WHAT KEEPS VISION-LANGUAGE MODELS LOOKING AT THE IMAGE?

Anonymous authors
Paper under double-blind review

ABSTRACT

When do vision-language models need direct access to the image while generating an answer? We study image dependence during answer generation by examining how the visual information needed for the current question becomes available in context. We intervene on direct image access while retaining previously computed states. Across real-image and synthetic tasks, we show that, depending on the generation process, direct access can continue to support accuracy after question processing. Supplying the required attributes as text in the context weakens this dependence. On synthetic tasks, we also examine how dependence changes as the model itself states the required attributes. Before attribute expression, severing access reduces accuracy, and replacing image-side states shifts answers toward the counterfactual content. After sufficient expression, both interventions have smaller effects. Even with an identical generated prefix, dependence differs according to whether image access was available during question processing. Thus, both the visible text and the preceding image access matter. Several of these patterns hold across model families, including Qwen2.5-VL-32B and InternVL3-14B. These findings offer a view of image dependence in terms of the information needed for the current question and the history of image access, beyond generation position alone. This perspective provides a basis for deciding when to reduce visual access during an answer and which visual information to retain for subsequent questions.

1 INTRODUCTION

Vision-language models (VLMs) answer questions about images, either directly or through intermediate reasoning. Chain-of-thought (CoT) methods generate intermediate reasoning before the final answer (Wei et al., 2022; Zhang et al., 2024). Consider a follow-up question. A model correctly answers "lower left" when asked which quadrant contains the only yellow star; when asked next how many green objects that quadrant contains, it needs colors that the first answer never stated. A correct earlier answer therefore need not contain what the next question needs. It remains unclear whether the model must read the image again or can rely on what it has already processed and written. The question bears on when image access can be reduced: efficiency methods reduce visual computation by selecting fewer visual tokens for input to the language model (Yang et al., 2025), pruning visual tokens after an early layer (Chen et al., 2024), freezing visual-token updates in selected layers (Yuan et al., 2025), or compressing and evicting visual key–value (KV) cache entries (Tu et al., 2025; Wan et al., 2024; Li et al., 2025). These approaches motivate the complementary question of how sensitivity to continued direct image access changes as answer generation proceeds. We ask when direct image access is needed during answer generation, focusing on how the visual information required by the current question becomes available in context. We show that sensitivity to blocking this access varies with the generation process, the attributes available in text, and the preceding image access.

Prior work has traced how visual information reaches the answer: it flows through direct attention to image tokens and through text-mediated pathways via question tokens (Kaduri et al., 2025; Zhang et al., 2025b; Yin et al., 2025; Wu et al., 2026), and under some interventions the text-mediated pathway can serve as a fallback when direct image access is disrupted (Salazar et al., 2026). Across dialogue turns, assistant-text states substitute for visual states for facts that were verbalized, less so for facts that were not (Chen et al., 2026). During long reasoning, attention to the image fades and visual contributions concentrate early (Sun et al., 2025; Chu et al., 2025; Sammani et al., 2026), and

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.

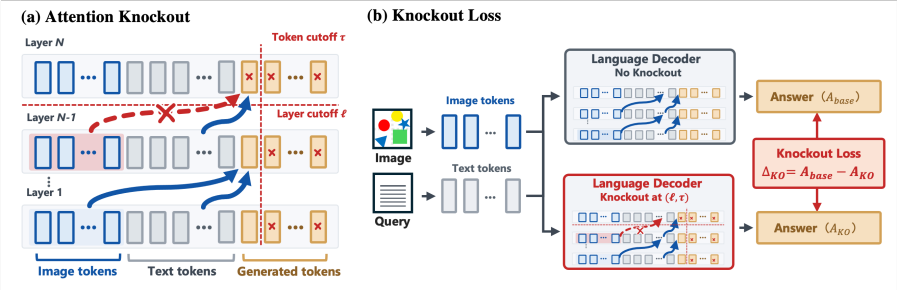


Figure 1: Attention-knockout intervention and measurement of knockout loss. (a) Answer-side knockout masks attention to image-token keys when either the layer or the token condition is met ($k \geq \ell$ or $t \geq \tau$). Cached keys and values and untargeted attention edges are retained, while the remaining attention weights are renormalized. Question-side knockout, not shown, targets the current question rows instead. **(b)** Knockout loss $\Delta_{\mathrm{KO}} = A_{\mathrm{base}} - A_{\mathrm{KO}}$ is the accuracy difference between a matched baseline and the knockout condition, with question-time access held fixed. Layers are numbered from 1 in the schematic and indexed from 0 in the formal definition.

re-completing a reasoning chain truncated midway without the image costs little accuracy (Sun et al., 2025). Whether attention weights alone can establish what a model uses has long been debated (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019), and a fall in attention to the image does not imply a fall in the accuracy cost of severing that attention. We therefore measure intervention sensitivity directly. These findings leave open how dependence on direct image access changes as information needed for the current question becomes available during a response. We address this with three complementary comparisons: question-access contrasts at a fixed visible prefix, knockout aligned to each response's attribute-expression stage, and a directional KV swap before and after that stage (Section 4).

We measure dependence with attention knockout (Geva et al., 2023; Kaduri et al., 2025). As Figure 1 illustrates, attention from chosen positions to image tokens is severed from a chosen layer and generation step onward, while previously computed states and the remaining attention connections are retained. The knockout loss, the accuracy drop against a baseline that differs only in the knockout, quantifies this sensitivity. Within each comparison, we hold the image set and scoring fixed and vary an input condition, access condition, or intervention stage. Synthetic counting tasks let us control which attributes are supplied in text and when image access is available. On real images, GQA tests the description and question-time access comparisons, and TallyQA tests the response-format comparison at answer onset. Fixed-prefix replay, stage-aligned knockout, and KV swaps on synthetic scenes further examine the roles of access history and generated content. Figure 2 maps the four comparisons reported in Section 3: response format, description type, question-time access, and attribute expression.

The response-format comparison shows that direct image access can continue to support accuracy after question processing. Direct answers are robust to answer-side knockout when question-time image access is available, whereas CoT responses lose accuracy. In an immediate-answer control, the CoT instruction alone does not reproduce this loss. Supplying the required attributes in text makes generation robust to knockout, whereas a correct history that omits them or a description with the attributes withheld does not. On synthetic follow-up counting, direct image access remains consequential before attribute expression even when image-conditioned question and prefix states are retained. Knockout placed before the model writes the required attributes reduces accuracy, and replacing the image KV states at that point shifts answers toward the donor's count. Both interventions have smaller effects after sufficient expression. Even when the initial generated prefix is held fixed, later knockout sensitivity still differs between question-time access conditions. The visible text and the preceding image access therefore both matter.

The main experiments use Qwen2.5-VL-32B (Bai et al., 2025; Qwen Team, 2025) on synthetic and real-image tasks. Additional experiments with InternVL3-14B (Zhu et al., 2025) support several patterns on synthetic tasks and also show differences across tasks and generation stages. Together, these findings relate image dependence to the information needed for the current question and the history of image access, beyond generation position alone. This perspective informs decisions on when to reduce visual access during an answer and which information to retain for subsequent questions.

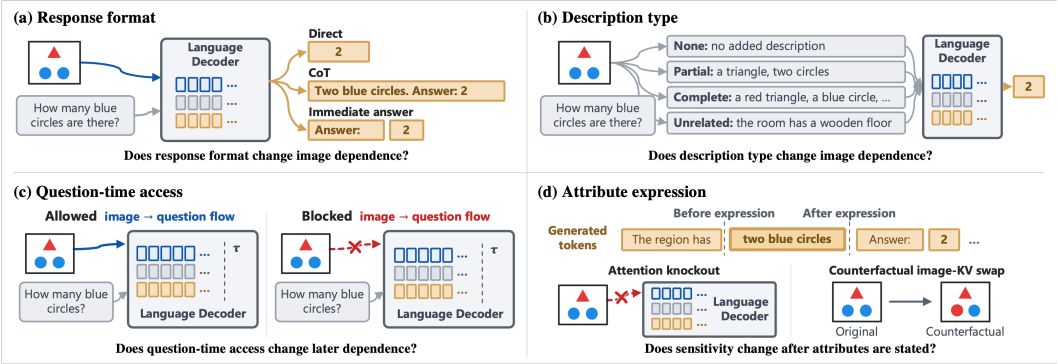


Figure 2: Experimental design of the four image-dependence comparisons. (a) Response format: Direct, CoT, and an immediate-answer control that retains the CoT instruction. **(b) Description type:** whether supplied text contains the attributes needed for the answer. **(c) Question-time access:** whether the current question can attend to image tokens. **(d) Attribute expression:** knockout and image-KV swap before and after sufficient attribute expression in the generated response.

2 Experimental Framework

An autoregressive VLM encodes the image as *image tokens*, processes the *question tokens* together with any dialogue history and attribute description, and generates the answer token by token. We distinguish *question processing* from *answer generation*. *Direct image access* is attention from question or answer positions to image tokens. We operationalize *image dependence* as *knockout sensitivity*, defined as the accuracy change caused by blocking specified attention edges to image tokens, relative to a baseline with the same input and question-access condition. This measure reflects sensitivity to removing direct access rather than how often the model accesses the image during unperturbed generation. The image-KV swap in Section 3.4 separately measures how answers respond to changes in image-side content. *Intervention sensitivity* refers to the effect of either intervention.

2.1 Attention Knockout

We use attention knockout to block selected attention edges (Geva et al., 2023), following its application to visual information flow in VLMs (Kaduri et al., 2025). The mask adds $-\infty$ to the pre-softmax logits from the selected positions to image tokens in every head and renormalizes the remaining weights; previously computed key–value states and text-to-text attention are retained (Appendix A.1).

Question-side and answer-side knockout are controlled independently. Question-time image access is *Allowed* when the rows of the current question can attend to image tokens and *Blocked* when these edges are masked at every layer. A preceding turn, image included, is encoded with full access and cached, so Blocked preserves the image information already in those states. On the answer side, knockout is active at layer k and generation step t when $k \geq \ell$ or $t \geq \tau$, where τ counts the answer tokens generated before blocking and $\ell = L$ imposes no layer restriction. Because a layer cutoff $\ell < L$ acts from the first answer token, statements about late generation rest on $\ell = L$.

