Dual-Pathway Circuits of Object Hallucination in Vision-Language Models

Jiaxin Liu[*]
University of Illinois Urbana-Champaign

Ding Zhong
University of Michigan, Ann Arbor

Yue Wang
Stanford University

Zhidong Yang
The Hong Kong University of Science and Technology

Zhaolu Kang
Peking University

Guangyuan Dong[*]
National University of Singapore

Qishi Zhan
Marquette University

Pengcheng Fang
University of Southampton

Aofan Liu
Peking University

Abstract

Vision-language models (VLMs) have demonstrated remarkable capabilities in bridging visual perception and natural language understanding, enabling a wide range of multimodal reasoning tasks. However, they often produce object hallucinations, describing content absent from the input image, which limits their reliability and interpretability. To address this limitation, we propose Dual-Pathway Circuit Analysis, a framework that identifies and characterizes hallucination-related circuits in VLMs for mechanistic understanding and causal probing. We first apply activation patching across five architecturally diverse VLMs to identify a visual grounding pathway that supports correct predictions and a hallucination pathway that drives erroneous outputs. We then introduce Conditional Pathway Analysis (CPA) to characterize pathway-level interactions, revealing that grounding components remain strongly redundant in both correct and hallucinating samples but undergo a consistent polarity flip, shifting from supporting the ground truth on correct samples to aligning with the hallucinated answer on erroneous ones. We further perform targeted suppression of hallucination-pathway components, showing that scaling these components reduces object hallucination by up to 76% with minimal accuracy cost, and validate that the same circuit selectively transfers to relational but not attribute hallucination. Evaluations on POPE-adversarial and AMBER show that the identified circuits are consistent across architectures, support causal intervention, and transfer selectively across hallucination types. Code is available at `https://github.com/jiaxin26/DualPath-VLM`.

1 Introduction

Vision-Language Models (VLMs) demonstrate strong visual reasoning [38, 12] and cross-modal understanding [6, 26]. Representative systems such as Qwen-VL [1, 2] and GPT-4 [22] consequently deliver strong performance on image captioning [40], visual grounding [3], and video understanding [39]. However, VLMs often produce object hallucinations, describing entities, attributes, or relations absent from the input image. Such outputs reflect over-reliance on text-centric priors rather than visual evidence, limiting model reliability and interpretability. Among hallucination types,

[*]Corresponding authors: Jiaxin Liu (`jiaxin26@illinois.edu`) and Guangyuan Dong (`guangyuan@u.nus.edu`).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

object-existence questions provide a particularly clean setting: their binary ground truth directly tests whether the model grounds its answer in visual evidence, while avoiding many confounds introduced by attribute or relational reasoning. This makes them a natural entry point for causal probing of hallucination-related computations. Benchmarks including POPE [15] and CHAIR [24] systematically document this problem. Existing mitigation strategies span training-time alignment [30, 37, 27], decoding-time constraints [9, 11], and post-hoc corrections [35]. These approaches primarily treat VLMs as black boxes, modifying input signals or decoding procedures without investigating the internal computations behind hallucination. As a result, these methods are often model-specific and generalize poorly across architectures.

Mechanistic interpretability offers a complementary perspective, asking not only whether hallucinations occur but how they arise within the network. Prior work has used activation patching to identify circuits underlying behaviors in language models, including indirect object identification [33], factual recall [18], and truthfulness [13]. Studying hallucination in VLMs presents additional challenges, as the relevant computations may span both visual representations and language-model components, and modern VLMs differ substantially in how visual information is integrated. This raises a fundamental question: are hallucination-related circuits architecture-specific, or do they reflect a more general property of VLM training?

To address this question, we propose a cross-model analysis strategy to fully investigate the mechanism of object-existence hallucination on POPE-adversarial across five architecturally diverse VLMs. Specifically, our analysis proceeds in three stages. We first apply activation patching to localize hallucination-related components, revealing a shared dual-pathway organization across all five models. We then introduce Conditional Pathway Analysis (CPA) to characterize how components within each pathway interact, exposing a polarity-flip signature invisible to standard component-level analysis. Finally, we use targeted suppression as a causal test of the circuit hypothesis and evaluate whether the identified pathway transfers across hallucination types. Together, these results are consistent with object hallucination in VLMs having a shared mechanistic structure that develops during training rather than being inherited from a specific architecture. In summary, our contributions are three-fold:

- **Cross-model dual-pathway discovery:** We apply activation patching across five architecturally diverse VLMs and identify a consistent dual-pathway organization, consisting of a visual grounding pathway and a hallucination pathway which may lead to errors (Section 5.1).
- **Pathway-level interaction analysis:** We propose Conditional Pathway Analysis (CPA) to quantify within-pathway interactions, discovering a consistent grounding-pathway IE polarity flip between correct and hallucinating samples that is invisible to single-component patching (Section 5.2).
- **Causal validation and benchmark transfer:** Component-wise suppression of the hallucination pathway reduces object hallucination by up to 76% with $\leq$2 pp accuracy cost. Matched static-direction baselines are consistently less effective than per-component scaling, and the identified circuit transfers to relational but not attribute hallucination on AMBER (Sections 5.5–5.6).

2 Related Work

Visual Hallucination The study of visual hallucination has been shaped largely by benchmark design. POPE [15] cast the problem as binary object-existence classification; CHAIR [24] measures hallucinated objects in free-form captions; AMBER [32] and MMHal-Bench [30] extend evaluation to attribute and relational errors; and HallusionBench [8] targets counter-commonsense scenarios (see Chen et al. [4] for a comprehensive survey). On the mitigation side, approaches span training-time alignment [37, 36], decoding-time contrast [9, 11], and post-hoc correction [35]. All of these treat the model as a black box. Our work is complementary: we characterize the internal circuits that produce hallucination, providing a mechanistic foundation for future mitigation.

Activation Patching Activation patching [18] has become a standard tool for circuit discovery in language models, with applications to indirect object identification [33], automated subgraph search [5], and edge-level information flow [7]. Recent work extends these techniques to VLMs: Neo et al. [20] applied logit lens and activation patching to visual information processing; Li et al. [14] proposed cross-modal causal tracing with an inference-time intervention; and Rudman et al. [25] identified attention heads responsible for prompt-induced hallucination in three VLMs. These studies operate at the level of individual components. Our analysis instead characterizes pathway-level

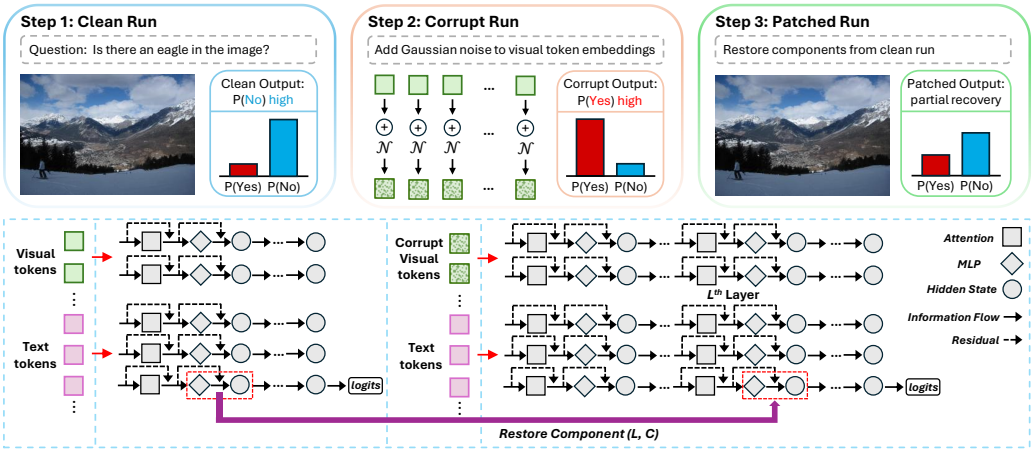

Figure 1: Overview of the activation-patching pipeline for identifying hallucination-related components in VLMs. For each image–question pair, we first run a clean pass, then corrupt visual token embeddings with Gaussian noise, and finally restore selected components from the clean run into the corrupt run. The recovery of the correct logit difference after restoration estimates the causal contribution of each component, which is later used to extract grounding and hallucination pathways.

interaction structure across architecturally diverse VLMs, which surfaces the cross-subset polarity flip in grounding components that single-component analysis cannot see.

Hallucination Mitigation A separate line of work modifies internal activations at inference time to reduce hallucination, including ITI [13], VISTA [16], DMAS [34], and SteerVLM [28] [see also 10, 29, 31]. These methods target mitigation and make varying assumptions about the geometry of the hallucination signal, from a single static linear direction to per-input or learned ones. Our work is not a mitigation contribution; we use ITI and mean-difference projection in Section 5.5 as a controlled probe of the static-direction assumption on the same components our circuit identifies.

3 Methodology

3.1 Activation Patching for Vision-Language Models

Activation patching [18] was developed for text-only models, where corruption is applied to token embeddings. Extending it to VLMs is non-trivial: visual information enters the network through architecture-specific mechanisms, including direct embedding concatenation (Qwen3-VL), learned MLP projectors (LLaVA, InternVL3), and dedicated cross-attention layers (Llama-3.2-Vision), each requiring a different patching interface. We therefore implement a unified patching framework with model-specific adapters for the four VLM families in our study, and corrupt visual rather than text tokens. Appendix A gives the full protocol.

For each image/question pair we run a clean pass, a corrupt pass (Gaussian noise at 3σ on all visual tokens at layer 0, sufficient to drive accuracy to chance), and a patch pass that restores a single component (L, C) to its clean value. The **indirect effect** of (L, C) is

$$\mathrm{IE}(L, C) = \Delta_{\text{patched}}(L, C) - \Delta_{\text{corrupt}}, \tag{1}$$

where $\Delta = \text{logit}(\text{correct token}) - \text{logit}(\text{incorrect token})$ at the final token position. A large positive IE indicates the component is a critical conduit for visual information. The **total effect** $\mathrm{TE} = \Delta_{\text{clean}} - \Delta_{\text{corrupt}}$ normalizes IE across samples.

