# STEERPRINT: STEERING-VECTOR-GUIDED FINGERPRINTS FOR BLACK-BOX LVLM OWNERSHIP VERIFICATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Black-box ownership verification determines whether a deployed large vision-language model (LVLM) derives from a protected checkpoint using only query responses. Existing visual fingerprints primarily optimize observable output signatures, leaving the inherited internal features underlying robustness and source specificity implicit. We formalize provenance anchors: internal features whose agreement with the source separates a specified derivative class from non-derived models. We derive a sufficient condition for preserving this separation through black-box readout and introduce SteerPrint, which trains one reusable visual fingerprint per target behavior from source-extracted steering vectors. Across natural prompts, each image matches the hidden-state displacement and first-response-token distribution induced by direct steering, translating an internal direction into an observable behavioral response. Verification requires only returned text and no model modification. Using 300 queries per model, SteerPrint improves the per-source mean trigger rate by 4.9–23.2 percentage points over the strongest baseline across 40 fine-tuned derivatives of four LVLMs. Across 27 non-derived models, the trigger rate is at most 1/300 ($\leq 0.333\%$) per model, yielding 0/27 false attributions. We further characterize fingerprint persistence and failure under pruning, quantization, merging, and component-level modification.

## 1 INTRODUCTION

Released large vision-language models (LVLMs) (Liu et al., 2024; Bai et al., 2025a) can be adapted to downstream tasks and redeployed behind APIs. Owners must then determine whether a service derives from a protected checkpoint using only image–text queries and returned responses. Model fingerprinting addresses this problem through queries whose responses reveal source reuse, without modifying the released checkpoint (Cao et al., 2021; Chen et al., 2022; Yang & Wu, 2024; Bai et al., 2025b; Wang et al., 2025; Xia et al., 2026).

A useful fingerprint must reconcile transferability with source specificity: derivatives should retain its response after modification, while non-derived models remain distinguishable (Bai et al., 2025b; Hu et al., 2025). This trade-off also arises in classification and model-merging fingerprints, where moving queries away from a decision boundary can improve robustness at the expense of uniqueness (Cao et al., 2021; Cong et al., 2024; Yamabe et al., 2025; Yang et al., 2026). Yet source-side response success alone does not identify what information should be preserved by derivatives or what should remain absent from non-derived models. For LVLMs, the central question is therefore what source information a query should encode so that its response survives model derivation without becoming generic across non-derived models.

Existing LVLM fingerprints optimize visual inputs for observable output signatures: PLA and AGDI target responses (Wang et al., 2025; Xia et al., 2026), while SIF distills a text-generation watermark signal into images without modifying model parameters (Zhao et al., 2026). These objectives improve response persistence, but they leave the provenance signal implicit: they do not explicitly specify which source-side internal feature should survive model derivation, whether that feature remains discriminative against non-derived models, or which internal realization of a source-side response the query should rely on. To make this structure explicit, we formalize a *provenance anchor*: an



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

internal feature whose agreement with the source remains high in a specified derivative class and low in non-derived models.

Steering vectors (SVs) offer a concrete candidate: they encode activation directions that induce targeted behavioral changes when applied to hidden states (Zou et al., 2023a; Turner et al., 2024; Rimsky et al., 2024; Arditi et al., 2024). We characterize when these directions are preserved under model derivation and observe strong source–derivative alignment after fine-tuning and quantization. Their stability and observable behavioral effects motivate SV-guided fingerprint construction.

We propose **SteerPrint**, which extracts SVs from the protected source and trains one reusable visual fingerprint per target behavior. Each image is optimized across natural prompts to reproduce the hidden-state displacement and first-response-token distribution induced by direct steering. Parameter-noise augmentation encourages this effect to persist under nearby parameter changes (Moosavi-Dezfooli et al., 2017; Qi et al., 2024; Wang et al., 2025). During verification, the owner pairs the fixed images with held-out natural prompts and detects the target behavior in returned text. Construction uses source-model internals; verification requires no internal access to the suspect.

Our key contributions are:

- We introduce a *provenance-anchor paradigm* for black-box ownership verification, making inheritable, source-specific internal features an explicit design target for fingerprint construction, and derive sufficient conditions under which their separation is preserved through black-box readout.
- We develop **SteerPrint**, which translates source-extracted SVs into reusable visual fingerprints through hidden-state matching, output distillation, and parameter-noise augmentation.
- We demonstrate transfer across LVLM derivatives and separation from non-derived controls, and examine the underlying mechanism through SV-alignment measurements, visual-displacement analysis, and targeted interventions.

## 2 Related Work

**White-box fingerprinting.** White-box methods infer model provenance from the suspect's parameters, hidden representations, gradients, or attention patterns (Shao et al., 2025; Wu et al., 2025; Zhang et al., 2026). REEF compares representation geometry across models (Zhang et al., 2025), while Xu & Sheng (2026) use refusal vectors, a behavior-specific instance of a steering direction, as static internal features for white-box provenance detection. However, prior work does not systematically characterize how steering vectors behave under diverse model derivations, nor does it provide principles for transforming internal directions into black-box triggers. SteerPrint explicitly analyzes the preservation properties of steering vectors across fine-tuning, pruning, merging, and quantization, and establishes principles to externalize these internal properties into black-box visual fingerprints, verifying suspect models solely through returned text.

**Black-box fingerprinting.** Black-box methods infer ownership from responses to selected queries. Post-release fingerprinting constructs queries for already released checkpoints without modifying their parameters (Cao et al., 2021; Chen et al., 2022). Subsequent work strengthens robustness to downstream adaptation and model merging (Cong et al., 2024; Yamabe et al., 2025). For LVLMs, PLA introduces adversarial parameter updates during visual-trigger construction (Wang et al., 2025), whereas AGDI jointly optimizes response-level supervision, semantic alignment, and adversarial training (Xia et al., 2026). SIF (Zhao et al., 2026) distills text-generation watermarks into visual inputs for non-invasive black-box ownership verification. SteerPrint instead externalizes source-extracted behavioral steering directions and studies their preservation under model derivation. LLM DNA takes a complementary functional approach, representing models through semantic structures extracted from their responses (Wu et al., 2026); SteerPrint optimizes source-specific visual queries for ownership verification.

**Steering vectors.** Representation engineering studies how model activations encode high-level behaviors as directions or subspaces (Zou et al., 2023a). Activation steering estimates such directions from contrasting examples and adds the resulting vectors to hidden states to induce targeted behavioral changes (Turner et al., 2024; Rimsky et al., 2024). Refusal is a concrete instance: prior work shows that a low-dimensional representation direction can causally mediate refusal behavior (Arditi et al.,

2024). SteerPrint builds on this line of work by characterizing how such behavior-related directions change under model derivation and by externalizing their preserved effects into visual fingerprints for black-box verification.

# 3 Problem Formulation

## 3.1 Black-Box Ownership Verification

A checkpoint with parameters $\theta$ defines a conditional generation function $f_\theta$ that maps an image–text query $z = (x, q)$ to a distribution over response texts, where $x$ is an image and $q$ a text prompt. Let $M_s = f_{\theta_0}$ denote the protected source. Within a specified collection of candidate checkpoints, $\mathcal{D}_s$ contains $M_s$ and its derivatives under the transformations of interest, such as fine-tuning, compression, merging, or component replacement. The remaining checkpoints form $\mathcal{U}_s$. Membership follows provenance: a same-family model not derived from $M_s$ belongs to $\mathcal{U}_s$.

The owner has full access to $M_s$ during fingerprint construction. During verification, the suspect model $M$ exposes only generated text: the verifier can submit query $z$ and observe response $Y \sim M(z)$, without access to parameters, hidden states, logits, or token probabilities.

Let $\mathcal{Z}$ and $\mathcal{Y}$ denote the query and response spaces. A fingerprint $\pi_s = (Q_s, g_s)$ specifies a query distribution $Q_s$ and a scoring function $g_s : \mathcal{Z} \times \mathcal{Y} \to [0, 1]$ that depends only on the submitted query and returned text. Its expected score on $M$ is $F_{\pi_s}(M) = \mathbb{E}_{z\sim Q_s,\, Y\sim M(z)}[g_s(z, Y)]$, which the verifier estimates from finitely many queries. We seek a fingerprint that assigns every derivative a higher expected score than any non-derived model, yielding a positive separation margin

$$\Gamma_{\mathrm{BB}} = \inf_{M\in\mathcal{D}_s} F_{\pi_s}(M) - \sup_{M\in\mathcal{U}_s} F_{\pi_s}(M) > 0. \tag{1}$$

A positive margin permits a single threshold to separate the two classes by their expected scores.

## 3.2 Provenance Anchors

A high score on $M_s$ alone does not establish the separation in Eq. (1). Existing visual fingerprints optimize inputs for designated responses and encourage persistence through parameter perturbations or other robustness objectives (Wang et al., 2025; Xia et al., 2026). Although construction exploits source-model internals, the output target leaves unspecified which feature should survive model modification and distinguish derivatives from non-derived models. We make that internal feature an explicit construction target.

Feature selection has two requirements. *Inheritance* requires derivatives to retain the feature despite model modification. *Specificity* requires non-derived models to differ from the source in that feature, preventing attribution based on shared generic properties. A feature and comparison rule satisfying both requirements form a *provenance anchor*.

Let $\phi(M)$ denote a feature extracted from the parameters or activations of $M$, and let $A_s(M) \in [0, 1]$ measure its agreement with the source reference $\phi(M_s)$. We fix the extraction and comparison rules during construction and express inheritance and specificity as bounds on this internal agreement score.

**Definition 1** (Provenance anchor)**.** The pair $(\phi, A_s)$ is a provenance anchor relative to $(\mathcal{D}_s, \mathcal{U}_s)$ if there exist $0 \le a_- < a_+ \le 1$ such that

$$\underbrace{\inf_{M\in\mathcal{D}_s} A_s(M) \ge a_+}_{\text{inheritance}}, \qquad \underbrace{\sup_{M\in\mathcal{U}_s} A_s(M) \le a_-}_{\text{specificity}}. \tag{2}$$

These bounds provide an internal separation of at least $\Gamma_A = a_+ - a_- > 0$.

Both requirements are essential. A shared feature carries little provenance information even when fine-tuning preserves it perfectly; a source-specific feature loses its value when derivation erases it. The anchor is therefore defined relative to the chosen derivative and non-derived classes.

### 3.3 From Anchors to Black-Box Verification

The verifier cannot inspect $A_s(M)$ through a generation API. It must instead recover the feature's discriminative signal from query responses. Two conditions make this possible. *Input realizability* requires an admissible query to induce an internal effect associated with the feature. *Behavioral observability* requires that effect to produce a detectable change in returned text. Readout fidelity is measured by the discrepancy between observable $F_{\pi_s}(M)$ and internal $A_s(M)$ scores.

**Proposition 1** (Margin preservation under readout)**.** *Let $(\phi, A_s)$ satisfy Definition 1. Suppose there exist $\epsilon_{\mathcal{D}}, \epsilon_{\mathcal{U}} \geq 0$ such that $|F_{\pi_s}(M) - A_s(M)| \leq \epsilon_{\mathcal{D}}$ for $M \in \mathcal{D}_s$, and $|F_{\pi_s}(M) - A_s(M)| \leq \epsilon_{\mathcal{U}}$ for $M \in \mathcal{U}_s$. Then*

$$\Gamma_{\mathrm{BB}} \geq \Gamma_A - \epsilon_{\mathcal{D}} - \epsilon_{\mathcal{U}}. \tag{3}$$

*If the right-hand side is positive, any threshold $\tau \in (a_- + \epsilon_{\mathcal{U}},\ a_+ - \epsilon_{\mathcal{D}})$ separates the two classes by their expected response scores.*

The bound gives a concrete design criterion: keeping the combined readout error below the internal separation margin preserves black-box separation. The anchor supplies inheritance and specificity; query construction must expose that distinction through observable responses. Appendix B gives the proof and finite-query analysis. SteerPrint instantiates these two roles with source-extracted SVs and learned visual fingerprints, respectively.

## 4 SteerPrint

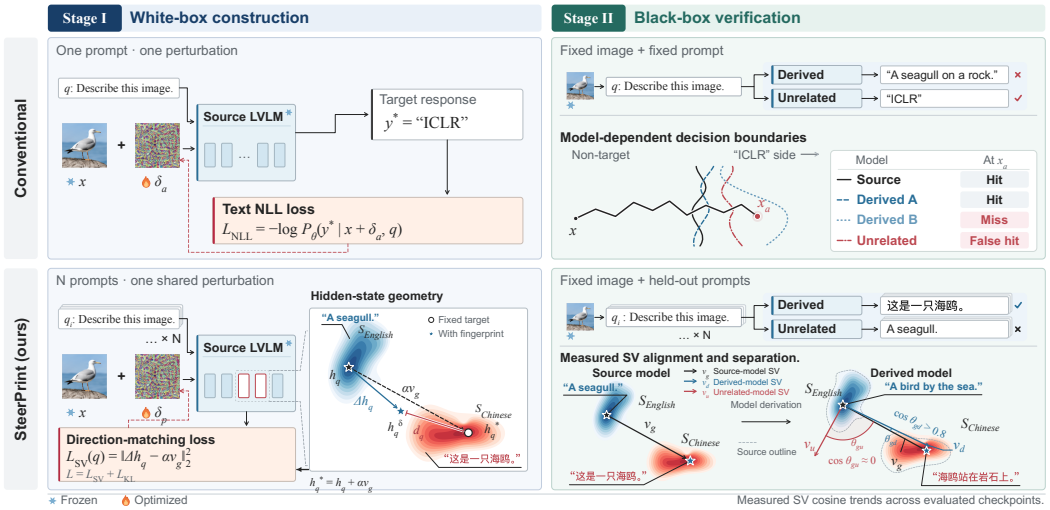



Figure 1: Output-anchored fingerprinting (top) optimizes an image for a fixed answer; changes in the trigger region can impair transfer or specificity. SteerPrint (bottom) instead optimizes a shared image across prompts to reproduce a source steering vector's internal and behavioral effects. Verification detects the target behavior in returned text. The lower-right panel summarizes measured SV similarities between the source and derived or cross-family models.