2.2 Knockout Loss

Let c denote an experimental condition (task, model, inputs, response format, and question-time access), $A_{\text{base}}(c)$ its accuracy without answer-side knockout, and $A_{\text{KO}}(c; \ell, \tau)$ its accuracy with knockout at (ℓ, τ). The knockout loss, in percentage points (pp), is

$$\Delta_{\text{KO}}(c; \ell, \tau) = A_{\text{base}}(c) - A_{\text{KO}}(c; \ell, \tau). \tag{1}$$

When question-time access is Blocked, the baseline blocks it too, so the loss isolates what answer-side knockout adds. For the question-access comparison, we define $D = \Delta_{\text{KO}}(\text{Blocked}) - \Delta_{\text{KO}}(\text{Allowed})$ with task, model, history, and cutoff fixed. A positive D indicates a larger knockout loss under Blocked access. We interpret this contrast alongside each arm's baseline and knockout

accuracy. A positive loss shows that direct image access contributed to accuracy under that condition, and a negative loss that severing it helped on those examples. A zero loss shows that severing this pathway at this position did not change aggregate accuracy on the evaluated examples. Because individual answers can still change in either direction and other pathways may compensate (Salazar et al., 2026), a zero loss does not show that the image information is absent or unused.

We report pointwise 95% confidence intervals (CIs) from a percentile bootstrap over images with pairing preserved (Efron, 1982). Every contrast is computed within one run because greedy CoT outputs can differ between runs. These intervals quantify image-sampling uncertainty conditional on the run, not variability across runs. Image sets, their overlap, execution environments, and pairing units are documented in Appendix A.3. Conclusions rest on cutoffs at which knockout is applied to every response. In the figures, bands and error bars show the CIs, and open markers indicate cutoffs where knockout is applied to only a subset of responses.

2.3 Tasks and Information Conditions

Synthetic tasks. Synthetic scenes come from a seeded generator, so every object's color, shape, and location is known (Appendix A.3). As in CLEVR (Johnson et al., 2017), the scenes let us control the attributes an answer requires and whether they are available in text. Single-turn counting asks for the number of objects of a given color among 6–7 objects, with five answer options. Follow-up counting first asks which quadrant contains a unique anchor object, then asks how many objects of a specified color that quadrant contains. The first reply is teacher-forced to control the dialogue history. The *required attributes* are the colors the current question needs, and text that supplies them is *answer-sufficient*. Description conditions vary the information supplied alongside the question and are *None*, *Complete* (every object's attributes, without the count), *Partial* (the same listing with the colors withheld), or *Unrelated* (token-matched text about something else). History conditions vary the information in the preceding reply and are *Label-only* (the correct quadrant), *Attribute-stating* (the quadrant and its objects by color and shape), or *Shape-only* (the same at equal length without the colors). For generated prefixes, *Sufficient* denotes the annotation category defined in Appendix A.5. Table A.4 gives the full matrix.

Real-image tasks. GQA tests the description and question-time access comparisons on real images. From GQA (Hudson & Manning, 2019), whose images come from Visual Genome (Krishna et al., 2017), we select relation-color questions that identify an object through a relation and ask for its color. We convert each question to a five-choice format for our evaluation. Complete descriptions are scene-graph listings that include the queried color and serve as an answer-revealing positive control. TallyQA tests the response-format comparison. We use its simple questions (Acharya et al., 2019), which ask how many instances of a noun are present, and balance the selected answers across 1–4. Both tasks introduce the image in a preceding turn so that question-time access can be controlled as in the follow-up task (Appendix A.3).

3 What Changes Image Dependence?

Unless otherwise stated, results use Qwen2.5-VL-32B-Instruct with greedy decoding, and Appendix B reports the corresponding analyses for InternVL3-14B; scoring is described in Appendix A.3. The four comparisons in Figure 2 examine how image dependence varies with response format, description type, question-time image access, and attribute expression during generation.

3.1 Response Format Changes Image Dependence

We first test whether image access during question processing is sufficient for subsequent answer generation across response formats. *Direct* answers give only the final answer, whereas *CoT* answers follow a zero-shot instruction to reason step by step (Kojima et al., 2022) in at most 120 words before giving the answer. An *immediate-answer control* keeps the CoT instruction verbatim but opens the answer at once; it shows no observed loss on either task (Appendix C.1), so the instruction alone does not explain the CoT sensitivity. In the matched answer-onset controls in Appendix C.1, all three formats use the same images and answer-side knockout within each task.

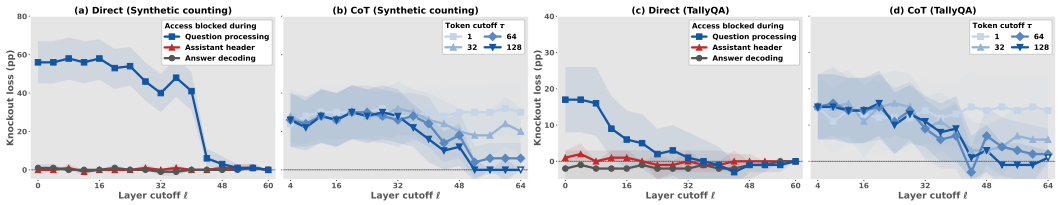


Figure 3: Comparison of knockout loss across response formats and tasks. Direct answers are sensitive to question-side knockout in earlier layers, whereas CoT answers also show answer-side loss. **(a, b)** Synthetic counting; **(c, d)** TallyQA. Direct panels separately block image access from the question, assistant-header, or decode rows at layers $k \geq \ell$. CoT panels combine answer-side layer and token cutoffs ($k \geq \ell$ or $t \geq \tau$), with question-time access Allowed.

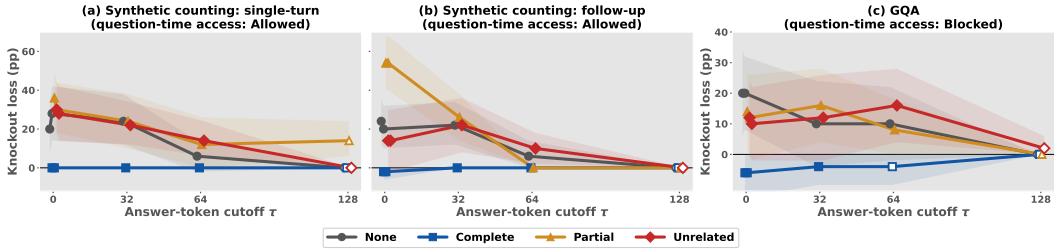


Figure 4: Comparison of knockout loss across description conditions. Complete descriptions yield little observed loss across the measured cutoffs; descriptions omitting the required attributes do not consistently provide the same robustness. **(a)** Single-turn synthetic counting and **(b)** synthetic follow-up counting with the Label-only history, both with question-time access Allowed. **(c)** GQA with question-time access Blocked. All panels use CoT and token cutoffs without a layer restriction.

On synthetic counting, Figure 3(a) shows the knockout loss of Direct answers when image access is blocked separately from the question, assistant-header, or decode rows. Blocking question rows at layers $k \geq \ell$ reduces accuracy for cutoffs below about 44, with no observed loss at higher cutoffs. Blocking assistant-header or decode rows produces no observed loss at any tested layer cutoff. Thus, Direct answers depend on image access during question processing but are robust to answer-side knockout when question-time access is Allowed. CoT answers remain sensitive during generation. With no layer restriction, the loss at $\tau = 32$ is about two thirds of that at $\tau = 1$ and becomes small at $\tau = 64$ (Figure 3(b)). Layer-only knockout likewise produces losses comparable to blocking all layers when the cutoff is $\ell \leq 40$. The same ordering holds on an independent count-balanced set of 100 images with decision rules fixed before the run (Appendix C.1). The Direct profile is consistent with the answer being formed from states computed during question processing (Salazar et al., 2026), although it does not show where the answer is stored.

The same response-format contrast holds on real images in TallyQA (Figure 3(c, d)). For Direct answers, blocking question rows reduces accuracy at layer cutoffs below about 24, while blocking header or decode rows produces no observed loss. CoT answers lose accuracy under answer-side knockout. The CoT loss has a layer profile similar to that on synthetic counting, but with no layer restriction it fades earlier, by $\tau = 32$. Without knockout, CoT is not more accurate than Direct on this task, consistent with reports that CoT does not always improve visual reasoning (Jin et al., 2026; Kancheti et al., 2026). Under knockout, CoT accuracy falls below the Direct baseline on TallyQA, but only to that baseline on the fresh synthetic set (Appendix C.1). Thus, direct image access can support CoT accuracy even when CoT offers no advantage over Direct.

Across synthetic counting and TallyQA, direct image access continues to support accuracy during CoT generation, while Direct responses are robust to answer-side knockout. The comparison therefore establishes a difference between generation protocols but does not isolate reasoning itself, since the protocols also differ in generated content, length, and computation (Pfau et al., 2024). We next examine whether supplying the required visual attributes in text reduces this dependence.

3.2 Description Type Changes Image Dependence

We test whether supplying the attributes needed for the current question reduces dependence on direct image access, and whether a correct prior answer provides the same robustness. We compare the four description types in single-turn and follow-up counting and the three history types in follow-up counting, with question-time access Allowed unless stated.

As Figure 4(a) shows for single-turn counting, supplying the required colors in a Complete description leaves aggregate accuracy unchanged under answer-side knockout at every measured cutoff. Without a description, the loss appears from answer onset and persists across later cutoffs. Complete serves as a positive control for answering from supplied visual attributes, as in systems that respond to visual questions from captions or tags (Yang et al., 2022; Zeng et al., 2022). Partial descriptions withhold the colors, while Unrelated descriptions provide token-matched text without the required information. Neither reproduces the robustness of Complete, and both show losses comparable to or larger than those under None through $\tau = 64$. Thus, the content of the description, beyond its added length, supports robustness in this setting.

Complete descriptions also remove the loss in follow-up counting, where the correct Label-only history omits the colors needed for the current question (Figure 4(b)). Under Partial descriptions, the loss at early cutoffs exceeds that of None, from a higher baseline, and vanishes by $\tau = 64$. Unrelated text yields smaller loss estimates than None at the first two cutoffs and the same loss at $\tau = 32$, with overlapping intervals. Thus, Complete provides robustness in both synthetic settings, whereas text without the required attributes can change the size of the loss without removing it.