3.2 Circuit Extraction

We run the patching procedure on n=1,000 POPE-adversarial samples per model. For each sample, the model either correctly answers "no" to a non-existent object or hallucinates by incorrectly answering "yes." We partition the samples accordingly and, for each component, compare the

normalized IE distributions between the two groups using Welch's t-test. All p-values are corrected for multiple comparisons via the Benjamini-Hochberg procedure at a false discovery rate of α=0.05. A component is included in the hallucination circuit if it satisfies two criteria: $p < 0.05$ after correction, and |Cohen's d| > 0.3.

The sign of Cohen's d determines the pathway assignment. Components with $d > 0$, meaning their indirect effect is larger when the model answers correctly, belong to the **visual grounding pathway**. Components with $d < 0$, meaning their indirect effect is larger during hallucination, belong to the **hallucination pathway**.

3.3 Pathway-Level Interaction Decomposition

Single-component patching identifies which components matter but leaves open whether a group of components carries overlapping information, cooperates, or contributes independently. We address this at the pathway level with a diagnostic we call Conditional Pathway Analysis (CPA).

If multiple components each carry the same evidence, removing any one of them still leaves the signal intact in the others, so the joint effect of restoring them all is smaller in magnitude than the sum of their individual effects (redundancy). If instead each component carries only a fragment and only their combination is sufficient to drive the prediction, the joint effect exceeds the sum in magnitude (synergy). The relationship between joint and summed effects thus indicates how the pathway is processing information.

We compute interaction within each pathway separately rather than over the full set of significant components, because grounding and hallucination components have opposite-signed causal effects, and a single joint-minus-sum term over the whole circuit would mix interactions of opposite sign.

For a pathway $\mathcal{P} = \{C_1, \ldots, C_k\}$, the **joint effect** $\mathrm{IE}(\mathcal{P})$ is the indirect effect obtained when all pathway components are restored simultaneously. The interaction term subtracts the sum of individual effects:

$$\mathrm{IE}(\mathcal{P}) = \Delta_{\text{patched}(\mathcal{P})} - \Delta_{\text{corrupt}}, \tag{2}$$

$$\mathrm{Interaction}(\mathcal{P}) = \mathrm{IE}(\mathcal{P}) - \sum_{i=1}^{k} \mathrm{IE}(C_i). \tag{3}$$

Because the polarity of $\mathrm{IE}(C_i)$ may differ between correct and hallucinating subsets, the sign of Eq. 3 alone conflates redundancy with the underlying IE polarity. We therefore additionally report a polarity-free magnitude metric:

$$\mathrm{MagDiff}(\mathcal{P}) = \big|\mathrm{IE}(\mathcal{P})\big| - \sum_{i=1}^{k} \big|\mathrm{IE}(C_i)\big|, \tag{4}$$

where negative values uniformly indicate magnitude-redundancy and positive values indicate magnitude-synergy. We use $\mathrm{MagDiff}$ as the primary diagnostic for within-subset interaction structure and report the per-subset mean of $\mathrm{IE}(C_i)$ separately to expose any polarity flip across subsets. We compute these per sample on n=400 POPE-adversarial samples per model, and test (i) whether $\mathrm{MagDiff}$ differs from zero (one-sample t-test) and (ii) whether the mean $\mathrm{IE}(C_i)$ shifts between correct and hallucinating samples (two-sample t-test).

3.4 Logit Lens Validation

To validate the pathway labels independently of the patching procedure, we ask whether layers assigned to the grounding vs. hallucination pathway by our patching-based criterion have differential layer-wise contributions to the correct answer. We answer this using logit lens analysis [21] as a decoding tool. At each layer l, we project the hidden state through the final layer norm and language model head to obtain intermediate logits: $\hat{y}_l = \text{lm_head}(\text{norm}(h_l))$. We compute the per-layer delta, the change in Δ relative to the preceding layer, which reveals each layer's marginal contribution to the final prediction. If the pathway labels are functionally accurate, grounding-pathway layers should have positive delta and hallucination-pathway layers negative delta on correct samples, with polarities reversing on hallucinating samples.

3.5 Cross-Architecture Comparison

Because the five models have different numbers of layers (28–48), direct layer-by-layer comparison is not meaningful. We normalize all layer indices to a relative depth in $[0, 1]$ and compare at two levels of granularity. At the macro level, we test whether the spatial distribution of pathway components is shared across models, using Fisher's exact test on 2×2 contingency tables. At the micro level, we compute Pearson correlations on depth-normalized effect-size profiles and apply TOST equivalence testing with a bound of $|r| < 0.3$.

3.6 Targeted Intervention

For models whose hallucination pathway is sufficiently concentrated, we suppress the identified components by scaling their outputs during inference:

$$h'_{\text{out}} = s \cdot h_{\text{out}}, \quad s \in \{0.0, 0.25, 0.5, 0.75\}. \tag{5}$$

This is applied at every forward pass to the output of each hallucination-pathway component ($d < 0$ in the circuit extraction). No retraining or additional data is required. We select the optimal s per model on a held-out selection set (see Section 4) and evaluate the selected configuration on a disjoint evaluation set.

For models with dispersed circuits (many hallucination-pathway components), we additionally introduce **top-k component selection**: we rank hallucination-pathway components by $|d|$ and suppress only the top-k strongest, evaluating $k \in \{3, 5, 8, 10, 15, 20, 30\}$. The optimal k is selected on the same selection set used for s, and reported on the disjoint evaluation set.

We compare against two linear-steering baselines from prior work [13, 42], applied to the same hallucination-pathway components selected by our circuit. **Mean-difference projection** projects out $\hat{d} = \text{normalize}(\mu_{\text{halluc}} - \mu_{\text{correct}})$ from each component's activations: $h' = h - \alpha(h \cdot \hat{d})\hat{d}$. **Probe-based ITI** [13] replaces $\hat{d}$ with a logistic-regression probe direction trained per attention head and adds it to the top-K heads. Both baselines use symmetric α grids; including negative α is essential because the probe direction points toward the hallucination class (Appendix D.3), so reducing hallucination requires subtracting it. As a further control, we include a **random control** that applies the uniform-scaling factor to the same number of randomly selected components. Full grids, probe-training details, and tie-breaking rules are in Appendix D.

4 Experimental Setup

Models. We select five VLMs that together cover three LLM backbone families, four vision–language integration strategies (embed-concat, projector-concat, ViT-MLP-LLM, cross-attention), and two model scales (Table 1). InternVL3-8B and InternVL3-14B share an architecture but use different LLM backbones, letting us test scale sensitivity within the same family. HuggingFace identifiers are listed in Appendix A.

Data and evaluation. All patching experiments use the adversarial split of POPE [15], which selects negative-sample objects based on co-occurrence frequency. We use 1,000 samples for circuit discovery, 400 for CPA, 200 for logit lens analysis, and 500 for intervention evaluation. These four sets are drawn independently from POPE-adversarial; the patching set and the intervention set are disjoint. Of the 500 intervention samples, 100 are reserved as a selection set for choosing s and k; the remaining 400 form the held-out evaluation set on which final numbers in Sections 5.5 and 5.6 are computed. For cross-benchmark evaluation, we test on the popular and random POPE splits and on AMBER [32], which covers existence, attribute, and relational hallucination types. Hyperparameters are selected only on the POPE-adversarial selection set and applied unchanged to all other benchmarks, making the cross-benchmark evaluations pure transfer tests.

Patching configuration. We corrupt visual embeddings with Gaussian noise at $3\times$ the standard deviation of clean visual tokens. All patching operates at the component level (one attention sublayer and one MLP sublayer per transformer block). All evaluations use greedy decoding to ensure reproducibility.

Table 1: **Models and experimental setup.** All patching experiments use POPE-adversarial (n=1,000). Corruption: Gaussian noise (3σ) on visual token embeddings at layer 0. Llama-3.2-Vision uses 8 cross-attention layers (of 40 total) for visual input.

Model	Arch. type	LLM backbone	Layers	Sig.	Grnd.	Hall.	Halluc. rate
Qwen3-VL-8B	Embed-concat	Qwen3 (8B)	36	26	14	12	13.1%
LLaVA-v1.6-7B	Projector-concat	Mistral (7B)	32	48	18	30	13.4%
Llama-3.2-V-11B	Cross-attention	Llama-3 (11B)	40	31	11	20	15.9%
InternVL3-8B	ViT-MLP-LLM	Qwen2.5 (8B)	28	27	21	6	12.6%
InternVL3-14B	ViT-MLP-LLM	Qwen2.5 (14B)	48	45	29	16	13.4%

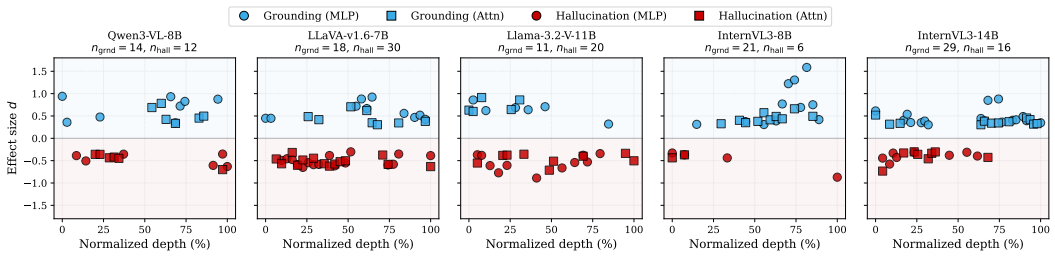

Figure 2: **Dual-pathway organization across all five VLMs.** Hallucination components (red) concentrate at early layers and at network boundaries; grounding components (blue) concentrate at mid-to-late depths. Each panel shows components satisfying the inclusion criteria ($p_{\text{adj}} < 0.05$, $|d| > 0.3$; $n = 1{,}000$) plotted by normalized depth (x-axis) and Cohen's d (y-axis). See Appendix Table 11 for the strongest components per model.

