5}$), though on POPE it does not help.

Prior work left open what the echo is made of and what it costs. Sections 6 and 7 take up that thread.

Table 2: The paradox is not a property of the three benchmarks it was found on: question-first hurts on 12 of these 13, and image echoing repays the loss on still images but not on video. Full ordering ladder for Qwen3-VL-8B, with the four video QA datasets below the rule. Bold marks the best of the four orderings per row, the echo column is SITIT with the second image copy at half resolution, and each mean row averages the Δ columns above it.

		Ordering				Echo $\frac{1}{2}$	Δ vs SIT	
Benchmark	N	SIT	STI	STIT	SITIT	SITIT	STI	SITIT
BLINK	793	53.09	44.51	52.21	**53.97**	50.44	−8.58	+0.88
CV-Bench	2,638	81.35	63.34	79.98	**83.55**	83.70	−18.01	+2.20
HR-Bench	1,600	67.81	59.25	63.94	**69.75**	68.44	−8.56	+1.94
MMStar	1,500	53.33	41.93	51.47	**55.13**	56.07	−11.40	+1.80
MMVP	300	73.00	64.00	72.33	**77.67**	76.00	−9.00	+4.67
RealWorldQA	765	69.28	63.92	67.19	**70.33**	70.33	−5.36	+1.05
WorldMedQA-V	568	53.35	54.40	56.87	**57.22**	57.57	+1.06	+3.87
TallyQA	38,589	79.14	74.15	77.85	**79.99**	79.36	−5.00	+0.84
VQAv2	214,354	**81.66**	78.72	80.49	81.30	81.38	−2.94	−0.36
Mean, 9 image datasets							−7.53	+1.88
MVBench	2,600	64.08	55.88	62.23	**64.62**	66.08	−8.19	+0.54
NExT-QA	8,564	**78.21**	69.30	76.05	77.95	78.11	−8.91	−0.26
MSVD-QA	13,157	**40.05**	35.52	36.61	39.27	39.51	−4.53	−0.78
TGIF-QA	25,751	**36.32**	34.59	33.62	35.26	35.50	−1.74	−1.06
Mean, 4 video datasets							−5.84	−0.39

6 Results: How Far the Paradox Reaches

Prior work tested three yes/no benchmarks, which leaves open whether the paradox belongs to yes/no visual question answering, to those three datasets, or to the prompt layout itself. We run the full ordering ladder on 13 further benchmarks spanning single-image and high-resolution perception, counting, medical VQA and four video QA datasets, and separately on RF20, a 24,449-item presence-detection benchmark built from 20 detection datasets across 7 categories.

6.1 The paradox holds on 13 new benchmarks

Moving the question from after the image to before it costs Qwen3-VL-8B 7.01 points on average and hurts on 12 of 13 (Table 2). The effect is far larger than the differences that usually separate published models: CV-Bench loses 18.01 points, MMStar 11.40, MMVP 9.00, NExT-QA 8.91. WorldMedQA-V is the single benchmark where question-first is not worse (+1.06); we report it rather than dropping it, and it is also the smallest in the set ($N = 568$).

The two echo orderings separate the cost of the repair from its size. Question echoing recovers 5.49 of the 7.01 points, or 78% of the deficit, for 13 extra tokens. Image echoing recovers all of it and exceeds question-last by 1.18 points, winning on 9 of 13.

Finding. The paradox belongs to the layout, not to yes/no visual question answering: across 13 previously untested benchmarks, question-first costs 7.01 points on average and hurts on 12 of 13, worst case 18.01 on CV-Bench, with prompt content held identical.

6.2 RF20-VL: the sign flips when the output is boxes

Every benchmark so far asks for a short answer read out at a single position. RF20-VL asks for something else: given all class names for an image, emit bounding boxes, scored by mAP over 20 detection datasets and 3,970 images. Its output is localization rather than a read-out, which makes it the sharpest test of how far the paradox extends.

It does not extend; the sign reverses (Table 3). For Qwen3-VL-8B, question-first is 2.80 mAP *better* than question-last (7.94 vs 5.14) and wins on 17 of 20 datasets, and Gemma-4-31B agrees at a

Table 3: Ask the model for boxes instead of a word and the paradox inverts: question-first now *beats* question-last, while question echoing stays the best ordering. RF20-VL detection, mean over 20 datasets and 3,970 images, multi-class prompting. Bold marks the best ordering per row; `SITIT_rev`, not run for Gemma-4-31B (n/a), costs nothing on yes/no tasks but is worst here.