The description comparison extends to real images in GQA relation-color questions (Figure 4(c)). With question-time access Blocked, answer-side knockout reduces accuracy from answer onset. With access Allowed, the observed loss is small at every measured cutoff (Appendix C.2). Complete descriptions remove the Blocked loss, whereas color-withholding Partial descriptions and unrelated text of matched length do not (Figure 4(c)). Because Complete includes the queried color, this result establishes robustness when answer-relevant information is supplied in text.

We then ask whether the same distinction holds for information supplied in the preceding reply. The history comparison in Figure 5 uses an independent 100-image set and separates correctness on the preceding question from sufficiency for the current one (InternVL3 in Appendix B). The Label-only history is correct, but the follow-up needs colors that it never states. The Attribute-stating history states them, and the Shape-only history has the same token count and structure without them. With the Attribute-stating history, knockout loss remains near zero in both question-time access conditions across the measured cutoffs, and the Blocked baseline rises to the level of the Allowed one. The Shape-only history does not reproduce this tolerance. Both arms lose accuracy under knockout, and the question-access difference remains. The Shape-only losses exceed the Label-only ones, but so do its baselines, so we read the result as a failure to reproduce the tolerance at matched length, not as harm from the added text.

In the main model, supplying the required attributes supports robustness to answer-side knockout across synthetic tasks and GQA. A correct prior answer or text that omits those attributes does not consistently provide the same robustness. InternVL3-14B shows the same distinction in single-turn counting and the history comparison, while its follow-up description conditions separate less clearly (Appendix B). These comparisons identify the relevance of the supplied information to the current question as a key factor in image dependence. We next hold the history fixed and examine how image access during question processing affects subsequent dependence.

3.3 Question-Time Access Changes Image Dependence

When the required attributes are not in the text, does image access while the question is processed change the later sensitivity to answer-side knockout? We hold the history fixed (Label-only, correct) and vary only question-time image access to test its effect on subsequent answer-side knockout sensitivity. Answer-side knockout is then applied at each τ, and each arm is compared with its own baseline, so the cost of blocking question-time access itself is separated from the later loss.

Free generation. With the Label-only history, answer-side knockout causes a larger loss under Blocked than Allowed at every measured cutoff on the independent synthetic follow-up set (Figure 5).

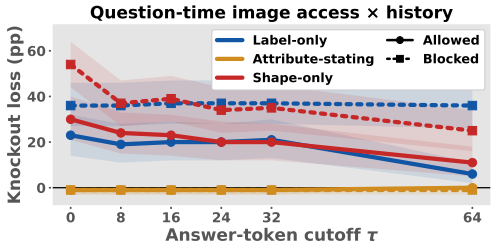


Figure 5: Knockout loss across history and question-time access. Attribute-stating histories yield near-zero knockout loss under both access conditions. Follow-up counting, CoT, independent set; each condition uses its own baseline.

Prefix source	Replay access	Accuracy (%)	Loss (pp)
Allowed	Allowed	87 → 81	6 [2, 11]
	Blocked	80 → 66	14 [8, 21]
	Difference: Blocked − Allowed		**8 [1, 15]**
Blocked	Allowed	83 → 67	16 [8, 25]
	Blocked	69 → 33	36 [26, 46]
	Difference: Blocked − Allowed		**20 [9, 31]**

Table 1: Fixed-prefix accuracy and knockout loss. The same 64-token prefix is replayed under Allowed and Blocked with states recomputed; both prefix sources retain a positive loss difference. Accuracy is baseline → knockout at $\tau = 64$; CoT, Label-only history.

The sustained Blocked loss accompanies a specific failure pattern: knocked-out responses answer "0" at early cutoffs on the profile set and mostly state that only the Turn 1 object is present on the independent set (Appendix C.3). The sustained Blocked loss therefore reflects a recurring failure of the continuation, rather than a graded measure of information retained during question processing.

Fixed-prefix replay. The access conditions also change the generated text: at the replay cutoff, some Allowed responses already state the answer-sufficient colors, whereas none of the Blocked responses do (Figure C.5). The smaller Allowed loss may therefore reflect differences in the visible prefix. To test whether this explains the contrast, we replay the same 64-token prefix under both access conditions, recompute the states, and apply knockout from token 65 (Appendix A.6). For either prefix source, knockout loss remains larger under Blocked than Allowed access (Table 1). With an Allowed-derived prefix, this difference persists while both knockout accuracies remain above the constant-answer reference; the Blocked-derived comparison includes a cell at that reference. Thus, differences in the initial visible text do not fully explain the question-access effect. The two prefix sources define separate comparisons, rather than a decomposition of the free-generation difference. Because prefix states are recomputed and subsequent text is generated freely, the experiment does not identify which retained states account for the difference.

In Qwen2.5-VL-32B, preceding image access affects subsequent knockout sensitivity even with the same visible prefix. In free generation, InternVL3-14B shows a question-access difference at answer onset but not at later cutoffs (Appendix B); the fixed-prefix result has not been established for this model.

3.4 Image Dependence Changes with Attribute Expression

The description comparison supplied the required attributes from outside. Does sensitivity also change when they appear in the model's own text during generation? A first indication does not depend on stage cutoffs: at token 64, Sufficient prefixes in the Allowed arm show no observed knockout loss, whereas Uninformative prefixes show loss in both access arms (Figure C.5).

Stage-aligned cutoffs. A fixed token cutoff falls at different stages of different responses. We therefore give each response its own cutoff, determined by a classifier whose rule was fixed after inspecting the 64-token prefixes with answer correctness hidden and before any cutoff run (Appendix A.5). *Before expression* places knockout at the prediction of the token at which the first non-anchor object color is written. *After expression* places it at the token after the prefix is first labeled Sufficient, meaning that the prefix states both required colors or gives a closed enumeration that implies the correct count. The two positions are content-defined stages separated by several descriptive steps, not the moments around one event; their absolute positions differ between arms, and a response enters a stage only if it reaches that position, so the stages cover different response subsets.

On the same images reaching both stages in the Allowed arm, knockout reduces accuracy before expression but produces no observed aggregate loss after expression (Figure 6, left). The Blocked arm and the all-eligible populations show the same qualitative pattern (Appendix C.4). Thus, direct access remains consequential before expression even though previously computed image-conditioned states are retained. The rule uses the ground-truth colors, an oracle annotation for analysis rather than

a deployment signal. It was developed on the 64-token prefixes of this same image set and has no independent annotator, and the uncertainty of the cutoff positions is not represented in the intervals (Appendix A.5).

Counterfactual image-KV swap. Knockout alone leaves open whether the before-expression loss reflects sensitivity to visual content or disruption caused by blocking access. We therefore replace the image-position KV states in every layer with those from a donor image at each stage, while retaining the recipient's text-position states. In the donor image, the target objects are recolored so that the correct count changes to one of the two alternative values. Generation continues from the recipient's prefix, a directional form of activation patching (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024) (Appendix A.7).

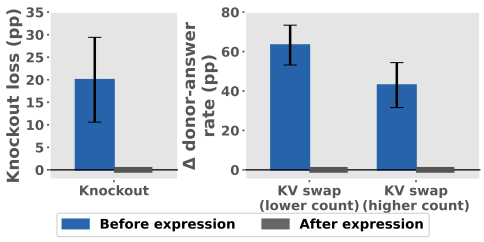


Figure 6: Intervention effects before and after expression. Both effects are smaller after expression. **Left:** knockout loss on the images reaching both stages. **Right:** change in donor-answer rate relative to no swap, for the donors with the smaller and larger alternative counts. Question-time access Allowed.

We measure the change in donor-answer rate relative to the no-swap continuation in the same run. Self swaps reproduce that continuation token for token on every checked item, while recoloring an object outside the target quadrant shows no clear accuracy loss (Appendix C.5). On the same images across stages, both donors increase donor-answer rates before expression but produce no observed increase after expression (Figure 6, right). These directional changes show that answers are sensitive to the inserted visual content before expression, arguing against a uniform-disruption account.

Together, the interventions show sensitivity to direct image access and image-side content before expression, with smaller effects after expression. This complements the supplied-description results: reduced sensitivity is also observed after the required information appears in the model's own text. The comparison does not locate the transition or establish whether verbalization itself, accumulated states, or answer commitment causes the change. These stage-aligned results are established on synthetic follow-up counting with Qwen2.5-VL-32B.

4 Related Work and Discussion

From visual pathways to dependence during generation. Three findings frame our question. Visual information can be transferred into question-token representations (Kaduri et al., 2025; Zhang et al., 2025b), and under some interventions models fall back on a text-mediated pathway when their usual pathway is ablated (Salazar et al., 2026). Across dialogue turns, facts stated in an earlier reply can support a later answer from assistant-text KV states, whereas unstated facts leave the answer dependent on the visual states (Chen et al., 2026). In long generations, the contribution of image tokens concentrates early in the response (Sammani et al., 2026). Our study connects these findings to how information needed for the current question becomes available during answer generation. Sammani et al. (2026) also remove image KV entries during generation at fixed token cutoffs and measure the resulting accuracy. Other nearby designs change the image (Shi et al., 2026), change question presentation (Abhinandan et al., 2026), re-complete a truncated reasoning chain without the image (Sun et al., 2025), or compare stated and unstated facts across dialogue turns (Chen et al., 2026). We add interventions aligned to attribute expression, directional image-KV replacement, and a comparison of question-time access under fixed visible prefixes.

Image dependence beyond visible text. The description and history comparisons show that supplying attributes needed for the current question supports robustness to knockout, whereas a correct prior answer that omits them does not. The stage comparisons extend this observation to the model's own generation: sensitivity to both knockout and image-KV replacement is smaller after expression. Fixed-prefix replay further shows that the same visible text can accompany different knockout sensitivity under different question-time access conditions. Together, these results relate image dependence to both the information expressed in context and the access history under which that context was computed.

Interpreting reduced dependence after expression. One explanation is that verbalizing relevant attributes makes visual evidence reusable for subsequent computation, as written intermediate steps can do in language models (Nye et al., 2021). This account is consistent with reports that VLMs rely on textual anchors and under-use available visual representations (Shahgir et al., 2026; Fu et al., 2025; Liu et al., 2025). However, attribute expression also accompanies changes in retained states and may follow an already committed answer; stated reasoning need not reveal what determined the answer (Turpin et al., 2023). Our results establish a difference in intervention sensitivity across the two stages, while leaving these mechanisms unresolved. Interventions on generated prefixes may help test the contribution of stated content (Lanham et al., 2023), with matched re-encoding controls to account for changes in prefix states.