5 Experiments

5.1 A Dual-Pathway Circuit for Visual Hallucination

Activation patching reveals a consistent dual-pathway organization across all studied models (Figure 2). The strongest components per model are summarized in Appendix Table 11. In each model, the set of significant components ($p < 0.05$ after BH correction, $|d| > 0.3$) splits into two groups that differ in sign, depth profile, and functional role.

The **visual grounding pathway** ($d > 0$) consists of components whose indirect effect is substantially higher on correct samples, with peak effect sizes reaching $d = +1.59$. The **hallucination pathway** ($d < 0$) is preferentially active during hallucination, with peak $|d|$ up to 0.89. In Llama-3.2-Vision, the strongest grounding component is one of the model's dedicated cross-attention layers (L3.attn, $d = +0.91$), providing an architectural sanity check on the patching-based classification. LLaVA exhibits the most dispersed hallucination pathway, with 30 of its 48 significant components having $d < 0$ and spanning the entire network depth.

Circuit sizes range from 26 to 48 significant components per model, with hallucination-pathway share varying from 22% (InternVL3-8B) to 65% (Llama-3.2). The dual-pathway organization is preserved across $|d|$ thresholds in $\{0.25, 0.30, 0.35\}$ for all five models, with both pathways always retaining at least 5 components (Appendix B).

5.2 CPA Reveals Consistent Redundancy and a Grounding-Pathway Polarity Flip

Both pathways are redundant within each subset; the grounding pathway additionally exhibits a consistent IE polarity flip between subsets. The polarity-free magnitude diagnostic $\mathrm{MagDiff}$ shows that, in every model and every subset, the joint pathway effect is substantially smaller in magnitude than the sum of individual effects: across the $5{\times}2{\times}2 = 20$ (model, pathway, subset) cells, the magnitude ratio $|\mathrm{IE}(\mathcal{P})|/\sum_i |\mathrm{IE}(C_i)|$ stays in $[0.07, 0.69]$ with a mean of 0.32, indicating strong within-subset redundancy throughout (Figure 3; full numerical values in Appendix Table 12). The shift in $\mathrm{MagDiff}$ between subsets is small and inconsistent in sign across models for both pathways (grounding $|d|$ from 0.13 to 0.92; hallucination $|d|$ from 0.14 to 0.59), giving no support for a redundancy-to-synergy transition under the polarity-free metric.

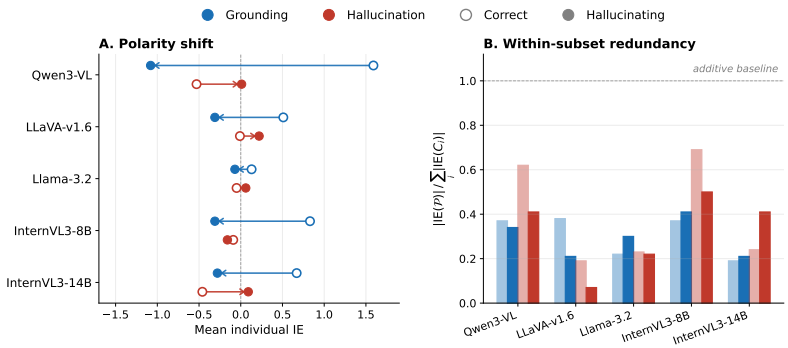

Figure 3: **Pathway-level CPA diagnostics across all five VLMs. Left:** mean individual indirect effect per pathway and subset; arrows point from correct (hollow) to hallucinating (filled) samples. The grounding pathway exhibits a consistent polarity flip from positive on correct samples to negative on hallucinating samples in every model; the hallucination pathway shows no consistent flip. **Right:** magnitude ratio $|\mathrm{IE}(\mathcal{P})|/\sum_i |\mathrm{IE}(C_i)|$, with light bars = correct and dark bars = hallucinating samples. All cells fall well below 1.0 (0.07–0.69, mean 0.32), indicating strong within-subset redundancy regardless of subset. Numerical values are in Appendix Table 12.

What does differ systematically between subsets, and only for the grounding pathway, is the polarity of the individual indirect effects. On correct samples, 61–83% of grounding-pathway component IEs are positive, with subset-mean IEs of $+0.13$ to $+1.59$ across the five models. On hallucinating samples, the same components flip: only 21–33% of IEs remain positive, and the subset-mean IE turns negative (-1.08 to -0.07) in every model (cross-subset Cohen's d from -1.60 to -2.45, all $p < 0.001$). The hallucination pathway shows no such consistent flip: the mean-IE shift is smaller in magnitude (mean $|d| = 0.59$ vs. 1.92 for grounding) and reverses direction across models, with InternVL3-8B showing no flip at all (mean IE remains negative in both subsets, -0.09 and -0.16).

The mechanistic reading is that on hallucinating samples, restoring a grounding-pathway component pulls the corrupt run toward the model's clean-run prediction, which on these samples is the hallucinated answer. The within-subset redundancy persists in both phases rather than dissolving when the model hallucinates, suggesting that the language-prior bias does not break apart the redundant copies carried by the grounding pathway but redirects what they encode. The failure on these samples is thus not that visual evidence is missing or weak, but that the "grounding" representations themselves have been entrained to the model's non-visual output.

5.3 Logit Lens Supports the Functional Roles

Logit lens analysis provides patching-independent evidence that the pathway labels track functional differences (full per-model values in Appendix F, Table 10). If grounding-pathway layers drive image-faithful predictions and hallucination-pathway layers push toward language-prior errors, their layer-wise contributions should shift direction between correct and hallucinating samples. In four of five models (Qwen3-VL, LLaVA, InternVL3-8B, InternVL3-14B), the differential is significant and in the predicted direction, with $d \in [-1.25, -0.65]$ (all $p < 0.05$).

Llama-3.2-Vision is the exception: its hallucinating-sample differential is not significant ($d = -0.18$, $p = 0.43$). Llama-3.2 injects visual information through dedicated cross-attention layers interleaved with self-attention, so grounding- and hallucination-pathway layers do not form a single layer-wise logit trajectory of the kind logit lens decodes. The pathway labels are nonetheless functionally validated for this model by its strongest hallucination-pathway concentration in our circuits (Table 1) and its cleanest response to the targeted intervention (Section 5.5).

5.4 Cross-Architecture Analysis: Shared Organization, Distinct Wiring

We compare circuits across architectures along two axes, normalizing layer indices to a relative depth in $[0, 1]$ to handle differing network depths (full pairwise statistics in Appendix E). At the macro level, hallucination-pathway components concentrate in early layers and at network boundaries while

grounding-pathway components concentrate in mid-to-late layers, and this spatial pattern is consistent across models (Fisher's exact test, all six pairs $p > 0.05$ against a shared-pattern null). At the micro level, layer-by-layer profile correlations are uniformly low ($|r| \leq 0.33$), and the within-family InternVL3-8B/14B comparison is no more similar than cross-family pairs. The combination is most consistent with the dual-pathway organization arising during VLM training rather than being inherited from a specific backbone or architecture.

5.5 Causal Validation: Suppressing Hallucination-Pathway Components

Figure 4 shows that uniform scaling of hallucination-pathway outputs substantially reduces hallucination in every model; full numerical values are in Appendix Table 13. All numbers reported in this section are computed on the held-out 400-sample test set, with the optimal s selected on the disjoint 100-sample selection set (setup in Section 4). Qwen3-VL responds best to s=0.25, while Llama-3.2, InternVL3-8B, and InternVL3-14B respond best to s=0.5. Relative hallucination reductions on POPE-adversarial range from 40% (LLaVA-v1.6) to 76% (Llama-3.2) at $\leq$2 pp accuracy cost.

Top-k ablation resolves LLaVA's dispersed circuit. LLaVA's hallucination pathway spans 30 components. Suppressing all of them (s=0.5) drops accuracy by 9.0 pp. We therefore evaluate suppressing only the k strongest components for $k \in \{3, 5, 8, 10, 15, 20, 30\}$, selecting k on the 100-sample selection set. The selected configuration is k=10, which yields a 40% relative hallucination reduction (5.0% $\rightarrow$ 3.0%) at no accuracy cost (Figure 4). Top-k selection thus identifies a compact core of strongly hallucination-driving components within an otherwise dispersed circuit, extending the intervention to models where the $|d| > 0.3$ threshold alone is too permissive.

Static-direction interventions are partially effective but bounded. To probe whether the hallucination pathway is well described by a single static linear direction, we apply mean-difference projection and probe-based ITI [13] to the same hallucination-pathway components, with hyperparameters selected under the same protocol as uniform scaling (full grids, per-model results, and probe-quality checks in Appendix D). ITI reduces hallucination on every evaluated model but consistently less than per-component scaling, while mean-difference projection fails to yield any reduction on the held-out test set across the three models evaluated. The per-model relative effect tracks the per-component activation geometry: the leading singular direction of the per-component hallucination signal captures only 45–69% of its variance (Appendix D.5), bounding what any single static direction can do. Per-component scaling sidesteps this bound by treating the pathway as a set of adjustable contributions rather than a single direction.

A random-component control at the same scaling factor produces no systematic reduction (mean $|\Delta\text{H}| = 3.0$ pp with the sign positive on all three evaluated models, indicating that random suppression tends to slightly increase rather than decrease hallucination; the largest such increase is $+7.2$ pp on Llama-3.2), confirming that the effect depends on which components are scaled, not on scaling alone. The probes are discriminative (per-model mean head-probe validation accuracy 0.62–0.67, hundreds of heads above 0.75), ruling out probe-fitting failure as the cause of ITI's bounded effect. As a specificity check, applying the same uniform-scaling protocol to the grounding pathway at s=0 collapses POPE accuracy to chance (0.50–0.51) on every model (Appendix D.4), establishing a double dissociation: hallucination-pathway suppression reduces hallucination at near-zero accuracy cost, while grounding-pathway ablation destroys task performance.

5.6 Cross-Benchmark Generalization

To test whether the hallucination pathways identified on POPE-adversarial generalize beyond object-existence questions, we evaluate the same targeted intervention on two additional settings: all three POPE splits and the AMBER benchmark [32], which covers three hallucination types. Circuits and hyperparameters are identical to those selected in Section 5.5. No tuning is done on these new benchmarks, and all of them are pure generalization evaluations.