Model	Metric	SIT	STI	STIT	SITIT	SITIT_rev
Qwen3-VL-8B	mAP	5.14	7.94	**8.48**	6.47	3.25
	mAP_{50}	11.15	14.97	**16.01**	12.07	7.67
Gemma-4-31B	mAP	13.52	14.21	**14.43**	14.36	n/a
	mAP_{50}	29.75	30.12	30.48	**30.58**	n/a
Qwen3-VL-8B Δ vs SIT (mAP)			+2.80	+3.34	+1.33	−1.89
datasets improved			17/20	16/20	16/20	1/20

smaller size (+0.69 mAP, 13 of 20). We read this as a scope limit rather than a contradiction: when the answer is a set of boxes rather than a token the decoder must read out, stranding the question before the image costs nothing, and the extra conditioning it gives every patch appears to help. Echoing survives the reversal: question echoing is the best ordering for both models (+3.34 mAP for Qwen3-VL-8B, +0.91 for Gemma-4-31B), so the recommendation we make from the yes/no benchmarks stands here even though its justification does not.

One control separates this task from the others. Reversing the echoed copy was free on the yes/no benchmarks (NaturalBench $p = 1.00$); on detection it costs 1.89 mAP and helps on 1 of 20 datasets. With the resolution sweep, where RF20 alone degraded at every step as the echo shrank (Section 7.1), this says the echo's content matters only when the task needs evidence localized in the image. We exclude Gemma-3-27B: it scores 0.06 mAP under every ordering, so differences there are not interpretable. RF20's yes/no presence variant (24,449 questions over the same images, Appendix A) behaves like the other yes/no benchmarks, with question-first hurting on 7 of 7 categories.

Finding. The paradox is a property of read-out, not of prompt order as such: on RF20-VL detection, where the output is boxes, question-first is 2.80 mAP *better* than question-last and wins on 17 of 20 datasets. Question echoing stays best (+3.34 mAP), but the reversed echo, free on yes/no tasks, now costs 1.89 mAP.

6.3 Where the repair stops: video

The four video QA datasets draw a boundary the original benchmarks could not. The paradox is if anything more reliable there, hurting on 4 of 4 by 5.84 points on average, but the repair splits by which thing is echoed. Question echoing still recovers 3.31 of those points. Image echoing does not: it beats question-last on 1 of 4 video datasets and loses 0.39 points on average, against +1.88 on the nine image benchmarks.

The reason is mechanical. For video the "image" is a frame sequence, so an image echo doubles a visual context that is already long, and the read-out surface it adds sits behind far more tokens than in the single-image case. Echo the question on video, the image on stills.

Finding. The paradox and its repair have different domains: on video question-first still hurts on 4 of 4 (−5.84), but image echoing helps on 1 of 4 (−0.39 on average against +1.88 on stills) while question echoing still recovers +3.31. Echo the question on video, the image on stills.

7 Ablations: Isolating the Active Ingredient

The diagnosis in Section 4 makes a strong claim about what the second image copy is for. It is not there to be re-perceived. It is there to put image tokens after the question so the read-out step has something local to attend to. That claim has a sharp consequence: the echoed copy should not need to be legible. We test it by degrading the echo, by pricing it, and by asking whether a prompt optimizer that can only edit text finds the same fix.

Table 4: The echoed image can be shrunk until it is barely an image without costing accuracy, except on RF20, where the evidence is small and accuracy falls at every step. Only the *second* occurrence is downscaled. Bold marks the best scale per row. Metrics: F1 for POPE and RF20, group accuracy for Winoground, pair accuracy for NaturalBench.

Benchmark	Echo 1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$
POPE	88.12	**88.35**	88.07	87.90
Winoground	40.25	**41.75**	41.00	**41.75**
NaturalBench	80.36	80.55	**80.57**	80.18
RF20 (presence)	**80.56**	80.06	78.80	78.68

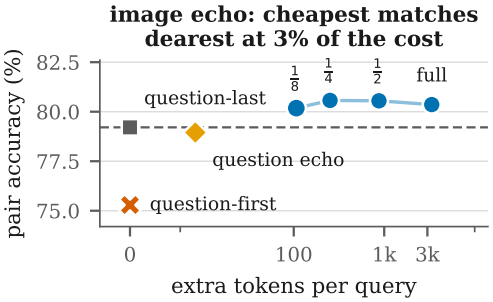


Figure 2: Accuracy is flat across a 31× range of echo cost, so the cheapest echo is the one to ship. NaturalBench; x is extra tokens per query against question-first, symmetric log so 0 shows. Every image echo clears the question-last baseline (dashed), and the $\frac{1}{8}$ echo (+107 tokens) is indistinguishable from the full one (+3,352).