Sufficiency is question-dependent. The correct quadrant label supports the first answer but omits the colors needed for the follow-up question. This illustrates two distinct decisions: whether direct image access can be reduced for the current answer, and which visual information should remain available for later questions. Our findings motivate access policies that consider the information needed by the current question and its availability in context. Reduced dependence for the current answer may justify restricting visual access temporarily, but it does not imply that the underlying visual information can be discarded permanently. Because a subsequent question may require attributes absent from the existing text (Chen et al., 2026), such policies also need to preserve or restore access to the image. Methods that select or prune visual tokens (Chen et al., 2024; Yang et al., 2025; Zhang et al., 2025a; Khaki et al., 2025), freeze visual-token updates in selected layers (Yuan et al., 2025), or compress and evict visual KV states (Tu et al., 2025; Wan et al., 2024; Li et al., 2025; Huang et al., 2025; Liu et al., 2026) face these decisions when reducing visual computation or memory. Other methods re-inject or re-attend to the image during generation (Sun et al., 2025; Chu et al., 2025; Zou et al., 2025), or explicitly control visual evidence use through dynamically triggered evidence insertion (Li & Ma, 2026), alternating looking and thinking steps coupled with compression (Cai et al., 2026), or perception tokens that the model generates to trigger additional perception (Yu et al., 2025). Our comparisons offer diagnostic conditions for evaluating whether such strategies preserve access to question-relevant information. They do not yet provide an online sufficiency detector or establish an accuracy–efficiency trade-off.

Scope and limitations. The response-format and information-condition comparisons provide evidence across synthetic and real-image tasks, with several patterns also observed in InternVL3-14B (Appendix B). The strength and timing of these patterns vary across models and tasks. The stage-aligned knockout and image-KV swap results are established only for Qwen2.5-VL-32B on synthetic follow-up counting. Common-set analyses align the images compared across stages but do not control for elapsed generation or difficulty (Appendix C.4, Appendix C.5). Finally, these interventions measure sensitivity under altered access conditions: preserved states, compensatory pathways (Salazar et al., 2026), and the out-of-distribution question-side mask limit what they imply about image use during unperturbed generation.

5 Conclusion

Image dependence during answer generation varies with how the information needed for the current question becomes available in context, rather than with generation position alone. The response-format comparison shows that direct image access can continue to support accuracy during CoT generation, whereas direct answers are robust to answer-side knockout when question-time access is available. Across real-image and synthetic tasks, supplying the required attributes in text reduces this dependence. On synthetic follow-up counting, direct image access remains consequential before the model states the required attributes, even though image-conditioned question and prefix states are retained. At this stage, knockout reduces accuracy and image-KV swaps move answers toward the donor's count; both effects are smaller once the attributes are stated. Fixed-prefix replay further shows that the same visible prefix can accompany different sensitivity depending on the preceding image access. Retaining image-conditioned states therefore does not by itself make direct image access dispensable. These findings motivate considering both question-relevant information and access history when deciding when to reduce visual access during an answer and which visual information to retain for later questions.

Reproducibility statement

Estimates use saved per-image outcomes, paired input identities, and condition-matched baselines; source records retain image and output hashes, settings, annotations, and intervention diagnostics. Appendix A specifies the knockout, generation, scoring, replay, and KV-swap protocols.

Ethics statement

This study uses synthetic scenes and benchmark images. Accurate attributes are experimental controls, not an assumption that reliable descriptions are available in deployment. Results do not establish safe access removal outside the evaluated conditions.

Statement on the use of large language models

In this work, we used generative AI tools to propose and refine hypotheses, to design and provide feedback on the research methodology and experiments, to implement methods (intervention code for attention knockout, image-KV swap, and fixed-prefix replay, and the analysis pipelines), to develop the rule that labels attribute expression in generated prefixes, to assist with translation of drafts written in Japanese, and to interpret results. We have not used generative AI tools to generate the synthetic data sets: all scenes are produced by a deterministic, seeded generator, whose code was written with AI assistance. Developing theoretical models, formulating mathematical claims, and writing proofs are not applicable to this work. Additionally, we used generative AI tools to create and edit software code, create and modify figures, draft parts of the paper, edit the paper for readability, identify related literature and check bibliographic metadata, format references, and suggest the structure and title of the paper. We have reviewed all AI-assisted work: the intervention and analysis code was checked with parity tests and against saved per-image outcomes (Appendix A.7), the attribute-expression rule was frozen before any stage intervention ran and its limits are reported (Appendix A.5), every cited work was checked against its source by the authors, and all numbers in the text were checked against the saved results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Rakshanda Hassan Abhinandan, John Galeotti, Deva Ramanan, and Gautam Rajendrakumar Gare. Ask Twice, Look Twice: Prompt Echoing Resolves the Question-First Paradox in Vision-Language Models, 2026. URL https://arxiv.org/abs/2607.15565. eXCV Workshop, ECCV 2026.

Manoj Acharya, Kushal Kafle, and Christopher Kanan. TallyQA: Answering Complex Counting Questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8076–8084, 2019. URL https://arxiv.org/abs/1810.12440.

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, 2025. URL https://arxiv.org/abs/2502.13923.

Xidi Cai, Junhao Zheng, Jingye Li, Boyuan Li, Shaowei Zhang, and Qianli Ma. Look and Think: Efficient Multimodal Reasoning via Modality-Decoupled Compression. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 24787–24806, 2026. doi: 10.18653/v1/2026.findings-acl.1241. URL https://aclanthology.org/2026.findings-acl.1241/.

Hong Chen, Kang Chen, Yuxuan Fan, Bo Wang, Yubo Gao, Yuanlin Chu, and Xuming Hu. Seen, Said, or Forgotten? A Causal Audit of Visual KV Memory Across Dialog Turns, 2026. URL https://arxiv.org/abs/2607.25467.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large

Vision-Language Models. In *European Conference on Computer Vision (ECCV)*, 2024. URL https://arxiv.org/abs/2403.06764. Oral.

Xu Chu, Xinrong Chen, Guanyu Wang, Zhijie Tan, Kui Huang, Wenyu Lv, Tong Mo, and Weiping Li. Qwen Look Again: Guiding Vision-Language Reasoning Models to Re-attention Visual Information, 2025. URL https://arxiv.org/abs/2505.23558.

Bradley Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics, 1982. doi: 10.1137/1.9781611970319.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: VLMs overlook their visual representations. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2506.08008.

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. URL https://arxiv.org/abs/2304.14767.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching, 2024. URL https://arxiv.org/abs/2404.15255.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020.

Kai Huang, Hao Zou, Bochen Wang, Ye Xi, Zhen Xie, and Hao Wang. AirCache: Activating Inter-modal Relevancy KV Cache Compression for Efficient Large Vision-Language Model Inference. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23958–23967, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Huang_AirCache_Activating_Inter-modal_Relevancy_KV_Cache_Compression_for_Efficient_Large_ICCV_2025_paper.html.

Drew A. Hudson and Christopher D. Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. URL https://arxiv.org/abs/1902.09506. Oral.

Alon Jacovi and Yoav Goldberg. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. URL https://arxiv.org/abs/2004.03685.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019. URL https://arxiv.org/abs/1902.10186.

Zhuoran Jin, Kejian Zhu, Hongbang Yuan, Yupu Hao, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Look light, think heavy: What multimodal chain-of-thought reasoning can and cannot do. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8567–8608, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.387. URL https://aclanthology.org/2026.acl-long.387/.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2901–2910, 2017. URL https://arxiv.org/abs/1612.06890.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the Image? A Deep-Dive into the Vision of Vision Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14558, 2025. URL https://arxiv.org/abs/2411.17491.

Sai Srinivas Kancheti, Aditya Sanjiv Kanade, Vineeth N. Balasubramanian, and Tanuja Ganu. Chain-of-Thought Degrades Visual Spatial Reasoning Capabilities of Multimodal LLMs. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 862–876, 2026. doi: 10.18653/v1/2026.acl-short.71. URL https://aclanthology.org/2026.acl-short.71/. Short paper.

Samir Khaki, Junxian Guo, Jiaming Tang, Shang Yang, Yukang Chen, Konstantinos N. Plataniotis, Yao Lu, Song Han, and Zhijian Liu. SparseVILA: Decoupling Visual Sparsity for Efficient VLM Inference. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. URL https://arxiv.org/abs/2510.17777.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2205.11916.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. doi: 10.1007/s11263-016-0981-7.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring Faithfulness in Chain-of-Thought Reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Kunxi Li, Zhonghua Jiang, Zhouzhou Shen, Zhaode Wang, Chengfei Lv, Shengyu Zhang, Fan Wu, and Fei Wu. MadaKV: Adaptive Modality-Perception KV Cache Eviction for Efficient Multimodal Long-Context Inference. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 13306–13318, 2025. doi: 10.18653/v1/2025.acl-long.652. URL https://aclanthology.org/2025.acl-long.652/.

Xiping Li and Jianghong Ma. AIM-CoT: Active Information-driven Multimodal Chain-of-Thought for Vision-Language Reasoning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 26656–26681, 2026. doi: 10.18653/v1/2026.acl-long.1227. URL https://aclanthology.org/2026.acl-long.1227/.

Benlin Liu, Amita Kamath, Madeleine Grunde-McLaughlin, Winson Han, and Ranjay Krishna. Visual Representations inside the Language Model. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2510.04819.

Sihao Liu, YuFan Xiong, Zhonghua Jiang, Zhaode Wang, Chengfei Lv, and Shengyu Zhang. RetentiveKV: State-Space Memory for Uncertainty-Aware Multimodal KV Cache Eviction. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 18717–18727, 2026. doi: 10.18653/v1/2026.findings-acl.934. URL https://aclanthology.org/2026.findings-acl.934/.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show Your Work: Scratchpads for Intermediate Computation with Language Models, 2021. URL https://arxiv.org/abs/2112.00114.

Jacob Pfau, William Merrill, and Samuel R. Bowman. Let's Think Dot by Dot: Hidden Computation in Transformer Language Models, 2024. URL https://arxiv.org/abs/2404.15758.

Qwen Team. Qwen2.5-VL-32B: Smarter and lighter, 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl-32b/.

Israfel Salazar, Stella Frank, Dan Oneata, Desmond Elliott, and Constanza Fierro. Pathways of Visual Information Flow in Vision-Language Models, 2026. URL https://arxiv.org/abs/2607.03358.