POPE splits. The intervention generalizes consistently across POPE difficulty levels (Appendix Table 14). Hallucination reduction is strongest on the adversarial split, where baseline hallucination is highest, and attenuates on the random split, where baseline hallucination is already low, consistent with a floor effect. Accuracy cost remains within 2 pp for most models.

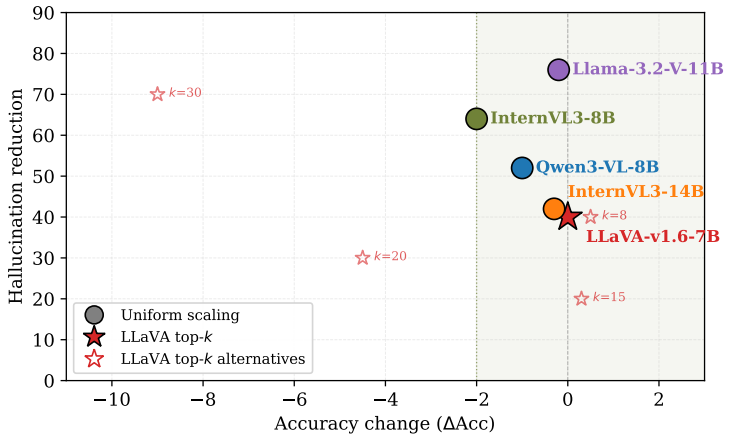

Figure 4: **Intervention Pareto front on POPE-adversarial** (n=400 **held-out).** Filled markers show each model's selected configuration; x-axis: accuracy change (pp), y-axis: relative hallucination reduction. All five models achieve 40–76% reduction at $\leq$2 pp accuracy cost. LLaVA uses top-k=10 at s=0.5; hollow stars show alternative k values. Full numerical values are in Appendix Table 13.

AMBER: type-dependent transfer. The intervention transfers to AMBER with a clear type-dependent pattern. On existence questions, the type semantically closest to POPE, four of five models show improvement (10–42% relative reduction). On relational hallucination, Qwen3-VL, InternVL3-8B, and Llama-3.2 achieve 36–47% relative reduction, suggesting substantial overlap between the object-existence and relational hallucination circuits. On attribute questions, only Llama-3.2 ($-$23%) and InternVL3-8B ($-$17%) show meaningful improvement; the remaining models are unaffected. This asymmetry suggests that attribute hallucination involves partially distinct circuits not captured by the POPE-derived pathway.

Model-specific patterns. Llama-3.2-Vision shows the strongest cross-benchmark generalization, with hallucination reduction across all six evaluation settings, consistent with its cross-attention architecture producing a more concentrated hallucination pathway. InternVL3-14B shows the opposite profile: strong POPE reductions (29–68%) but weak AMBER transfer; we discuss possible mechanistic correlates (the largest cross-subset polarity-flip magnitude, Section 5.2) in Appendix E.

6 Discussion

What generalizes, and implications for intervention design. The macro-level dual-pathway organization recurs despite differences in backbone, fusion strategy, and scale, while micro-level wiring differs even within the same backbone family (Section 5.4), pointing to training rather than architecture as the origin of the organization. Across hallucination types, the intervention transfers well to relational hallucination in three of five models but leaves attribute hallucination largely unaffected. One interpretation is that relational hallucination, like object-existence hallucination, is driven primarily by language priors such as co-occurrence statistics, whereas attribute hallucination may involve a different failure mode such as imprecise visual feature extraction. The pathway asymmetry also constrains intervention design. The hallucination pathway does not show a consistent polarity flip, static-direction baselines are bounded by per-component activation geometry, and hallucination-signal variance is spread across many singular directions (Appendix D.5). These results suggest that no single static direction captures the full pathway contribution. Per-component scaling sidesteps this bound for our causal test, but does not constitute a deployable mitigation method; designing one under this constraint is a separate engineering problem.

Broader impact. A mechanistic account of object hallucination can support more reliable VLMs. The same pathway-level handles can also act in the harmful direction: applying a probe direction with the opposite sign raises hallucination from 20% to 74% (Appendix D.3), so such interventions should be evaluated carefully before deployment.

Conclusion. We identified a shared dual-pathway organization for object-existence hallucination across five architecturally diverse VLMs, and showed that pathway-level analysis exposes a consistent IE polarity flip in grounding components between correct and hallucinating samples, a signature invisible to single-component patching. Causal validation by component scaling, together with the bounded effect of matched static-direction interventions, indicates that the hallucination pathway acts through multiple distinct activation directions rather than a single shared one. The selective transfer to relational but not attribute hallucination points to partially distinct circuits underlying different hallucination types; direct circuit extraction on attribute- and relation-focused benchmarks is a natural next step. We release our code to support further work.

Acknowledgments and Disclosure of Funding

This work received no third-party funding. The authors declare no competing interests.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*, 2025.

[4] Zhiyuan Chen, Yuecong Min, Jie Zhang, Bei Yan, Jiahao Wang, Xiaozhen Wang, and Shiguang Shan. A survey of multimodal hallucination evaluation and detection. *International Journal of Computer Vision*, 2025.

[5] Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36, 2023.

[6] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

[7] Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

[8] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusion-Bench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2023.

[9] Qidong Huang et al. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*, 2024.

[10] Sai Akhil Kogilathota, Sripadha Vallabha E. G, Luzhe Sun, and Jiawei Zhou. HALP: Detecting hallucinations in vision-language models without generating a single token. In *EACL*, 2026.

[11] Sicong Leng et al. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, 2024.

[12] Boyi Li, Yifan Shen, Yuanzhe Liu, Yifan Xu, Jiateng Liu, Xinzhuo Li, Zhengyuan Li, Jingyuan Zhu, Yunhan Zhong, Fangzhou Lan, et al. Toward cognitive supersensing in multimodal large language model. *arXiv preprint arXiv:2602.01541*, 2026.

[13] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 2024.

[14] 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. *arXiv preprint arXiv:2511.05923*, 2025.

[15] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023.

[16] Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *ICML*, 2025.

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35, 2022.

[19] Meta AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog*, 2024.

[20] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *ICLR*, 2025.

[21] nostalgebraist. Interpreting gpt: the logit lens. *LessWrong*, 2020.

[22] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

[23] Qwen Team. Qwen3-vl. `https://qwenlm.github.io/blog/qwen3-vl/`, 2025. Accessed: 2025.

[24] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, 2018.

[25] William Rudman, Michal Golovanevsky, Dana Arad, Yonatan Belinkov, Ritambhara Singh, Carsten Eickhoff, and Kyle Mahowald. Mechanisms of prompt-induced hallucination in vision-language models. *arXiv preprint arXiv:2601.05201*, 2026.

[26] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

[27] Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in vlms. *Advances in Neural Information Processing Systems*, 38:17929–17960, 2026.

[28] Anushka Sivakumar, Andrew Zhang, Zaber Ibn Abdul Hakim, and Chris Thomas. SteerVLM: Robust model control through lightweight activation steering for vision language models. *arXiv preprint arXiv:2510.26769*, 2025.

[29] 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 *ACL*, 2025.

[30] Zhiqing Sun et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

[31] Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte Pres. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2024.

[32] Junyang Wang et al. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

[33] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in gpt-2 small. In *ICLR*, 2023.

[34] Jianghao Yin et al. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models. *arXiv preprint arXiv:2602.21704*, 2026.

[35] Shukang Yin et al. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.

[36] Qifan Yu et al. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. *arXiv preprint arXiv:2311.13614*, 2023.

[37] Tianyu Yu et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*, 2024.

[38] Zhenlong Yuan, Xiangyan Qu, Chengxuan Qian, Rui Chen, Jing Tang, Lei Sun, Xiangxiang Chu, Dapeng Zhang, Yiwei Wang, Yujun Cai, et al. Video-star: Reinforcing open-vocabulary action recognition with tools. *arXiv preprint arXiv:2510.08480*, 2025.

[39] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024.

[40] Fei Zhao, Chengcui Zhang, Runlin Zhang, Tianyang Wang, and Xi Li. Mitigating image captioning hallucinations in vision-language models, 2025. URL `https://arxiv.org/abs/2505.03420`.

[41] Jinguo Zhu et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

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

A Implementation Details

Models and HuggingFace identifiers. The five VLMs evaluated are: Qwen3-VL-8B [23] (`Qwen/Qwen3-VL-8B-Instruct`), which concatenates visual tokens directly into the embedding sequence; LLaVA-v1.6-7B [17] (`llava-hf/llava-v1.6-mistral-7b-hf`), which uses a learned MLP projector into the Mistral-7B token space; Llama-3.2-Vision-11B [19] (`meta-llama/Llama-3.2-11B-Vision-Instruct`), which injects visual information through dedicated cross-attention layers interleaved with standard self-attention; and InternVL3-8B / 14B [41] (`OpenGVLab/InternVL3-8B-hf`, `OpenGVLab/InternVL3-14B-hf`), which route vision-encoder outputs through an MLP bridge into Qwen2.5-8B / 14B backbones.

Activation patching protocol. Our patching implementation adapts the procedure of Meng et al. [18] to handle the diversity of VLM architectures. We implement model-specific adapters for each VLM family (Qwen, LLaVA, InternVL, Llama) that handle differences in visual token placement, attention layer structure (self-attention vs. cross-attention), and embedding concatenation strategy. The core patching loop is architecture-agnostic: it operates on cached activations and patched forward passes through the language-model backbone.

For the corrupt run, we add Gaussian noise with standard deviation $3 \times \sigma_{\text{visual}}$ to all visual token positions at layer 0, where σ_{visual} is the empirical standard deviation of the clean visual token embeddings computed per sample. We verified that this noise level is sufficient to destroy visual information: the accuracy of the corrupt run drops to near chance level (50%) for all five models.

CPA computation. The joint patching required for CPA is implemented by simultaneously restoring all pathway components in a single patch run. For a pathway with k components, this requires one additional forward pass per sample (beyond the $2k$ passes for individual component patching), making CPA computationally efficient. We run CPA on 400 POPE-adversarial samples per model.