7.1 The echoed copy does not need to be legible

We downscale only the second image occurrence in SITIT, leaving the first at full resolution, and sweep the echo scale over 1, $\frac{1}{2}$, $\frac{1}{4}$ and $\frac{1}{8}$ (Table 4). On three of four benchmarks a quarter-resolution echo matches or beats a full-resolution one: POPE moves by -0.05 F1, Winoground by $+0.75$ group accuracy, and NaturalBench by $+0.21$ pair accuracy. An eighth sharpens the point. On NaturalBench, an echo carrying $\frac{1}{64}$ of the original pixels scores 37.32 group accuracy against the full echo's 37.37, a difference of 0.001 with $p = 1.00$, while still beating question-last by $+0.023$ ($p = 0.005$). The echo can be destroyed as an image and keep its entire benefit.

RF20 is the exception, and the informative one: its accuracy falls at every step (-0.50 F1 at half, -1.76 at quarter, -1.88 at an eighth). It asks whether a specific, often small object class is present, so the second look does real perceptual work rather than only re-anchoring. The split says the echo has two separable jobs and only one of them needs pixels.

Finding. The echoed image does not need to be legible: an eighth-resolution echo carrying $\frac{1}{64}$ of the pixels is statistically indistinguishable from a full one on NaturalBench (37.32 vs 37.37, $p = 1.00$) while still beating question-last ($p = 0.005$). Only RF20, whose targets are small, needs the pixels (-1.76 F1 at quarter).

7.2 What echoing costs

Prompt-level interventions are usually reported without their bill. We count exact tokens for each ordering with the Qwen3-VL-8B processor over 20 images spanning RF20's constituent datasets, separating vision from text tokens (Table 11, Appendix G).

The two echo mechanisms are three orders of magnitude apart. Repeating the question costs 13 tokens; repeating the image at full resolution costs 3,352, roughly doubling the prompt. Downscaling collapses that cost quickly because vision-token count falls with area: 875 added tokens at half, 252 at a quarter, 107 at an eighth. Figure 2 puts the two axes together, and the accuracy curve is flat across a 31× range of cost, so the operating point to ship is the cheapest one.

Finding. Echo cost is a tunable knob, not a fixed doubling of the prompt: a full-resolution echo adds 3,352 tokens per query, a quarter adds 252 (7.5%) and an eighth adds 107 (3.2%) for indistinguishable accuracy, while question echoing adds 13.

7.3 Prompt optimization finds part of the fix, and only where the mechanism says it should

A reviewer will reasonably ask whether careful prompt engineering would do the same work as echoing, making the ordering story incidental. We test this with GEPA (Agrawal et al., 2026), a reflective prompt optimizer, used as a prediction test rather than a baseline to beat. The mechanism predicts an asymmetry: GEPA may rewrite the system message but cannot move the image, so it should find recoverable headroom under `STI`, where read-out is broken, and little or nothing under `SIT`, where it is not. That is what happens (Table 12, Appendix H). Under `STI`, GEPA improves all three benchmarks: POPE $86.41 \rightarrow 87.25$, VQAv2 $86.11 \rightarrow 88.13$, NaturalBench $74.74 \rightarrow 75.79$. Under `SIT`, the same optimizer with the same budget returns the seed prompt unchanged on POPE and VQAv2, both exact ties, and regresses NaturalBench by 0.41 by overfitting its validation split.

Two further numbers matter. `SIT`'s baseline is higher than `STI`'s on every benchmark (88.27 vs 86.41, 88.55 vs 86.11, 78.80 vs 74.74), so the asymmetry is headroom rather than a quirk of the optimizer; and GEPA's best `STI` prompt closes only 26% to 83% of the gap, never reaching the plain `SIT` baseline. Its prompts add 26 to 71 tokens per query, less than an eighth-resolution echo, but its search costs up to $982{,}343$ tokens once plus a separate reflection model. Rewording recovers part of what bad ordering costs; moving or echoing the image recovers the rest.

Finding. Prompt optimization behaves as the read-out diagnosis predicts: GEPA gains under `STI` (POPE $+0.84$, VQAv2 $+2.02$, NaturalBench $+1.04$) and finds nothing under `SIT` (two exact ties and -0.41). It closes 26% to 83% of the gap and never reaches the `SIT` baseline, so rewording is not a substitute for fixing token order.