Fawaz Sammani, Tzoulio Chamiti, and Nikos Deligiannis. On Test-Time Scaling for Vision-Language Models. In *European Conference on Computer Vision (ECCV)*, 2026. URL https://arxiv.org/abs/2606.28864.

Haz Sameen Shahgir, Xiaofu Chen, Yu Fu, Erfan Shayegani, Nael Abu-Ghazaleh, Yova Kementchedjhieva, and Yue Dong. VLMs Need Words: Vision Language Models Ignore Visual Detail In Favor of Semantic Anchors. In *Conference on Language Modeling (COLM)*, 2026. URL https://arxiv.org/abs/2604.02486.

Chufan Shi, Cheng Yang, Yaokang Wu, Linghao Jin, Bo Shui, Taylor Berg-Kirkpatrick, and Xuezhe Ma. Are VLMs Seeing or Just Saying? Uncovering the Illusion of Visual Re-examination. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2605.15864. Oral.

Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. Mitigating Visual Forgetting via Take-along Visual Conditioning for Multi-modal Long CoT Reasoning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. URL https://arxiv.org/abs/2503.13360.

Dezhan Tu, Danylo Vashchilenko, Yuzhe Lu, and Panpan Xu. VL-Cache: Sparsity and Modality-Aware KV Cache Compression for Vision-Language Model Inference Acceleration. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.23317.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.04388.

Zhongwei Wan, Ziang Wu, Che Liu, Jinfa Huang, Zhihong Zhu, Peng Jin, Longyue Wang, and Li Yuan. LOOK-M: Look-Once Optimization in KV Cache for Efficient Multimodal Long-Context Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4065–4078, 2024. doi: 10.18653/v1/2024.findings-emnlp.235. URL https://aclanthology.org/2024.findings-emnlp.235/.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. URL https://arxiv.org/abs/2201.11903.

Sarah Wiegreffe and Yuval Pinter. Attention is not not Explanation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. URL https://arxiv.org/abs/1908.04626.

Hongxuan Wu, Yukun Zhang, and Xueqing Zhou. How Vision Becomes Language: A Layer-wise Information-Theoretic Analysis of Multimodal Reasoning, 2026. URL https://arxiv.org/abs/2602.15580.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is Better but Not Necessary in Vision Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802, 2025. URL https://arxiv.org/abs/2412.04467.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022. URL https://arxiv.org/abs/2109.05014.

Hao Yin, Guangzong Si, and Zilei Wang. Lifting the Veil on Visual Information Flow in MLLMs: Unlocking Pathways to Faster Inference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9382–9391, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yin_Lifting_the_Veil_on_Visual_Information_Flow_in_MLLMs_Unlocking_CVPR_2025_paper.html.

Runpeng Yu, Xinyin Ma, and Xinchao Wang. Auto-Controlled Image Perception in MLLMs via Visual Perception Tokens. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21822–21831, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Yu_Auto-Controlled_Image_Perception_in_MLLMs_via_Visual_Perception_Tokens_ICCV_2025_paper.html.

Qianhao Yuan, Qingyu Zhang, Yanjiang Liu, Jiawei Chen, Yaojie Lu, Hongyu Lin, Jia Zheng, Xianpei Han, and Le Sun. ShortV: Efficient Multimodal Large Language Models by Freezing Visual Tokens in Ineffective Layers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. URL https://arxiv.org/abs/2504.00502.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language, 2022. URL https://arxiv.org/abs/2204.00598.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.16042.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM: Visual Token Sparsification for Efficient Vision-Language Model Inference. In *International Conference on Machine Learning (ICML)*, 2025a. URL https://arxiv.org/abs/2410.04417.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal Information Flow in Multimodal Large Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b. URL https://arxiv.org/abs/2411.18620.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal Chain-of-Thought Reasoning in Language Models. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2302.00923.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models, 2025. URL https://arxiv.org/abs/2504.10479.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. Look Twice Before You Answer: Memory-Space Visual Retracing for Hallucination Mitigation in Multimodal Large Language Models. In *International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2410.03577.

APPENDIX CONTENTS

A. Experimental Details 16

A.1 Attention Knockout Implementation 16

A.2 Metrics and Uncertainty 17

A.3 Evaluation Data and Information Conditions 18

A.4 Prompts 21

A.5 Attribute-Expression Annotation and Intervention Stages 23

A.6 Fixed-Prefix Replay 24

A.7 Counterfactual Image-KV Swap and Implementation Checks 25

B. Model Family Ablations (InternVL3-14B) 26

B.1 Response Format 26

B.2 Description Type 26

B.3 History Content and Question-Time Access 27

C. Supplementary Controls 28

C.1 Immediate-Answer Control and Independent Confirmation 28

C.2 Additional Information Conditions 28

C.3 The Conditions behind the Fixed-Prefix Replay 29

C.4 Attribute Stages and Common Sets 30

C.5 Counterfactual Swap Controls 31

A Experimental Details

A.1 Attention Knockout Implementation

Table A.1 lists the attention edges severed by the interventions of Section 2.1 and by the comparisons of Figure 3. Every intervention adds $-\infty$ to the pre-softmax logits, in every head, from the selected query rows to the image-token key columns, and the remaining attention weights are renormalized. Image tokens themselves, the attention edges that are not targeted, and the key–value (KV) states computed before the intervention are retained.

Table A.1: Targets and onset of attention knockout. Decoder layers are indexed $k = 0, \ldots, L-1$. Ordinary answer-side knockout severs the edges whenever either the layer condition or the token condition holds; the row-group sweeps are a separate comparison that severs one kind of row only.

Intervention	Query rows	Layer condition	Where it applies	Retained computation
Question-side knockout	rows of the current user question	all layers	while the current question is processed	KV of earlier turns; edges from the assistant-header and decode rows
Answer-side knockout	assistant-header rows and the decode rows that follow	$k \geq \ell$, or all layers once the token condition holds	from the position set by the token cutoff τ; the layer condition also acts before it	KV computed before the intervention; text-to-text attention
Question-row sweep	rows of the current user question only	$k \geq \ell$	while those rows are processed	image access from the header and decode rows
Header-row sweep	assistant-header rows only	$k \geq \ell$	while the header is processed	image access from the question and decode rows
Decode-row sweep	decode rows only	$k \geq \ell$	while the decode rows are processed	image access from the question and header rows

The last assistant-header row predicts answer token 1. $\tau = 0$ severs access from every assistant-header row; $\tau = m > 0$ keeps image access for the first m answer tokens and severs it from the prediction of token $m + 1$. $\tau = \infty$ applies no token-based knockout, and $\ell = L$ no layer-based knockout. Because the two conditions combine by *or*, a layer cutoff $\ell < L$ acts from the first answer token whatever τ is; statements about late generation therefore use $\ell = L$, and layer effects are read at $\tau = \infty$.

In follow-up and real-image inputs, Turn 1, which contains the image, the first question, and the fixed reply, is computed once with image access, and its KV is kept. Question-side knockout applies to the rows of the Turn 2 question, and the assistant header is treated as the answer side. The row-group sweeps run at batch size 1.

Knockout counts as *applied* when, in the forward passes actually executed, at least one answer-side edge to an image token is severed. When a response ends before the cutoff, the token condition never applies. Curves with such cutoffs report the application rate, and the aggregate loss, which includes responses without the intervention, is distinguished from the loss over the responses that were actually knocked out.

A.2 METRICS AND UNCERTAINTY

Table A.2 summarizes the knockout loss, the question-access difference, and the KV-swap measures of the main text. Each comparison keeps the pairing by image and uses a baseline from the same execution condition.

Table A.2: Measures and the conditions each pairs. A denotes accuracy and p the rate of a given answer; rate differences are in percentage points (pp).

Measure	Definition	Held fixed in the pairing	Unit
Knockout loss Δ_{KO}	$A_{\mathrm{base}} - A_{\mathrm{KO}}$	image, input, model, response format, question-time access	pp
Question-access difference D	$\Delta_{\mathrm{KO}}^{\mathrm{Blocked}} - \Delta_{\mathrm{KO}}^{\mathrm{Allowed}}$	image, history, cutoff; in replay also the prefix source	pp
Donor-answer rate	fraction of answers equal to the donor image's correct count	the specified donor count	%
Change in donor-answer rate	$p_{\mathrm{target\ swap}} - p_{\mathrm{no\ swap}}$	image, prefix, execution path, donor count	pp

When question-time access is Blocked, the baseline also blocks question-side access, so the knockout loss measures the change that answer-side knockout adds to the effect of question-side blocking. A zero loss means that aggregate accuracy did not change; it does not mean that individual answers are unchanged or that the retained states carry no image information.

The 95% confidence intervals (CIs) are percentile bootstrap intervals over images (Efron, 1982). Resampling keeps the pairing of the compared conditions, and the several donors of one image are not counted as independent samples. We draw 20,000 resamples, with seed 20260915 for the September 2026 sweeps and 20260923 for the additional campaigns. An image missing one condition of a paired contrast is excluded from that contrast. Intervals on curves are pointwise, not simultaneous over the curve.

When no answer changes in any example, the bootstrap interval degenerates to $[0, 0]$, which does not show that the population change rate is exactly zero. With zero changes in n independent examples, the upper limit of the two-sided 95% exact binomial interval for the change rate is $1 - 0.025^{1/n}$.

A.3 Evaluation Data and Information Conditions

Table A.3 lists the evaluation sets and settings of Section 2.3 and Section 3.1–3.3, and Table A.4 the information supplied as text. Conditions compared within one experiment use the same images, and no set is filtered by model correctness.

Runs use batch size 1 with prefill chunk 128, or batch size 4 without chunking. These settings change some greedy CoT outputs (the single-turn exploration set loses 24 pp at $\tau = 0$ in one setting and 28 pp in the other), so the baseline and the intervention of every comparison run in the same setting. Image-KV swaps run at batch size 1 and are compared with the no-swap continuation of the same run.

Answers are scored on the final `Answer:` x line: an option letter for single-turn and GQA, a nonnegative integer for follow-up and TallyQA. An answer is correct when exactly one such line exists, it is the last line, x is a valid answer, and x matches the ground truth. Format violations and responses that reach the output cap count as incorrect and stay in the denominator.