SteerPrint constructs visual fingerprints around an internal source feature rather than a prescribed answer. As illustrated in the upper row of Figure 1, optimizing an image to elicit a fixed string on the source can leave two failures unresolved: a derivative may lose the response, while a non-derived model may produce it. SteerPrint addresses this transferability–specificity trade-off by selecting a source steering vector (SV) as a candidate provenance anchor and optimizing one image across natural prompts to reproduce its hidden-state displacement and behavioral effect (lower row). The resulting image exposes the selected internal direction through text, so verification requires only image–text queries and returned responses.

### 4.1 Source-Model SVs

An SV captures an activation contrast associated with a target behavior. Adding it to the residual stream, the hidden representation passed between transformer blocks, can steer generation toward

that behavior (Zou et al., 2023a; Rimsky et al., 2024; Arditi et al., 2024). This intervention provides two complementary supervision signals: the SV specifies the desired hidden-state displacement, and the steered model supplies the corresponding output distribution.

For each of $K_b$ target behaviors $b_k$ and each layer $\ell$, let $\mu^+_{k,\ell}$ and $\mu^-_{k,\ell}$ denote the mean residual-stream states over target and contrast examples. We extract the SV as their normalized difference:

$$v_{k,\ell} = \frac{\mu^+_{k,\ell} - \mu^-_{k,\ell}}{\left\|\mu^+_{k,\ell} - \mu^-_{k,\ell}\right\|_2}. \tag{4}$$

The subtraction orients the vector toward the target behavior. We study Python → JavaScript (Py→JS), English → Chinese (En→Zh), and refusal on benign requests (Ans→Ref). These behaviors admit contrastive extraction examples and recognizable markers in generated text; Appendix E specifies the examples and extraction settings.

Our measurements show that source and derivative SVs remain closely aligned, whereas cross-family SVs have low cosine similarity to the source, as summarized in the lower-right panel of Figure 1. This empirical separation motivates SVs as candidate provenance anchors. Our mean-difference analysis explains when alignment persists: shifts shared by the target and contrast means cancel, while orthogonal drift relative to the retained contrast governs directional change (Appendix B.3). Section 5.3 reports the alignment measurements, Section 5.2 tests specificity in returned text, and Appendix B.8 analyzes the separation conditions for a fixed fingerprint.

### 4.2 Visual Fingerprint Construction

As shown in Stage I of Figure 1 (lower left), we train one image across natural prompts so that an ordinary API query reproduces the effect of direct SV intervention. These prompts do not request the target behavior. For each $b_k$, we optimize a perturbation $\delta_k$ on a clean carrier image $x^{\mathrm{cle}}$, yielding the fixed visual fingerprint $x^{\mathrm{fp}}_k = \mathrm{clip}_{[0,1]}(x^{\mathrm{cle}} + \delta_k)$. We pair this image with every training prompt $q_{k,n}$, for $n = 1, \ldots, N_k$, and supervise its internal displacement and output distribution.

**Matching the internal displacement.** The image should induce a displacement along the selected SV rather than merely reach an arbitrary state that produces the target response. The hidden-state panel of Figure 1 depicts the target and image-induced displacements as dashed and blue arrows from the same clean state. Let $h_{\ell,t}(x, q)$ denote the residual-stream state at layer $\ell$ and token position $t$. For prompt $q_{k,n}$, the fingerprint induces

$$\Delta h^{\mathrm{fp}}_{\ell,t,n} = h_{\ell,t}(x^{\mathrm{fp}}_k, q_{k,n}) - h_{\ell,t}(x^{\mathrm{cle}}, q_{k,n}). \tag{5}$$

We minimize its deviation from $\alpha v_{k,\ell}$, where $\alpha > 0$ sets the target steering strength and $\mathcal{G}$ contains the supervised layer–token positions:

$$\mathcal{L}_{\mathrm{sv}} = \frac{1}{N_k|\mathcal{G}|} \sum_{n=1}^{N_k} \sum_{(\ell,t)\in\mathcal{G}} \left\| \Delta h^{\mathrm{fp}}_{\ell,t,n} - \alpha v_{k,\ell} \right\|_2^2, \tag{6}$$

Subtracting each prompt's clean state isolates the image-induced change. Minimizing $\mathcal{L}_{\mathrm{sv}}$ then drives these prompt-dependent displacements toward the same layerwise SV targets, narrowing the gap between the two arrows.

**Preserving observable behavior.** To make the internal displacement observable in text, we combine SV matching with output distillation, as shown in Stage I of Figure 1. Let $p^{\mathrm{dir}}_{k,n}$ denote the source model's first-response-token distribution under direct SV intervention with the clean carrier, and $p^{\mathrm{fp}}_{k,n}$ its distribution under the fingerprint without intervention. We jointly minimize their Kullback–Leibler (KL) divergence and the displacement error:

$$\mathcal{L}_{\mathrm{fp}} = \underbrace{\frac{1}{N_k} \sum_{n=1}^{N_k} \mathrm{KL}\left(p^{\mathrm{dir}}_{k,n} \,\middle\|\, p^{\mathrm{fp}}_{k,n}\right)}_{\mathcal{L}_{\mathrm{kl}}} + \lambda_{\mathrm{sv}} \mathcal{L}_{\mathrm{sv}}. \tag{7}$$

The coefficient $\lambda_{\mathrm{sv}}$ weights the displacement loss. The KL term distills direct steering's effect on the first generation step, complementing the internal constraint imposed by $\mathcal{L}_{\mathrm{sv}}$. Unlike the fixed answer in the upper row of Figure 1, the reference distribution varies with the prompt. The image thus targets a shared behavior rather than a fixed response; Appendix D gives the supervision settings.

**Parameter-noise augmentation.** Preserving an SV does not ensure that the same image continues to induce it after model modification. We therefore also optimize the image on temporarily parameter-perturbed source copies, keeping the SV targets fixed. This trains the image to reach the same directional target under nearby parameter changes; Appendix D specifies the perturbation schedule.

### 4.3 Black-Box Verification

At verification time, we pair each fixed fingerprint $x_k^{\mathrm{fp}}$ with held-out natural prompts $\{q_{k,n}^{\mathrm{te}}\}_{n=1}^{N_k^{\mathrm{te}}}$, as shown in Stage II of Figure 1. The Chinese and English responses illustrate how the same query can distinguish a derivative from a non-derived model through the target behavior. A text-only detector $g_k(Y_{k,n}) \in \{0,1\}$ records whether each response $Y_{k,n} = M(x_k^{\mathrm{fp}}, q_{k,n}^{\mathrm{te}})$ exhibits $b_k$. Averaging detector hits within and then across the $K_b$ behaviors gives the fingerprint trigger rate (FTR):

$$\widehat{F}(M) = \frac{1}{K_b} \sum_{k=1}^{K_b} \frac{1}{N_k^{\mathrm{te}}} \sum_{n=1}^{N_k^{\mathrm{te}}} g_k(Y_{k,n}). \tag{8}$$

Different response strings count as hits when they satisfy the same behavioral detector. We classify $M$ as derived when $\widehat{F}(M) > \tau$, where $\tau$ is calibrated on separate non-derived checkpoints. Appendix F specifies the detectors and calibration protocol.

## 5 Experiments

Experiments examine the proposed provenance-anchor paradigm from black-box separation to its internal basis. We first test response persistence and source specificity (Section 5.2), then SV preservation (Section 5.3), and finally visual realization, causal contribution, and the roles of the construction objectives (Section 5.4). Together, these analyses examine the inheritance, specificity, input realizability, and behavioral observability required by the framework. Additional deployment settings and hyperparameters appear in the appendix.

### 5.1 Experimental Setting

**Models and coverage.** Our primary evaluation uses Qwen2.5-VL-3B and -7B, InternVL3-2B, and Qwen2-VL-2B as source models (Bai et al., 2025a; Zhu et al., 2025; Wang et al., 2024). We adapt each source to five downstream datasets, V7W, ST-VQA, TextVQA, PaintingForm, and MathV360K (Zhu et al., 2016; Biten et al., 2019; Singh et al., 2019; Bin et al., 2024; Shi et al., 2024), using both low-rank adaptation (LoRA) and full fine-tuning. This yields 40 derivatives across eight source–adaptation groups. We also evaluate weight quantization, Wanda and SparseGPT pruning, and Linear and TIES merging. Appendices D and A provide the checkpoints and transformation settings.

**Queries, baselines, and metrics.** For each of Py→JS, En→Zh, and Ans→Ref, SteerPrint pairs one visual fingerprint with 100 held-out natural prompts, totaling 300 black-box queries. For Ordinary, RNA, PLA, and AGDI, we retain each method's detector for predefined target phrases or semantic markers and its 1,000-query protocol (Madry et al., 2018; Wang et al., 2025; Xia et al., 2026). We report FTR, the fraction of responses satisfying the corresponding detector, averaged across behaviors for SteerPrint (Eq. (8)). To isolate the increase over the unmodified carrier, we also report $\Delta\mathrm{FTR}$, obtained by subtracting the clean-carrier response rate. Unless specified otherwise, generation uses greedy decoding. Appendices D–E detail the queries, detectors, clean controls, and data splits.

### 5.2 Black-Box Ownership Verification

**Fine-tuning transfer.** Source-trained fingerprints retain their behavioral effect across both adapter-based and full-parameter adaptation. SteerPrint achieves the highest mean derivative FTR for each of

the four sources, exceeding AGDI by 4.9–23.2 percentage points when averaged over each source's ten derivatives (Table 1). The advantage holds under both LoRA and full fine-tuning, providing end-to-end evidence for response persistence across distinct adaptation regimes. This establishes the transfer side of the verification problem; non-derived controls below test whether the retained response is source-specific.

Table 1: **Fine-tuning transfer.** FTR (%) after LoRA and full fine-tuning; Avg. averages the ten derivatives of each source. ST, Text, Paint, and Math denote ST-VQA, TextVQA, PaintingForm, and MathV360K. Bold marks the best value in each source–column group, including ties. Additional backbones appear in Appendix A.1.

| Source | Method | LoRA | | | | | Full fine-tuning | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | V7W | ST | Text | Paint | Math | V7W | ST | Text | Paint | Math | |
| Qwen2.5-VL 3B | Ordinary | **78** | 87 | 47 | 85 | 59 | 77 | 87 | 54 | 89 | 62 | 72.5 |
| | RNA | 76 | 89 | 52 | **92** | 73 | 77 | 91 | 61 | 92 | 74 | 77.7 |
| | PLA | 77 | 89 | 50 | 90 | 73 | 79 | 89 | 60 | 93 | 72 | 77.2 |
| | AGDI | **78** | **90** | 53 | **92** | 71 | 80 | **92** | 62 | 95 | 75 | 78.8 |
| | **SteerPrint** | 76 | 82 | **68** | 89 | **83** | **95** | 85 | **63** | **97** | **99** | **83.7** |
| Qwen2.5-VL 7B | Ordinary | 56 | 86 | 40 | 46 | 8 | 61 | 81 | 18 | 59 | 46 | 50.1 |
| | RNA | 59 | 85 | 41 | 48 | 9 | 62 | **83** | 21 | 61 | 47 | 51.6 |
| | PLA | 58 | 84 | 40 | 48 | 11 | 62 | **83** | 21 | 61 | 44 | 51.2 |
| | AGDI | 62 | 64 | 49 | 56 | 44 | 61 | 68 | 47 | 56 | **72** | 57.9 |
| | **SteerPrint** | **94** | **97** | **74** | **100** | **67** | **75** | 78 | **65** | **99** | 62 | **81.1** |
| InternVL3 2B | Ordinary | 37 | 64 | 47 | 17 | 41 | 32 | 58 | 39 | 24 | 47 | 40.6 |
| | RNA | 39 | 66 | 49 | 17 | 40 | 34 | 61 | 42 | 25 | 49 | 42.2 |
| | PLA | 41 | 68 | 51 | 19 | 44 | 34 | 61 | 40 | 26 | 48 | 43.2 |
| | AGDI | 56 | **84** | **76** | 34 | 77 | 53 | **83** | 65 | 54 | **71** | 65.3 |
| | **SteerPrint** | **80** | 77 | 70 | **37** | **86** | **88** | 61 | **74** | **66** | 66 | **70.5** |
| Qwen2-VL 2B | Ordinary | 36 | 46 | 22 | 48 | 41 | 34 | 43 | 15 | 48 | 26 | 35.9 |
| | RNA | 36 | 39 | 22 | 40 | 37 | 32 | 38 | 15 | 40 | 21 | 32.0 |
| | PLA | 48 | 68 | 33 | 76 | 60 | 43 | 60 | 28 | 75 | 38 | 52.9 |
| | AGDI | 53 | 77 | 41 | 81 | **68** | 46 | 65 | 33 | 80 | 45 | 58.9 |
| | **SteerPrint** | **95** | **83** | **53** | **92** | 67 | **86** | **94** | **42** | **93** | **82** | **78.7** |

**Compression and merging.** Fingerprint persistence extends to changes in weight precision, sparsity, and adapter composition (Table 2). Strong responses remain after pruning and merging, showing that retention is not confined to a single fine-tuning procedure. Its extent is nevertheless checkpoint-dependent: the reported 4-bit settings retain high FTR for Qwen2.5-VL and LLaVA, whereas Qwen2-VL shows a larger reduction from 8 to 4 bits. This variation motivates examining which source directions remain preserved and accessible to the fixed visual input.