8 Discussion

Three results constrain what the echo can be. It survives being reduced to $\frac{1}{64}$ of its pixels, it survives being turned upside down, and it fails where the visual context is already long. None of that fits an echo that supplies a second, better look; all of it fits an echo that supplies a read-out surface, whose content barely matters and whose position matters entirely. The one place content does matter, RF20, is the one whose evidence is a small region rather than a whole scene.

Implications beyond prompt ordering. An unreported prompt-layout choice is a confound the size of the model differences benchmarks exist to detect: it moves CV-Bench by 18.01 points and one RF20 category by 47.35, so two implementations of "the same" prompt can differ by more than the gap being measured. For deployed and agentic systems, pipelines that prepend a goal to a growing context and then append retrieved images produce question-first layouts by construction and pay for it silently; the fix costs 13 tokens on video and about 107 on stills. For training recipes, the failure lives in read-out rather than perception, which suggests instruction tuning that varies section order rather than fixing one canonical template; the Gemma models, which show no paradox, suggest it is trainable away.

Limitations. The diagnosis is established on Qwen3-VL-8B and does not hold for either Gemma model, so the read-out account describes a common failure mode rather than a property of the architecture; image echoing still helps the Gemma models (Appendix C), but the mechanism we recap cannot be the reason there. The repair is also not uniformly right: it does not beat question-last on POPE (88.12 against 88.35 F1), and on video it loses on three of four datasets, which is why we recommend question echoing there. Finally, the knockout result depends on restricting the query side to the answer position; the broader scope does not reproduce the question-side half of the dissociation, but its span-size control is large, so we treat it as uninformative rather than contrary (Appendix J).

Reproducibility, AI Use, and Ethics Statements

Reproducibility. Models, benchmarks and protocol: Section 3.1. Per-run detail, including the two configuration choices that change a result if missed: Appendices A–J. Every accuracy is read from a per-run result file, and significance tests use paired McNemar with a paired bootstrap over $5{,}000$

resamples at a fixed seed. Code and the figure scripts, which regenerate every figure from those result files, are provided in the supplementary material together with the prior workshop paper this work extends, and will be released publicly. **AI use.** AI tools helped with code and editing; the authors reviewed all of it and take full responsibility. **Ethics.** We audit existing, publicly released models on public research benchmarks, collecting no new data and involving no human subjects. RF20 includes surveillance-adjacent detection domains, so the usual dual-use risk of detection systems applies. Our intervention improves accuracy without fine-tuning but does nothing about the biases these models already carry; a more accurate answer is not a fairer one. We report where it fails, on video and POPE, so that it is not applied where it does not help.

REFERENCES

Manoj Acharya, Kushal Kafle, and Christopher Kanan. TallyQA: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Anonymous. Title withheld to preserve anonymity, 2026. Prior workshop paper by the present authors. The full paper is included in the supplementary material for reviewers.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, 2024.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

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

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. DSPy: Compiling declarative language model calls into self-improving pipelines. In *International Conference on Learning Representations (ICLR)*, 2024.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. NaturalBench: Evaluating vision-language models on natural adversarial samples. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

João Matos, Shan Chen, Siena Placino, Yingya Li, Juan Carlos Climent Pardo, Daphna Idan, Takeshi Tohyama, David Restrepo, Luis F Nakayama, Jose M M Pascual-Leone, Guergana Savova, Hugo Aerts, Leo A Celi, A Ian Wong, Danielle S Bitterman, and Jack Gallifant. WorldMedQA-V: A multilingual, multimodal medical examination dataset for multimodal language models evaluation. *arXiv preprint arXiv:2410.12722*, 2024.

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

nostalgebraist. Interpreting GPT: The logit lens. LessWrong, 2020. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`.

Peter Robicheaux, Matvei Popov, Anish Madan, Isaac Robinson, Joseph Nelson, Deva Ramanan, and Neehar Peri. Roboflow100-VL: A multi-domain object detection benchmark for vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025.

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)*, 2022.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025.

xAI. RealWorldQA. Benchmark released with the Grok-1.5 Vision preview, 2024. `https://x.ai/news/grok-1.5v`.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM International Conference on Multimedia*, 2017.

Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, Jian-Guang Lou, and Shuai Ma. Re-reading improves reasoning in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Appendix

A RF20 PRESENCE TASK: THE YES/NO VARIANT

Besides the detection task of Section 6.2, RF20's images support a yes/no presence question ("is a <class> in the image?"), 24,449 questions over 20 datasets in 7 categories. This variant has the same answer format as NaturalBench, POPE and Winoground, and it behaves like them: question-first hurts on 7 of 7 categories, by 11.36 points on average. On Lab Imaging it takes accuracy from 86.29 to 38.94, a 47.35 point loss, and both echo orderings recover the category past the question-last baseline. The contrast with Table 3, where the same images under a detection objective show the opposite sign, is the cleanest evidence we have that the paradox tracks the answer format rather than the imagery.

Table 5: Asked as a yes/no question rather than as detection, the same RF20 images give the opposite result: question-first hurts on 7 of 7 categories. Accuracy by category, Qwen3-VL-8B, 24,449 questions. Bold marks the best ordering in each row. Compare Table 3, where the sign flips.

Category	N	SIT	STI	STIT	SITIT	Δ STI	Δ SITIT
Document	10,200	80.94	76.20	80.30	**84.19**	−4.74	+3.25
Misc	4,114	85.71	82.50	**85.90**	85.05	−3.21	−0.66
Industrial	2,958	**82.45**	68.19	81.54	82.18	−14.26	−0.27
Lab Imaging	2,822	86.29	38.94	**91.71**	90.93	−47.35	+4.64
Sport	2,654	49.06	48.61	42.99	**51.96**	−0.45	+2.90
Flora/Fauna	1,592	86.56	82.54	**87.44**	86.75	−4.02	+0.19
Aerial	109	**94.50**	88.99	89.91	**94.50**	−5.51	+0.00
Mean over 7						−11.36	+1.44

B ATTENTION KNOCKOUT IN FULL

Qwen3-VL-8B, 250 NaturalBench pairs, query side restricted to the answer position. `ko_question` severs the attention edges from the answer position to the question span, `ko_image` to the image span, and `ko_random` to an equal-length random span of pre-image text as a span-size control.

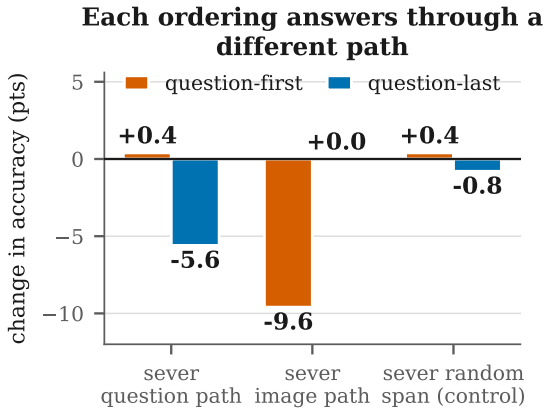


Figure 3: Each ordering answers through a different path, and cutting the other one costs it nothing. Bars are the change in accuracy against the unmodified run (57.6 for question-first, 58.8 for question-last), so downward is damage. Severing answer-to-question hurts only question-last; severing answer-to-image hurts only question-first; severing a random pre-image span of equal length, the control for simply removing attention mass, moves both by at most 0.8 points.

The paired bootstrap on the question-last question knockout gives $\Delta = 0.056$ with a 95% CI of $[+0.0199, +0.096]$, so the interval excludes zero.

C CROSS-MODEL LADDERS IN FULL

Table 6: Four of five model families lose accuracy when the question moves before the image, so the paradox is not one model's quirk. NaturalBench, 1,900 groups. Bold marks the best ordering in each row; n/a marks combinations that were not run. LLaVA-1.5's question-first group accuracy is exactly zero, meaning it never gets all four items of any group right.

Model	Metric	SIT	STI	STIT	SITIT	Δ (STI−SIT)
Qwen3-VL-8B	pair acc.	79.21	75.29	78.95	**80.36**	−3.92
	group acc.	35.05	26.95	35.00	**37.37**	−8.10
Qwen2.5-VL-7B	pair acc.	**76.05**	64.95	71.83	75.71	−11.10
	group acc.	**27.74**	10.16	19.11	27.63	−17.58
InternVL3-8B	pair acc.	79.63	75.61	79.76	**80.80**	−4.02
	group acc.	35.53	27.63	36.47	**39.32**	−7.90
LLaVA-1.5	pair acc.	**66.87**	50.26	66.29	n/a	−16.61
	group acc.	**12.32**	0.00	12.05	n/a	−12.32
Gemma-3-27B	pair acc.	73.34	73.21	**74.64**	74.61	−0.13
	group acc.	22.63	23.21	25.26	**25.53**	+0.58

Table 7: The repair outlives the diagnosis: neither Gemma model shows the paradox, yet image echoing is still their best ordering on 7 of 8 model-benchmark pairs. The four benchmarks run for all three models. Bold marks the best ordering in each row; the mean rows average the Δ columns above them.