Scene generation and image selection. Single-turn scenes place 6–7 non-touching objects on a white 256×256 canvas; colors are drawn from blue, red, yellow, and green, shapes from star, circle, triangle, and square, radii from 18, 24, 30, and 36 pixels, and angles from $0°$, $45°$, $90°$, and $135°$. Each question asks for the number of objects of one color. As in the task design of CLEVR (Johnson et al., 2017), controlled attributes and relations make the required information explicit. Follow-up scenes use 512×512 images with fixed candidate slots in each quadrant. The anchor quadrant and one decoy quadrant hold three objects each, the anchor's color–shape pair is unique, and the asked color differs from the anchor's color. The anchor quadrant holds $c \in \{0, 1, 2\}$ objects of the asked color and the decoy $3 - c$, so the image-wide count never equals the local answer; quadrant and answer are balanced within blocks of 12 items. GQA questions ask for a color together with a spatial, wearing, or holding relation; yes/no and counting questions are excluded, scene graphs have 2–40 objects, one question per image is chosen by a seeded hash, and the options are the answer color and four distractors from a 15-color list. We select simple object-count questions from TallyQA (Acharya et al., 2019) on Visual Genome images (Krishna et al., 2017). For our evaluation, we retain one question per image and exclude questions about body parts and abstract nouns.

Table A.3: Evaluation sets and execution settings. Set names in Panel A denote their roles in separate experiments, and several sets of one task are never pooled into one sample; the GQA 100-image set is a subset of the 300-image set. No paired contrast crosses the setting differences of Panel B.

Panel A: evaluation sets

Set	Task and source	N	Selection and answers
Single-turn exploration	synthetic generator, validation split	50	256×256, 6–7 objects; counts 1–5: 14/21/12/2/1
Single-turn fresh	synthetic generator, same style	100	counts 1–5, 20 each; generated after exploration, decision rules fixed before the run
Single-turn training-style	synthetic generator, training split	100	items 0–99; five options (Direct row-group sweep of Figure 3(a))
Follow-up exploration	synthetic quadrant scenes	50	counts 0/1/2: 16/17/17
Follow-up independent	same synthetic generator	100	counts 0/1/2: 33/34/33; generated after exploration
Follow-up profile	same synthetic generator	100	counts 0/1/2
GQA main	GQA relation-color, validation split	50	seeded-hash ranks 51–100; five options
GQA confirmation	GQA relation-color, training-split candidates	300	five options
GQA matched subset	subset of the confirmation set	100	the same images evaluated on both models
TallyQA simple	simple questions on Visual Genome images	100	counts 1–4, 25 each

Panel B: models and execution settings

Item	Qwen2.5-VL-32B[a]	InternVL3-14B[b]
Decoder layers	64	48
Precision, attention	bfloat16, eager attention	bfloat16, eager attention
Decoding	greedy, repetition penalty 1.05	greedy, no added repetition penalty
Output caps	CoT 512 tokens, Direct 24 tokens	as left
Image processing	resized to the visual-token budget	448-pixel tiles, 256 visual tokens per tile
Visual-token budget	single-turn 256; follow-up, GQA, TallyQA 1280	tiling under the same budgets

[a] https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct
[b] https://huggingface.co/OpenGVLab/InternVL3-14B-Instruct

Table A.4: Information conditions supplied as text. "The current answer" is the answer to the Turn 2 question in follow-up counting. Conditions that supply the required attributes are distinct from the label Sufficient given to generated prefixes (Appendix A.5).

Task	Condition	Position	Information supplied	States the current answer	Length control
Single-turn	None	–	none	no	–
	Complete	user text before the question	color, shape, and location of every object	the count itself is not stated	–
	Partial	as above	the same listing with the colors withheld	no	listing format shared except color
	Unrelated	as above	text unrelated to the image	no	token count matched to Complete
Follow-up	**Complete**	Turn 1 user text	color, shape, and quadrant of the objects	the count itself is not stated	–
	Partial	as above	the attribute listing with the colors withheld	no	listing format shared except color
	Unrelated	as above	text unrelated to the image	no	token count matched to Complete line by line
	Label-only history	Turn 1 assistant reply	the correct quadrant name	only the Turn 1 answer	–
	Attribute-stating history	Turn 1 question and reply	the quadrant and the color and shape of its objects	the count itself is not stated	matched to Shape-only at 95 tokens
	Shape-only history	as above	the quadrant and the shapes of its objects, without colors	no	matched to Attribute-stating at 95 tokens
GQA	None	–	none	no	–
	Complete	Turn 1 user text	scene-graph listing, including the relation that identifies the target and the asked color	yes: contains the answer color (positive control)	–
	Partial	as above	the scene information with the colors withheld	no	control without color information
	Unrelated	as above	text unrelated to the image	no	matched to Complete per tokenizer

A.4 PROMPTS

Boxes A.1–A.4 give the inputs used in the experiments; the English text of the prompts is reproduced verbatim, and square brackets mark task-dependent parts.

Box A.1: Answer instructions (end of every evaluated question)

CoT

Inspect the image and reason step by step in at most 120 words using its visible evidence. End with a plain line containing Answer: followed by exactly one *[option letter | nonnegative integer]*. Do not include *[option text | extra text]* or Markdown on that line.

Direct

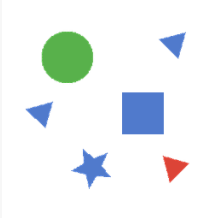
[Choose one listed option.] Give only the final answer without explanation. End with a plain line containing Answer: followed by exactly one *[option letter | nonnegative integer]*. Do not include *[option text | extra text]* or Markdown on that line.

Immediate-answer control

The CoT instruction verbatim; the assistant turn is teacher-forced to open with Answer: and decoding continues from there.

Square brackets mark the task-dependent alternatives. Each input is wrapped in the model's native chat template.

Box A.2: Single-turn input (image `multi_val_00000100`; answer C)



User

[image at left] *[description; omitted under None]*

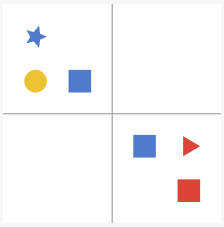
How many blue objects are there in the image?

Options:
A. 3
B. 2
C. 4
D. 1
E. 5

[answer instruction from Box A.1]

Complete lists `color`, `shape`, `size`, `location`, and `angle` for every object; Partial uses the same list with `color=not provided`; Unrelated is text unrelated to the image, matched to the token count of Complete. The repeated attribute list is omitted here.

Box A.3: Follow-up input and histories (image arc-0012; the anchor is the red triangle in lower_right; answer 1)



User, Turn 1

[image at left] `[Additional information: description; omitted under None]`

```
The image is divided into four quadrants:
upper_left, upper_right, lower_left, lower_right.
Which quadrant contains the only red triangle?
Reply with only the quadrant name -- one of
upper_left, upper_right, lower_left, lower_right --
and nothing else. Do not explain.
```

Assistant, Turn 1 (teacher-forced)

```
lower_right
```

User, Turn 2

```
In that same quadrant, how many blue objects are
there?
```

[answer instruction from Box A.1]

Attribute-stating history: Turn 1 also asks for the objects of the quadrant by color and shape, and the fixed reply is

```
lower_right. It contains a blue square, a red triangle, and a red
square.
```

Shape-only history: Turn 1 asks for the shapes in order, and the fixed reply is

```
lower_right. It contains a visible square, a visible triangle, and a
visible square.
```

Box A.4: Real-image input (GQA image 2379795, answer C; TallyQA image 2355775)





User, Turn 1

[image (top left)] `Keep this image available for my next question. Reply only Ready.`

[Additional information: description; omitted under None]

Assistant, Turn 1 (fixed)

```
Ready.
```

User, Turn 2 (GQA example)

```
What color is the rope that the skier is holding?

Options:
A. red
B. brown
C. orange
D. white
E. purple
```

[answer instruction from Box A.1]

The GQA Complete description contains the relation "the skier is holding the rope" and the attribute "rope: orange". For TallyQA no description is added, and Turn 2 asks, for image 2355775, `How many laptops in the photo?` *with a nonnegative-integer answer.*

A.5 Attribute-Expression Annotation and Intervention Stages

The stages of Section 3.4 are defined by a fixed rule that judges how far the required attributes have been expressed in a generated prefix. The rule consults the ground-truth attributes of the scene, so it is an oracle annotation for analysis, not a signal for access control at deployment. It classifies visible text, and does not show that the model causally uses that text (Jacovi & Goldberg, 2020).

Table A.5: Prefix categories and stage definitions. Categories are applied from top to bottom. The follow-up answer is determined by the colors of the two non-anchor objects, and the anchor itself never has the asked color. Objects are matched by color, and wrong shape names are tolerated.

Panel A: prefix categories

Category	Criterion	Note
Incorrect	mentions a nonexistent object or a wrong color or count, or a closed enumeration implies a wrong count	judged before the other categories
Sufficient	states the colors of both non-anchor objects, or completes a closed enumeration that implies the correct count	a closed enumeration is judged at the step that follows the last object mention
One color	states the color of exactly one non-anchor object	the other required color is not yet expressed
Uninformative	none of the above	includes mentions of the anchor alone

Panel B: intervention positions

Stage	Intervention position	Responses included
Before expression	the prediction of the token at which the first non-anchor color is written	responses in which this position is defined
After sufficient expression	the token after the prefix is first labeled Sufficient	responses that state at least one non-anchor color and reach a Sufficient prefix

The rule was developed on the 64-token prefixes of the independent set under both question-access conditions and fixed before any stage intervention ran. The prefixes were inspected with answer correctness hidden, although correctness was visible for six images while the first version of the rule was written. The judgment applied at each position has not been validated by an independent annotator, and the uncertainty of the cutoff positions is not included in the CIs of the results.

The before and after positions are not the moments just before and after one event, and several descriptive steps may separate them. The comparison measures sensitivity at two content stages; it does not identify a switch at the moment of verbalization or a causal effect of verbalization.

A.6 FIXED-PREFIX REPLAY

The replay of Table 1 uses the same initial text for each prefix source and changes only question-time access, and compares the additional knockout loss.