Logit lens. At each layer, we apply the final layer norm and language model head to the hidden state at the last token position (the position where the model generates "Yes" or "No"). We record $P(\text{Yes})$, $P(\text{No})$, and the logit difference at each layer. The per-layer delta is computed as $\Delta_l = \Delta(\text{logit_diff})_l - \Delta(\text{logit_diff})_{l-1}$. We use 200 POPE-adversarial samples per model.

Hardware and runtime. All experiments were conducted on NVIDIA H200 144GB GPUs. Component-level patching on 1,000 samples takes approximately 2–18 hours per model depending on model size and sequence length (Llama-3.2 is fastest at $\sim$1.8s/sample; InternVL3-14B is slowest at $\sim$18s/sample). CPA adds roughly 2–3 hours per model for 400 samples. Logit lens analysis on 200 samples completes in under 1 hour per model. The full pipeline for one model runs in under 24 hours on a single GPU.

B Threshold Sensitivity for Circuit Extraction

The circuit-extraction criterion in Section 3.2 uses two thresholds: BH-corrected $p < 0.05$ and |Cohen's d| > 0.3. The p-value cutoff follows standard FDR practice. To evaluate sensitivity to the effect-size cutoff, we re-extract circuits at $|d| \in \{0.25, 0.30, 0.35\}$ from the same patching results (no additional model runs required), and measure overlap with the reference $|d| > 0.3$ circuit using Jaccard similarity computed separately for each pathway.

Results are reported in Table 2. The dual-pathway organization is preserved at every threshold for all five models: each model retains at least 5 grounding-pathway and 5 hallucination-pathway components throughout the ± 0.05 window, and 16 of the 20 (model, pathway, perturbation) cells have Jaccard ≥ 0.80 (minimum 0.60). The qualitative finding, that both pathways exist and that the grounding pathway dominates in size, is preserved at every threshold.

Table 2: **Circuit composition across $|d|$ thresholds.** For each model, we report the number of grounding-pathway and hallucination-pathway components at three thresholds, plus Jaccard similarity of pathway membership relative to the paper's $|d| > 0.3$ circuit. Computed post-hoc on the same patching results used in the main paper; no additional GPU runs were performed.

	$\lvert d\rvert > 0.25$		$\lvert d\rvert > 0.30$ (paper)		$\lvert d\rvert > 0.35$	
Model	$n_{\text{grnd}}/n_{\text{hall}}$	$J_{\text{grnd}}/J_{\text{hall}}$	$n_{\text{grnd}}/n_{\text{hall}}$	J (ref)	$n_{\text{grnd}}/n_{\text{hall}}$	$J_{\text{grnd}}/J_{\text{hall}}$
Qwen3-VL-8B	20/16	0.70/0.75	14/12	1.00	12/12	0.86/1.00
LLaVA-v1.6-7B	20/32	0.90/0.94	18/30	1.00	15/28	0.83/0.93
Llama-3.2-V-11B	15/22	0.73/0.91	11/20	1.00	10/18	0.91/0.90
InternVL3-8B	22/10	0.96/0.60	21/6	1.00	18/5	0.86/0.83
InternVL3-14B	32/20	0.91/0.80	29/16	1.00	19/9	0.66/0.56

C Statistical Testing Details

For CPA (Section 3.3), we perform two planned tests per pathway per model (one-sample t-test and correct-vs-hallucinating two-sample t-test), yielding 4 tests per model or 20 across all five models. For logit lens (Section 3.4), we perform one differential test per model (5 tests). Because these are a small number of pre-specified, hypothesis-driven comparisons rather than exploratory screening, we report uncorrected p-values but note that all key findings (grounding-pathway differential d, all $p < 0.001$) survive Bonferroni correction at any conventional α level. BH-corrected and TOST

equivalence-test details for circuit extraction and cross-architecture comparison are in Sections 3.2 and 5.4; full pairwise TOST values are in Table 9.

D Linear-Steering Baselines: Full Numerical Results

This appendix gives the complete numerical results underlying Section 5.5 for both linear-steering baselines: basic directional suppression (mean-difference projection) and probe-based ITI [13].

D.1 Held-Out Test-Set Comparison

Table 3 reports the held-out test-set comparison for all four interventions side by side. Directional suppression projects out the per-component mean-difference direction $h' = h - \alpha(h \cdot \hat{d})\hat{d}$ at $\alpha \in \{\pm 0.5, \pm 1.0, \pm 1.5\}$, with α selected on the same selection set and accuracy budget as uniform scaling. Directional suppression was evaluated on three models (Qwen3-VL-8B, Llama-3.2-Vision-11B, InternVL3-8B) covering two LLM backbone families and three vision-integration strategies; ITI and the random control follow the same setup on the models indicated. Across the five models, ITI achieves a mean $|\Delta\text{H}|$ of 4.1 pp under the symmetric grid (ranging from 0.5 pp on LLaVA and InternVL3-14B to 11.0 pp on InternVL3-8B), versus 7.4 pp for component scaling on the same components. Mean-difference projection achieves 0.8 pp on its three evaluated models, smaller than ITI on every directly comparable cell. The random control produces no systematic reduction, confirming that the effectiveness of uniform scaling depends on the component selection and not the scaling alone. ITI on Qwen3-VL is the only model where the selected configuration uses positive α; the remaining four models select negative α, consistent with the probe-trained direction pointing toward the hallucination class (Appendix D.3, Table 5).

Table 3: **Linear-steering baselines vs. uniform scaling on the held-out test set** (n=400; hyperparameters selected on a disjoint 100-sample selection set under a $\leq$2 pp accuracy budget). ΔH is in pp; negative = reduction. Random scales an equal number of randomly chosen components by the uniform-scaling factor (mean over 5 seeds). Directional and Random cover three models spanning two backbone families and three vision-integration strategies; "–" = not evaluated. Selected hyperparameters are given in Appendices D.2–D.3. Baselines are measured under the directional/ITI pipeline; uniform-scaling ΔH is taken from Table 13, whose baselines agree with these to within 2 pp.

Model	Baseline halluc.	Random ΔH	Directional ΔH	ITI ΔH	Uniform scaling ΔH
Qwen3-VL-8B	7.5%	+1.1	+0.0	−1.5	−4.4
LLaVA-v1.6-7B	5.5%	—	—	−0.5	−2.0
Llama-3.2-V-11B	16.5%	+7.2	+1.0	−7.0	−14.0
InternVL3-8B	14.5%	+0.7	+1.5	−11.0	−10.0
InternVL3-14B	15.5%	—	—	−0.5	−6.5
mean $\lvert\Delta\text{H}\rvert$ (across evaluated models)		**3.0 pp**	**0.8 pp**	**4.1 pp**	**7.4 pp**

D.2 α-Saturation: Directional Suppression

Across $\alpha \in \{\pm 0.5, \pm 1.0, \pm 1.5\}$, the mean-difference projection $h - \alpha(h \cdot \hat{d})\hat{d}$ saturates on every model tested: per-sample decisions stop changing after a small number of samples shift, despite the projection magnitude scaling linearly with $|\alpha|$. The mean-difference direction therefore behaves similarly to the probe direction analyzed in Appendix D.3: the symmetric grid is needed to find the sign that reduces (rather than amplifies) hallucination, and once that sign is found the achievable reduction remains bounded below per-component scaling.

D.3 Probe-Based ITI

We implement ITI [13] adapted to VLMs. For each attention head at every layer, we capture pre-W_O activations at the last input-token position on a disjoint 1000-sample probe-training split of POPE-adversarial (non-overlapping with the 500-sample intervention split). We label samples as correct or hallucinating by the model's own answer, train a logistic-regression probe per head with class-weighting to handle label imbalance, and rank heads by 20%-holdout validation accuracy.

At inference, we add $\alpha \cdot \sigma \cdot \hat{w}$ to each selected head's pre-W_O activation, where $\hat{w}$ is the unit probe weight and σ is the standard deviation of $\hat{w} \cdot X$ on the training split. We grid-search over $K \in \{12, 24, 48, 96\}$ and $\alpha \in \{\pm 5, \pm 10, \pm 15, \pm 20\}$ (32 configurations) on the 100-sample selection set, breaking ties at the same hallucination rate by maximum selection-set accuracy, and select the configuration minimizing hallucination rate subject to the $\leq$2 pp accuracy budget, applied to the disjoint 400-sample test split. Including negative α is essential because, as Table 5 below shows, increasing positive α pushes the model toward the hallucination class; reducing hallucination requires moving in the opposite direction. We run ITI on all five models.

Probe quality rules out an estimator-failure explanation. Table 4 shows that probes learn discriminative information on every model: mean per-head validation accuracy is 0.62–0.67 (vs. 0.5 chance), with hundreds of heads per model achieving validation accuracy above 0.75 and maxima reaching 0.835–0.895. The bounded effect of ITI at inference is therefore not explicable by probes failing to learn; the learned directions discriminate well, and even in the negative-α direction they translate into a behavioral change strictly smaller than per-component scaling on every model.

Table 4: **ITI probe quality: held-out validation accuracy per head.** One logistic-regression probe per (layer, head) trained on 1000 POPE-adversarial samples with 20% held-out validation. Chance accuracy is 0.5. All five models have probes that clearly discriminate correct from hallucinating activations, with hundreds of heads exceeding 0.75 validation accuracy. This rules out "probes failed to learn" as an alternative explanation for the bounded effect ITI achieves at intervention time.