			Ordering				Δ vs SIT	
Model	Benchmark	N	SIT	STI	STIT	SITIT	STI	SITIT
Gemma-3-27B	BLINK	793	44.51	43.51	42.75	**44.89**	−1.01	+0.38
	MMVP	300	72.67	72.33	72.00	**74.67**	−0.33	+2.00
	TallyQA	38,589	63.00	67.60	68.93	**69.97**	+4.60	+6.96
	VQAv2	214,354	**64.77**	62.07	63.04	62.07	−2.70	−2.70
	Mean						+0.14	+1.66
Gemma-4-31B	BLINK	793	54.48	56.12	56.12	**56.49**	+1.64	+2.02
	MMVP	300	84.67	86.00	86.67	**87.33**	+1.33	+2.67
	TallyQA	2,000	80.65	80.50	81.20	**82.60**	−0.15	+1.95
	VQAv2	3,000	73.50	74.66	75.65	**76.00**	+1.16	+2.50
	Mean						+1.00	+2.28

D SIGNIFICANCE TESTS IN FULL

Paired on group index (NaturalBench) or example index (Winoground), exact two-sided McNemar plus a paired bootstrap 95% CI over 5,000 resamples, seed 0. Accuracies are group accuracy.

E READ-OUT PROBE AND DECISION LAYER IN FULL

Qwen3-VL-8B, 36 layers. The disagreement set is the 150 NaturalBench pairs where STI is wrong and SIT is right; the decision-layer analysis uses 300 pairs. a_q and a_{img} are attention mass from the answer position onto the question and image spans, averaged over heads.

$P(\text{correct})$ over the {Yes, No} decision on the 300-pair disagreement set. All four orderings sit at ≈ 0.50 through layer 25, so only layers 26 to 35 are shown.

Table 8: The paradox is significant wherever it appears, and its absence in Gemma is also clear rather than merely underpowered. Consecutive rungs of the ladder plus the headline comparisons, paired on group or example index. Δ is B minus A with a paired bootstrap 95% interval; p is an exact two-sided McNemar test. `STI` versus `SIT` is highly significant for Qwen3-VL-8B on both benchmarks and absent for Gemma-3-27B.

Benchmark	Model	Comparison	acc A	acc B	Δ (95% CI)	McNemar p
NaturalBench	Qwen3-VL-8B	`STI` vs `SIT`	0.269	0.351	$+0.081$ $[+0.058, +0.103]$	1.3×10^{-12}
		`SIT` vs `STIT`	0.351	0.350	-0.001 $[-0.022, +0.020]$	1.00
		`STIT` vs `SITIT`	0.350	0.374	$+0.024$ $[+0.005, +0.043]$	1.6×10^{-2}
		`SITIT` vs `SITIT_rev`	0.374	0.374	$+0.000$ $[-0.010, +0.011]$	1.00
	Gemma-3-27B	`STI` vs `SIT`	0.232	0.226	-0.006 $[-0.026, +0.015]$	0.62
		`SIT` vs `STIT`	0.226	0.253	$+0.026$ $[+0.006, +0.046]$	1.4×10^{-2}
		`STIT` vs `SITIT`	0.253	0.255	$+0.003$ $[-0.015, +0.021]$	0.82
Winoground	Qwen3-VL-8B	`STI` vs `SIT`	0.223	0.318	$+0.095$ $[+0.050, +0.140]$	7.7×10^{-5}
		`SIT` vs `STIT`	0.318	0.375	$+0.057$ $[+0.015, +0.102]$	1.2×10^{-2}
		`SIT` vs `SITIT`	0.318	0.403	$+0.085$ $[+0.045, +0.125]$	2.4×10^{-5}
		`STI` vs `SITIT_rev`	0.223	0.410	$+0.187$ $[+0.140, +0.235]$	1.2×10^{-13}
	Gemma-3-27B	`STI` vs `SIT`	0.212	0.207	-0.005 $[-0.052, +0.043]$	0.92
		`SIT` vs `STIT`	0.207	0.253	$+0.045$ $[+0.000, +0.087]$	6.3×10^{-2}
		`SIT` vs `SITIT`	0.207	0.318	$+0.110$ $[+0.070, +0.152]$	8.1×10^{-7}