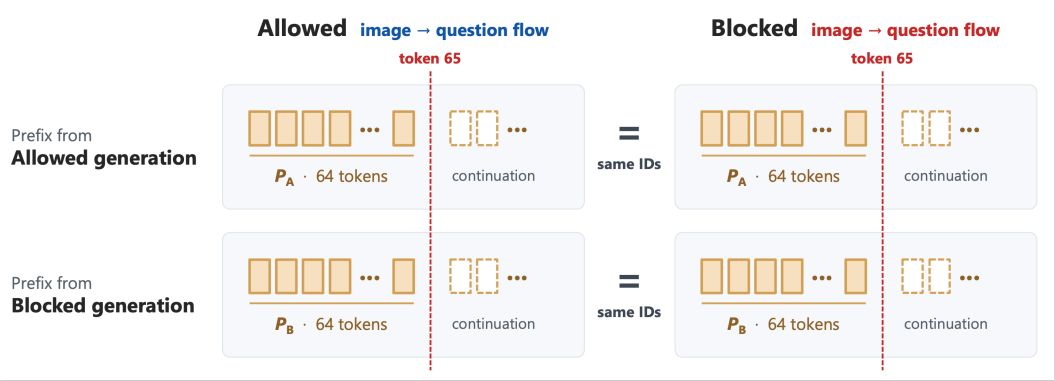
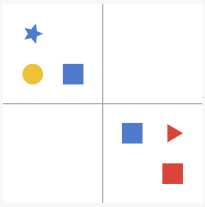


Figure A.1: Fixed-prefix replay design. Rows indicate the source of the initial 64-token prefix; columns indicate question-time image access during replay. Within each row, the prefix token IDs are identical across Allowed and Blocked conditions; prefixes in different rows need not match. The image, Turn 1 cache, and current question are held fixed, while question-side and prefix states are recomputed under the replay condition. Continuations may differ. The dashed lines mark the onset of answer-side knockout at the prediction of token 65, with no layer restriction. Each cell is compared against its own baseline without answer-side knockout under the same question-time access condition.

Replaying a prefix under the condition that generated it reproduced the original generation in all 400 checked continuations. The comparison controls the initial visible text, but it fixes neither the internal states nor the text after token 64.

Box A.5: The same prefix, two continuations (image arc-0012; follow-up question with answer 1)



End of the shared prefix, generated with Allowed

```
3. Examine the contents of the lower_right quadrant:
```

Continuation replayed with Allowed (excerpt): final answer 1

```
   - There is one red triangle.
   - There is also one blue object, which appears to
be a square.
```

Continuation replayed with Blocked (excerpt): final answer 0

```
   - There is a red triangle.
   - No other objects are present in this quadrant.
```

Excerpts for readability; the replay uses the original 64 token IDs. The example illustrates the operation; the results are in Table 1 and Figure C.3.

A.7 Counterfactual Image-KV Swap and Implementation Checks

In Figure 6 (right), the prefix computed from the recipient image is kept, and the KV at the image positions is replaced by KV computed from another image. The donor KV comes from a separate full-access prefill of the same Turn 1, and the keys and values at the image positions of every layer are replaced. The KV at the text positions of Turn 1, the current question, and the generated prefix keeps the values computed from the original image. The operation therefore does not re-read an image and update every state; it combines states derived from different images.

Table A.6: Swap conditions and implementation checks. In Panel A, c is the original correct count; donors A and B have the smaller and the larger value of $\{0, 1, 2\} \setminus \{c\}$ as their correct count, which need not be smaller or larger than c.

Panel A: interventions and controls

Condition	Image change	Image KV	Text KV	Purpose
No swap	none	original KV kept	kept	baseline of the forced-prefix continuation
Self swap	the same image	replaced by KV computed from the same image	kept	consistency of the replacement itself
Target A	objects in the anchor quadrant recolored	replaced by donor A's KV	from the original image	directional response to another correct count
Target B	objects in the anchor quadrant recolored	replaced by donor B's KV	from the original image	response to the other correct count
Unrelated swap	one object in the decoy quadrant recolored	replaced by the changed image's KV	from the original image	control for an image change that keeps the count

Panel B: implementation and execution-path checks

Check	Runs compared	Items	Result
Cache parity	cached vs uncached full-access generation	211	token sequences identical in 211/211
Fixed-prefix parity	fixed-prefix vs native generation	6	identical in 6/6
Own-condition replay	own-condition replay vs free generation	400 continuations	identical in 400/400
Stage-prefix match	prefix up to the stage intervention vs free prefix	before: Allowed 96, Blocked 93; after: Allowed 85, Blocked 72	the pre-intervention prefix matches in all
Self-swap parity	self swap vs no swap	Allowed 30, Blocked 29	token sequences identical in all, both conditions
No-swap vs free path	no swap vs free generation of the same run	Allowed 96, Blocked 93	token sequences identical 68/96 and 74/93; final answers identical 93/96 and 92/93

Swap comparisons use the no-swap continuation of the same arm and run as the reference, and do not assume exact identity with free generation. Each run checks the application position, the prefix reproduction, the replaced KV, the planned item set, format violations, and output caps. Target swaps are evaluated before expression in both question-access conditions and after sufficient expression in the Allowed condition.

B Model Family Ablations (InternVL3-14B)

To check the correspondence with Qwen2.5-VL-32B in the main text, we compare response formats, description content, and histories with question-time access on the synthetic tasks in InternVL3-14B (Zhu et al., 2025). All results in this appendix use InternVL3-14B on the synthetic counting tasks, with the same intervention definitions and outcome measures as in the main text. A change of model family also changes size, training, and implementation, so these comparisons do not isolate a single factor.

B.1 Response Format

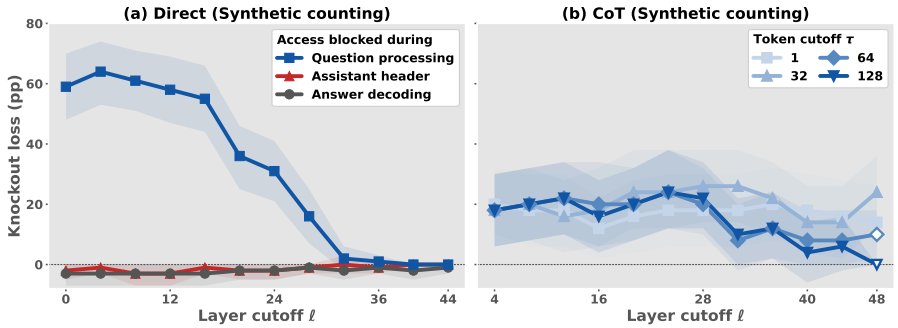


Figure B.1: Comparison of knockout loss across response formats. Direct answers lose accuracy when the question rows are severed but little otherwise, whereas CoT answers also lose accuracy under answer-side knockout. **(a)** Direct: question, assistant-header, or decode rows severed separately at layers $k \geq \ell$. **(b)** CoT: answer-side layer and token cutoffs combined, with question-time access Allowed.

For Direct answers, severing image access of the question rows from shallow layers causes a loss, whereas severing the header or decode rows produces little observed loss. For CoT, answer-side knockout causes a loss. This contrast supports the main-text result for Qwen2.5-VL-32B that Direct and CoT answers are sensitive through different access routes.

B.2 Description Type

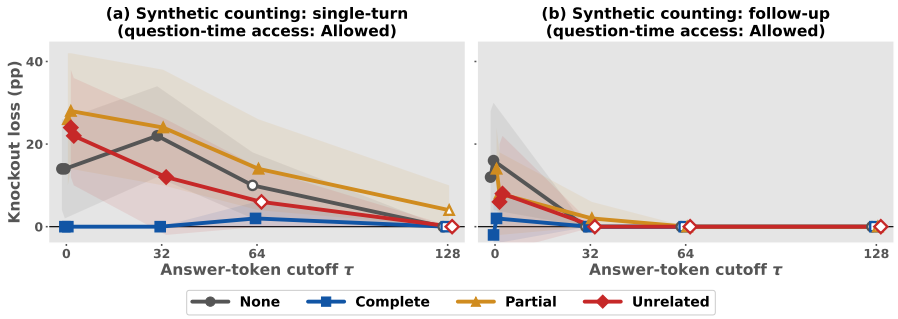


Figure B.2: Comparison of knockout loss across description conditions. Complete descriptions keep loss small in single-turn counting, where Partial and Unrelated do not; in follow-up counting the conditions separate less. **(a)** Single-turn and **(b)** follow-up counting with the Label-only history; CoT, question-time access Allowed, no layer restriction.

In single-turn counting, Complete, which supplies the required colors, shows a small loss, whereas Partial, which withholds the colors, and unrelated text do not remove the loss at answer onset. This supports the main-text result that containing the attributes the answer needs matters, not merely the presence of added text. In follow-up counting (Figure B.2(b)), the differences between conditions are small, and the separation seen in single-turn counting does not appear.

B.3 History Content and Question-Time Access

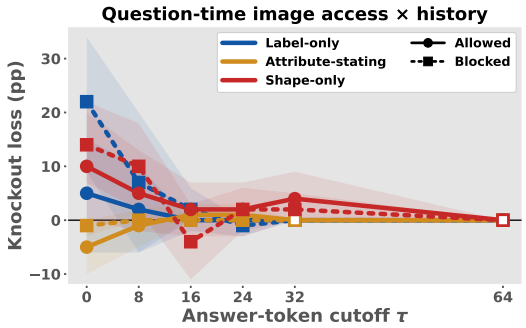


Figure B.3: Comparison of knockout loss across histories and question-time access. Attribute-stating histories keep loss near zero; the Label-only access difference appears at answer onset but not at later cutoffs. Follow-up counting, CoT, no layer restriction; colors give histories and line styles access (solid: Allowed; dashed: Blocked), each against its own baseline.

With the Attribute-stating history, which contains the required attributes, the loss is small in both access conditions, and the answer-onset loss seen with the Label-only and Shape-only histories is suppressed. This supports the main-text comparison that separates the correctness of an earlier answer from the sufficiency of the information for the current question.

With the Label-only history, the answer-onset loss is larger under Blocked than under Allowed, so the question-access difference has the same direction as in Qwen2.5-VL-32B. The difference becomes small at later cutoffs, however, and the persistent difference seen in Qwen2.5-VL-32B does not reproduce. The effect of question-time access on the sensitivity of later generation therefore depends on the model and on the generation position.

C SUPPLEMENTARY CONTROLS

C.1 IMMEDIATE-ANSWER CONTROL AND INDEPENDENT CONFIRMATION

The immediate-answer control keeps the CoT instruction and tests whether the format difference of Section 3.1 arises from the instruction text alone.