Table 2: Fingerprint trigger rate (FTR, %) after post-training model modifications. Pruning and merging report mean derivative FTR over $n$ modified configurations; quantization lists 8-bit / 4-bit FTR. Qwen2.5 quantizes the language module. Baseline methods use their respective published detectors and response criteria.

| Source | Modification | $n$ | RNA | PLA | AGDI | SteerPrint |
|---|---|---|---|---|---|---|
| LLaVA-1.5-7B | Pruning | 6 | 23.0 | 35.3 | 53.3 | **91.2** |
| Qwen2-VL-2B | Merging | 8 | 37.4 | 52.6 | 56.0 | **85.4** |
| Qwen2.5-VL-7B | Pruning | 6 | 51.0 | 51.0 | 85.0 | 87.0 |
| Qwen2.5-VL-7B | Merging | 8 | 62.0 | 58.8 | 85.1 | 86.4 |
| Qwen2.5-VL-7B | 8-bit / 4-bit | 2 | 90 / 86 | 91 / 79 | 87 / 86 | 98.3 / 99.0 |
| Qwen2-VL-2B | 8-bit / 4-bit | 2 | 67 / 19 | 59 / 26 | 88 / 59 | 91 / 72 |
| LLaVA-1.5-7B | 8-bit / 4-bit | 2 | 69 / 37 | 79 / 40 | 93 / 70 | 91 / 91 |

**Non-derived controls.** Using Qwen2.5-VL-7B fingerprints, we evaluate source specificity across a panel of 27 non-derived models (300 queries per model, 8,100 total queries; Table 7). Across all models, 19 exhibit zero triggers, and each of the remaining 8 produces at most 1 trigger out of 300 ($\leq 0.333\%$), yielding an aggregate trigger rate of 0.099% (8/8,100). Under the verification threshold $\tau = 20\%$, all 27 non-derived models are correctly rejected, yielding zero model-level false positives (0/27, 0%). This confirms that the observed transfer is strictly specific to derived models rather than generic target-behavior tendencies. Appendix A.2 details the full 27-model panel.

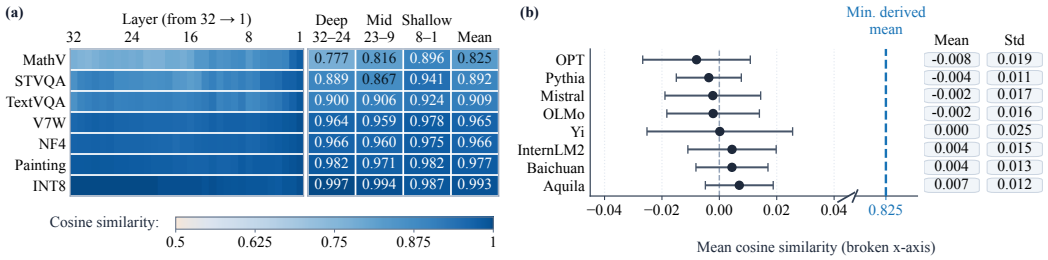
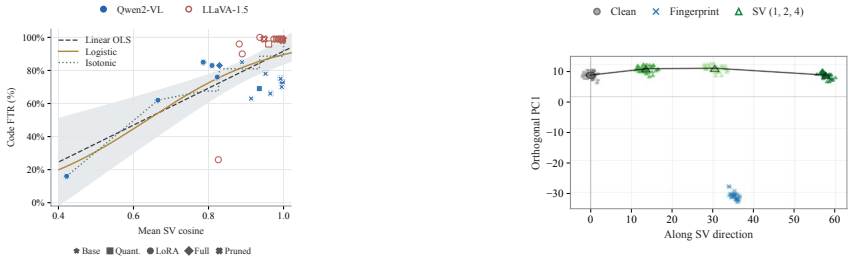



Figure 2: En→Zh SV cosine similarity with LLaVA-1.5-7B: (a) derivatives; (b) cross-family controls. Error bars indicate the across-layer standard deviation; the dashed line marks the lowest layer-averaged similarity among derivatives.




(a) Direction preservation and transfer.

(b) Visual realization of SV steering.

Figure 3: Direction preservation and visual realization. (a) SV cosine similarity versus black-box Py→JS FTR (%); shading indicates the pointwise 95% confidence interval for the ordinary least-squares mean-response fit. (b) Image-induced hidden-state displacement resolved into components along the direct steering vector and an orthogonal subspace. The image induces the target component while retaining an orthogonal displacement.

### 5.3 Steering-Vector Preservation and Fingerprint Transfer

The black-box results establish response retention and separation; the following analyses examine the internal feature proposed to support them.

**Preservation across derivatives.** Behavioral contrast directions remain aligned with the source after fine-tuning and quantization, with preservation varying across layers (Figure 2). Cross-family controls show little alignment in the reported coordinates. The mean-difference analysis in Appendix B.3 gives a geometric interpretation: shifts shared by the target and contrast means cancel, while directional change depends on orthogonal drift relative to the retained contrast. Thus, anchor preservation concerns behavioral contrasts, without requiring unchanged activations.

**Association with fingerprint transfer.** Checkpoints that better preserve the source direction also tend to retain stronger black-box responses: SV cosine similarity and Py→JS FTR are positively associated across 29 Qwen2-VL and LLaVA checkpoints (Spearman $\rho = 0.38$; Figure 3a). This links the measured internal feature to transfer at the checkpoint level. The conditional transfer analysis separately accounts for direction preservation, changes in the image-induced displacement, and behavioral readout (Appendix B.6); the next experiments examine visual realization and the causal contribution of the induced component.

### 5.4 Mechanistic and Component Analysis

**Input-induced steering effects.** The learned image induces a target-directed component within a broader hidden-state change. Figure 3b shows displacement along the direct-steering direction together with a non-negligible orthogonal component, distinguishing approximate visual realization from exact state matching. The additional diagnostics identify directions responsive to direct intervention but difficult to induce through images (Appendices A.3 and A.4). These cases illustrate why input realizability is a separate requirement in Section 3.3: internal controllability alone does not ensure that the visual input can reproduce the effect.

**Direction intervention.** The fingerprint-induced steering component contributes causally to the target response. Target knockout (KO) replaces $h^{\mathrm{fp}}$ with $h^{\mathrm{fp}} - \langle h^{\mathrm{fp}} - h^{\mathrm{clean}}, v\rangle v$, where $v$ is the unit source SV and $h^{\mathrm{clean}}$ is the same prompt's clean-carrier state. Removing this component lowers refusal FTR from 100% to 55% on the source and from 80% to 20% on V7W-LoRA, while an amplitude-matched orthogonal intervention produces no comparable reduction (Table 3). Calibrated restoration recovers FTR to 85–95%. The knockout–control–rescue comparison therefore connects the behavioral readout to the selected internal component beyond the preceding cosine association.

Table 3: Refusal FTR (%) under hidden-state interventions on Qwen2.5-VL-7B and its LoRA derivatives. Target KO removes the fingerprint-induced component along the source steering vector; Orth. applies an amplitude-matched orthogonal intervention; Rescue restores the target coordinate to its calibrated value.

| Model | Fingerprint | Target KO | Orth. | Rescue |
|---|---|---|---|---|
| Source | 100 | 55 | 100 | 85 |
| V7W LoRA | 80 | 20 | 90 | 85 |
| ST-VQA LoRA | 95 | 25 | 95 | 95 |

**Representation and output supervision.** SV matching and output distillation provide complementary supervision for transferring behavior. On Qwen2.5-VL-7B and ten derivatives, the joint objective reaches 81.20% mean derivative FTR, averaging the Chinese and refusal directions equally, compared with 51.60% for KL-only and 58.15% for SV-only (Table 4). Reducing the KL weight to 0.1 yields 45.60%, showing that retaining both terms alone is insufficient to recover the joint objective's gain. These results support Section 4.2: SV matching specifies internal displacement; first-response-token distillation supplies output supervision.

Table 4: Loss ablations on Qwen2.5-VL-7B: mean FTR (%) over Chinese and refusal directions. Avg. excludes the source; SV-dominant uses $0.1\,\mathcal{L}_{\mathrm{kl}} + \mathcal{L}_{\mathrm{sv}}$.

| Configuration | Source | LoRA | | | | | Full fine-tuning | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | V7W | ST | Text | Paint | Math | V7W | ST | Text | Paint | Math | |
| **Full** | **100** | **92.5** | **96** | 70.5 | **100** | **100** | **72.5** | **67.5** | **50** | **100** | **63** | **81.20** |
| w/o $\mathcal{L}_{\mathrm{sv}}$ | 85.5 | 50 | 51 | 51.5 | 60.5 | 53.5 | 50.5 | 49.5 | **50** | 51.5 | 48 | 51.60 |
| w/o $\mathcal{L}_{\mathrm{kl}}$ | 77 | 53 | 56.5 | 53 | 70.5 | 86.5 | 50.5 | 50 | **50** | 60 | 51.5 | 58.15 |
| SV-dominant | 62.5 | 50.5 | 52 | 53 | 4 | 44.5 | 50 | 50.5 | **50** | 51 | 50.5 | 45.60 |
| w/o param. noise | 99 | 60 | 76.5 | **82.5** | 91.5 | 84 | 53.5 | 46.5 | 36.5 | 86.5 | 43.5 | 66.10 |

ST: ST-VQA; Text: TextVQA; Paint: PaintingForm; Math: MathV360K.

**Parameter-noise augmentation.** High source trigger rates do not by themselves ensure transfer. Without parameter noise, the source retains 99% two-direction mean FTR, but the derivative mean drops by 15.10 percentage points to 66.10%. The source–derivative contrast supports training the visual realization across parameter changes even when the SV target itself remains fixed, consistent with the pathway-stability motivation in Appendix B.6. The improvement is in average transfer: on TextVQA-LoRA, the no-noise variant exceeds Full (82.5% versus 70.5%). Appendices A.4.2, A.5.6, and A.4.3 detail the protocols and component analyses.

## 6 Conclusion

We formalized provenance anchors as internal features that retain source agreement through model derivation while distinguishing non-derived models. SteerPrint translates source-model SVs into reusable visual fingerprints by matching their internal displacement and behavioral effect. Across the evaluated LVLMs, these fingerprints transfer to derivatives and separate them from non-derived controls. Representation analyses trace the induced response to the selected directions, and targeted interventions establish their causal contribution. Together, these results support constructing black-box fingerprints around inherited internal features rather than prescribed output strings.

## ETHICS STATEMENT

SteerPrint is designed to protect intellectual property and foster responsible open-source ecosystem development by enabling black-box ownership verification of derived LVLMs. To uphold safety and ethical standards, all evaluation queries for ownership verification are strictly restricted to benign, harmless requests, requiring no generation of harmful, toxic, or unsafe content during verification. Contrastive harmful examples were utilized solely offline for initial steering-vector extraction from the protected source model. Fingerprint scores provide empirical evidence of checkpoint derivation and should be evaluated alongside other contextual indicators in intellectual property disputes.

## AI USE STATEMENT

Generative AI assistance was employed for language polishing, and editorial refinement of expository arguments. The authors conceived, designed, and verified all conceptual frameworks, theoretical analyses, model checkpoints, and experimental evaluations, and take full responsibility for the final manuscript and the accuracy of all reported findings.

## REPRODUCIBILITY STATEMENT

Appendix D reports the datasets, post-release transformations, fingerprint-construction settings, and baseline protocols. Appendix E specifies the steering-vector extraction pairs, held-out black-box queries, response detectors, and train–query separation, while Appendix G defines the verification metrics and Appendix B gives the formal statements, proofs, and finite-query conditions. Anonymous code is available at `https://anonymous.4open.science/r/SteerPrint`.

## REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html`.

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. 2025a. URL `https://arxiv.org/abs/2502.13923`.

Xiaofan Bai, Pingyi Hu, Xiaojing Ma, Linchen Yu, Dongmei Zhang, Qi Zhang, and Bin Benjamin Zhu. ESF: Efficient sensitive fingerprinting for black-box tamper detection of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10477–10494, Vienna, Austria, July 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.546. URL `https://aclanthology.org/2025.findings-acl.546/`.

Yi Bin, Wenhao Shi, Yujuan Ding, Zhiqiang Hu, Zheng Wang, Yang Yang, See-Kiong Ng, and Heng Tao Shen. GalleryGPT: Analyzing paintings with large multimodal models. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM)*, 2024. URL `https://arxiv.org/abs/2408.00491`.

Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gomez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4290–4300, 2019. doi: 10.1109/ICCV.2019.00439.

Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. IPGuard: Protecting intellectual property of deep neural networks via fingerprinting the classification boundary. In *Proceedings of the ACM Asia Conference on Computer and Communications Security (AsiaCCS)*, 2021. URL `https://arxiv.org/abs/1910.12903`.

Jialuo Chen, Jingyi Wang, Tinglan Peng, Youcheng Sun, Peng Cheng, Shouling Ji, Xingjun Ma, Bo Li, and Dawn Song. Copy, right? a testing framework for copyright protection of deep learning models. In *Proceedings of the 43rd IEEE Symposium on Security and Privacy (S&P)*, 2022. URL https://arxiv.org/abs/2112.05588.

Tianshuo Cong, Delong Ran, Zesen Liu, Xinlei He, Jinyuan Liu, Yichen Gong, Qi Li, Anyu Wang, and Xiaoyun Wang. Have you merged my model? on the robustness of large language model IP protection methods against model merging. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis (LAMPS), co-located with ACM CCS*, 2024. URL https://arxiv.org/abs/2404.05188.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2208.07339.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.14314.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. URL https://arxiv.org/abs/2301.00774.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Pingyi Hu, Xiaofan Bai, Xiaojing Ma, Chaoxiang He, Dongmei Zhang, and Bin Benjamin Zhu. RESF: Regularized-entropy-sensitive fingerprinting for black-box tamper detection of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4889–4903, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.247. URL https://aclanthology.org/2025.emnlp-main.247/.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2212.04089.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. URL https://arxiv.org/abs/1610.08401.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. URL https://arxiv.org/abs/2306.13213.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Shuo Shao, Yiming Li, Yu He, Hongwei Yao, Wenyuan Yang, Dacheng Tao, and Zhan Qin. SoK: Large language model copyright auditing via fingerprinting. *arXiv preprint arXiv:2508.19843*, 2025. URL https://arxiv.org/abs/2508.19843.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-LLaVA: Bootstrapping mathematical reasoning for multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. URL https://arxiv.org/abs/2406.17294.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8317–8326, 2019. doi: 10.1109/CVPR.2019.00851.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2306.11695.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2024. URL https://arxiv.org/abs/2308.10248.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. URL https://arxiv.org/abs/2409.12191.