Model	#probes	Mean val acc.	Max val acc.	#(>0.65)	#(>0.75)
Qwen3-VL-8B	1152	0.628	0.835	540	125
LLaVA-v1.6-7B	1024	0.632	0.825	467	109
Llama-3.2-V-11B	1280	0.619	0.850	509	41
InternVL3-8B	784	0.652	0.855	420	162
InternVL3-14B	1920	0.666	0.895	1113	589

Held-out test-set numbers under the symmetric grid. Held-out test-set numbers for the selected configuration on each model are reported in Table 3. The selected (K, α) are $(12, +10)$ for Qwen3-VL, $(12, -5)$ for LLaVA, $(48, -20)$ for Llama-3.2, $(96, -15)$ for InternVL3-8B, and $(24, -20)$ for InternVL3-14B; four of five models thus select a negative α. ITI's test-set $|\Delta\text{H}|$ ranges from 0.5 pp (LLaVA, InternVL3-14B) to 11.0 pp (InternVL3-8B), with mean 4.1 pp. On the four models where ITI achieves a non-trivial reduction, the relative effect (ITI $|\Delta\text{H}|$ divided by component-scaling $|\Delta\text{H}|$ on the same model) is 34% for Qwen, 25% for LLaVA, 50% for Llama-3.2, and 110% for InternVL3-8B; InternVL3-14B's $|\Delta\text{H}| = 0.5$ pp is below the 1 pp resolution at which we report the relative quotient. The InternVL3-8B value above 100% is consistent with that model's hallucination pathway being unusually concentrated (only 6 components, the lowest count of any model in our study), so $K = 96$ heads spans much of the relevant subspace; the other four models lie at or below the per-component leading-direction variance fraction reported in Appendix D.5 (45–69%).

The positive-α regime: probe direction points toward hallucination. Restricting attention to the positive-α slice of the grid (Table 5, illustrated on InternVL3-8B as the representative model) clarifies why the symmetric grid was needed. At small K, α-saturation makes the injected direction behaviorally inert (e.g., $K = 12$, $\alpha \in \{5, 10, 15\}$ produce identical TP/FP/TN/FN counts), and at larger (K, α) the effect emerges with the wrong sign: at $K = 96, \alpha = 20$, the same probe-trained direction drives hallucination from a 20% baseline to 74% (a 3.7$\times$ increase). The probe direction therefore points toward the hallucination class rather than away from it, which is why a positive-α-only grid cannot mitigate hallucination. The other four models show qualitatively similar saturation-then-amplification patterns of varying severity (InternVL3-14B is fully grid-saturated; Llama-3.2 reaches 54% hallucination at $K = 96, \alpha = 20$ from a 26% baseline).

The negative-α regime: bounded reduction tracking activation geometry. Mirroring the grid to negative α (Table 6) reverses the sign of the effect: the probe direction now reduces hallucination, with the strongest effect concentrated at large K and large $|\alpha|$. Two failure modes constrain the achievable mitigation: at the most aggressive settings, recall collapses (e.g., on InternVL3-8B at $K = 96, \alpha = -20$, hallucination reaches 6% but selection-set recall drops from 0.94 to 0.78),

revealing an over-correction toward the "no" bias; and at smaller K the effect saturates at only a few percentage points of reduction. The selection rule (min hallucination subject to the $\leq$2 pp accuracy budget, ties broken by maximum selection-set accuracy) avoids the over-correction regime, yielding the bounded test-set reductions in Table 3.

The conjunction of (i) probes learning real discriminative information (Table 4), (ii) saturation at small $|\alpha|$ in both signs, (iii) over-correction toward the "no" bias at large negative $|\alpha|$, and (iv) per-model $|\Delta\text{H}|$ values bounded by the leading singular direction's variance fraction (Appendix D.5) is consistent with the probe direction capturing only one component of a multi-directional hallucination signal: enough to produce a partial reduction when applied with the correct sign, but not enough to match per-component scaling, which addresses the full multi-directional structure.

Table 5: **ITI on the positive-α slice of the grid (selection set, n=100, representative model: InternVL3-8B; baseline halluc. 20.0%).** Entries show (Halluc., TP/FP/TN/FN). At $K = 12$, $\alpha \in \{5, 10, 15\}$ produce identical decisions despite the linear scaling of the injected direction; at $K = 96, \alpha = 20$, the probe-trained direction increases hallucination by 3.7$\times$ from baseline. The probe direction therefore points toward the hallucination class, motivating the symmetric grid (Table 6).

K	$\alpha = 5$	$\alpha = 10$	$\alpha = 15$	$\alpha = 20$
12	20.0% (47/10/40/3)	20.0% (47/10/40/3)	20.0% (47/10/40/3)	24.0% (47/12/38/3)
24	20.0% (47/10/40/3)	24.0% (47/12/38/3)	24.0% (47/12/38/3)	24.0% (48/12/38/2)
48	24.0% (48/12/38/2)	32.0% (49/16/34/1)	40.0% (49/20/30/1)	50.0% (49/25/25/1)
96	24.0% (48/12/38/2)	36.0% (49/18/32/1)	50.0% (49/25/25/1)	**74.0% (49/37/13/1)**

Table 6: **ITI on the negative-α slice of the grid (selection set, n=100, same model and baseline as Table 5).** Entries show (Halluc., TP/FP/TN/FN). With the sign reversed, the probe direction reduces hallucination, with the strongest selection-set effects concentrated at large K and large $|\alpha|$. The selected configuration on this model is $K = 96, \alpha = -15$ (val: 6.0% halluc., 0.92 acc.; test: -11.0 pp, -1.0 pp acc., Table 3). At $K = 96, \alpha = -20$ hallucination drops further to 6.0% on val but recall collapses to 0.78 (test recall 0.92$\rightarrow$0.73), revealing an over-correction toward the "no" bias.

K	$\alpha = -5$	$\alpha = -10$	$\alpha = -15$	$\alpha = -20$
12	20.0% (47/10/40/3)	20.0% (47/10/40/3)	20.0% (47/10/40/3)	18.0% (47/9/41/3)
24	20.0% (47/10/40/3)	20.0% (47/10/40/3)	18.0% (47/9/41/3)	18.0% (47/9/41/3)
48	18.0% (47/9/41/3)	14.0% (47/7/43/3)	10.0% (47/5/45/3)	8.0% (46/4/46/4)
96	16.0% (47/8/42/3)	10.0% (47/5/45/3)	**6.0% (45/3/47/5)**	6.0% (39/3/47/11)†

† Same val hallucination rate as K=96, α= $-$ 15, but with markedly lower recall (0.78 vs. 0.90); rejected by the max-accuracy tie-break.

D.4 Grounding-Pathway Suppression: Specificity Control

As a specificity control complementing the random-component, directional, and ITI baselines above, we apply the uniform-scaling protocol of Section 5.5 to the grounding pathway ($d > 0$) instead of the hallucination pathway. The hyperparameter sweep is identical: $s \in \{0, 0.25, 0.5, 0.75\}$ on the 100-sample selection set. Table 7 reports accuracy and hallucination rate at each scale.

At s=0, full grounding ablation, accuracy collapses to chance (0.50–0.51) in every one of the five models, with Δacc ranging from -0.31 to -0.39 despite their differing grounding-pathway sizes (11–29 components) and depth profiles. The graded effect across s is also informative. At intermediate s, models retain partial accuracy but degrade toward different default biases. Llama-3.2, InternVL3-14B, and LLaVA collapse toward "always no" (halluc. near zero with acc. near chance), while InternVL3-8B collapses toward random guessing (halluc. ≈ 0.48 with acc. ≈ 0.51). The fact that grounding ablation produces these qualitatively different collapse modes across models, while hallucination-pathway suppression (Table 13) produces consistent reductions, confirms that the two pathways carry functionally distinct information rather than reflecting a shared yes-bias axis. Partial suppression (s=0.5) leaves InternVL3 accuracies nearly unchanged, consistent with the within-subset redundancy that CPA identifies in the grounding pathway (Section 5.2): under partial suppression, multiple grounding components carry overlapping information and the residual signal suffices.

Table 7: **Grounding-pathway suppression at varying scaling factors (selection set, n=100).** Same uniform-scaling protocol as Section 5.5, applied to $d > 0$ components. At s=0 (full ablation) accuracy collapses to chance in every model. Compare with Table 13 (hallucination-pathway suppression on the same models).

Model	Metric	baseline	s=0	s=0.25	s=0.5	s=0.75
Qwen3-VL-8B	acc.	0.860	**0.500**	0.880	0.870	0.860
	halluc.	0.120	0.000	0.160	0.140	0.120
LLaVA-v1.6-7B	acc.	0.890	**0.500**	0.790	0.880	0.880
	halluc.	0.100	0.000	0.300	0.080	0.120
Llama-3.2-V-11B	acc.	0.810	**0.500**	0.600	0.790	0.800
	halluc.	0.260	0.000	0.020	0.320	0.320
InternVL3-8B	acc.	0.870	**0.510**	0.750	0.880	0.860
	halluc.	0.200	0.480	0.500	0.160	0.220
InternVL3-14B	acc.	0.870	**0.500**	0.520	0.890	0.890
	halluc.	0.200	0.000	0.000	0.080	0.140

D.5 Direct Geometric Evidence on the Dimensionality of the Hallucination Signal

Section 5.5 reports that under the symmetric α grid, single-direction interventions (ITI, mean-difference projection) produce a bounded effect on hallucination (mean $|\Delta\text{H}|$ of 4.1 pp for ITI and 0.8 pp for mean-difference projection, against 7.4 pp for per-component scaling on the same components), with the per-model effect tracking the per-component activation geometry. This appendix provides a direct geometric test of the multi-directional claim, independent of any intervention.

Procedure. For each hallucination-pathway component, we collect last-token pre-output activations on n=1,000 POPE-adversarial samples (the same protocol used to extract mean-difference directions in Appendix D) and partition them by whether the model answered correctly or hallucinated. We then form $\Delta X = X_{\text{halluc}} - \mu_{\text{correct}}$, the per-sample deviation of hallucinating activations from the correct-class centroid, and compute its singular spectrum. Two summary numbers per component:

- Participation ratio $\text{PR} = (\sum_i \sigma_i^2)^2 / \sum_i \sigma_i^4$, an effective dimensionality that equals 1 for a one-dimensional signal and grows as variance spreads over multiple comparable directions.
- Top-1 variance fraction $\sigma_1^2 / \sum_i \sigma_i^2$, the share of variance captured by the leading singular direction (the only direction a single-direction intervention can target).

We additionally report rank_{90}, the number of singular directions needed to capture 90% of the variance. All metrics are averaged over the hallucination-pathway components of each model.

