CAUSAL READOUT DEPENDENCE IN MULTIMODAL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A multimodal model can identify an object when choosing among named alternatives yet fail to mention the same object when asked to describe an image. We ask whether this difference arises only from competition among possible outputs, or whether image-conditioned internal states themselves have different causal effects across readouts. Using Qwen, Mistral, and Llama vision–language models, we identify hidden states at text-token positions that change substantially with the image while preserving the concept named in text. We then transplant these image-conditioned states into otherwise image-absent runs and measure their effects across readouts. Under forced choice, late-layer transplantation recovers 89.1% of the image effect in Qwen, 82.8% in Mistral, and 48.0% in Llama. At the next-token position of a generation prompt, the same intervention recovers only 14.8%, 0.3%, and 7.4% of the image-induced increase in probability for the depicted object. Transplanting states across layers 0–31 increases recovery but leaves a large gap, and the depicted object is detected in none of 464 neutral transplanted captions per model family. Controls show that this difference is not explained by answer-space size, candidate naming, scoring procedure, or transplant coverage. Conversely, removing continued access to the image sharply disrupts generation while preserving much of the relative structure of scored decisions. We call this **causal readout dependence**: an internal intervention can have a large effect under one readout and a much smaller effect under another. The gap is already present before the first token is generated, and competition during generation widens it. Causal effects established under one multimodal readout therefore should not be assumed to transfer to another.

1 INTRODUCTION

Multimodal language models can be asked what they see in very different ways. A model may score candidate answers, respond to a yes/no question, or freely describe an image. These outputs are often treated as different measurements of the same visual capability, but they place different demands on the model. A scored decision compares alternatives that have already been supplied. Free generation must decide what is worth saying, then produce an answer token by token. Differences between the two could therefore arise because less visual evidence reaches the output in one setting, because the evidence faces different competition, or both.

This question is closely related to a familiar limitation of representation analysis. Information can be recoverable from a hidden state without being used to produce the model's output (Elazar et al., 2021). Causal interventions address this limitation by changing internal representations and measuring the downstream consequences (Geiger et al., 2022; Meng et al., 2022). In vision–language models, related methods have been used to trace visual information and alter multimodal predictions or hallucinations (Palit et al., 2023; Yang et al., 2025; Li et al., 2026).

Recent work further suggests that representation and expression can come apart. Visual information can remain internally available when a model's final answer is wrong or incomplete (Liu et al., 2026; Ashok et al., 2025; Nooralahzadeh et al., 2026), can migrate into representations at text-token positions during multimodal processing (Lin et al., 2025; Zhang et al., 2025), and can support visible objects that nevertheless go unmentioned in generated descriptions (Ko et al., 2026). What remains

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.

unclear is whether a representation shown to causally affect one kind of output has a comparable effect when the same model is queried differently.

We test this using matched activation transplants across scored and generative readouts. Image-conditioned hidden states from one run are transplanted into an otherwise image-absent run, allowing us to measure how much of the image's behavioral effect those states can restore. We call a change in the effect of this intervention across ways of querying the model causal readout dependence.

We first identify hidden states at text-token positions that change substantially with the image while retaining the concept named in text. Using THINGS concepts and their human behavioral similarity structure (Hebart et al., 2019; 2020; 2023), we find that images reshape relationships among concept representations without simply replacing the text-named concept with the object depicted in the image. This gives us image-sensitive states in which the text-named concept remains preserved, allowing us to test what those states causally contribute to different outputs.

The resulting difference across readouts is large. At layer 31, transplanting these states into image-absent runs recovers 89.1% of the image-induced forced-choice effect in Qwen, 82.8% in Mistral, and 48.0% in Llama. At a generation position, recovery of the depicted-object token probability falls to 14.8%, 0.3%, and 7.4%. The gap remains even when states are donated across layers 0–31.

Crucially, the difference is already visible before any word is generated. Transplantation increases the depicted object's probability and improves its vocabulary rank, but much less than presenting the image itself. The remaining signal then faces competition from the rest of the vocabulary as generation unfolds. The depicted object is detected in none of 464 neutral transplanted captions per model family, despite appearing in roughly two-thirds of image-present captions. Controls show that this is not simply a closed- versus open-vocabulary effect: once the transplant includes the position whose output probabilities are scored, open-vocabulary recovery becomes substantial. Conversely, removing continued access to the image sharply disrupts generation while preserving much of the relative ordering of scored decisions.

If causal effects depend on readout, evaluations that query visual capabilities differently may also disagree. On the same images, yes/no object-presence questions and free-form hallucination measures produce different model orderings and weak item-level agreement, and an intervention can substantially affect one readout without measurably affecting another. We do not claim that the transplant or continued-visual-access results explain this benchmark disagreement. Rather, these experiments show why transfer across readouts should be tested rather than assumed.

Our contributions are threefold:

1. We identify text-token hidden states whose relationships are reshaped by visual input while the concept named in text remains preserved.

2. We show that matched activation transplantation restores substantially more image-conditioned evidence under forced choice than at a generation position, with the remaining signal failing to produce detectable source mentions in neutral captions.

3. We show that discriminative and generative evaluations can expose different failures and intervention effects, motivating explicit tests of whether findings transfer across readouts.

2 Related Work

2.1 Causal analysis of internal representations

Probing can show that information is decodable from a hidden state, but not whether that information contributes to behavior (Elazar et al., 2021). Causal methods address this distinction more directly. Interchange interventions replace internal representations across computations (Geiger et al., 2022), while causal tracing restores selected activations and measures their effect on model output (Meng et al., 2022). However, conclusions from activation patching can themselves depend on the intervention site, corruption procedure, and metric used (Zhang & Nanda, 2024).

2.2 Visual evidence and behavioral output in multimodal models

Mechanistic studies increasingly show that visual information reaches hidden states at text-token positions. Causal tracing has been adapted to vision–language models (Palit et al., 2023) adapted causal tracing to a vision-language model; visual tokens can become redundant after their information migrates into text tokens (Lin et al., 2025); cross-modal information can move through successive stages from visual input to linguistic representations and output positions (Zhang et al., 2025); and object information can be manipulated directly through causal tracing and representation injection (Li et al., 2026).

A related line of work shows that internal visual information can coexist with wrong or incomplete output. Models can process relevant visual content yet answer incorrectly (Liu et al., 2026), exhibit internal-state signatures associated with reliable versus unreliable visual-reference responses (Ashok et al., 2025), and retain visual attributes even when language priors determine the response (Nooralahzadeh et al., 2026). Internally supported visible objects can also remain absent from generated descriptions (Ko et al., 2026). These findings establish that representation and expression can diverge. They do not, however, establish whether a representation that causally affects one readout has a comparable effect under another, or whether cross-readout differences arise before output selection, during it, or both.

2.3 Multimodal evaluation across readouts

Multimodal failures are evaluated through different output formats. CHAIR measures hallucinated objects in free-form captions (Rohrbach et al., 2018), whereas POPE uses discriminative presence/absence questions (Li et al., 2023); newer benchmarks such as ODE explicitly combine generative and discriminative settings (Tu et al., 2025). Hallucination interventions are likewise evaluated across heterogeneous response formats, including Visual Contrastive Decoding, OPERA, and hidden-state or attention interventions (Leng et al., 2024; Huang et al., 2024; Yang et al., 2025; Su et al., 2025). Whether effects established under one readout transfer to another is rarely tested directly.

3 Experimental Setup

3.1 Models, concepts, and conditions

We study Qwen3-VL-8B-Instruct, Mistral-Small-3.1-24B-Instruct, and Llama-3.2-11B-Vision-Instruct, all frozen at inference time. Our main concept set contains 1,854 THINGS object concepts. Each concept is paired with one fixed matched image and, through a fixed mismatch map, with an image depicting a different concept. Within each model, the text is held fixed while the image is varied among matched, mismatched, degraded, and blank conditions. Full checkpoint identifiers, prompts, preprocessing, and degradation procedures are given in Appendix A.

The causal experiments use the same 464 fixed target–source pairs in all three model families. We call the concept named in the text the target and the object depicted by the mismatched image the source. No items are filtered based on baseline performance or intervention outcome.

3.2 Language-state representations

We analyze hidden states from the language backbone rather than the vision encoder. For each input, we average the hidden states across the complete token span corresponding to the named concept. In multimodal conditions, these states are measured after visual information has entered the language-model computation.