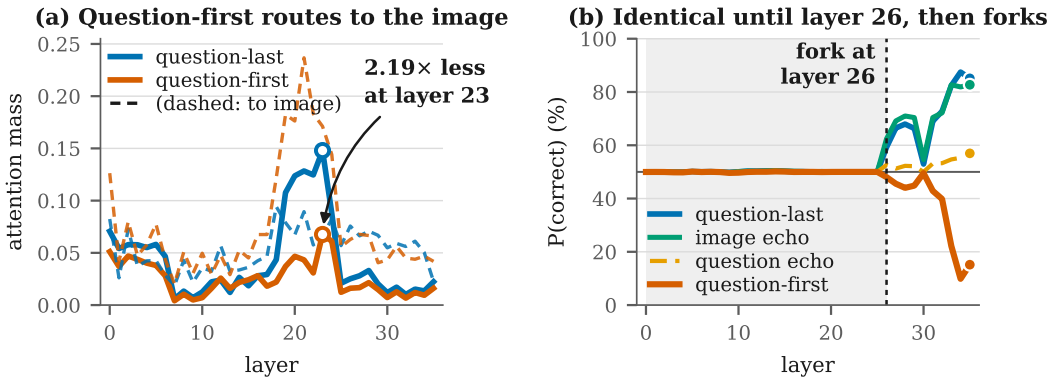


Figure 4: Question-first answers by reading the image instead of the question, and the two orderings only part company in the last quarter of the network. Measured on the 150 NaturalBench pairs where question-first is wrong and question-last is right. **(a)** Attention mass leaving the answer position, to the question (solid lines) and to the image (dashed): at layer 23, where question attention peaks for both, question-first sends $2.19\times$ less of it to the question. **(b)** Probability of the correct answer at each layer. Both baselines sit at chance through layer 25 (shaded), fork at layer 26 and never re-cross; both echo orderings track question-last, not question-first.

Layer	26	27	28	29	30	31	32	33	34	35
`STI`	0.480	0.456	0.440	0.449	0.495	0.429	0.398	0.222	0.099	0.152
`SIT`	0.593	0.665	0.680	0.663	0.529	0.691	0.728	0.824	0.875	0.851
`SITIT`	0.624	0.691	0.710	0.704	0.546	0.704	0.722	0.827	0.819	0.827
`STIT`	0.526	0.513	0.523	0.521	0.500	0.532	0.533	0.547	0.553	0.570

F PER-PATCH COSINE TO AN IMAGE-ONLY FORWARD PASS

Layer 18, images resized to 448×448, $14 \times 14 = 196$ patches, Qwen3-VL-8B. Mean per-patch cosine between each ordering's image-token hidden states and a forward pass on the image alone.

Sanity checks: $\cos(\text{base}, \texttt{IST}) = 0.999997$ and $\cos(\texttt{STI}, \texttt{STIT}) = 0.999992$ at $n = 7{,}600$.

Table 9: Question-first reads the image where question-last reads the question, and the ratio between the two inverts. Read-out probe on the disagreement set; a_q and a_{img} are attention mass from the answer position onto the question and image spans, and the layer of each peak is in brackets.

Ordering	probe acc	$\max a_q$ (layer)	$\max a_{img}$ (layer)	a_q:a_{img} @L24	final P(gt)
STI	0.073	0.0676 (23)	0.2367 (21)	0.424	0.1346
SIT	0.887	0.1479 (23)	0.0942 (18)	1.065	0.8485
STIT	0.600	0.1437 (23)	0.1448 (21)	0.986	0.5912
SITIT	0.847	0.1286 (23)	0.0647 (26)	1.544	0.8374
SITIT_rev	0.867	0.1218 (23)	0.0617 (26)	1.624	0.8503

Table 10: The question rewrites the image tokens themselves, and a later copy of it provably cannot. Mean cosine between each ordering's layer-18 image tokens and a forward pass on the image alone, so 1.0 means untouched. IST is that baseline by construction; STIT is identical to STI to five decimals because causal masking prevents a later question copy from reaching earlier patches.

Ordering	n	mean cos	median	min
IST	7,600	0.999998	1.000000	0.999919
SIT	30	0.911017	0.921399	0.722478
STI	7,600	0.862723	0.873233	0.614870
STIT	7,600	0.862727	0.873237	0.614835

G Exact token cost per ordering

Table 11: Repeating the question is nearly free; repeating the image is not, but its price falls with echo area because vision tokens scale with pixels. Exact counts per query, mean over 20 images spanning RF20's constituent datasets, Qwen3-VL-8B. Δ is against the STI baseline, which is why that row reads n/a. STI and SIT contain the same tokens in a different order, so their totals are identical and only STI is listed.