Results. Table 8 reports the per-model averages. The leading singular direction captures 45–69% of the variance across the five models, and the participation ratio ranges from 2.4 to 6.8. No model produces a hallucination signal close to one-dimensional. The geometric implication is direct: even a perfectly estimated direction $\hat{d}$ can directly modify at most $\sigma_1^2 / \sum_i \sigma_i^2$ of the per-component activation variance, with the remainder propagating through directions orthogonal to $\hat{d}$. This bound is consistent with the per-model behavior of ITI under the symmetric grid (Table 3, Appendix D.3): on four of five models, ITI's relative effect (versus per-component scaling) lies at or below the leading-direction variance fraction, with InternVL3-8B as the exception, in which only six hallucination-pathway components carry the signal and an ITI configuration with K=96 heads spans much of the relevant subspace.

The cross-model variation in PR also tracks cross-model variation in the hallucination-pathway CPA signature (Section 5.2): InternVL3-8B has both the lowest PR (2.40) and the weakest hallucination-pathway behavior under CPA (no cross-subset polarity flip, mean IE remaining negative in both subsets at -0.09 and -0.16), consistent with a small set of approximately independent bias sources, while LLaVA has the highest PR (6.78) and operates with the most dispersed pathway (30 components). Llama-3.2 is intermediate on both axes. The internal consistency of two independent measurements, pathway-level behavior (CPA) and per-component activation geometry (PR), supports

the interpretation that hallucination-pathway components contribute biases along multiple activation directions.

We deliberately limit this analysis to hallucination-pathway components. The corresponding measurement on grounding-pathway components reflects within-component activation spread driven by per-image visual variation, which is a different geometric object than the between-component redundancy quantified by CPA in Section 5.2; we do not interpret it here.

Table 8: **Singular-spectrum geometry of the hallucination signal.** For each model, mean over hallucination-pathway components (k components per model, matching Table 1) of: participation ratio PR (effective dimensionality), top-1 variance fraction $\sigma_1^2/\sum\sigma_i^2$, and rank_{90} (number of directions needed to capture 90% of variance). n=1,000 POPE-adversarial samples per model; per-component activations partitioned by model answer, $\Delta X = X_{\text{halluc}} - \mu_{\text{correct}}$.

Model	k	mean PR	mean $\sigma_1^2/\sum\sigma_i^2$	mean rank_{90}
Qwen3-VL-8B	12	3.07	0.634	12.5
LLaVA-v1.6-7B	30	6.78	0.452	28.6
Llama-3.2-V-11B	20	4.85	0.564	31.9
InternVL3-8B	6	2.40	0.688	7.8
InternVL3-14B	16	3.16	0.643	13.2

E Cross-Architecture Circuit Comparison

Table 9 gives the full pairwise comparison supporting Section 5.4. We report the six unique pairwise comparisons among Qwen3-VL, LLaVA, Llama-3.2, and InternVL3-8B, and treat InternVL3-14B as a within-family scale comparison analyzed in the main text. Fisher's exact test on 2×2 contingency tables is non-significant for all six pairs, consistent with a shared macro-level spatial pattern (hallucination components concentrated at early layers/boundaries, grounding components in mid-to-late layers). TOST equivalence testing at an equivalence bound of $|r| < 0.3$ positively confirms negligibility for 2/6 pairs; the remaining pairs are inconclusive (the data are compatible with either small non-zero correlation or negligibility) rather than significantly correlated.

Table 9: **Cross-architecture circuit comparison: pairwise statistics.** Fisher's exact test on macro-level spatial distribution; Pearson correlation on depth-normalized effect-size profiles; TOST equivalence test with bound $|r| < 0.3$. "Micro" summarizes the TOST verdict.

Pair	**Fisher** p	r	**TOST** p	**95% CI**	**Micro**
Qwen–LLaVA	0.715	−0.05	**0.005**	[−.23, .13]	Negligible
Qwen–Llama	0.713	−0.20	0.149	[−.37, −.03]	Inconclusive
Qwen–InternVL3-8B	0.615	+0.24	0.244	[.04, .42]	Inconclusive
LLaVA–Llama	0.229	+0.06	**0.007**	[−.15, .28]	Negligible
LLaVA–InternVL3-8B	1.000	−0.21	0.164	[−.39, −.02]	Inconclusive
InternVL3-8B–Llama	0.352	−0.33	0.631	[−.47, −.18]	Mod. negative

F Logit Lens Validation: Per-Pathway Layer Contributions

Table 10 reports the full per-model results underlying Section 5.3. For each model and sample type (correct vs. hallucinating), we report the mean per-layer Δlogit_diff averaged separately over layers assigned to the grounding pathway and the hallucination pathway by our patching-based criterion (Section 3.2), together with the between-pathway effect size and significance. In four of five models, grounding-pathway layers contribute positively to the correct answer on correct samples and negatively on hallucinating samples, with hallucination-pathway layers showing the opposite polarity, providing patching-independent support for the functional roles assigned by the circuit-extraction procedure. Llama-3.2-Vision is the exception discussed in the main text.

Table 10: **Logit lens validation.** Mean Δlogit_diff per layer, grouped by pathway membership. Positive Δ: layer increases correct-answer probability. The functional role reverses during hallucination in four of five models. n=200 per model.

Model	Sample type	Grounding Δ	Halluc. Δ	d	p
Qwen3-VL	Correct	+1.623	−0.675	1.80	<0.001
	Hallucinating	−0.786	+0.464	−1.15	<0.001
LLaVA-v1.6	Correct	+0.328	+0.011	0.46	<0.001
	Hallucinating	−0.349	+0.128	−0.65	0.029
Llama-3.2	Correct	−0.140	−0.290	0.48	<0.001
	Hallucinating	+0.228	+0.290	−0.18	0.430
InternVL3-8B	Correct	+0.934	−0.998	2.91	<0.001
	Hallucinating	−0.429	+0.103	−0.76	0.007
InternVL3-14B	Correct	+0.309	−0.060	1.20	<0.001
	Hallucinating	−0.216	+0.047	−1.25	<0.001

G Full Numerical Results

This appendix collects the full numerical tables referenced from the main paper: per-model strongest components (Table 11), pathway-level CPA values (Table 12), per-model intervention configurations and the LLaVA top-k ablation (Table 13), and cross-benchmark generalization (Table 14).

Table 11: **Strongest grounding and hallucination components per model.** For each model, we report the three grounding components with the largest positive Cohen's d and the three hallucination components with the most negative Cohen's d. Components are selected from the significant circuit ($p_{\text{adj}} < 0.05$, Benjamini–Hochberg corrected; $|d| > 0.3$; $n = 1{,}000$ per model). "Comp." denotes the component identifier (layer.type), and "Depth" denotes normalized layer position.

	Top grounding (+d)			Top hallucination (−d)		
Model	Comp.	Depth	d	Comp.	Depth	d
Qwen3-VL-8B	L0.mlp	0%	+0.940	L34.attn	97%	−0.700
	L23.mlp	66%	+0.932	L35.mlp	100%	−0.631
	L33.mlp	94%	+0.875	L32.mlp	91%	−0.606
LLaVA-v1.6-7B	L20.mlp	65%	+0.922	L7.mlp	23%	−0.651
	L18.mlp	58%	+0.880	L31.attn	100%	−0.632
	L17.mlp	55%	+0.718	L12.attn	39%	−0.623
Llama-3.2-V-11B	L3.attn	8%	+0.912	L16.mlp	41%	−0.889
	L12.attn	31%	+0.859	L7.mlp	18%	−0.772
	L1.mlp	3%	+0.858	L19.attn	49%	−0.713
InternVL3-8B	L22.mlp	81%	+1.586	L27.mlp	100%	−0.870
	L20.mlp	74%	+1.306	L9.mlp	33%	−0.438
	L19.mlp	70%	+1.223	L0.attn	0%	−0.436
InternVL3-14B	L35.mlp	74%	+0.880	L2.attn	4%	−0.733
	L32.mlp	68%	+0.850	L4.mlp	9%	−0.577
	L0.mlp	0%	+0.614	L15.attn	32%	−0.458

Table 12: **Pathway-level CPA: full numerical values underlying Figure 3.** For each (model, pathway, subset) we report the mean individual indirect effect ($\overline{\text{IE}}$), the fraction of positive IEs (f_+), the magnitude ratio $|\text{IE}(\mathcal{P})|/\sum_i |\text{IE}(C_i)|$ (values <1 indicate within-subset redundancy), and the polarity-free magnitude diagnostic $\text{MagDiff} = |\text{IE}(\mathcal{P})| - \sum_i |\text{IE}(C_i)|$. $\Delta\overline{\text{IE}}$ is the cross-subset shift in mean IE (Cohen's d, two-sample t-test); $\Delta\text{MagDiff}$ similarly tests the cross-subset shift in within-subset redundancy. n=400 per model. The grounding pathway flips IE polarity between subsets in every model while remaining redundant in both subsets; the hallucination pathway shows no comparable polarity flip.

			Correct samples		**Halluc. samples**		**Cross-subset shift**	
Model	**Pathway**	k	$\overline{\text{IE}}$ / f_+	$\|\text{IE}\|/\sum$	$\overline{\text{IE}}$ / f_+	$\|\text{IE}\|/\sum$	$\Delta\overline{\text{IE}}$: d / p	ΔMagDiff: d / p
Qwen3-VL	Grounding	14	+1.59 / 0.83	0.37	−1.08 / 0.24	0.34	−2.45 / $<.001$	−0.25 / .120
	Halluc.	12	−0.53 / 0.32	0.62	+0.01 / 0.52	0.41	−0.14 / .385	+0.42 / .009
LLaVA-v1.6	Grounding	18	+0.51 / 0.68	0.38	−0.31 / 0.27	0.21	−1.60 / $<.001$	−0.13 / .370
	Halluc.	30	−0.01 / 0.43	0.19	+0.22 / 0.63	0.07	+0.59 / $<.001$	+0.35 / .019
Llama-3.2	Grounding	11	+0.13 / 0.61	0.22	−0.07 / 0.26	0.30	−1.72 / $<.001$	−0.45 / .001
	Halluc.	20	−0.05 / 0.25	0.23	+0.06 / 0.49	0.22	+0.99 / $<.001$	+0.14 / .302
InternVL3-8B	Grounding	21	+0.83 / 0.74	0.37	−0.31 / 0.33	0.41	−1.83 / $<.001$	−0.88 / $<.001$
	Halluc.	6	−0.09 / 0.42	0.69	−0.16 / 0.41	0.50	+0.20 / .206	+0.42 / .002
InternVL3-14B	Grounding	29	+0.67 / 0.70	0.19	−0.28 / 0.21	0.21	−2.02 / $<.001$	−0.92 / $<.001$
	Halluc.	16	−0.46 / 0.28	0.24	+0.09 / 0.50	0.41	+1.04 / $<.001$	−0.59 / $<.001$

Table 13: **Full intervention numerical values underlying Figure 4.** Top: per-model baseline and selected intervention configuration on POPE-adversarial, computed on the 400-sample held-out set. Bottom: LLaVA-v1.6 top-k ablation (s=0.5), with k selected on the disjoint 100-sample selection set. The selected k for LLaVA is k=10. These numbers are consistent with the POPE-adv column of Table 14, which reports the same intervention applied alongside other benchmarks.