To measure how the relationships among concepts change, we compute pairwise cosine distances between concept representations and compare the resulting distance structure with human THINGS similarity judgments using Spearman representational similarity analysis (RSA). Aggregate analyses use the final half of the available layers. Layer-resolved results, additional reference spaces, and geometry controls are reported in Appendix B.

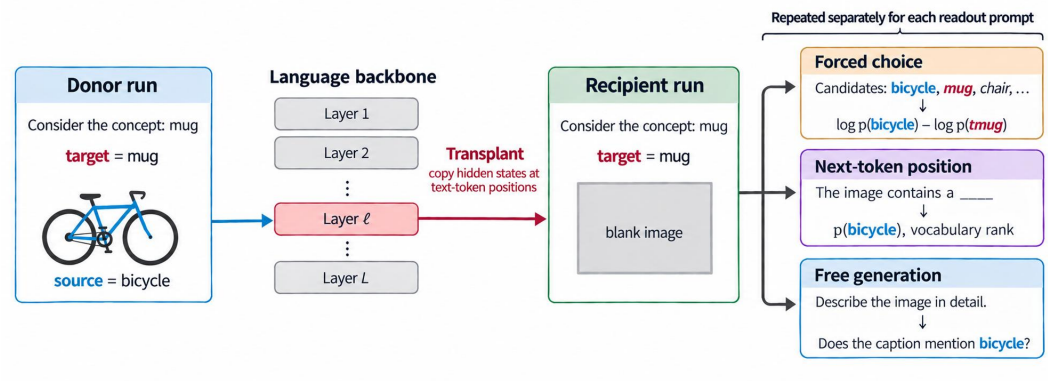

Figure 1: The donor contains the text-named target paired with a mismatched image depicting the source; the recipient contains the same text paired with a blank image. Hidden states at text-token positions are copied from donor to recipient. The procedure is performed separately under forced choice, the fixed next-token assay, and free generation. The intervention rule is matched across readouts, while donor activations are specific to the prompt used for each readout.

3.3 Activation transplantation

The donor run contains text naming the target together with a mismatched image depicting the source. The recipient contains the same text paired with a blank image. At transformer layer ℓ, the corresponding donor hidden states replace the recipient states:

$$h^{(\ell)}_{\text{recipient}}[p] \leftarrow h^{(\ell)}_{\text{donor}}[p].$$

Only language-backbone hidden states are replaced; image-token states and attention caches are left unchanged. The default intervention is performed at layer 31 and replaces all text-token states after the visual block. A concept-span-only condition restricts the intervention to the tokens naming the target. In the multi-layer condition, the corresponding donor states are inserted at every layer from 0 through 31. Across readouts, we match the item set, donor and recipient conditions, intervention layer, span, and replacement rule. Because the prompts differ, donor activations are captured separately under the prompt used for each readout.

The mismatch design isolates image-conditioned changes from the concept supplied in text. Donor and recipient contain identical text; their relevant difference before transplantation is therefore the image. The donor sees the source image, whereas the recipient sees a blank image. A post-transplant shift toward the source can thus arise from image-conditioned changes carried by the donated hidden states, not from introducing the source through the textual prompt.

3.4 Readouts and recovery

Forced choice presents the source, target, and four fixed decoys. Each candidate is scored by the sum of its token log-probabilities. The primary endpoint is

$$Y = \log p(\text{source}) - \log p(\text{target}).$$

Neutral free generation uses the prompt "Describe the image in detail." with greedy decoding, a 96-token maximum, and EOS stopping. Source mentions are detected using alias-aware matching.

For any endpoint Y, we summarize transplantation strength as

$$\text{Recovery} = \frac{\overline{Y}_{\text{transplant}} - \overline{Y}_{\text{blank}}}{\overline{Y}_{\text{image}} - \overline{Y}_{\text{blank}}}.$$

Recovery is computed from condition means; it is undefined when the denominator is zero.

3.5 Measuring source evidence before generation

Forced-choice margins and caption mentions are not directly comparable: the former is a continuous score, whereas the latter is the outcome of an autoregressive sequence. We therefore introduce an intermediate assay that measures evidence for the depicted object *before* any new token is generated. We score the next token after the generation-style frame:

> Describe the image in detail. The image contains a

At this position, the source token competes against the full vocabulary. We measure its log-probability and vocabulary rank under blank, image-present, layer-31 transplant, and multi-layer (0–31) transplant conditions. Generation-position recovery is defined as

$$R_{\text{gen}} = \frac{\overline{\log p_{\text{transplant}}(\text{source})} - \overline{\log p_{\text{blank}}(\text{source})}}{\overline{\log p_{\text{image}}(\text{source})} - \overline{\log p_{\text{blank}}(\text{source})}}.$$

This assay measures how much image-induced evidence for the source reaches a generation position without conflating that evidence with the subsequent decoding trajectory.

3.6 Controls, continued visual access, and statistics

Prompt controls vary whether candidate names are supplied, whether the model is explicitly asked to choose, whether the target name is present, and how the output is scored. A separate coverage control ensures that the transplant includes the position from which output probabilities are computed.

To test whether generation continues to depend on the image after prompt processing, we first allow the model to process the image normally. During continuation, we then prevent newly generated text from accessing the visual input while preserving the language states already computed during the prompt. Architecture-specific implementations are described in Appendix G.

The unit of analysis is the target–source pair ($n = 464$ per model family). Continuous effects use paired bootstrap resampling over items with 95% percentile intervals; zero caption counts use exact binomial intervals. Model families are treated as replications rather than pooled observations. Primary causal tests and the main prompt, scoring, continued-access, and coverage controls were preregistered or registered before the corresponding results were inspected; the analysis timeline is reported in Appendix J.

4 Results

4.1 Visual grounding reorganizes text-token representations while preserving concept identity

Matched images increased alignment with the human THINGS similarity structure in Qwen and Mistral: late-layer RSA rose from 0.1958 to 0.2723 in Qwen and from 0.1800 to 0.3389 in Mistral. Llama showed a smaller increase, from 0.2133 to 0.2289. Increased alignment with the external THINGS space is therefore not a uniform consequence of adding an image.

A more consistent cross-family pattern appeared in the models' own representational geometry. In later layers, a mixture analysis comparing image-conditioned and prompt-conditioned geometry assigned image weights of 0.87–0.92. At the same time, the identity of the text-named concept was retained at greater than 0.99, and representations almost never switched to the concept depicted by the mismatched image.

Mismatched images make this separation especially clear. The text-named target remained almost always the nearest concept representation from the matched-image condition (median rank 1), while the depicted source moved closer than a random non-target. Images therefore reshape relationships among concept representations without simply replacing the concept supplied by text. The same broad pattern appeared for imSitu and MIT-States (Appendix B). We use these image-sensitive, text-anchored states in the causal experiments that follow.

4.2 IMAGE-CONDITIONED TEXT-TOKEN STATES CAUSALLY SUPPORT SCORED DECISIONS

At layer 31, transplanting image-conditioned states into blank-image runs recovered 89.1% of the image effect in Qwen (95% CI 87.7–90.4%), 82.8% in Mistral (81.1–84.5%), and 48.0% in Llama (45.4–50.5%).

The transplant also preserved item-level differences in the image effect. In Qwen, transplant and image-present forced-choice margins were strongly correlated across items ($\rho = 0.9643$); the association remained high after controlling for the blank-image margin ($\rho = 0.9468$). Items strongly affected by the image therefore tended also to be strongly affected by transplantation.

Recovery emerged late in the network. In Qwen, it rose from 20.9% at layer 16 to 85.6% at layer 28, 89.1% at layer 31, and 97.3% at layer 34. The effect was also distributed across the text context rather than confined to the concept name: donating only the concept span recovered 18.9% of the image effect at layer 31, compared with 89.1% when all text-token states after the visual block were donated. Full depth sweeps, direct activation edits, and additional consistency analyses are reported in Appendix C.

Thus, these states do more than contain decodable visual information: transplanting them restores a large fraction of the image's effect on scored decisions.

4.3 MATCHED TRANSPLANTS HAVE A MUCH SMALLER EFFECT AT A GENERATION POSITION

Forced-choice recovery and caption mentions are not directly comparable: one is a continuous score and the other is the outcome of an autoregressive sequence. We therefore first ask how much image-induced evidence for the depicted object reaches a generation position before any new token is emitted.