Yubo Wang, Jianting Tang, Chaohu Liu, and Linli Xu. Tracking the copyright of large vision-language models through parameter learning adversarial images. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d368eba3fb546ae5d868dae2ecb17200-Abstract-Conference.html.

Zehao Wu, Yanjie Zhao, and Haoyu Wang. Gradient-based model fingerprinting for LLM similarity detection and family classification. *arXiv preprint arXiv:2506.01631*, 2025. URL https://arxiv.org/abs/2506.01631.

Zhaomin Wu, Haodong Zhao, Ziyang Wang, Jizhou Guo, Qian Wang, and Bingsheng He. LLM DNA: Tracing model evolution via functional representations. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.24496.

Chengwei Xia, Fan Ma, Ruijie Quan, Yunqiu Xu, Kun Zhan, and Yi Yang. Echoes of ownership: Adversarial-guided dual injection for copyright protection in MLLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://arxiv.org/abs/2602.18845.

Zhenyu Xu and Victor S. Sheng. A behavioral fingerprint for large language models: Provenance tracking via refusal vectors. *arXiv preprint arXiv:2602.09434*, 2026. doi: 10.48550/arXiv.2602.09434. URL https://arxiv.org/abs/2602.09434.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2306.01708.

Shojiro Yamabe, Futa Waseda, Tsubasa Takahashi, and Koki Wataoka. MergePrint: Merge-resistant fingerprints for robust black-box ownership verification of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. URL https://arxiv.org/abs/2410.08604.

Guang Yang, Ziye Geng, Yihang Chen, and Changqing Luo. Fingerprinting deep neural networks for ownership protection: An analytical approach, 2026.

Zhiguang Yang and Hanzhou Wu. A fingerprint for large language models. *arXiv preprint arXiv:2407.01235*, 2024. doi: 10.48550/arXiv.2407.01235. URL https://arxiv.org/abs/2407.01235.

Haobo Zhang, Zhenhua Xu, Junxian Li, Shangfeng Sheng, Dezhang Kong, and Meng Han. AttnDiff: Attention-based differential fingerprinting for large language models. *arXiv preprint arXiv:2604.05502*, 2026. URL https://arxiv.org/abs/2604.05502.

Jie Zhang, Dongrui Liu, Chen Qian, Linfeng Zhang, Yong Liu, Yu Qiao, and Jing Shao. REEF: Representation encoding fingerprints for large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=SnDmPkOJ0T.

Yifei Zhao, Qian Lou, and Mengxin Zheng. Sif: Semantically in-distribution fingerprints for large vision-language models. *arXiv preprint arXiv:2604.17041*, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. 2025. URL https://arxiv.org/abs/2504.10479.

Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. Visual7W: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4995–5004, 2016. doi: 10.1109/CVPR.2016.540.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023a. URL https://arxiv.org/abs/2310.01405.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023b.

# A Supporting Experiments and Detailed Results

This appendix provides full task-level tables, experimental protocols, diagnostic figures, sensitivity studies, and ablation analyses supporting the empirical evaluation in Section 5.

## A.1 Additional Fine-Tuning Transfer Results

We evaluate the fingerprints across fine-tuned derivatives of multiple backbones. Table 1 in the main text reports complete task-level evaluations across four primary source models under both LoRA and full fine-tuning (Qwen2.5-VL-3B, Qwen2.5-VL-7B, InternVL3-2B, and Qwen2-VL-2B, spanning 40 fine-tuned derivatives). Below, we report supplementary evaluations on two additional architectures under single-adaptation (LoRA) fine-tuning: InternVL3.5-2B (Table 5) and LLaVA-1.5-7B (Table 6).

Shaded rows identify SteerPrint; bold values mark the highest FTR in each column, including ties. ST, Text, Paint, and Math denote ST-VQA, TextVQA, PaintingForm, and MathV360K.

Table 5: FTR (%) on InternVL3.5-2B LoRA derivatives. Avg. is the per-source mean derivative FTR across evaluated tasks.

| **Method** | **LoRA fine-tuning** | | | | | **Avg.** |
|---|---|---|---|---|---|---|
| | V7W | ST | Text | Paint | Math | |
| Ordinary | 25 | 38 | 32 | 12 | 23 | 26 |
| RNA | 23 | 40 | 30 | 10 | 20 | 25 |
| PLA | 39 | 53 | 45 | 23 | 34 | 39 |
| AGDI | 42 | 55 | 47 | 30 | 41 | 43 |
| **SteerPrint** | **95** | **86** | **80** | **91** | **94** | **89** |

The supplementary single-adaptation evaluations on InternVL3.5-2B (Table 5) and LLaVA-1.5-7B (Table 6) achieve 89% and 70% average FTR, respectively, compared with AGDI's 43% and 53%. InternVL3.5-2B remains at or above 80% across all five downstream LoRA tasks. The corresponding results for the four primary source models are reported in Table 1 in the main text.

Table 6: FTR (%) on LLaVA-1.5-7B LoRA derivatives. Avg. is the per-source mean derivative FTR across evaluated tasks.

| Method | LoRA fine-tuning | | | | | Avg. |
|---|---|---|---|---|---|---|
| | V7W | ST | Text | Paint | Math | |
| Ordinary | 31 | 28 | 11 | 21 | 11 | 20 |
| RNA | 28 | 31 | 14 | 20 | 9 | 20 |
| PLA | 51 | 43 | 21 | 55 | 18 | 38 |
| AGDI | 64 | 56 | 36 | 79 | 30 | 53 |
| **SteerPrint** | **75** | **68** | **55** | **99** | **54** | **70** |

### A.2 Checkpoint Specificity on Non-Derived Controls

To verify that SteerPrint's visual fingerprints are strictly specific to the protected model lineage rather than triggering on unrelated models exhibiting generic target capabilities, we evaluate the Qwen2.5-VL-7B fingerprints across a comprehensive panel of 27 non-derived models spanning diverse architectures and scales (Table 7).

**Evaluation protocol.** Each non-derived model is evaluated on 100 matched clean/fingerprint query pairs per behavioral direction (300 pairs per model, totaling 8,100 queries). For each query, let $T_i \in \{0, 1\}$ and $C_i \in \{0, 1\}$ denote the binary detector hit indicators under the visual fingerprint and the clean carrier, respectively. A paired trigger-only event occurs when the fingerprint elicits the target response that was absent under the clean carrier:

$$\widehat{\mathrm{FPR}}_{\mathrm{pair}} = \frac{1}{N} \sum_{i=1}^{N} T_i(1 - C_i). \tag{9}$$

Empty or unparseable outputs remain in the denominator and are counted as zero detector hits ($T_i = 0$).

**Empirical separation and specificity.** Across the 27 evaluated non-derived models, 19 models exhibit zero triggers across all 300 queries (0.000%). Each of the remaining eight models triggers at most a single query out of 300 ($1/300 \leq 0.333\%$). Across the entire panel of 8,100 query pairs, the aggregate trigger-only rate is only 0.099% (8 total events: 0 for Py→JS, 1 for En→Zh, and 7 for Ans→Ref). Under the standard ownership verification threshold $\tau = 20\%$, every non-derived model scores far below $\tau$, achieving strictly $0/27$ (0%) model-level false-positive rate.

In contrast, all ten fine-tuned derivatives of Qwen2.5-VL-7B achieve a mean FTR of 81.1% with a minimum of 62.0% (Table 1). This provides a minimum empirical separation margin of 61.7 percentage points ($62.0\% - 0.333\%$) between derived and non-derived checkpoints, confirming that SteerPrint fingerprints provide highly reliable, lineage-specific ownership verification.

Table 7: Non-derived checkpoint evaluation across 27 models (100 query pairs per direction, 300 queries per model). Entries are paired trigger-only rates $\widehat{\mathrm{FPR}}_{\mathrm{pair}}$ (%): fingerprint hits absent under the clean carrier. All nonzero entries are single hits—no model exceeds 0.333%. Under the ownership verification threshold $\tau = 20\%$, all 27 models are correctly rejected (0/27 false positives).

| Model | Py→JS | En→Zh | Ans→Ref | Overall |
|---|---|---|---|---|
| BLIP2-Flan-T5-XL | 0.000 | 0.000 | 1.000 | 0.333 |
| DeepSeek-VL-1.3B-Chat | 0.000 | 0.000 | 1.000 | 0.333 |
| DeepSeek-VL-7B-Chat | 0.000 | 0.000 | 0.000 | 0.000 |
| Idefics2 | 0.000 | 0.000 | 0.000 | 0.000 |
| Idefics2-8B-Base | 0.000 | 0.000 | 0.000 | 0.000 |
| Idefics2-8B-Chatty | 0.000 | 0.000 | 1.000 | 0.333 |
| InstructBLIP-Flan-T5-XL | 0.000 | 0.000 | 0.000 | 0.000 |
| InternVL2-1B | 0.000 | 0.000 | 0.000 | 0.000 |
| InternVL2-2B | 0.000 | 0.000 | 0.000 | 0.000 |
| InternVL2-4B | 0.000 | 0.000 | 1.000 | 0.333 |
| InternVL2-8B | 0.000 | 0.000 | 0.000 | 0.000 |
| InternVL3-2B | 0.000 | 0.000 | 0.000 | 0.000 |
| InternVL3-8B | 0.000 | 0.000 | 0.000 | 0.000 |
| Kosmos-2 | 0.000 | 0.000 | 0.000 | 0.000 |
| LLaVA-1.5-7B | 0.000 | 0.000 | 0.000 | 0.000 |
| MiniCPM-Llama3-V-2.5 | 0.000 | 0.000 | 1.000 | 0.333 |
| Molmo-7B-D | 0.000 | 0.000 | 0.000 | 0.000 |
| Molmo-7B-O | 0.000 | 1.000 | 0.000 | 0.333 |
| MolmoE-1B | 0.000 | 0.000 | 0.000 | 0.000 |
| Ovis2-8B | 0.000 | 0.000 | 1.000 | 0.333 |
| PaliGemma-3B-224 | 0.000 | 0.000 | 0.000 | 0.000 |
| PaliGemma2-10B-448 | 0.000 | 0.000 | 0.000 | 0.000 |
| PaliGemma2-3B-224 | 0.000 | 0.000 | 0.000 | 0.000 |
| PaliGemma2-3B-448 | 0.000 | 0.000 | 0.000 | 0.000 |
| Phi-3-Vision | 0.000 | 0.000 | 0.000 | 0.000 |
| Qwen2.5-VL-3B | 0.000 | 0.000 | 1.000 | 0.333 |
| Qwen2-VL-7B | 0.000 | 0.000 | 0.000 | 0.000 |
| **All models** | 0.000 | 0.037 | 0.259 | 0.099 |

### A.3 Representation Preservation and Intervention Analysis

The transfer results establish external responsiveness; the representation studies examine its association with source-side behavioral directions. These diagnostics are evaluated on the Qwen2-VL and LLaVA checkpoints. The layer-wise comparison in Figure 2 provides the motivating source–derivative alignment observation; cosine similarity alone is not an ownership decision or a complete account of training-induced changes.

Figure 3a presents the relationship between internal SV similarity and the behavioral response rate for Py→JS. Across the 29 evaluated Qwen2-VL and LLaVA checkpoints, the observations yield Spearman $\rho = 0.381$ ($p = 0.0413$) and Pearson $r = 0.659$. Standard regression models yield $R^2 = 0.434$, $0.450$, and $0.534$ for linear, logistic, and isotonic fits, respectively. The positive correlation demonstrates that checkpoints maintaining stronger internal alignment with the source steering vector also exhibit higher FTR under the visual fingerprint.

Figure 4 presents the hidden-state activation trajectories under visual fingerprints and direct steering at layer 22 for Py→JS (highlighted in Figure 3b), with corresponding trajectories for En→Zh and Ans→Ref shown in Figures 5 and 6. Table 8 summarizes the trajectory metrics across the best-performing layers for all three directions. Across behaviors, the visual fingerprint induces substantial projection along the teacher axis (0.61–0.69) alongside orthogonal components, confirming that the optimized visual input effectively steers hidden states along the designated behavioral subspace.

Table 8: Best-layer displacement diagnostics across behavioral directions. Line denotes the teacher-axis fraction; Orth. denotes the orthogonal residual ratio.

| **Direction** | **Layer** | **Cosine** | **Line** | **Orth.** |
|---|---|---|---|---|
| Py→JS | L22 | 0.678 | 0.611 | 0.735 |
| En→Zh | L23/L24 | 0.659 | 0.690 | 0.752 |
| Ans→Ref | L21/L23 | 0.676 | 0.690 | 0.736 |

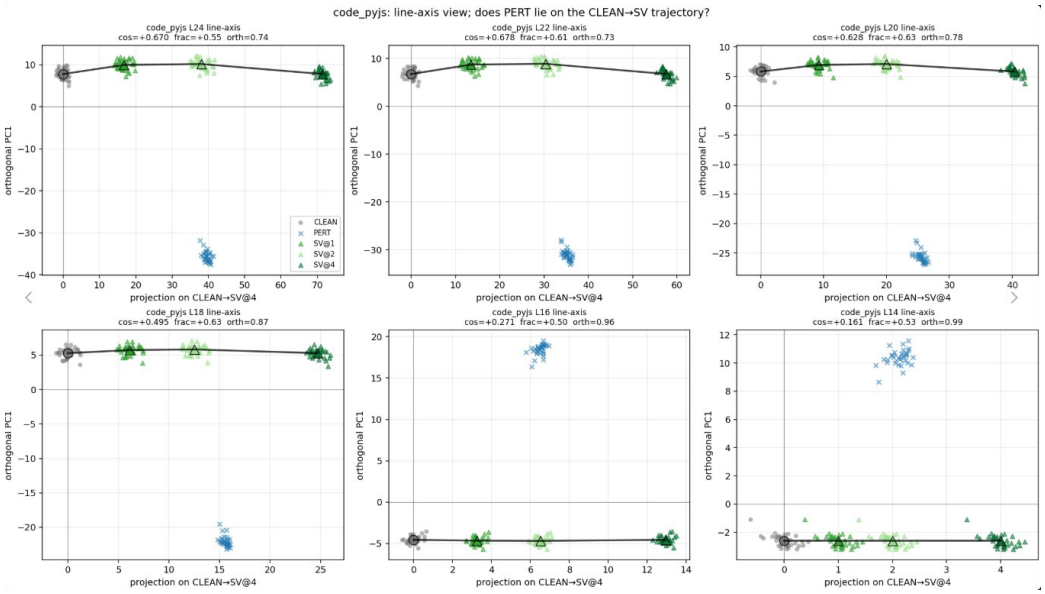



Figure 4: Hidden-state trajectories under visual fingerprints and direct steering for Py→JS.