Prompt	Total	Image	Text	Δ total	Δ image	Δ text
STI (baseline)	3385.8	3335.8	50.0	n/a	n/a	n/a
STIT (question echo)	3398.8	3335.8	63.0	+13.0	+0.0	+13.0
SITIT, echo 1	6737.5	6671.5	66.0	+3351.8	+3335.8	+16.0
SITIT, echo $\frac{1}{2}$	4260.7	4194.7	66.0	+874.9	+858.9	+16.0
SITIT, echo $\frac{1}{4}$	3637.6	3571.6	66.0	+251.8	+235.8	+16.0
SITIT, echo $\frac{1}{8}$	3492.6	3426.6	66.0	+106.8	+90.8	+16.0

H GEPA configuration and training cost

GEPA (Agrawal et al., 2026) optimizes the system message only. The task model is Qwen3-VL-8B; the reflection model is Gemma-3-27B on a separate GPU, since reusing the task model as its own reflector produced proposals that never passed GEPA's strict-improvement acceptance test. Training and validation examples are disjoint from the held-out evaluation set.

I Logit-lens word divergence between STI and IST

To ask whether question-first and image-first prompting diverge in the words they are internally predicting, we decode the top-1 token at every (layer, generation step) position under both orderings on 100 samples, discard positions whose two words are identical or WordNet synonyms, and discard any position where either side is not a dictionary word, since intermediate layers frequently project onto subword fragments.

Table 12: A prompt optimizer improves only the ordering our diagnosis says has headroom, and never catches the ordering it is competing with. GEPA run under both orderings with the same budget, Qwen3-VL-8B, scored on held-out evaluation sets of size N. Each ordering gives the seed prompt's accuracy and the optimized prompt's, with the change in parentheses. This is the table Section 7.3 reads its numbers from. "Gap closed" is the fraction of the STI-to-SIT accuracy gap the optimized STI prompt recovers.

		Question-first (STI)			Question-last (SIT)	
Benchmark	N	Seed	GEPA	Gap closed	Seed	GEPA
POPE	8,800	86.41	87.25 (+0.84)	45%	88.27	88.27 (+0.00)
VQAv2	10,000	86.11	88.13 (+2.02)	83%	88.55	88.55 (+0.00)
NaturalBench	7,000	74.74	75.79 (+1.04)	26%	78.80	78.39 (−0.41)

Table 13: GEPA's cost is dominated by a one-time search, not by inference. Search budget under the STI ordering; the recurring per-query cost is the prompt-length delta in Table 14.

Benchmark	train	val	metric calls	training tokens	wall clock (s)
POPE	100	60	400	157,725	124.3
VQAv2	150	80	1,008	396,253	232.4
NaturalBench	250	120	1,540	982,343	528.7

Genuine word changes are rare and concentrated: 8 across 100 samples, of which 6 occur at layer 16, one at layer 8 and one at layer 20, with none at layers 0, 4, 12, 24 or 27. Divergence at the level of decoded output words is therefore sparse and mid-network, consistent with the read-out account in Section 4.3, where the two orderings are indistinguishable at the decision level until late and differ mainly in routing rather than in vocabulary.

J Reproduction notes

Every accuracy in this paper is read from a per-run result file that records the model, the ordering, the system prompt, the decoding budget and the sample count. Two details are worth flagging for anyone re-running the analysis.

The attention knockout must restrict the query side to the answer position. With the query side widened to every position after the question, the question-side half of the dissociation does not reproduce (severing answer-to-question raises SIT by 2.8 points, CI $[-0.084, +0.028]$), and the random-span control is itself worth 10 points on STI, so that configuration is uninformative rather than contradictory.

The Gemma-3-27B RF20 run is degenerate and we exclude it from all RF20 claims. It answers "yes" to roughly 96% of items under every ordering (yes_ratio 0.955 to 0.964, recall 0.979 to 0.991), so its accuracy sits at the positive base rate and the spread across orderings is 0.003.

Table 14: The optimized prompts are cheap to run, and under question-last the optimizer had nothing to add. Prompt length of GEPA's system messages in tokens per query. A zero delta means the optimizer returned the seed prompt unchanged, which is what happened under `SIT` on POPE and VQAv2.

Benchmark	Order	seed	optimized	Δ
POPE	`STI`	315	341	$+26$
	`SIT`	316	316	$+0$
VQAv2	`STI`	313	345	$+32$
	`SIT`	314	314	$+0$
NaturalBench	`STI`	216	287	$+71$
	`SIT`	217	381	$+164$