At layer 31, recovery of the image-induced increase in source-token log-probability was 14.8% in Qwen (95% CI 14.3–15.4%), 0.3% in Mistral (0.2–0.4%), and 7.4% in Llama (7.1–7.8%). These values are far below the corresponding forced-choice recoveries of 89.1%, 82.8%, and 48.0%. Donating states across layers 0–31 increased generation-position recovery to 16.7% (15.1–18.2%), 33.1% (30.5–35.7%), and 12.8% (12.1–13.4%), respectively, but did not close the gap.

Vocabulary rank showed the same pattern. Transplantation improved the source rank in every model family, but substantially less than presenting the image itself (Appendix D). In Qwen, for example, mean source-token log-probability was -18.03 with a blank image, -16.64 after the layer-31 transplant, and -8.63 with the image present. The transplant therefore carries measurable source evidence to the generation position, but recovers only a small fraction of the image-present effect.

This partial recovery did not translate into spontaneous naming. The depicted source was detected in 64.7–70.5% of image-present captions but in 0/464 neutral transplanted captions in each model family, after either layer-31 or multi-layer donation (exact 95% upper bound 0.79%). An auxiliary target-named generation control confirms that transplantation can alter generated output, although this control uses a different item inventory and is treated as an implementation check rather than a quantitative recovery estimate.

The three measurements therefore reveal a progression that captioning alone would obscure: strong causal recovery when the model compares supplied alternatives, substantially weaker recovery before generation begins, and no detected source mentions during neutral captioning. We call this **causal readout dependence**. The cross-readout difference is therefore already present before the first generated token; subsequent token competition can widen that gap.

4.4 CONTROLS RULE OUT SIMPLER EXPLANATIONS

Several alternatives could produce an apparent readout difference without reflecting a difference in the effect of the transplanted states. Table 1 summarizes the controls that address these explanations.

The coverage control substantially changed the interpretation of the initial open-vocabulary result. The original experiment showed only single-digit recovery across model families, suggesting a possible restricted- versus open-vocabulary boundary. However, the transplant stopped before the

Table 1: Controls for simpler explanations of causal readout dependence.

Alternative explanation	Control	Outcome
Restricted candidate sets drive high recovery	Open-vocabulary scoring with the scored position included in the transplant	Recovery: 73.0% / 81.0% / 73.5% in Qwen / Mistral / Llama
Candidate names or explicit choice instructions create the effect	Vary candidate names, choice instructions, target-name availability, and comparison wording	Cross-readout gap remains
Candidate scoring itself creates the effect	Candidate-scored versus free-position recovery under the same prompt	90.42 vs 90.95% / 84.81 vs 86.30% / 49.94 vs 49.77%
Layer 31 is simply unfavorable for generation	Donate layers 0–31	Generation-position recovery rises to 16.7% / 33.1% / 12.8%, but remains well below forced-choice recovery
Transplantation cannot affect generation	Target-named generation positive control	Transplantation changes generated output; treated as an implementation control

position from which output probabilities were read. Once that position was included, recovery rose to 73.0%, 81.0%, and 73.5%. The initial collapse was therefore largely an intervention-coverage artifact rather than evidence that vocabulary size itself eliminates the transplanted signal.

A second set of controls separates prompt composition from scoring. Holding the prompt fixed while changing only the scoring procedure produced nearly identical candidate-scored and free-position recovery in all three families. Candidate names and explicit choice instructions likewise did not account for the remaining difference. The supported claim is therefore narrower than a closed- versus open-vocabulary distinction: image-conditioned states can strongly influence scored comparisons while having a substantially smaller effect at the generation position used for neutral captioning. Full prompt-ladder and coverage results are reported in Appendix F.

4.5 Free generation depends more strongly on continued visual access

The transplant experiments ask whether image-conditioned text-token states can substitute for the image when it is absent. We next ask the complementary question: once those states have formed normally, what happens when the model can no longer access the image during continuation?

Removing continued image access reduced source-object generation by 82–93% across the three model families. Scored decisions also changed, but their relative structure was more stable. Correlations between full-access and access-removed forced-choice margins were 0.908 in Qwen, 0.986 in Mistral, and 0.693 in Llama; the sign of the source-versus-target margin was preserved for 83.6%, 92.5%, and 76.9% of items.

These results do not imply that scored decisions are independent of the image: their score levels also change. Rather, much of their itemwise structure survives. Items that strongly favored one alternative before image access was removed often continued to favor that alternative afterward. Free generation, in contrast, lost most source mentions.

The transplant and access-removal experiments therefore provide complementary evidence. Late text-token states can carry enough image-conditioned information to reproduce much of a scored decision, but they do not fully substitute for continued visual access during generation. The visual stream may refresh or amplify evidence as decoding proceeds, although these experiments do not isolate a single mechanism. Reverse-transplant and attention-based follow-ups are reported in Appendix G as architecture-specific diagnostics.

4.6 Readout dependence extends to multimodal evaluation

We finally asked whether differences across readouts also appear in standard multimodal evaluation. On the same 500 COCO images, yes/no object-presence questions (POPE) and free-caption hallucination scoring (CHAIR) produced different model orderings. POPE accuracy was 0.891 for Qwen,

0.837 for Mistral, and 0.814 for Llama. CHAIR(_i), where lower hallucination rates are better, was 0.064 for Llama, 0.081 for Mistral, and 0.086 for Qwen.

The disagreement also appeared at the item level. Cross-readout coupling was weak ($\phi \approx 0.10$–0.13). In a 1,336-item adversarial comparison, 10 items failed both readouts, compared with 153 POPE-only failures and 10 caption-only failures. The association was greater than expected by chance, but the overlap between failure sets remained small.

Intervention effects also differed across readouts. On the adversarial Qwen set, Visual Contrastive Decoding changed the scored POPE endpoint by +0.084 but changed absent-object caption mentions by only +0.001. Only two items changed under both readouts. Full confidence intervals and contingency analyses are reported in Appendices H and I. This is not a general evaluation of VCD; it demonstrates that an intervention effect measured under one output format need not transfer to another.

We separately tested the stronger hypothesis that the transplant and continued-visual-access results explain the POPE–CHAIR disagreement. The preregistered predictions were mixed or negative (Appendix H), so we do not make that mechanistic claim. The benchmark result is narrower: discriminative and generative readouts can expose related but non-interchangeable aspects of multimodal behavior.

5 Discussion

The experiments separate three questions that are easy to collapse into one: whether visual information is represented, whether it causally affects behavior, and whether that effect becomes visible in the model's output. The image-conditioned text-token states pass the first two tests under forced choice. Yet matched transplantation restores far less evidence at a generation position, and the depicted object is not detected in any of the neutral transplanted captions. A causal effect demonstrated under one readout therefore does not automatically transfer to another.

The next-token assay shows that part of this difference arises before free-running generation begins. The zero-caption result alone could reflect a thresholding effect: the transplant might restore nearly all of the relevant evidence without making the source win the eventual discrete choice. Instead, transplantation improves both source probability and vocabulary rank while leaving them much closer to the blank condition than to the image-present condition. The cross-readout gap is therefore already present in the evidence reaching the first output position.

What happens afterward places additional demands on that evidence. A scored decision needs enough evidence to shift a comparison among supplied alternatives. Free generation must make a concept competitive with the rest of the vocabulary, select it at an appropriate point, and continue along a trajectory in which that choice remains coherent. One interpretation of the observed pattern is therefore a two-stage process: less image-conditioned evidence reaches the generation position, and the evidence that remains then faces a harder selection problem. Our results are consistent with this account, but they do not identify its underlying mechanism. Prompt-dependent transformations, propagation of image-conditioned information, language priors, gating, repeated visual access, or some combination of these could contribute.

The continued-visual-access experiment provides a complementary view. Prior work shows that visual information can migrate into text-token states during multimodal processing (Lin et al., 2025; Zhang et al., 2025), consistent with the strong forced-choice recovery obtained by transplanting late text-token states. But those states do not fully substitute for the image. When newly generated text can no longer access the visual input, source generation falls sharply, whereas much of the relative structure of scored decisions survives. Continued visual input may refresh or amplify evidence as decoding unfolds, but the present experiments do not isolate that process directly.