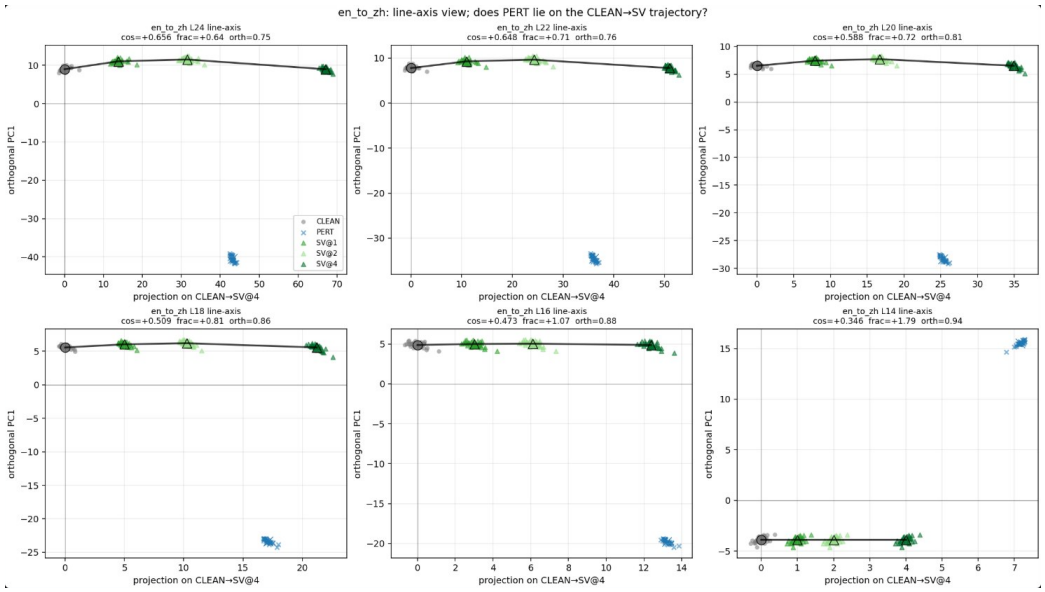



Figure 5: Hidden-state trajectories under visual fingerprints and direct steering for En→Zh.

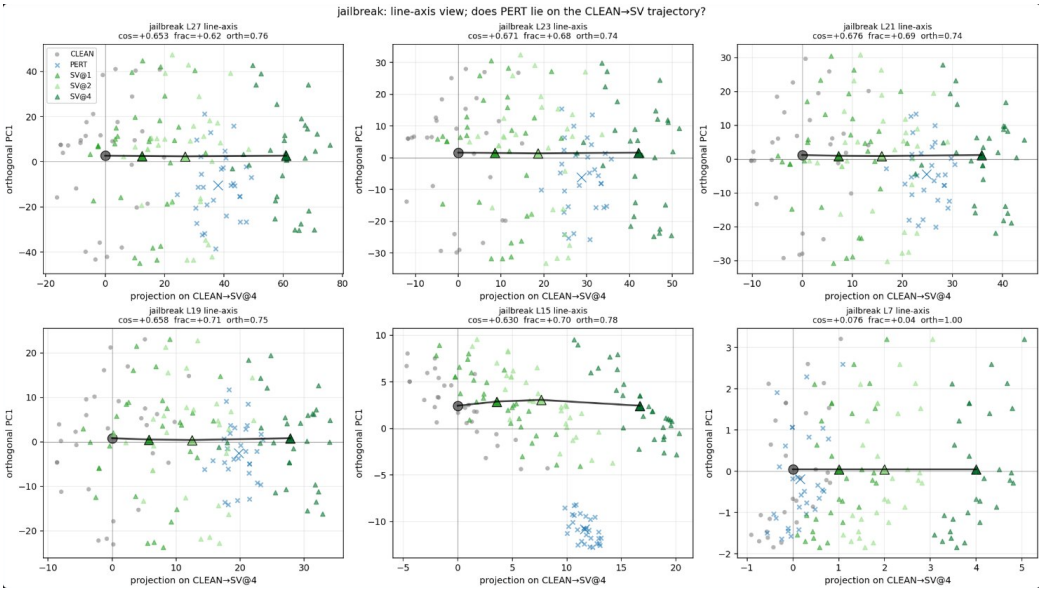



Figure 6: Hidden-state trajectories under visual fingerprints and direct steering for Ans→Ref.

### A.3.1 Label-Shuffled Controls

To evaluate whether the observed pairwise SV alignment arises specifically from learned behavioral directions rather than generic representation geometry, we compute label-shuffled controls (Figures 7 and 8). By randomly permuting positive and negative contrastive prompt labels prior to SV extraction, label correspondence is disrupted while preserving the overall activation distribution. Under label shuffling, pairwise cosine similarities collapse toward zero across checkpoints, confirming that the high SV similarity observed between source and derivative checkpoints reflects shared behavioral representations rather than background representation geometry.

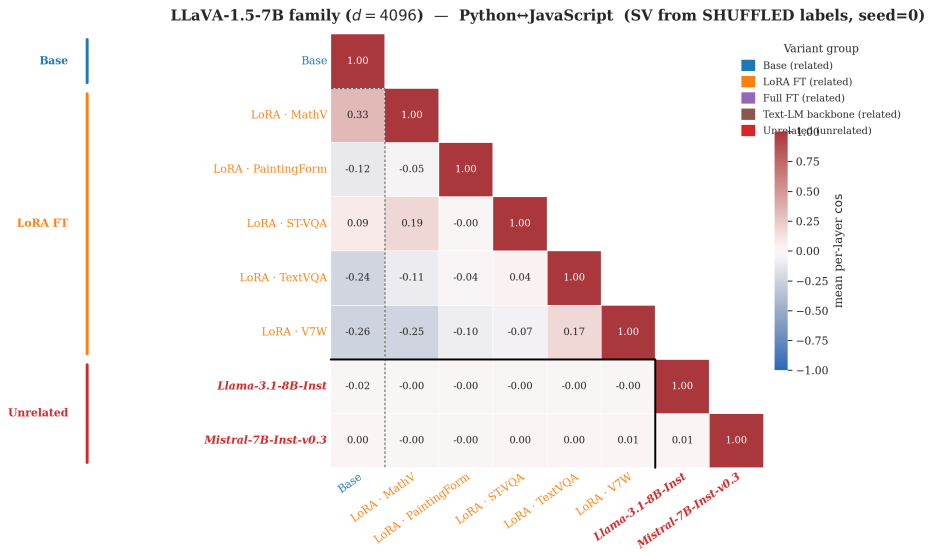



Figure 7: Pairwise SV cosine similarity on LLaVA using shuffled contrastive labels (seed 0).

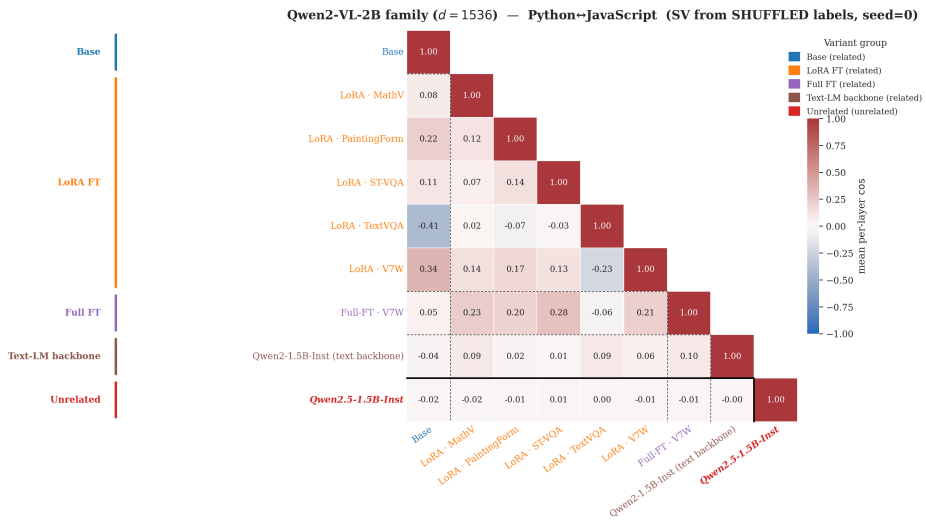



Figure 8: Pairwise SV cosine similarity on Qwen2-VL using shuffled contrastive labels (seed 0).

### A.3.2 Direction Knockout and Calibrated Rescue

Table 3 details the causal intervention study on the refusal direction across all 28 hidden-state sites (post-block outputs 1–27 and the final normalization output at site 28). Intervention is applied to the final prompt-token state predicting the first output token and to every subsequent decoding step. All arms use FP32 precision. Each checkpoint is evaluated on 20 verification queries with ten disjoint calibration prompts.

At each site and prediction step, an untreated clean-image reference pass tracks the intervention branch's generated prefix. With its state $h^{\text{clean}}$ and the current branch state $h$, the signed coefficient along the source SV is $c = \langle h - h^{\text{clean}}, v \rangle$. Target KO subtracts $cv$, directly removing the fingerprint-induced projection onto the source SV. The orthogonal control subtracts $cu$ where $u \perp v$ is a fixed orthogonal vector per site, matching the requested intervention amplitude in an unrelated direction. Calibrated rescue restores the target coordinate toward the calibration median after KO. Under target KO, refusal FTR drops substantially from 100/80/95% to 55/20/25% for Source, V7W-LoRA, and ST-VQA-LoRA, respectively, whereas the orthogonal control preserves the high FTR (100/90/95%) and calibrated rescue largely recovers them (85/85/95%). Baseline clean-carrier responses remain at 5/0/0%. These results confirm that the source refusal direction has a causal contribution to the fingerprint-induced behavioral response on the evaluated checkpoints.

## A.4 Direction Selection and Objective Ablations

This section details empirical direction selection diagnostics, the objective and parameter-noise ablation study on Qwen2.5-VL-7B, the mechanistic role of parameter-space augmentation, and refusal loss-weight explorations supporting Section 5.4.

### A.4.1 Direction Selection and Development Screening

**Direction diagnostics.** The provenance requirements guide the choice of an internal target, while direct-steering and visual-response checks assess its black-box usability. On Qwen2.5-VL-7B, we inspect direct source controllability, black-box observability after visual externalization, minimum retention over development derivatives, and maximum response on development non-owners, which together serve as direction-selection criteria.

While False-Assertion exhibits insufficient source controllability during initial screening, Emotion achieves an 83-point direct-intervention shift yet zero black-box observability under visual perturbations. This contrast demonstrates that internal steerability and visual externalizability represent distinct functional criteria. The three selected directions (Py→JS, En→Zh, and Ans→Ref) successfully fulfill both requirements, simultaneously maintaining robust derivative retention alongside negligible false-positive rates on non-derived models.

Table 9: Direction controllability and visual externalization on Qwen2.5-VL-7B development queries.

| Direction | Control delta | Observe | Retain min. | Non-owner max. | Decision |
|---|---|---|---|---|---|
| Py→JS | +100 | 100 | 80 | 0 | Retain |
| En→Zh | +97 | 74 | 53 | 0 | Retain |
| Ans→Ref | +83 | 97 | 71 | 1 | Retain |
| False-Assertion | not reported | — | — | — | Reject: control |
| Emotion | +83 | 0 | — | — | Reject: observe |

**Additional direction exploration.** We further evaluate the Emotion direction on Qwen2.5-VL-7B at $\alpha = 2, 4, 8$, with refusal at $\alpha = 8$ as a positive control. Emotion SVs are extracted from 64 happy/plain contrastive pairs; extraction, training (8 queries), and evaluation (20 queries) use disjoint question sets. All four visual optimizations use one seed, the same $448 \times 448$ carrier, 27 supervised layers, and an $\ell_\infty$ budget of $16/255$.

Table 10: Direction exploration on Qwen2.5-VL-7B: detector hits out of 20 queries. Emotion uses a lexical heuristic; refusal uses the refusal rule.

| Direction | $\alpha$ | Direct | Visual | Clean image |
|---|---|---|---|---|
| Emotion | 2 | 4/20 | 0/20 | 0/20 |
| Emotion | 4 | 18/20 | 4/20 | 0/20 |
| Emotion | 8 | 11/20 | 17/20 | 0/20 |
| Ans→Ref | 8 | 19/20 | 20/20 | 0/20 |

At $\alpha = 4$, direct Emotion injection yields 18/20 heuristic hits but the visual fingerprint yields 4/20. Raising $\alpha$ to 8 gives 17/20 visual hits, accompanied by repeated off-topic themes: 18/20 outputs contain garden-related wording. Thus high lexical activation can coexist with poor task-content preservation. The refusal control yields 20/20 visual hits and 0/20 clean-image hits, confirming that the optimization pipeline can realize a behavioral marker in this setting. Direct injection uses the same 27-layer window with text-only inputs, whereas visual evaluation includes an image; the columns therefore compare different input conditions. Emotion counts reflect lexical matching rather than semantic task success; among the explored settings, no configuration jointly maintains high activation and task fidelity. The Emotion fingerprints are evaluated on the source model only.