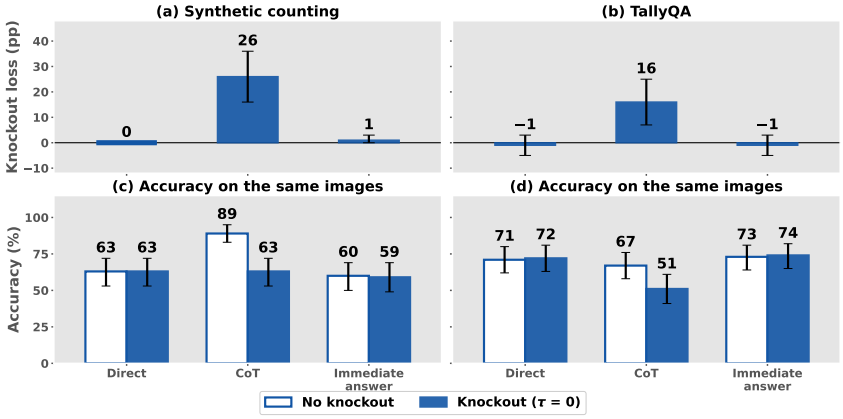


Figure C.1: Comparison of knockout loss and accuracy across response formats, including the immediate-answer control. The immediate-answer control, which keeps the CoT instruction, shows no CoT-like loss. **(a, c)** Count-balanced single-turn fresh set and **(b, d)** TallyQA; **(a, b)** knockout loss and **(c, d)** baseline and knockout accuracy. Question-time access Allowed, $\tau = 0$, no layer restriction.

The immediate-answer control does not show the CoT loss, so the instruction text alone does not explain the format difference. The three formats also differ in response length, so the comparison does not separate the content of reasoning from the effect of intervening on a long generation.

C.2 ADDITIONAL INFORMATION CONDITIONS

Question-access conditions not shown in Figure 4 complement the relation between description content and loss.

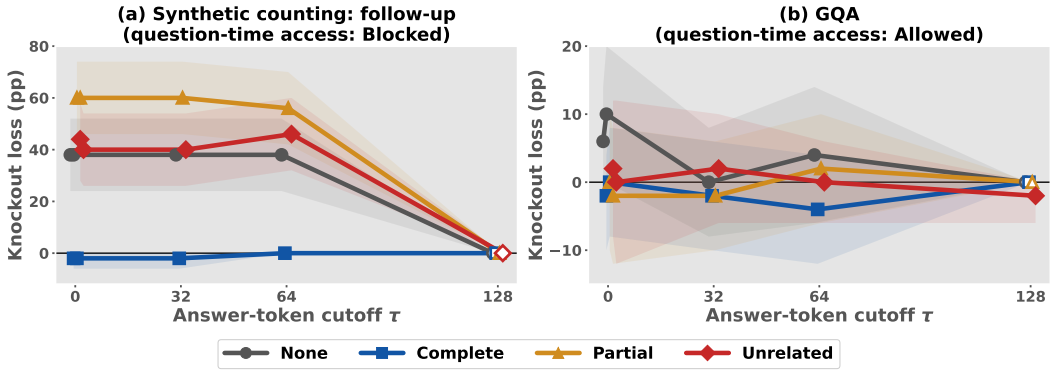


Figure C.2: Comparison of knockout loss across description conditions in the complementary access conditions. Complete keeps loss small in follow-up counting under Blocked; on GQA under Allowed the loss is small even without a description. **(a)** Follow-up counting with the Label-only history, question-time access Blocked; **(b)** GQA, question-time access Allowed; CoT, no layer restriction.

In follow-up counting under Blocked, Complete also keeps the loss small, whereas descriptions that withhold the required colors or are unrelated do not. On GQA under Allowed, the loss is small even without a description, unlike the Blocked condition of the main text.

C.3 The Conditions behind the Fixed-Prefix Replay

The differences in Table 1 are split here into the loss and accuracy of the four underlying replay conditions.

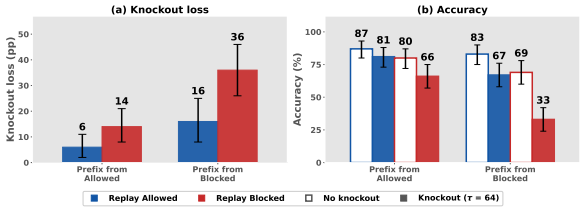


Figure C.3: Knockout loss and accuracy in the four fixed-prefix replay conditions. For both prefix sources, replay with question-time access Blocked loses more, but the baselines also differ. **(a)** Knockout loss and **(b)** baseline and knockout accuracy, grouped by the source of the 64-token prefix and colored by replay access. Independent set, Label-only history, CoT; within a source the same 64 token IDs are replayed with states recomputed, and knockout starts at the prediction of token 65 with no layer restriction.

For both prefix sources the replay with question-time access Blocked loses more. The baseline accuracy also differs between conditions, so the difference in final knocked-out accuracy is not read as the question-access difference in loss. The two source-specific differences are separate comparisons, not an additive decomposition of the free-generation difference. The knocked-out Blocked continuations show a recurring failure pattern: on the profile set every knocked-out Blocked response at $\tau \leq 32$ answers "0", and on the independent set 43 of the 67 errors at $\tau = 64$ state that only the Turn 1 object is present.

C.4 Attribute Stages and Common Sets

Table C.1 gives the responses that enter the before and after comparisons of Figure 6. Because the stages are selected from the generated content, conditioning on reaching them can change the evaluated population (Hernán & Robins, 2020); a common set aligns the images but not the elapsed generation or the difficulty.

Table C.1: Populations of the stage analyses. Paired comparisons within a row use the same images; different rows need not share a population.

Analysis	Question access	Selection	N
Stage knockout (main text)	Allowed	both before and after positions are defined	85
Stage swap (main text)	Allowed	Uninformative prefix at the before position, and the after position is evaluable	79
Stage knockout, both arms	Allowed / Blocked	before and after are defined in both arms	66
Fixed-prefix replay	both	64-token prefixes comparable under both replay conditions	100
Token-64 content categories	both	64-token prefixes of generation without knockout	100 each

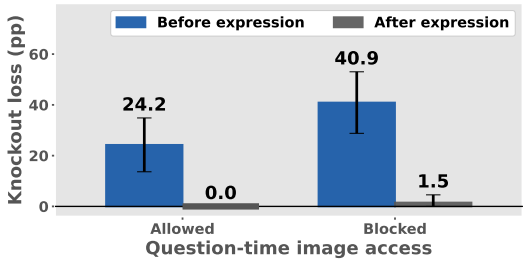


Figure C.4: Comparison of knockout loss before and after sufficient expression on the common set. The before–after difference holds on the images that reach both stages in both access conditions. Independent set, Label-only history, CoT; each response's intervention positions come from its own free generation.

The before–after difference survives the restriction to the same images in both access conditions, and all eligible responses give the same ordering (Allowed 16.7 vs 0.0 pp; Blocked 29.0 vs 1.4 pp). What the restriction cannot remove is that the two stages also differ in elapsed generation and answer commitment, which is why the comparison does not establish a causal effect of writing the attributes.

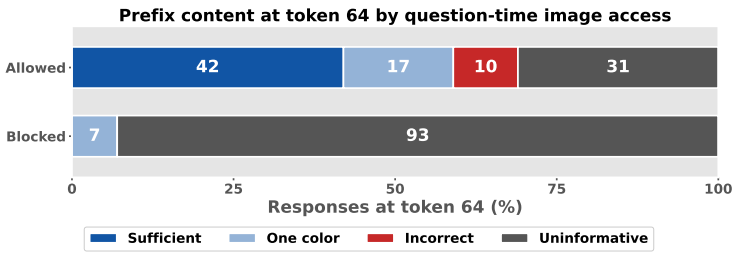


Figure C.5: Share of responses by prefix category at token 64. At the same token position, Allowed and Blocked responses have expressed different content. Free generations on the independent set, classified by Table A.5; a descriptive comparison, not a manipulation of attribute expression.

The same token position thus falls at different content stages under the two access conditions, and the knockout loss at that position follows the content already expressed: Sufficient prefixes keep their accuracy (Allowed 42 $\rightarrow$ 42 correct), whereas Uninformative prefixes lose it (Allowed 27 $\rightarrow$ 23; Blocked 62 $\rightarrow$ 29). This is why the stage analyses align the cutoff to content rather than to token position.

C.5 Counterfactual Swap Controls

The change in donor-answer rate of Figure 6 (right) is split into the raw rates without a swap and the response to an image change unrelated to the task.

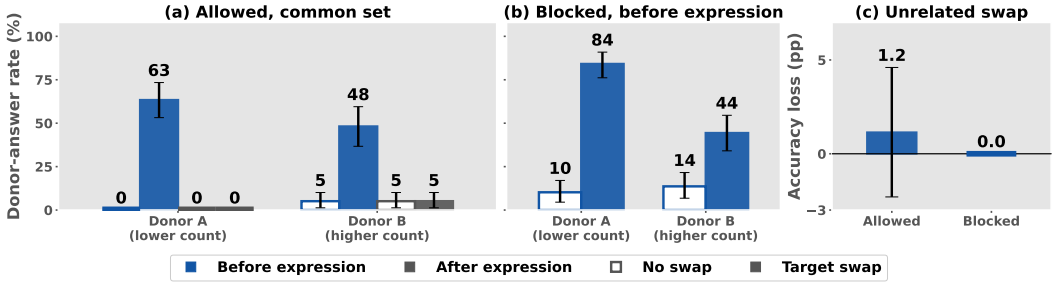


Figure C.6: Donor-answer rates and unrelated-swap accuracy loss under counterfactual image-KV swap. Target swaps raise the donor-answer rate before expression but not after, and the unrelated swap costs little accuracy. **(a)** Donor-answer rate under no swap and target swap on the common set, Allowed, before and after expression; **(b)** the same before expression under Blocked, on images with an Uninformative prefix; **(c)** accuracy loss under the unrelated swap. Rates are in %, accuracy differences in pp; the self-swap checks are in Table A.6.

After expression, donor B's answer also occurs without a swap, so the rate after a target swap alone does not measure the effect of the replacement; the main text reports the difference from no swap on the same images. The unrelated swap changes accuracy little (Allowed 1.2 [−2.3, 4.6] pp; Blocked 0.0 pp), so the directional response to target swaps before expression is hard to capture with an account in which any image-KV replacement disrupts answers uniformly.

The self and unrelated swaps do not exclude every effect of a mixed KV state. The results are read as sensitivity to the specified replacement, not as the share of a pathway used in ordinary generation or as evidence that image information is lost from the internal states after expression.