The controls are important because they changed the interpretation of an initially stronger result. Our first open-vocabulary experiment appeared to show that transplanted evidence worked only when the answer space was restricted. That conclusion did not survive the coverage-matched control: once the transplant included the position actually used for scoring, open-vocabulary recovery became substantial in all three model families. Candidate names, explicit choice instructions, and the scoring procedure likewise did not explain the remaining difference. The supported result is therefore

narrower, but also more precise: strong causal recovery in a scored context does not imply comparable recovery at the generation position used for neutral captioning.

This distinction matters for multimodal evaluation. A yes/no benchmark asks whether a model selects the correct answer when a particular proposition is presented to it; a captioning benchmark asks whether the model volunteers the corresponding content during free generation. These are related questions, but our POPE–CHAIR comparison shows that their outcomes are not interchangeable. The same caution applies to interventions: VCD substantially changes the scored POPE endpoint without a detectable corresponding change in absent-object caption mentions on the matched test. Claims about grounding, hallucination, or mitigation should therefore remain tied to the readout on which they were established unless transfer is tested directly.

We stop short of claiming a single mechanism for these differences. Preregistered attempts to explain POPE–CHAIR disagreement through the distinction between transplanted states and continued visual access were mixed or negative, while reverse-transplant and attention-based follow-ups did not produce a common explanation across architectures. What generalizes is the causal readout dependence itself: internal states that strongly affect one behavior need not have an equivalent effect when the model is asked to express the same visual information differently. **What a model represents, what that representation can causally support, and what the model ultimately says are related but distinct properties.**

6 LIMITATIONS

Our causal comparison matches the item set, donor and recipient conditions, intervention layer, span, and replacement rule across readouts, but the two readouts do not branch from a single shared hidden state. Because their prompts differ, donor activations are captured separately under each readout-specific prompt. A stronger test would transplant the same hidden state and vary only the downstream readout.

The endpoints also differ. Forced choice measures the log-probability difference between the complete source and target strings, whereas the generation-position assay measures the log-probability of the depicted object's first name token. Same-prompt scoring controls reduce this concern, but an ideal comparison would use the same evidence metric across readouts. The fixed generation frame, `The image contains a`, may also differ from the context in which a free-running caption would naturally mention the source. Prefix-clamped analyses could more directly distinguish attenuation of source evidence from divergence of the subsequent generation trajectory.

Finally, the three model families differ in architecture and effect size, and neutral captioning represents only one form of free generation. The continued-visual-access intervention also requires architecture-specific implementations because the models route visual information differently. Our results therefore support a cross-readout phenomenon rather than a single universal circuit mechanism. Nor do they establish that this mechanism explains the POPE–CHAIR disagreement: the registered tests of that stronger hypothesis were mixed or negative.

7 CONCLUSION

Image-conditioned text-token states can recover most of an image-induced forced-choice effect while restoring far less evidence for the depicted object at a generation position. This gap is already visible before the first token is generated and becomes larger during free generation. Continued access to the image also matters more for generation than for the relative structure of scored decisions.

We call this **causal readout dependence**: the behavioral effect of an internal intervention can change substantially with the way a model is queried. For multimodal evaluation, the implication is that evidence established under one output format should not be assumed to transfer to another. What a model represents, what that representation can causally support, and what the model ultimately expresses are related but distinct questions.

AI USE STATEMENT

Generative AI coding agents were used to assist with implementing and executing experimental and analysis pipelines, debugging code, LaTeX formatting, proofreading, and figure and table preparation. The authors specified the research questions and experimental design, reviewed the analysis code and outputs, and independently evaluated the resulting scientific claims and conclusions. Generative AI was not used to generate experimental data. All AI-assisted analyses and manuscript edits were reviewed and verified by the authors.

REFERENCES

Dhananjay Ashok, Ashutosh Chaubey, Hirona Jacqueline Arai, Jonathan May, and Jesse Thomason. Can VLMs recall factual associations from visual references? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15691–15708, 2025. doi: 10.18653/v1/2025.findings-emnlp.850.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338, 2022.

Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):e0223792, 2019. doi: 10.1371/journal.pone.0223792.

Martin N. Hebart, Charles Y. Zheng, Francisco Pereira, and Chris I. Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11):1173–1185, 2020. doi: 10.1038/s41562-020-00951-3.

Martin N. Hebart, Oliver Contier, Lina Teichmann, Adam H. Rockter, Charles Y. Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I. Baker. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, 2023. doi: 10.7554/eLife.82580.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024. doi: 10.1109/CVPR52733.2024.01274.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):31645–31653, 2026. doi: 10.1609/aaai.v40i37.40431.

Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5334–5342, 2025. doi: 10.1609/AAAI.V39I5.32567.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Hanqing Lu, Benoit Dumoulin, and Hanghang Tong. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in VLMs. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.17771.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards vision-language mechanistic interpretability: A causal tracing tool for BLIP. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 2856–2861, 2023.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 12964–12974, 2025. doi: 10.18653/v1/2025.acl-long.634.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *International Conference on Learning Representations*, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=Hf17y6u9BC`.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19781–19791, 2025. doi: 10.1109/CVPR52734.2025.01842.

A FULL EXPERIMENTAL DETAILS

This appendix specifies the implementation choices required to reproduce the experiments in the main paper. The causal experiments use the multimodal checkpoints listed below; paired text-only checkpoints are used only for representation and reference-space analyses.

A.1 MODELS, CHECKPOINTS, HARDWARE, AND PRECISION

The three multimodal families are <Qwen/Qwen3-VL-8B-Instruct>, <mistralai/Mistral-Small-3.1-24B-Instruct-2503>, and <meta-llama/Llama-3.2-11B-Vision-Instruct>. Their paired text checkpoints are <Qwen/Qwen3.5-9B>, <mistralai/Mistral-Small-24B-Instruct-2501>, and <meta-llama/Llama-3.1-8B-Instruct>. Hidden-state extraction uses deterministic forward passes, with bfloat16 inference where supported. Representation extraction was run on NVIDIA A40 GPUs; intervention, continued-visual-access, and benchmark experiments used NVIDIA L40S or A100 GPUs. The 24B Mistral model was sharded across two GPUs when required. No model parameters are updated.

Language-backbone layer indices are zero-based. The extracted multimodal layer ranges are 0–36 for Qwen and 0–40 for both Mistral and Llama. The default transplant is performed at layer 31 in all three model families. The Qwen depth diagnostic additionally evaluates layers 8, 16, 28, and 34, while the multi-layer intervention inserts donor states at every layer from 0 through 31. Exact model resolution, processor, tokenizer, and run metadata are retained with the corresponding machine-readable outputs.

A.2 IMAGES AND CONCEPTS

The main representation analysis contains 1,854 THINGS object concepts. Concepts are selected from the THINGS behavioral concept list and retained when a matched image is available in the fixed image manifest. The manifest assigns one matched image to each concept and records its source and preprocessing status. All analyses use a fixed concept ordering.

For the causal experiments, a fixed mismatch map assigns each target concept a different source concept. Self-pairs and pairs without available images are removed. For each retained target–source pair, two target-neighbor and two source-neighbor decoys are selected from the THINGS behavioral similarity matrix, excluding the target and source and preventing duplicate candidates. This procedure yields the same 464 fixed target–source pairs for each model family. The target is the concept named in the text; the source is the concept depicted in the mismatched image.

Different analyses use separate prespecified inventories. The primary transplant and neutral-caption analyses use 464 pairs per model family. Open-vocabulary controls use 368 Qwen pairs, 369 Mistral pairs, and 368 Llama pairs, selected in advance to avoid first-token overlap. Direct activation-edit analyses use the full 1,854-concept inventory. Benchmark analyses use 500 COCO images, while the adversarial POPE–caption and VCD analyses use 1,336 ground-truth-absent object items. These inventories are analyzed separately and are not treated as replications of one another.

Images are converted to RGB and resized with aspect ratio preserved to a maximum side length of 384 pixels before being passed through the checkpoint-specific Hugging Face processor. Degraded images are first downsampled to 25% of their original linear resolution, converted to grayscale, upsampled, and Gaussian-blurred with radius 2. Blank images are uniform gray images passed through the same processor. Matched conditions use the target's fixed image, whereas mismatched conditions use the source's fixed image. Image-token states are never transplanted.