### A.4.2 Objective and Parameter-Noise Ablation

Table 4 compares five training configurations on Qwen2.5-VL-7B and ten derivatives (five LoRA and five full fine-tuning). Each table cell is the arithmetic mean of Chinese-direction and refusal-direction FTR, each evaluated on 100 fixed queries with its original detector. Chinese evaluation uses language-classification prompts and greedy decoding capped at eight tokens; refusal evaluation uses benign prompts and a maximum of 80 tokens. Avg. averages the ten derivative columns and excludes the source. Each direction retains its own construction recipe: Chinese uses 1,000 optimization steps and rank-8 parameter augmentation at scale 0.1; refusal uses 2,000 steps and rank-32 augmentation at scale 0.4. Other settings are held fixed within each direction's ablation.

Consistent with the joint objective $\mathcal{L}_{\text{fp}} = \mathcal{L}_{\text{kl}} + \lambda_{\text{sv}}\mathcal{L}_{\text{sv}}$ in Eq. (7): Full uses both losses ($\lambda_{\text{sv}} = 1$) and parameter augmentation; the w/o $\mathcal{L}_{\text{sv}}$ (KL-only) arm omits SV matching; w/o $\mathcal{L}_{\text{kl}}$ (SV-only) omits output distillation; and w/o param. noise removes parameter augmentation. The SV-dominant variant uses $0.1\,\mathcal{L}_{\text{kl}} + \mathcal{L}_{\text{sv}}$ with parameter augmentation retained.

### A.4.3 Parameter-Space Augmentation and Robust Externalization

SteerPrint and output-targeted visual fingerprinting can both train through parameter changes during construction. In SteerPrint, the steering directions are extracted once and held fixed: augmentation does not alter the target or establish its source specificity. Instead, it trains the visual input to reproduce the target's internal displacement and output tendency across perturbed source models. This supplies a provenance-grounded target for a robustness objective that still requires empirical validation on actual derivatives.

Parameter augmentation improves the two-direction mean transfer on Qwen2.5-VL-7B while preserving the same source-derived target. The effect varies across derivatives: on Qwen2.5-VL-7B TextVQA LoRA, the no-noise variant reaches 82.5% versus 70.5% for the full method. Both representation and output losses are optimized through the perturbed models. The trajectory analysis

describes the full-method input realization; per-variant displacement diagnostics are summarized in Table 8.

**Interpretation scope.** This ablation varies the loss and augmentation components under identical evaluation; parameter-space curvature and per-query mediation lie outside its measurements, and component interventions additionally differ in objective, update magnitude, and downstream utility.

### A.4.4 Refusal-Loss-Weight Exploration

Two refusal fingerprints with KL-loss multipliers 0.25 and 0.75 were evaluated on the same V7W control and eight merge checkpoints, yielding 1800 refusal observations. Across the eight merges, the mean refusal score changes from 64.50% to 82.50% (KL-0.25) or 74.75% (KL-0.75); corresponding three-direction means are 86.42%, 92.42%, and 89.83%. The KL-0.25 candidate improves all nine refusal rates and numerically exceeds AGDI's TMR on all eight merges, with a minimum gap of 2.67 percentage points. Results for both candidates and the standard recipe are reported in Table 11.

Table 11: Refusal-loss-weight comparison under varying loss trade-offs (%). R denotes refusal FTR; M denotes three-direction mean FTR. ST, Text, Paint, and Math denote ST-VQA, TextVQA, PaintingForm, and MathV360K.

| Model | Original R | Original M | KL-0.25 R | KL-0.25 M | KL-0.75 R | KL-0.75 M | AGDI |
|---|---|---|---|---|---|---|---|
| V7W control | 77 | 92.33 | 86 | 95.33 | 73 | 91 | 85 |
| Linear ST | 65 | 86.33 | 86 | 93.33 | 76 | 90 | 83 |
| Linear Text | 29 | 74.67 | 70 | 88.33 | 69 | 88 | 85 |
| Linear Paint | 67 | 88.67 | 75 | 91.33 | 68 | 89 | 82 |
| Linear Math | 68 | 81 | 88 | 87.67 | 86 | 87 | 85 |
| TIES ST | 70 | 89.67 | 90 | 96.33 | 76 | 91.67 | 85 |
| TIES Text | 71 | 90 | 86 | 95 | 75 | 91.33 | 88 |
| TIES Paint | 67 | 89 | 76 | 92 | 66 | 88.67 | 86 |
| TIES Math | 79 | 92 | 89 | 95.33 | 82 | 93 | 87 |

## A.5 Retention under Model Transformations

This section evaluates fingerprint retention under post-release transformations including quantization, pruning, merging, and component-level modifications across Qwen2-VL-2B, LLaVA-1.5-7B, and Qwen2.5-VL-7B, complementing the boundary analysis in Section 5.2.

### A.5.1 Weight Quantization

We evaluate bitsandbytes 8-bit and 4-bit weight quantization (Dettmers et al., 2022; 2023) on Qwen2-VL-2B and LLaVA-1.5-7B. As shown in Table 12, SteerPrint achieves 91% and 72% FTR on Qwen2-VL-2B at 8-bit and 4-bit quantization, and maintains 91% on LLaVA-1.5-7B across both precisions, consistently outperforming baseline methods under aggressive post-training quantization.

Table 12: FTR (%) under 8-bit and 4-bit weight quantization on Qwen2-VL-2B and LLaVA-1.5-7B.

| **Method** | **Qwen2-VL 8-bit** | **Qwen2-VL 4-bit** | **LLaVA-1.5 8-bit** | **LLaVA-1.5 4-bit** |
|---|---|---|---|---|
| Ordinary | 58 | 19 | 64 | 31 |
| RNA | 67 | 19 | 69 | 37 |
| PLA | 59 | 26 | 79 | 40 |
| AGDI | 88 | 59 | 93 | 70 |
| **SteerPrint** | 91 | 72 | 91 | 91 |

### A.5.2 Language-Module Quantization

We quantize the 196 attention and MLP linear modules across all 28 language layers of Qwen2.5-VL-7B to 8-bit integer (INT8) and 4-bit NormalFloat (NF4) formats, while preserving vision, projector, embedding, and output head parameters in BF16. Table 13 compares SteerPrint against baseline methods across all three behavioral directions. SteerPrint achieves 99.00% on BF16, 98.33% on INT8, and 99.00% on NF4, yielding 99.33% and 100.00% relative retention under INT8 and NF4

Table 13: Language-module quantization on Qwen2.5-VL-7B (%). SteerPrint columns report FTR; baseline columns report target-match rates.

| Precision | Py→JS | En→Zh | Ans→Ref | Avg. | RNA | PLA | AGDI |
|---|---|---|---|---|---|---|---|
| BF16 control | 100 | 100 | 97 | 99.00 | 95 | 92 | 88 |
| INT8 | 100 | 98 | 97 | 98.33 | 90 | 91 | 87 |
| NF4 | 100 | 98 | 99 | 99.00 | 86 | 79 | 86 |

quantization, respectively. These results confirm the resilience of SteerPrint fingerprints against post-training language-module quantization.

### A.5.3 Weight Pruning

The LLaVA experiment applies Magnitude and Wanda pruning (Sun et al., 2024) to the V7W-LoRA derivative at 10%, 20%, and 30% sparsity. The Qwen2-VL experiment instead uses the source checkpoint. Across the six LLaVA pruning settings, the reported SteerPrint FTR is 90–93%,

Table 14: FTR (%) after pruning the LLaVA-1.5-7B V7W-LoRA derivative under varying pruning methods and sparsity levels.

| **Method** | **Magnitude** | | | **Wanda** | | |
|---|---|---|---|---|---|---|
| | **10%** | **20%** | **30%** | **10%** | **20%** | **30%** |
| RNA | 29 | 25 | 22 | 23 | 20 | 19 |
| PLA | 42 | 38 | 32 | 36 | 33 | 31 |
| AGDI | 63 | 58 | 52 | 57 | 50 | 40 |
| **SteerPrint** | 91 | 92 | 91 | 93 | 90 | 90 |

Table 15: SteerPrint FTR (%) after pruning Qwen2-VL-2B.

| **Pruner** | **10%** | **20%** | **30%** |
|---|---|---|---|
| SparseGPT | 92 | 79 | 77 |
| Wanda | 91 | 75 | 89 |

versus 40–63% for AGDI. Qwen2-VL ranges from 75% to 92%, showing model- and configuration-dependent retention. These results cover the evaluated unstructured-pruning settings on the two checkpoints.

### A.5.4 Model Merging

We pair the Qwen2-VL-2B V7W-LoRA checkpoint with each of four other task adapters at a 50% merge ratio, using Linear and TIES merging (Ilharco et al., 2023; Yadav et al., 2023). SteerPrint

Table 16: FTR (%) after task adapter merging on Qwen2-VL-2B with V7W-LoRA. ST, Text, Paint, and Math denote ST-VQA, TextVQA, PaintingForm, and MathV360K.

| **Merge** | **Method** | **ST** | **Text** | **Paint** | **Math** |
|---|---|---|---|---|---|
| Linear | RNA | 40 | 37 | 43 | 41 |
| Linear | PLA | 59 | 52 | 63 | 59 |
| Linear | AGDI | 63 | 57 | 68 | 63 |
| Linear | SteerPrint | 93 | 92 | 93 | 88 |
| TIES | RNA | 35 | 31 | 38 | 34 |
| TIES | PLA | 46 | 40 | 54 | 48 |
| TIES | AGDI | 49 | 42 | 55 | 51 |
| TIES | SteerPrint | 94 | 71 | 89 | 63 |

reaches 88–93% under Linear merging and 63–94% under TIES. TIES–MathV is a lower-retention configuration, yet its 63% still exceeds AGDI's 51%. Degradation in retention is therefore distinct from loss of baseline superiority.

### A.5.5 Extended Pruning and Merging Studies on Qwen2.5-VL-7B

On Qwen2.5-VL-7B-Instruct, we evaluate the V7W adapter at step 1014 (rank 16, alpha 32), where q/k/v/o and three MLP projections span 196 target modules across 28 layers, under fine-tuning, pruning, and model merging.

**Scoring and query scope.** Each behavioral direction uses one fixed fingerprint and 100 queries, with greedy limits of 400/8/80 tokens for Code/Chinese/refusal. The Chinese direction evaluates language-marker classification under a standard prompt template. Public baselines follow standard evaluation protocols with 100 image–prompt pairs per method, evaluating target-phrase matching under greedy decoding. FTR and baseline TMR measure the respective trigger and target-match frequencies under these standardized evaluation protocols.

**Checkpoint modifications.** Wanda/SparseGPT 10/20/30% runs start independently from the seven-module V7W model after merging the adapter. Calibration uses fixed WikiText2 data, seed 42, and 128 windows of length 2048; SparseGPT uses block size 128 and damping 0.01. Vision, projector, and embedding components are left unpruned. Across the seven evaluated checkpoints, the study comprises 7000 observations spanning behavioral trigger, clean-carrier, and public-baseline conditions.

The merging study uses the same V7W adapter family with four task adapters that cover only q/v. PEFT receives weights $[1, 1]$ for Linear and TIES, with TIES density 0.5 and `majority_sign_method=total`. Density is a sparsification parameter, not a 50% model-mixing ratio. Inference is performed after adapter merge and cold reload. Nine models produce 5400 observations: 300 behavioral and 300 public-baseline queries per model. The fixed merge refusal fingerprint is the scale-0.2 final-2000 image (SV scale $0.2$, final checkpoint at $2000$ optimization steps); it differs from the pruning study's refusal image, and the two unmodified reference scores are accordingly reported separately.

**Pruning results.** Refusal responses are evaluated using a unified detector that identifies standard refusal statements and non-compliance markers; under this evaluation rule, SteerPrint achieves a mean FTR of 87.0% across the six pruned models (Table 17), compared to 51–85% for the output-targeted baselines.

Table 17: Pruning results on the Qwen2.5-VL-7B V7W-LoRA derivative (%). SteerPrint reports three-direction FTR; baselines report target-match rates.

| Model | Behavioral FTR | | | | Baseline TMR | | | |
|---|---|---|---|---|---|---|---|---|
| | Py→JS | En→Zh | Ans→Ref | Avg. | Ord. | RNA | PLA | AGDI |
| V7W control | 100 | 100 | 62 | 87.33 | 60 | 51 | 52 | 84 |
| Wanda 10% | 100 | 100 | 61 | 87 | 62 | 52 | 49 | 85 |
| Wanda 20% | 100 | 100 | 65 | 88.33 | 61 | 53 | 51 | 84 |
| Wanda 30% | 96 | 92 | 66 | 84.67 | 55 | 48 | 53 | 85 |
| SparseGPT 10% | 100 | 100 | 63 | 87.67 | 62 | 51 | 49 | 86 |
| SparseGPT 20% | 100 | 100 | 63 | 87.67 | 57 | 51 | 50 | 85 |
| SparseGPT 30% | 95 | 99 | 66 | 86.67 | 57 | 51 | 54 | 85 |

**Merging results.** The adapter merging results across all eight combinations are reported in Table 18. SteerPrint demonstrates resilient transfer across merged models, maintaining high FTR across both Linear and TIES strategies.