Model	**Config**	**Acc.**	**Halluc.**	**ΔAcc.**	**ΔHall.**	**Rel. ↓**
Qwen3-VL	Baseline	88.2%	8.4%	—	—	—
	s=0.25, all 12	87.2%	**4.0%**	−1.0	−4.4	52%
LLaVA-v1.6	Baseline	88.5%	5.0%	—	—	—
	s=0.5, top-10	88.5%	**3.0%**	+0.0	−2.0	40%
Llama-3.2	Baseline	84.8%	18.4%	—	—	—
	s=0.5, all 20	84.6%	**4.4%**	−0.2	−14.0	76%
InternVL3-8B	Baseline	88.4%	15.6%	—	—	—
	s=0.5, all 6	86.4%	**5.6%**	−2.0	−10.0	64%
InternVL3-14B	Baseline	88.0%	15.5%	—	—	—
	s=0.5, all 16	87.8%	**9.0%**	−0.3	−6.5	42%
LLaVA-v1.6 top-k ablation (s=0.5, evaluated on the 400-sample held-out set):						
LLaVA-v1.6	k = — (baseline)	88.5%	5.0%	—	—	—
	k = 3	89.3%	6.5%	+0.8	+1.5	—
	k = 5	89.0%	6.5%	+0.5	+1.5	—
	k = 8	89.0%	3.0%	+0.5	−2.0	40%
	k = **10**	**88.5%**	**3.0%**	+0.0	−2.0	**40%**
	k = 15	88.8%	4.0%	+0.3	−1.0	20%
	k = 20	84.0%	3.5%	−4.5	−1.5	30%
	k = 30 (all)	79.5%	1.5%	−9.0	−3.5	70%

Table 14: **Cross-benchmark generalization: full numerical values.** We evaluate the same targeted intervention (circuits and scale factors selected as in Figure 4) across three POPE splits and three AMBER hallucination types, on the 400-sample held-out set per condition (with the corresponding 100-sample selection sets used only for hyperparameter selection on POPE-adversarial). ΔH: absolute change in hallucination rate (percentage points, negative = reduction). Rel.↓: relative reduction. Dashes indicate negligible change ($|\Delta\text{H}| < 1$ pp). AMBER uses only negative-ground-truth questions, so hallucination rate equals false-positive rate.

	POPE-adv		**POPE-pop**		**POPE-rand**		**AMBER-exist**		**AMBER-attr**		**AMBER-rel**	
Model	ΔH	Rel.↓	ΔH	Rel.↓	ΔH	Rel.↓	ΔH	Rel.↓	ΔH	Rel.↓	ΔH	Rel.↓
Qwen3-VL-8B	−4.4	52%	−2.5	45%	–	–	−2.2	10%	–	–	−16.5	36%
LLaVA-v1.6-7B	−2.0	40%	−1.0	40%	–	–	−3.0	15%	+1.5	–	−2.3	6%
Llama-3.2-V-11B	−14.0	76%	−7.0	74%	−4.0	80%	−10.8	42%	−9.5	23%	−6.5	47%
InternVL3-8B	−10.0	64%	−6.0	65%	−2.4	60%	−3.4	15%	−8.6	17%	−15.8	39%
InternVL3-14B	−6.5	42%	−6.5	68%	−1.0	29%	–	–	−1.5	3%	+1.0	–

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state three contributions—identifying a shared dual-pathway organization across five VLMs (Section 5.1, Figure 2), introducing Conditional Pathway Analysis and surfacing a consistent grounding-pathway IE polarity flip between correct and hallucinating samples (Section 5.2, Figure 3), using targeted suppression as a causal test under which static-direction baselines produce a bounded effect tracking the per-component activation geometry (Section 5.5), and demonstrating selective transfer to relational but not attribute hallucination (Section 5.6).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 notes that per-component scaling serves as a causal test but does not constitute a deployable mitigation, and that the intervention transfers to relational but not attribute hallucination, indicating partially distinct circuits. Section 5.3 discusses the Llama-3.2-Vision logit-lens exception, and Appendix B reports threshold sensitivity of circuit extraction.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is empirical. The methodological objects (indirect effect in Eq. 1, the CPA interaction term in Eqs. 2–3, and the scaling intervention in Eq. 5) are operational definitions rather than theoretical claims.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 specifies all five models with HuggingFace identifiers, the POPE-adversarial split sizes for each phase (1,000 / 400 / 200 / 500), the disjoint selection (100) and held-out evaluation (400) sets used for hyperparameter selection, the corruption protocol (3σ Gaussian noise on visual tokens at layer 0), and greedy decoding. Appendix A details the model-specific patching adapters, CPA implementation, logit-lens computation, and the 3σ corruption verification. Appendix D–D.3 provide the full hyperparameter grids for both linear-steering baselines.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is publicly available at `https://github.com/jiaxin26/DualPath-VLM` under the MIT license. It contains the unified patching framework with model-specific adapters for all four VLM families (`patching/`), the circuit-extraction and cross-architecture analysis (`analysis/`), per-phase run scripts (`experiments/run_patching.py`, `run_pathway_patching.py`, `run_logit_lens.py`, `run_intervention.py`), a top-level `reproduce.sh` pipeline, `requirements.txt`, and a README describing the full ten-phase reproduction. POPE and AMBER are publicly available; the five VLMs are loaded from public HuggingFace checkpoints.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

Justification: Section 4 (Table 1) lists all five models with their HuggingFace identifiers, layer counts, hallucination rates, and sample sizes for each phase. The intervention scaling factor $s \in \{0, 0.25, 0.5, 0.75\}$, the LLaVA top-$k$ grid $k \in \{3, 5, 8, 10, 15, 20, 30\}$, and the linear-steering baseline grids ($\alpha \in \{\pm 0.5, \pm 1.0, \pm 1.5\}$ for mean-difference; $K \in \{12, 24, 48, 96\}$ and $\alpha \in \{\pm 5, \pm 10, \pm 15, \pm 20\}$ for ITI, with ties broken by maximum selection-set accuracy) are all reported, with hyperparameters selected only on the disjoint 100-sample selection set and applied unchanged to all other benchmarks (Section 4). Appendix B–D.3 provide further sensitivity analysis.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Circuit extraction (Section 3.2) uses Welch's t-test with Benjamini–Hochberg FDR correction at $\alpha = 0.05$ and an effect-size threshold $|d| > 0.3$. CPA (Section 5.2) reports Cohen's d and p-values for both within-condition (one-sample t-test) and between-condition (two-sample t-test) comparisons; full numerical values are in Appendix C and Table 12. Cross-architecture comparison (Section 5.4) uses Fisher's exact test plus TOST equivalence with bound $|r| < 0.3$. Threshold sensitivity is quantified in Appendix B via Jaccard similarity at $|d| \in \{0.25, 0.30, 0.35\}$.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix A reports that all experiments were conducted on a single NVIDIA H200 144 GB GPU; component-level patching on 1,000 samples takes approximately 2–18

hours per model depending on size and sequence length (Llama-3.2 fastest at $\sim$1.8 s/sample, InternVL3-14B slowest at $\sim$18 s/sample), CPA adds 2–3 hours per model, and logit lens completes in under 1 hour per model. The full pipeline for one model runs in under 24 hours on a single GPU, for a total of approximately 5 GPU-days across the five models.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses publicly available VLMs and benchmarks (POPE [15], AMBER [32]). It does not involve human subjects, personal data, or scraped content, and the released artifacts are an interpretability analysis pipeline rather than a generative model with misuse potential.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 6 (Broader impact) notes that a mechanistic account of hallucination supports more reliable VLMs, and that pathway-level interventions can increase hallucination when applied in the opposite direction (Appendix D.3, Table 5). The released code is an analysis pipeline on existing public checkpoints and introduces no new generative capability.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released artifact is an interpretability and analysis codebase. The paper does not release pretrained generative models, scraped datasets, or other components with elevated misuse risk; the studied VLMs are pre-existing public checkpoints loaded from HuggingFace.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All studied VLMs are cited and used under their respective public licenses: Qwen3-VL-8B (Apache 2.0) [23], LLaVA-v1.6-7B (Apache 2.0) [17], Llama-3.2-Vision-11B (Llama 3.2 Community License) [19], and InternVL3-8B / 14B (MIT) [41]. The benchmarks POPE (MIT) [15] and AMBER (Apache 2.0) [32] are used under their published terms. Activation patching [18], the ITI baseline [13], and the logit-lens technique [21] are credited to their original authors.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released codebase is licensed under MIT and shipped with a README documenting the four-family patching adapters, the per-phase run scripts, an end-to-end `reproduce.sh` pipeline, and a `requirements.txt` pinning supported library versions.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects; all experiments are conducted on public benchmarks (POPE, AMBER) and public model checkpoints.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human-subjects research, so IRB approval (or equivalent) is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The five VLMs in this study are the *objects* of mechanistic analysis, not tools used in the research methodology itself. The core methods (activation patching, Conditional Pathway Analysis, statistical testing, targeted suppression) do not depend on LLMs as a methodological component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.