A.3 PROMPTS AND READOUTS

The primary forced-choice analysis uses six candidates: the source, target, two target-neighbor decoys, and two source-neighbor decoys. These are inserted into:

```
The named object is "target." Which concept best
matches the object being represented?

A. candidate_1 ...  F. candidate_6

Answer:
```

Each candidate completion is scored by the sum of its token log-probabilities. The primary endpoint is the source-minus-target margin. A separate two-option template is used only for auxiliary pairwise or control analyses where explicitly stated:

```
The named object is "target." Based on the visual
evidence, which option better matches the object?  A.
option_a B. option_b Answer:
```

The primary neutral free-generation prompt is exactly:

```
Describe the image in detail.
```

Decoding is greedy with `max_new_tokens=96` and EOS stopping. One caption is generated per target–source pair and condition. The target concept is not named in this prompt.

The fixed next-token assay uses the generation context followed by:

```
The image contains a
```

Before any continuation is generated, we score the first token of the source name at this position against the full vocabulary. This assay measures source evidence at a generation position without including the subsequent autoregressive trajectory.

A separate concept-named descriptive prompt,

```
In 10–20 words, describe the visible appearance,
texture, and sensory details of "concept." Do not
define it.  Do not include reasoning.
```

is used only for representation-oriented or auxiliary analyses where explicitly stated. It is not used for the primary neutral-caption result.

The POPE prompt is:

```
Is there a name in the image?  Answer with Yes or No.
```

The POPE endpoint is the Yes-minus-No log-probability margin.

A.4 Span matching and transplantation

The named concept is tokenized without special tokens using the corresponding model tokenizer. Its token-ID sequence is located as a contiguous subsequence of the fully rendered prompt or chat-template input. We use the first complete occurrence corresponding to the explicit concept mention. If this span cannot be resolved, the item is excluded rather than assigned a broader heuristic span.

The donor contains the target-naming text paired with the mismatched source image. The recipient contains identical text paired with a blank image. At layer ℓ, donor language-backbone block outputs replace the corresponding recipient text-token states:

$$h^{(\ell)}_{recipient}[p] \leftarrow h^{(\ell)}_{donor}[p].$$

The transplant uses direct replacement without interpolation, scaling, or averaging. Only text-token positions are replaced; image-token states, attention masks, and key–value caches are unchanged. The default intervention replaces the full text suffix after the visual block at layer 31. The concept-span condition restricts replacement to the tokens corresponding to the named target. The multi-layer condition performs the corresponding donor replacement independently at every layer from 0 through 31. Sham same-image donations and concept-span interventions serve as implementation and localization controls.

For forced choice, each candidate's completion log-probability is scored and the primary endpoint is $Y = \log p(source) - \log p(target)$. Recovery is calculated from aggregate condition means as the

transplanted change relative to blank divided by the image-present change relative to blank. For the generation-position assay, the first token of the source name is scored against the full vocabulary at the fixed generation frame. This assay is distinct from free-running caption generation.

A.5 Mention detection and statistics

Source mentions are detected using case-insensitive, alias-aware lexical matching after normalization of punctuation and whitespace. The matcher includes singular and plural variants, common irregular forms, and multiword object names. Matches must respect token boundaries and cannot occur solely as substrings of unrelated words. The same detector is applied to blank, image-present, transplant, and access-removed captions.

The primary mention counts are outputs of this automated lexical detector. We have not yet used independent blinded annotation to estimate its sensitivity or specificity. Paraphrastic references not covered by the alias set may therefore produce false negatives, a limitation that is particularly relevant to the 0/464 source-mention result. Appendix E describes the prepared blinded validation procedure.

Continuous effects use paired bootstrap resampling over target–source items with percentile 95% confidence intervals. Zero caption counts use exact binomial intervals. Prespecified causal grids use Benjamini–Hochberg correction within the corresponding stated family of tests. Model families are analyzed as replications rather than pooled observations.

Recovery is computed from aggregate condition means rather than by averaging item-level ratios. Values above 100% are therefore possible when the transplant mean exceeds the image-present mean, and recovery is undefined when the aggregate denominator is zero. Geometry analyses use a separate bootstrap procedure in which concepts are resampled and the corresponding dissimilarity matrices are reconstructed for each bootstrap sample.

A.6 Architecture-specific continued-visual-access intervention

The continued-visual-access intervention allows the model to process the image and prompt normally during prefill, then prevents newly generated text from accessing the image during continuation while preserving the language states already computed from the image-conditioned prompt.

For Qwen and Mistral, image-token key positions are blocked in the continuation self-attention mask while the precomputed language states and key–value cache are retained. Explicit position indices are supplied to preserve the multimodal rotary-position scheme used during standard generation. The implementation is validated by teacher-forcing the model's standard greedy sequence through the manual continuation loop.

Llama routes visual information through dedicated cross-attention decoder layers rather than through the same token-stream pathway. Its continuation intervention therefore removes the relevant cross-attention contribution using forward hooks after normal image-conditioned prefill while preserving the prefilled language states. The Llama implementation is validated using corresponding prefill–continuation and teacher-forcing checks. Neither implementation transplants a key–value cache.

B Full geometry analysis

B.1 THINGS RSA across conditions

The geometry analysis compares text-only, sensory-prompt, matched-image, mismatched-image, degraded-image, and blank-image conditions. Concept representations are obtained by mean-pooling hidden states across the text-token span corresponding to the named concept. For each condition and layer, pairwise cosine distances between concept representations form a representational dissimilarity matrix (RDM). We compare these RDMs with external reference RDMs using Spearman representational similarity analysis (RSA).

Figure B1 shows layer-resolved trajectories for the sensory-prompt, VLM text-only, and matched-image conditions against four reference spaces: THINGS, controlled THINGS, SigLIP2, and Lancaster perceptual similarity. The complete condition-by-layer results, including mismatched, degraded, and blank-image controls, are provided with the supplementary machine-readable outputs.

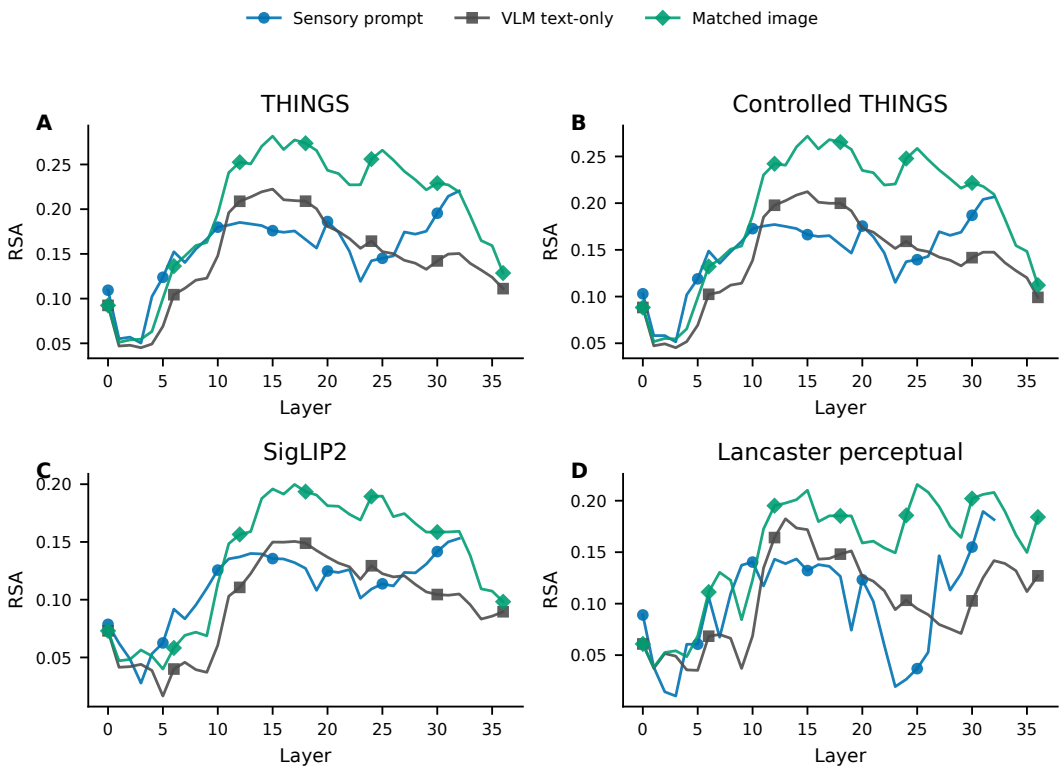

Figure 2: Layer-resolved RSA trajectories for sensory-prompt, VLM text-only, and matched-image conditions across THINGS, controlled THINGS, SigLIP2, and Lancaster perceptual reference spaces.