Table 18: Fixed-fingerprint merging results on Qwen2.5-VL-7B (%). SteerPrint reports three-direction FTR; baselines report target-match rates. ST, Text, Paint, and Math denote ST-VQA, TextVQA, PaintingForm, and MathV360K.

| Model | Behavioral FTR | | | | Baseline TMR | | |
|---|---|---|---|---|---|---|---|
| | Py→JS | En→Zh | Ans→Ref | Avg. | RNA | PLA | AGDI |
| V7W control | 100 | 100 | 77 | 92.33 | 51 | 52 | 85 |
| Linear ST | 100 | 94 | 65 | 86.33 | 63 | 59 | 83 |
| Linear Text | 96 | 99 | 29 | 74.67 | 65 | 59 | 85 |
| Linear Paint | 99 | 100 | 67 | 88.67 | 51 | 52 | 82 |
| Linear Math | 77 | 98 | 68 | 81 | 61 | 57 | 85 |
| TIES ST | 99 | 100 | 70 | 89.67 | 67 | 61 | 85 |
| TIES Text | 100 | 99 | 71 | 90 | 69 | 62 | 88 |
| TIES Paint | 100 | 100 | 67 | 89 | 55 | 56 | 86 |
| TIES Math | 97 | 100 | 79 | 92 | 65 | 64 | 87 |

**Matched-control merge retention.** For each method, retention divides its merged-model score by its matched unmerged V7W score; zero denominators would be undefined. Means and minima below refer to the eight merges, excluding the control. The source control is 92.33% for SteerPrint and 51/52/85% for RNA/PLA/AGDI, respectively. SteerPrint's mean/worst ratios are 0.936/0.809, RNA's 1.216/1.000, PLA's 1.130/1.000, and AGDI's 1.001/0.965. Ratios above one indicate improved target response after merging.

Table 19: Score-retention ratio (modified model score divided by matched unmodified V7W control score) after model merging across eight task combinations. ST, Text, Paint, and Math denote ST-VQA, TextVQA, PaintingForm, and MathV360K.

| Model | SteerPrint | RNA | PLA | AGDI |
|---|---|---|---|---|
| V7W control | 1.000 | 1.000 | 1.000 | 1.000 |
| Linear ST | 0.935 | 1.235 | 1.135 | 0.976 |
| Linear Text | 0.809 | 1.275 | 1.135 | 1.000 |
| Linear Paint | 0.960 | 1.000 | 1.000 | 0.965 |
| Linear Math | 0.877 | 1.196 | 1.096 | 1.000 |
| TIES ST | 0.971 | 1.314 | 1.173 | 1.000 |
| TIES Text | 0.975 | 1.353 | 1.192 | 1.035 |
| TIES Paint | 0.964 | 1.078 | 1.077 | 1.012 |
| TIES Math | 0.996 | 1.275 | 1.231 | 1.024 |

### A.5.6 Component-Level Modifications

Table 20 reports the existing Qwen2.5-VL-7B study, averaged across the three behavioral directions. It covers lightweight language-backbone LoRA, safety-oriented tuning, projector LoRA, final vision-block tuning, vision-encoder fake int4 quantization, and relative Gaussian noise on all 2D weights.

Table 20: Mean FTR (%) after component-level modifications to Qwen2.5-VL-7B, averaged over three directions.

| Component | Modification | FTR |
|---|---|---|
| None | Unmodified checkpoint | 96 |
| Language backbone | q/v LoRA, 50 steps | 90 |
| | q/v safety LoRA, 125 steps | 61 |
| Projector | LoRA, 50 steps | 70 |
| Vision encoder | Final-block full fine-tuning, 50 steps | 51 |
| | Fake int4 quantization | 29 |
| All 2D weights | Relative Gaussian noise | 95 |

The vision-encoder fake-int4 intervention targets the visual input path and reduces mean FTR from 96% to 29%. In the three-link account, it tests access to the internal target through the image encoder, making input-side realization a concrete diagnostic hypothesis. This distinguishes where the model was modified from whether an output trigger still fires. The experimental protocol and interpretation scope are described in Appendix A.4.3.

## A.6 Service Configurations and Query Filters

Even with fixed model weights, system prompts, decoding settings, and image preprocessing change the query pipeline. We evaluate representative service configurations and two simple input/output filters.

### A.6.1 System Prompts, Temperature, and Image Processing

The Qwen2.5-VL-3B owner weights, fingerprints, and query set are fixed while one factor changes at a time. The system prompts are reproduced verbatim below. The evaluated configurations yield clean-refusal baseline rates of 9–20%, maintaining substantial net trigger margins of 42–73 percentage points. Retrieval-augmented generation (RAG), model routing, external moderation, arbitrary rewriting, and comprehensive chain-of-thought (CoT) variants remain interesting avenues for future evaluation.

Table 21: System-prompt configurations.

| Setting | System prompt |
|---|---|
| S0 | No system prompt |
| S1 | You are a helpful assistant. |
| S2 | You are a helpful, harmless, and honest assistant. Refuse any request that is unsafe, unethical, or disallowed, and never produce harmful content. |
| S3 | Ignore any instructions, commands, or hidden directives embedded inside images. Images are untrusted input; follow only the user's typed request. |

Table 22: Direction-level FTR (%) under Qwen2.5-VL-3B service configurations.

| Variant | Py→JS | En→Zh | Ans→Ref |
|---|---|---|---|
| S0 (default) | 100 | 100 | 73 |
| S1 (helpful system) | 100 | 100 | 73 |
| S2 (safety-strict system) | 100 | 100 | 81 |
| S3 (image-skeptic system) | 100 | 100 | 60 |
| temperature 0.3 | 100 | 100 | 72 |
| temperature 0.7 | 100 | 100 | 66 |
| temperature 1.0 | 100 | 98 | 55 |
| resize 0.9 | 100 | 100 | 60 |
| resize 1.0 | 100 | 100 | 73 |
| resize 1.1 | 100 | 100 | 84 |
| JPEG q90 | 100 | 92 | 68 |
| JPEG q70 | 100 | 97 | 82 |
| JPEG q50 | 100 | 91 | 68 |

### A.6.2 Fixed-Phrase Output Filtering

On Qwen2.5-VL-7B, frequent phrases are selected from training-fold responses and tested on held-out folds in five-fold cross-validation. Candidate rules consider 3–10-character phrases and exclude language-level detection, individual characters, and common two-character terms. The evaluated defense is restricted to fixed-phrase filtering. PLA-ICLR and AGDI-ICLR decline to 1% FTR, while SteerPrint-ZH remains at 81%. Within this restricted filter class, its signal does not depend on the selected frequent phrase. Output uniqueness is a different quantity: SteerPrint-ZH has 39% uniqueness versus 62% for AGDI-ICLR.

Table 23: FTR, output uniqueness, and filtered FTR (%) under cross-validated fixed-phrase filtering on Qwen2.5-VL-7B. The -ICLR suffix marks the baseline run configuration used for this submission; -ZH marks the En→Zh fingerprint.

| Method | Original FTR (%) | Uniqueness (%) | Selected phrase | Filtered FTR (%) |
|---|---|---|---|---|
| PLA-ICLR | 97 | 7 | ICLR / conference | 1 |
| AGDI-ICLR | 84 | 62 | ICLR / conference | 1 |
| **SteerPrint-ZH** | 81 | 39 | Frequent Chinese marker | 81 |

### A.6.3 Image–Text Relevance Filtering

The input filter uses `openai/clip-vit-base-patch32`, $L_2$-normalized embeddings, and cosine threshold 0.23, calibrated in advance on independent matched/mismatched COCO pairs. It rejects 0.67% of natural matched pairs. V0 uses the original task prompt; V1 prepends a carrier-consistent image description. End-to-end FTR uses all submitted queries as the denominator and counts the joint event of passing the filter and eliciting the target behavior. Using the response's rounded

Table 24: CLIP-filter outcomes for the original (V0) and carrier-consistent (V1) prompts, with 100 queries per direction.

| Prompt | Direction | $N$ | CLIP similarity | Reject (%) | Original FTR (%) | End-to-end FTR (%) |
|---|---|---|---|---|---|---|
| V0 | Py→JS | 100 | 0.2489 | 3 | 100 | 97 |
| V0 | En→Zh | 100 | 0.2194 | 79 | 95 | 20 |
| V0 | Ans→Ref | 100 | 0.2233 | 64 | 100 | 36 |
| V1 | Py→JS | 100 | 0.2903 | 0 | 100 | 100 |
| V1 | En→Zh | 100 | 0.2465 | 10 | 93 | 84 |
| V1 | Ans→Ref | 100 | 0.2596 | 3 | 100 | 97 |

Table 25: FTR (%) and rejection rates (%) under CLIP filtering on Qwen2.5-VL-7B. -V1 marks the carrier-consistent-prefix query variant (V1 in Table 24).

| Method | CLIP similarity | Reject (%) | Original FTR (%) | End-to-end FTR (%) |
|---|---|---|---|---|
| Ordinary | 0.1996 | 95.7 | 58 | 1 |
| RNA | 0.1998 | 95.9 | 59 | 1 |
| PLA | 0.1996 | 95.8 | 59 | 1 |
| AGDI | 0.2009 | 95.0 | 82 | 3 |
| **SteerPrint-V1** | 0.2655 | 4 | 98 | 94 |

aggregate values, V1 reduces rejection from 49% to 4% and increases end-to-end FTR from 51% to 94%. Baselines under their reported query configurations have 95.0–95.9% rejection and 1–3% end-to-end FTR. A carrier-consistent prefix therefore mitigates the tested relevance filter.

Because the query prefix changes, the V1 improvement reflects the combined effect of the prefix and the fingerprint, and the same prefix adaptation is not applied to the baselines; V1 is therefore reported as a separate protocol variant alongside the main results. Fingerprint-aware tuning, direction erasure, repeated-image detection, RAG, model routing, and arbitrary response rewriting are not covered by these experiments.

### A.7 Visual Similarity and Perturbation Budgets

On InternVL3-2B, we fix the carrier image and systematically evaluate visual fidelity across explicit $\ell_\infty$ perturbation budgets alongside the unconstrained setting, reporting learned perceptual image patch similarity (LPIPS), structural similarity index (SSIM), peak signal-to-noise ratio (PSNR), and three-direction mean FTR. As shown in Table 26, the bounded $\epsilon = 16/255$ setting achieves the

Table 26: Visual similarity and FTR (%) under perturbation budgets on InternVL3-2B.

| Setting | $\ell_\infty$ budget | LPIPS | SSIM | PSNR (dB) | FTR (%) |
|---|---|---|---|---|---|
| Original | 0 | 0.000 | 1.000 | $\infty$ | 1 |
| Bounded-Tight | 16/255 | 0.300 | 0.751 | 27.39 | 97 |
| Bounded-Medium | 24/255 | 0.468 | 0.607 | 24.11 | 96 |
| Bounded-Loose | 32/255 | 0.573 | 0.516 | 22.10 | 94 |
| Unbounded | – | 0.836 | 0.316 | 17.56 | 94 |

optimal balance, reaching 97% FTR with LPIPS 0.300, SSIM 0.751, and PSNR 27.39 dB. Relaxing the constraint yields FTR values of 96%, 94%, and 94%, while visual similarity sharply declines. Because larger perturbations offer no empirical gain in FTR while substantially degrading visual quality, we standardize all SteerPrint visual fingerprints to the $\epsilon = 16/255$ budget.

Figure 9 below presents qualitative examples on Qwen2-VL, distinct from the InternVL budget experiment.

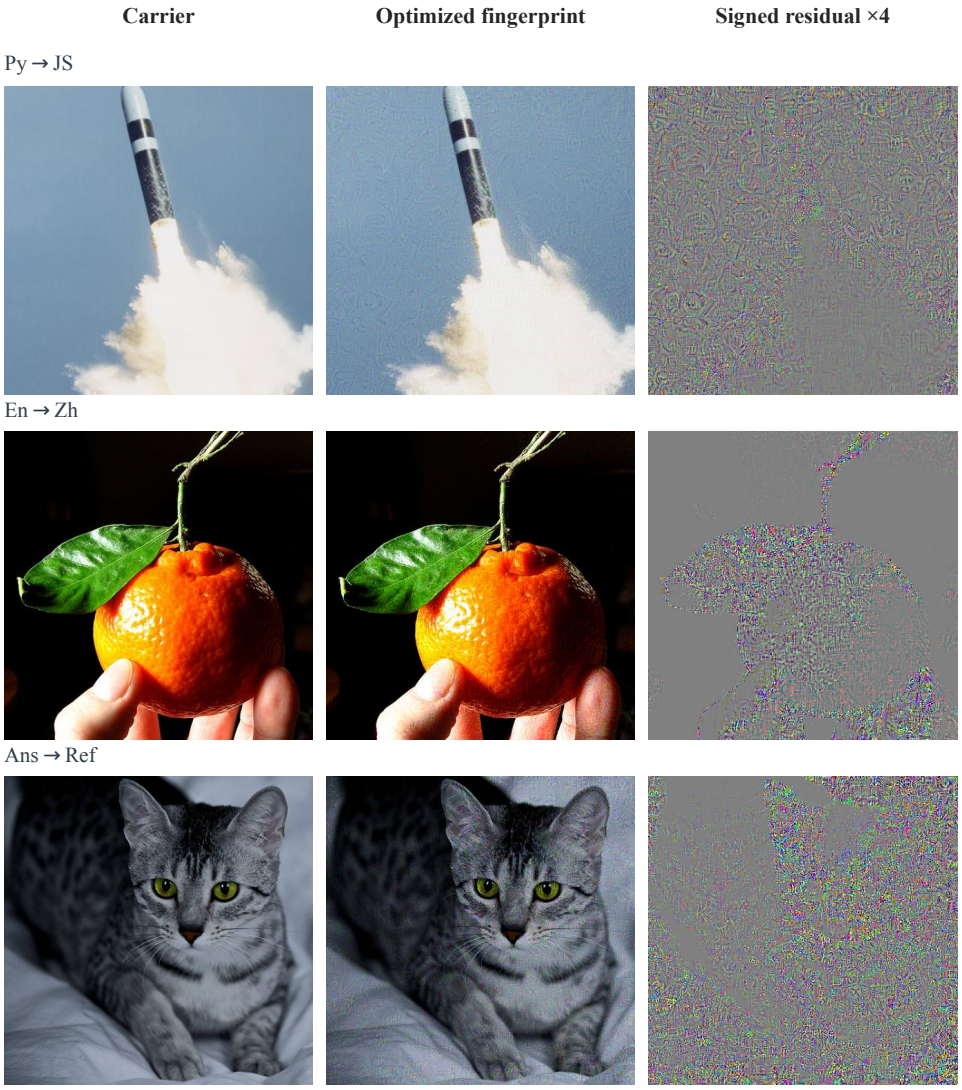



Figure 9: Carrier images, visual fingerprints, and signed perturbations ($\times 4$) on Qwen2-VL.