Table 2: Layerwise reference-space summary. The first positive layer is the first extracted layer at which the matched-minus-prompt RSA contrast becomes positive. The late–early gap is the difference between the corresponding late- and early-layer contrasts.

Family	Reference space	First positive layer	Late–early gap
Qwen	THINGS	10	+0.0745
Qwen	Controlled THINGS	10	+0.0740
Qwen	SigLIP2	3	+0.0569
Qwen	Lancaster perceptual	2	+0.0811
Mistral	THINGS	3	+0.0312
Mistral	Controlled THINGS	3	+0.0299
Mistral	SigLIP2	3	+0.0158
Mistral	Lancaster perceptual	3	+0.0418
Llama	THINGS	4	+0.0193
Llama	Controlled THINGS	4	+0.0172
Llama	SigLIP2	4	+0.0145
Llama	Lancaster perceptual	4	-0.0046

The Lancaster-perceptual contrast in Llama is the only late–early contrast that is not positive. We therefore treat increased alignment with external reference spaces as strongest in Qwen and Mistral and weaker and reference-dependent in Llama.

Table 3: Position-control contrasts from the 200-concept analysis.

Contrast	Concept span	Length-matched control span
Matched minus text-only	-0.0708	-0.0045
Matched minus blank	-0.0621	-0.0009

B.2 Layer-resolved geometry

The layer-resolved analysis computes RSA separately at every extracted RSA is computed separately at every extracted layer using a fixed concept ordering and reference RDM within each analysis. Bootstrap intervals are obtained by resampling concepts and reconstructing the corresponding RDMs for each bootstrap sample. The primary geometric conclusion is based on the prespecified late-layer pattern rather than selection of an individual layer after inspection.

B.3 Image–prompt decomposition

We additionally characterize how image-conditioned geometry relates to image- and prompt-conditioned reference geometries. At each layer, the mixture analysis estimates an image weight and prompt weight describing the relative geometric resemblance of the image-conditioned representation to these reference structures. Separate measures quantify retention of the text-named concept and reassignment toward the concept depicted by the mismatched image. These quantities describe representational geometry; they should not be interpreted as attention weights, causal contributions, or literal mixture coefficients in the model's computation.

Across later layers, the estimated geometry is predominantly image-associated, with image weights of approximately 0.87–0.92 in the primary analysis. At the same time, local identity of the text-named concept remains above 0.99 and reassignment to the depicted source remains near zero. Full layerwise estimates are included in the supplementary machine-readable outputs.

B.4 Mismatch identity and generalization

The mismatch analysis examines the rank distributions of the text-named target, depicted source, and random non-target concepts relative to the matched-image reference representations. Under mismatched images, the target remains the nearest matched-image concept for nearly all items, while the depicted source moves closer than a random non-target. The image therefore alters the surrounding representational geometry without generally replacing the identity supplied by text.

We repeat the analysis on imSitu and MIT-States as generalization checks rather than independent replications of the causal transplant experiment. For imSitu ($n = 504$ event concepts), the prompt-plus-image mixture yields a matched-image weight of 0.8947, prompt weight of 0.0647, and $R^2 = 0.8690$; text retention is 0.9960 and source assignment is 0. For MIT-States ($n = 2{,}207$ compositional concepts), the corresponding values are 0.9085, 0.0772, and, $R^2 = 0.9493$, with text retention of 0.9583 and source assignment of 0. These extensions support the image-associated, text-retaining geometric pattern but do not imply improved alignment with every external reference space.

The substantially larger changes at the named-concept span than at a length-matched control span argue against a generic token-position or image-prefix effect as an explanation for the representation result.

C Transplant localization and depth

C.1 Depth profile

The default layer-31 result is complemented by a Qwen depth analysis at layers 16, 28, 31, and 34. Forced-choice transplant recovery increases sharply in later layers, from 20.9% at layer 16 to 85.6% at layer 28, 89.1% at layer 31, and 97.3% at layer 34. Layer 31 is used for the primary

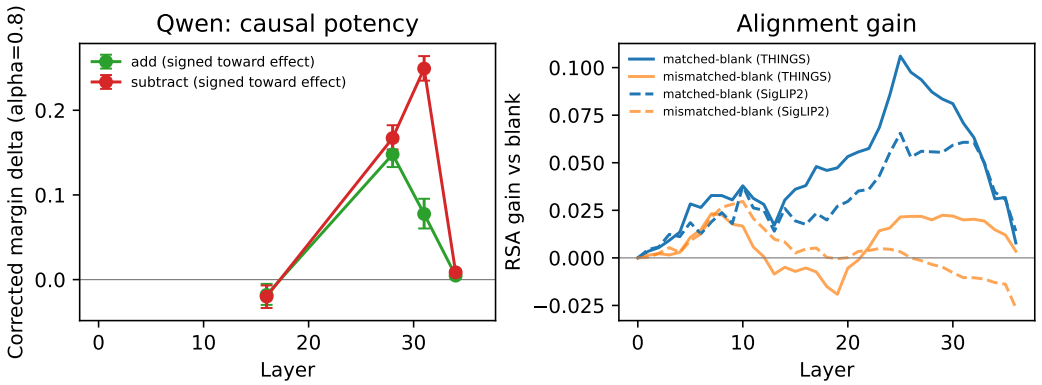

Figure 3: Layerwise intervention profile in Qwen. Forced-choice transplant recovery and control-corrected direct-edit effects are concentrated in later layers. Generation-position effects are reported separately in Appendix D.

Table 4: Qwen control-corrected direct-edit effects across measured layers. Values are additive or subtractive edit effects relative to the corresponding norm-matched control, not full-state transplant recoveries.

Layer	Positive edit	Negative edit
16	-0.0180 [-0.0452, +0.0084]	+0.0200 [-0.0106, +0.0502]
28	+0.1479 [+0.1122, +0.1849]	-0.1672 [-0.2000, -0.1352]
31	+0.0776 [+0.0386, +0.1213]	-0.2492 [-0.2809, -0.2173]
34	+0.0047 [-0.0016, +0.0106]	-0.0086 [-0.0147, -0.0032]

cross-family comparison because it provides a common late-layer intervention point across all three model families.

The direct-edit analysis uses the 1,854-concept inventory and is mechanistically distinct from the 464-pair full-state transplant. Its effect sizes should therefore not be interpreted as independent estimates of transplant recovery.

C.2 Concept-span versus full-suffix transplantation

The concept-span condition replaces only the tokens corresponding to the named target, whereas the full-suffix condition replaces all text-token states after the visual block. In Qwen at layer 31, concept-span transplantation recovers 18.9% of the forced-choice image effect, compared with 89.1% for full-suffix transplantation. The causal effect is therefore distributed across the text context rather than localized to the tokens naming the concept.

C.3 Direct activation edits

vector at the edited text-token span. This vector is defined as the mismatched-minus-blank state difference after removal of the mean generic-image difference and is norm-matched across source-specific, unrelated-source, generic-image, and random control conditions. Primary doses are $\alpha \in \{0, 0.2, 0.4, 0.8\}$, with additional escalation runs where specified. At $\alpha = 0.8$, the Qwen layer-31 subtraction effect relative to the norm-matched random control is -0.2492 [-0.2809, -0.2173], compared with -0.0086 [-0.0147, -0.0032] at layer 34. Positive and negative edits, dose-response analyses, image-effect fractions, promotion/veto decompositions, and identity controls are treated as supplementary mechanistic analyses rather than estimates of the primary transplant effect.

Table 5: Recovery of the image-induced increase in source-token log-probability at the fixed generation position. Values are percentages of the aggregate image-present-minus-blank effect.

Family	Layer 31	Layers 0–31	Forced-choice layer 31
Qwen	14.8%	16.7%	89.1%
Mistral	0.3%	33.1%	82.8%
Llama	7.4%	12.8%	48.0%

Table 6: Mean source-token log-probabilities and full-vocabulary ranks at the fixed generation position.

Family	Blank log p	Image log p	Layer-31 log p	Layers 0–31 log p	Blank rank	Image rank	Layer-31 rank	Layers 0–31 rank
Qwen	-18.026100	-8.631875	-16.635148	-16.457668	7953.679	718.623	7276.155	5178.720
Mistral	-18.553633	-8.205472	-18.525417	-15.130053	13660.300	356.981	12950.747	4896.655
Llama	-14.590970	-5.341178	-13.903259	-13.410891	7498.134	286.550	5385.280	4394.127

C.4 Itemwise transplant agreement

In Qwen, image-present and transplanted forced-choice margins are strongly correlated across items $\rho = 0.9643$. The association remains high after controlling for the blank-image margin ($\rho = 0.9643$l main text).This analysis shows that transplantation preserves much of the itemwise ordering of the image effect; it does not imply that transplantation reproduces the image-conditioned computation in every hidden dimension.

D Generation-position analyses

D.1 Source-token probability and recovery

The generation-position assay scores the first token of the depicted source name before any continuation is generated.

The corresponding absolute log-probabilities and vocabulary ranks are shown below.

Ranks are computed over the full next-token vocabulary, with rank 1 indicating the highest-probability token.

D.2 Source-rank distributions

Across all three model families, transplantation improves the vocabulary rank of the source relative to the blank condition but remains far from the image-present distribution.

D.3 Layer-31 versus layers 0–31

Multi-layer transplantation increases generation-position recovery from 14.8% to 16.7% in Qwen, from 0.3% to 33.1% in Mistral, and from 7.4% to 12.8% in Llama. Despite the substantial increase in Mistral, recovery remains below the corresponding layer-31 forced-choice values in all three families. The generation-position result therefore cannot be explained solely by the choice of a single unfavorable intervention layer.

D.4 Target-named generation control

An auxiliary Qwen control tests whether transplantation can alter generated output when the relevant concept is already named in the prompt. This analysis uses a different inventory ($n = 580$) from the primary 464-pair neutral-caption experiment. We therefore treat it as an implementation control rather than report a normalized recovery estimate alongside the primary results.

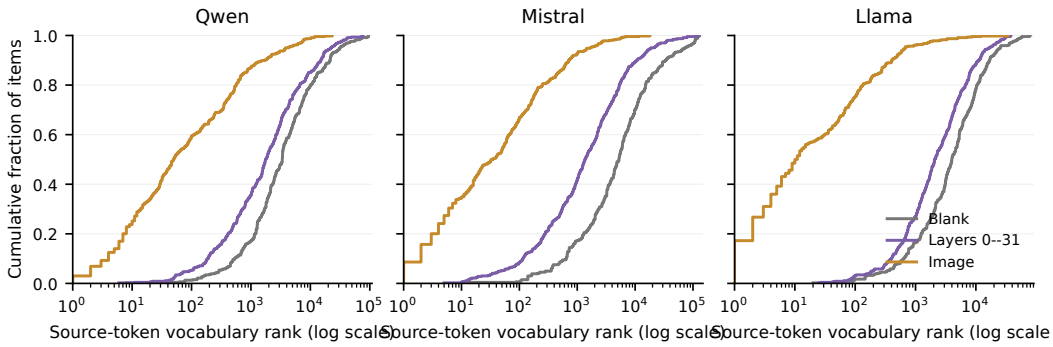

Figure 4: Empirical cumulative distributions of source-token vocabulary rank under blank, image-present, and layers-0–31 transplant conditions. The rank axis is logarithmic because the distributions are strongly right-skewed.

Table 7: Source-token vocabulary-rank quartiles. Values give the 25th percentile / median / 75th percentile.

Family	Condition	Rank quartiles
Qwen	Blank	1304.75 / 3357.00 / 7813.00
Qwen	Image	10.00 / 50.00 / 432.50
Qwen	Layer 31	1205.75 / 2906.50 / 6938.75
Qwen	Layers 0–31	582.50 / 1769.50 / 5019.25
Mistral	Blank	1909.25 / 5268.00 / 12151.00
Mistral	Image	5.00 / 34.00 / 176.25
Mistral	Layer 31	1879.25 / 5009.50 / 11272.00
Mistral	Layers 0–31	383.50 / 1363.50 / 4340.25
Llama	Blank	1684.00 / 4010.00 / 9246.50
Llama	Image	2.00 / 10.50 / 96.75
Llama	Layer 31	1148.75 / 2724.50 / 6652.00
Llama	Layers 0–31	875.25 / 2070.50 / 5224.00

E Free-generation validation

E.1 Prompts and decoding

The primary free-generation prompt is exactly: `Describe the image in detail.` The target concept is not named. Decoding is greedy with `max_new_tokens=96` and EOS stopping. One caption is generated for each target–source pair and condition using the same model, processor, and image conditions as the corresponding causal experiment. Raw generated text is stored before lexical mention detection.

The separate concept-named 10–20-word descriptive prompt is used only in auxiliary analyses and does not contribute to the primary neutral-caption result.

E.2 Mention detector

The mention detector lowercases text, normalizes punctuation and whitespace, and searches for canonical object names and predefined aliases, including singular/plural variants, common irregular forms, and multiword names. Matches must respect word boundaries. The same detector is applied to every experimental condition without condition-specific tuning.

The reported 0/464 counts are therefore lexical-detector results. Paraphrastic references not represented in the alias set may be missed.

Table 8: Manual validation of the lexical mention detector and transplanted-caption result. Detector sensitivity is calculated among image-present captions judged to contain a semantic source mention.

Model	Semantic image-present mentions	Lexical detector hits	Estimated sensitivity	Semantic transplant mentions
Qwen	18/20	13/20	72.2%	0/20
Mistral	19/20	14/20	73.7%	0/20
Llama	19/20	14/20	73.7%	0/20
Pooled	**56/60**	**41/60**	**73.2% (41/56)**	**0/60**

Table 9: Prespecified qualitative examples. These examples are illustrative only and are not used as quantitative evidence.

Target	Mismatched source	Image-present caption	Layers-0–31 caption	Transplant forced-choice margin	Preference
aardvark	whale	Detailed photograph of an aardvark walking across dry terrain.	Uniform gray field with no discernible objects.	4.7832	source
barbell	exerciser	Detailed photograph of a barbell with weight plates.	Solid gray rectangle with no discernible features.	1.1719	source
baton	bacon	Photograph of a wooden walking stick or cane on a gray carpet.	Uniform gray field with no discernible objects.	1.7552	source
blanket	playpen	Detailed photograph of a large woven blanket draped over furniture.	Uniform gray field with no discernible objects.	-1.7614	target
bowl	baton	Detailed photograph of a wide-mouthed ceramic bowl.	Solid gray rectangle with no discernible objects.	-1.5924	target
butterfly	berry	Close-up photograph of a monarch butterfly on a flower.	Uniform gray field with no discernible objects.	4.3125	source

E.3 Manual validation of source mentions

Because lexical matching can miss paraphrastic references to an object, we manually audited 20 image-present and 20 transplanted captions per model family. Captions were evaluated for whether they semantically referred to the relevant source object, independently of whether the canonical name or a predefined alias appeared.

The lexical detector therefore has incomplete sensitivity for semantic source mentions. Several detector misses nevertheless described the depicted object clearly using paraphrases or related surface forms, such as "baby cradle" for *bassinet* and "air mover" for *blower*. Accordingly, the primary 0/464 results are reported as lexical-detector counts rather than literal absence of all semantic references.

Importantly, manual annotation independently showed the same separation between conditions. Semantic source mentions were present in 56/60 (93.3%) audited image-present captions but in 0/60 audited transplanted captions. Thus, although lexical matching underestimates semantic mention rates, its imperfect sensitivity does not appear to explain the absence of detected source mentions after transplantation in this audit.

Six deterministic Qwen examples are also provided for illustration. They were selected at evenly spaced positions after alphabetically sorting the eligible items and were not selected based on intervention outcome.

The image-present captions shown here correspond to the matched target image. The mismatched source identifies the donor image used for transplantation. These examples are illustrative only and are not used as quantitative evidence.

F Alternative explanations and controls

F.1 Open-vocabulary scoring and intervention coverage

Our initial open-vocabulary experiment withheld candidate names and scored the source against the full next-token vocabulary. At layer 31, recovery was only 3.3% in Qwen, 0.6% in Mistral, and 6.9% in Llama. This initially suggested that transplantation might be effective only when the answer space was restricted.

Inspection of the intervention revealed a coverage mismatch: the donated region ended before the position from which output probabilities were scored. We therefore repeated the experiment with the transplant extended to include the scored position. Recovery increased to 73.0% in Qwen, 81.0% in Mistral, and 73.5% in Llama.

F.2 Prompt and scoring controls

We next varied candidate naming, explicit choice instructions, availability of the target name, and comparison wording while holding the image condition, target–source pair, intervention layer, span, and replacement rule fixed.

To separate prompt composition from the scoring procedure itself, we also compared candidate-scored and free-position endpoints under the same prompt. Recovery was nearly identical: 90.42% versus 90.95% in Qwen, 84.81% versus 86.30% in Mistral, and 49.94% versus 49.77% in Llama. The large difference between forced-choice and neutral-generation-position recovery therefore cannot be attributed simply to the algorithm used to score candidate completions.

F.3 Gain-ladder analyses

Additional gain-ladder experiments vary intervention dose and control condition. These analyses include full-image, generic-image, unrelated-source, norm-matched random, concept-span, and full-suffix conditions. Candidate order, intervention span, dose, control arm, and item count are recorded for each analysis. Order-balanced estimates are used when candidate position affects the magnitude of generic-image or unrelated-source controls.

G Continued visual access

G.1 Intervention

The continued-visual-access experiment complements transplantation by asking whether image-conditioned states formed during normal prompt processing are sufficient once direct access to the image is removed.

The image and prompt are processed normally during prefill. During continuation, Qwen and Mistral are prevented from attending to image-token keys through the continuation self-attention mask. Llama routes visual information through dedicated cross-attention layers, so the corresponding visual cross-attention contribution is removed using forward hooks. In all cases, language states already computed during image-conditioned prefill are retained. Implementation and validation details are given in Appendix A.

G.2 Generation and forced-choice effects

Removing continued image access reduced source-mention rates across all three model families.

The corresponding relative reductions in source mentions are approximately 93%, 82%, and 90% for Qwen, Mistral, and Llama. Forced-choice margins also change in level, with architecture-specific direction, so we do not interpret the intervention as preserving the absolute scored effect. Instead, the cross-family result is the persistence of substantial itemwise structure: full-access and access-removed margins remain correlated, and the sign of the source-versus-target margin is preserved for most items.

Table 10: Continued-visual-access intervention. Source-mention rates are proportions. Forced-choice (ρ) is the item-level correlation between full-access and access-removed source-minus-target margins.

Family	Full access	Access removed	Paired change [95% CI]	Forced-choice ρ	Sign preserved
Qwen	0.1552	0.0108	-0.1444 [-0.1789, -0.1121]	0.9083	83.6%
Mistral	0.0366	0.0065	-0.0302 [-0.0474, -0.0151]	0.9857	92.5%
Llama	0.4677	0.0453	-0.4224 [-0.4698, -0.3772]	0.6929	76.9%

Table 11: POPE and CHAIR on the matched 500-image benchmark set. Lower CHAIR values indicate fewer hallucinated objects. The caption-retest ceiling summarizes the attainable item-level coupling given repeated caption-side outcomes.

Family	CHAIR_s	CHAIR_i	POPE accuracy	Cross-readout ϕ	Caption-retest ceiling
Qwen	0.430	0.0864	0.8912	0.0966	0.5768
Mistral	0.394	0.0812	0.8367	0.1324	0.5554
Llama	0.332	0.0643	0.8136	0.1011	0.6644

G.3 Reverse transplantation and preregistered attention analysis

Reverse-transplant experiments in Qwen and Llama provide architecture-specific diagnostics but do not support a common cross-family mechanism.

A preregistered position-controlled analysis testing the proposed relationship between live visual attention and ordinary caption hallucinations was also not supported. The controlled difference was -0.0020 (95% CI $[-0.0056, +0.0015]$). We therefore do not interpret the corresponding raw attention difference as evidence that hallucinated mentions systematically arise from a low-visual-access state.

These negative results constrain the mechanistic interpretation of the continued-access experiment: continued visual input matters for generation, but the present analyses do not identify a universal attention-based mechanism for that dependence.

H Benchmark disagreement

H.1 POPE and CHAIR

We evaluate POPE and CHAIR on the same 500 COCO images. POPE uses yes/no object-presence questions, whereas CHAIR measures object hallucination in freely generated captions. The two benchmarks therefore share visual inputs but differ in how model behavior is elicited and scored.

The model ordering differs across the two evaluation formats: Qwen has the highest POPE accuracy, whereas Llama has the lowest CHAIR hallucination rates. Item-level coupling is weak ($\phi = 0.097$–0.132) relative to the caption-retest ceilings.

H.2 Item-level coupling and contingency analysis

A matched adversarial analysis further compares failures at the item level.

The two failure indicators are associated (odds ratio of 7.60; Fisher exact $p = 3.49 \times 10^{-5}$), but the association is modest in magnitude (Cramer's $V = 0.142$). Only 10 items fail both readouts, compared with 153 POPE-only failures and 10 caption-only failures. The readouts are therefore related, but their failure sets overlap only partially.

H.3 Signal-detection and threshold-matching analyses

We next asked whether the weak item-level correspondence could be explained primarily by different response thresholds.

Table 12: Adversarial POPE–caption contingency table.

	Caption correct	Caption failure
POPE correct	1163	10
POPE failure	153	10

Table 13: Signal-detection and threshold-matching summaries.

Family	Caption d'	Caption c	POPE d'	POPE c	Matched POPE threshold	Matched false-positive rate	Matched ϕ
Qwen	1.79 [1.62, 1.99]	1.27 [1.19, 1.36]	2.26 [2.15, 2.39]	0.05 [-0.01, 0.11]	5.625	0.0142	0.2455
Mistral	1.61 [1.47, 1.76]	0.99 [0.92, 1.07]	1.69 [1.59, 1.81]	-0.27 [-0.33, -0.22]	5.565	0.0359	0.2653
Llama	1.65 [1.52, 1.80]	0.98 [0.92, 1.06]	1.94 [1.82, 2.09]	-0.79 [-0.86, -0.72]	6.064	0.0352	0.2282

Matching false-positive rates increases cross-readout coupling to $\phi = 0.228$–0.265, but the correspondence remains incomplete. Different response thresholds therefore contribute to the benchmark disagreement but do not fully account for it. This analysis does not imply that the two readouts rely on separate neural pathways.

H.4 Preregistered mechanism predictions

We separately tested whether the transplant and continued-visual-access findings explain the POPE–CHAIR disagreement. The preregistered predictions did not provide consistent support for that stronger hypothesis: the continued-access ranking prediction was only partly supported, and the position-controlled ordinary-caption attention prediction was not supported.

We therefore treat the benchmark result as a behavioral cross-readout finding rather than evidence for a specific shared mechanism. Discriminative and generative evaluations can expose different failures, but the present experiments do not establish why their failure sets differ.

I VCD transfer across rollouts

We finally test whether an intervention effect measured under one readout transfers to another. The analysis uses Qwen on 1,336 ground-truth-absent adversarial items with Visual Contrastive Decoding (VCD) strength ($\alpha = 1.0$), contrast coefficient ($\beta = 0.1$), and 500 noise samples. Because all items are ground-truth absent, both endpoints measure false-positive behavior rather than overall discrimination performance.

VCD changes the POPE endpoint substantially, increasing the false-positive rate by 0.0838, while the corresponding change in absent-object caption mentions is +0.0007 with a confidence interval spanning zero. The caption-side estimate is also small relative to its minimum detectable effect.

VCD changes the POPE outcome on 112 items and the caption outcome on 23 items. Only two items change under both readouts.

The expected overlap under independence is 1.9 items, close to the observed overlap of two (Fisher exact $p = 1$). This result illustrates that intervention effects measured under one readout need not produce corresponding item-level changes under another.

Table 14: VCD effects across discriminative and generative readouts. MDE denotes the minimum detectable absolute change for the corresponding binary endpoint.

Endpoint	Baseline	VCD	Paired change [95% CI]	MDE	Change/MDE
POPE false-positive decoded	0.0913	0.1751	+0.0838 [+0.0696, +0.0981]	0.0212	3.94
Caption mentions absent object	0.0157	0.0165	+0.0007 [-0.0060, +0.0082]	0.0101	0.07

Table 15: Items changed by VCD in the two readouts.

	Caption unchanged	Caption changed
POPE unchanged	1199	23
POPE changed	112	2