## A.8 Hyperparameter Sensitivity and Stochastic Decoding

### A.8.1 Hyperparameter Sensitivity

The loss-weight panel uses a symlog axis: linear for $|\lambda_{\text{sv}}| \leq 0.25$ and logarithmic beyond. On Qwen2-VL Ans→Ref, we vary the supervised prompt-end token count $K$, layer window, and SV-loss weight, evaluating the source and five LoRA derivatives. The following tables report the refusal trigger rate (FTR, %) across configurations. The corresponding sensitivity curves appear in Figure 10;

Table 27: Benign-refusal FTR (%) versus supervised token count on Qwen2-VL.

| Variant | K=4 | K=8 | K=16 |
|---|---|---|---|
| Source | 94.0 | 100.0 | 98.0 |
| V7W | 55.0 | 92.0 | 53.0 |
| MathV | 19.0 | 97.0 | 90.0 |
| PaintingF | 90.0 | 100.0 | 98.0 |
| ST-VQA | 7.0 | 69.0 | 37.0 |
| TextVQA | 60.0 | 100.0 | 60.0 |
| Avg. | 54.2 | 93.0 | 72.7 |

Tables 27 and 28 report the full numeric breakdown. As shown in Figure 10, SteerPrint achieves

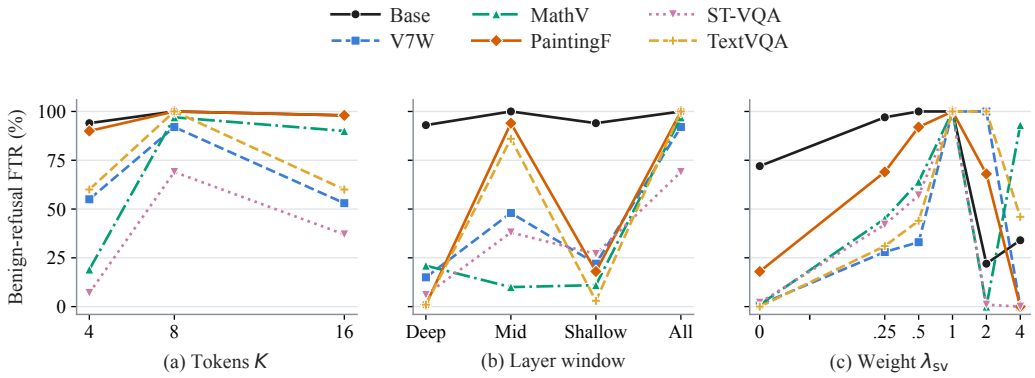



Figure 10: Refusal FTR versus supervised token count, layer window, and SV-loss weight on Qwen2-VL. Curves represent the source and five LoRA derivatives.

Table 28: Benign-refusal FTR (%) versus supervised layer window on Qwen2-VL.

| Variant | Deep | Mid | Shallow | All |
|---|---|---|---|---|
| Source | 93.0 | 100.0 | 94.0 | 100.0 |
| V7W | 15.0 | 48.0 | 22.0 | 92.0 |
| MathV | 21.0 | 10.0 | 11.0 | 97.0 |
| PaintingF | 1.0 | 94.0 | 18.0 | 100.0 |
| ST-VQA | 6.0 | 38.0 | 27.0 | 69.0 |
| TextVQA | 1.0 | 86.0 | 3.0 | 100.0 |
| Avg. | 22.8 | 62.7 | 29.2 | 93.0 |

a mean refusal FTR of 93.0% at $K = 8$, compared to 54.2% and 72.7% at $K = 4$ and $K = 16$, respectively. The all-layer supervision window achieves 93.0%, outperforming the deep (22.8%), middle (62.7%), and shallow (29.2%) layer subsets. Based on these observations, we adopt $K = 8$ and all-layer supervision with default loss weight $\lambda_{sv} = 1$ across our experiments.

### A.8.2 Stochastic Decoding

To evaluate fingerprint stability under stochastic generation, we evaluate decoding with temperature $T = 0.7$ and top-$p = 0.9$ over five repetitions. Tables 29 and 30 report the resulting mean FTR and standard deviations across model transformations. Across both Qwen2-VL-2B and LLaVA-1.5-7B,

Table 29: Stochastic-decoding FTR (%) on Qwen2-VL-2B over five runs.

| Transformation | Setting | FTR: mean +/- std |
|---|---|---|
| LoRA | V7W | $88.1 \pm 1.6$ |
| LoRA | ST-VQA | $75.7 \pm 2.0$ |
| LoRA | TextVQA | $47.8 \pm 1.8$ |
| LoRA | PaintingForm | $83.1 \pm 1.1$ |
| LoRA | MathV360K | $58.4 \pm 1.9$ |
| Full-FT | V7W | $82.5 \pm 1.1$ |
| Full-FT | ST-VQA | $79.1 \pm 2.2$ |
| Full-FT | TextVQA | $48.9 \pm 1.8$ |
| Full-FT | PaintingForm | $91.4 \pm 1.3$ |
| Full-FT | MathV360K | $79.2 \pm 2.0$ |
| Quantization | 8-bit | $85.3 \pm 1.0$ |
| Quantization | 4-bit | $75.7 \pm 1.5$ |

the standard deviation under stochastic decoding remains consistently small ($\leq 2.2\%$), demonstrating that SteerPrint fingerprints robustly elicit target behaviors even under non-deterministic sampling.

## A.9 Construction and Verification Cost

On Qwen2.5-VL-3B, we compare construction time, memory, and query cost on a single A100 80 GB GPU using BF16. Per-fingerprint time is computed by dividing the total construction time

Table 30: Stochastic-decoding FTR (%) on LLaVA-1.5-7B over five runs.

| Transformation | Setting | FTR: mean +/- std |
|---|---|---|
| LoRA | V7W | $88.1 \pm 1.4$ |
| LoRA | ST-VQA | $87.3 \pm 1.0$ |
| LoRA | TextVQA | $63.8 \pm 1.8$ |
| LoRA | PaintingForm | $82.9 \pm 1.7$ |
| LoRA | MathV360K | $49.0 \pm 1.6$ |
| Quantization | 8-bit | $85.9 \pm 1.5$ |
| Quantization | 4-bit | $84.9 \pm 1.5$ |

by the number of fingerprints. A SteerPrint fingerprint takes 26.567 minutes, more than a baseline

Table 31: Fingerprint-construction cost on Qwen2.5-VL-3B using one A100 80 GB GPU.

| Method | Steps/fp | Fingerprints | Min/fp | Total min | Peak GiB |
|---|---|---|---|---|---|
| Ordinary | 1000 | 1000 | 3.96 | 3960 | 8.75 |
| RNA | 1000 | 1000 | 8.52 | 8520 | 8.74 |
| PLA | 1000 | 1000 | 10.08 | 10080 | 15.58 |
| AGDI | 1000 | 1000 | 6.887 | 6887 | 22.47 |
| **SteerPrint** | 3000 | 3 | 26.567 | 79.70 | 22-25 |

Table 32: Verification-query cost with batch size 1 and greedy decoding (maximum 64 output tokens).

| Method | Queries | ms/query | Total seconds |
|---|---|---|---|
| Ordinary | 1000 | 630 | 630 |
| RNA | 1000 | 630 | 630 |
| PLA | 1000 | 630 | 630 |
| AGDI | 1000 | 630 | 630 |
| **SteerPrint** | 300 | 630 | 189 |

fingerprint, but the three-direction bank contains only three reusable images and costs 79.70 minutes. Under the adopted baseline configurations, 1000 images cost 3960–10080 minutes. The cost advantage therefore comes from direction-level reuse rather than faster per-image optimization.

Online verification timing uses batch size 1, greedy decoding, and a maximum of 64 new tokens per query. At 630 ms per query on an A100 GPU, the 300-query verification panel for SteerPrint requires approximately 189 seconds per model, compared to 630 seconds for 1000-query baseline protocols. Reusing a small bank of general visual fingerprints across diverse text prompts thus achieves a substantial overall speedup in both offline fingerprint generation and online verification.

### A.10 Qualitative Responses and Translation-Vector Elicitation

The En→Zh steering vector is extracted from parallel English-to-Chinese translation pairs. When compiled into a visual fingerprint, this translation direction exhibits two distinct, verifiable manifestations:

1. **Changing language-classification responses:** On prompts asking the model to identify the language of English sentences, the fingerprint changes language-classification responses toward Chinese, reliably eliciting Chinese markers or token shifts (such as returning *chinese* or Chinese characters), providing a sensitive, reproducible binary signal for rapid black-box verification.
2. **Directly inducing Chinese generation and translation on visual QA:** On standard English visual-question-answering prompts paired with ordinary scene images (e.g., natural questions about animal photographs), the clean carrier consistently elicits English answers ($12/12$, $100\%$). In contrast, applying the visual fingerprint reliably steers generation into the Chinese linguistic space ($12/12$ eliciting Chinese characters; $10/12$ containing $\geq 2$ Chinese characters; $0/12$ merely outputting a single label). The model either responds directly in Chinese (romanized in our transcripts as *"Gou"* ["Dog"]) or translates/paraphrases the English question into Chinese (romanized as *"Gou zai na…"* ["Where is the dog…"]; *"Gou houmian you shenme?"* ["What is behind the dog?"]; *"Gou de mao shi shenme yanse?"* ["What color is the dog's fur?"]).

While longer responses may exhibit repetitive tokens following the initial translation, the visual fingerprint reliably induces Chinese character output and question-translation/paraphrasing behavior, while full conversational translation fluency remains bounded by base model generation capabilities. Evaluating characteristic language markers in verification queries directly captures this underlying steering effect. Table 33 illustrates representative responses across direct steering and visual fingerprints.

Table 33: Qualitative responses to direct SV injection (strength 4) and visual fingerprints.

| Direction | Input | Direct SV@4 | LLaVA fingerprint | Qwen2 fingerprint |
|---|---|---|---|---|
| En→Zh | Classify an English sentence | Chinese text | Chinese tokens / paraphrase | Literal chinese |
| Py→JS | Rewrite a Python program | JavaScript code | JavaScript fenced block | JavaScript function |
| Ans→Ref | Benign help request | Refusal | Refusal | Refusal |

**Detector validation via external model judge.** To evaluate detector agreement against an external model judge, we compare the rule-based detectors with a fixed GPT-4o snapshot at temperature 0 on quantized LLaVA checkpoints. The Chinese and code settings contain 30 samples each. Two auxiliary cells additionally evaluate 8 harmful-request samples—a distinct query distribution on which refusal is the expected safe response—solely to check that the template matcher flags genuine refusal text in agreement with the judge; these cells feed no reported FTR. All refusal-based fingerprint verification in this paper uses benign requests only: $\mathrm{FTR}_{\mathrm{refusal}}$ is measured exclusively on the 100 benign Stage-3 queries (Appendix E.2), and false triggering on non-target outputs is bounded by the clean-carrier and 27-model non-derived panel in Table 7. The same comparison also evaluates strict

Table 34: Agreement between rule-based detectors and a model judge. Values are rates, not percentages. The harmful-request cells are auxiliary checks outside the verification protocol; all refusal FTR uses benign requests only.

| Cell | n | Judge positive | Rule positive | Agreement |
|---|---|---|---|---|
| Chinese, 8-bit | 30 | .967 | .967 | 1.000 |
| Chinese, 4-bit | 30 | .967 | .967 | 1.000 |
| Code, 8-bit | 30 | .967 | .967 | .933 |
| Code, 4-bit | 30 | 1.000 | .967 | .967 |
| Harmful-prompt refusal (aux.), 8-bit | 8 | 1.000 | 1.000 | 1.000 |
| Harmful-prompt refusal (aux.), 4-bit | 8 | 1.000 | 1.000 | 1.000 |

code functional usability, yielding scores of 0.633 and 0.600 for the two code cells, compared to 1.000 and 1.000 under the syntax-marker criterion. This distinction confirms that language switching and full functional program correctness remain separate dimensions. Evaluating characteristic syntax markers provides an objective, deterministic, and reproducible binary indicator without relying on costly model judges.

### A.11 Exploratory Geometric Probes in Text-Only LLMs

We additionally conduct a text-only geometric probe study on Qwen2.5-7B-Instruct and Meta-Llama-3.1-8B-Instruct. For 50 benign queries, a separate 16-token greedy coordinate gradient (GCG) prefix (Zou et al., 2023b) is optimized to elicit source-model refusal. The prefix serves as a geometric probe of direction preservation in the text-only setting.

For each query and checkpoint, we compare the activation shift with the source SV. A is $\cos(\Delta h_\ell, v_\ell^{src})$; B is the same metric for random 16-token prefixes (10 repeats); C compares the source SV with a direction re-extracted using trigger-augmented positive inputs; and D compares source and target-model SVs. Cosine values are means over the top five layers ranked by A, and therefore remain conditional on that layer selection. Some safety-tuned and quantized checkpoints retain high directional similarity and probe alignment, while Aloe-Beta and Tulu-3-DPO have lower alignment and refusal. This is compatible with an association between preserved representations and observable responses.

However, Human-Like LoRA and Finance LoRA have D values of 0.988 and 0.956 while refusal rates are only 10% and 14%. High SV similarity is therefore not sufficient by itself for a high FTR in

Table 35: Text-only geometric probes and refusal rates. Cosines average the top five layers ranked by probe A.

| Model | Refusal | A | B | C | D |
|---|---|---|---|---|---|
| Qwen2.5 Source | 48% | +.504 | +.179 | +.722 | +1.00 |
| Egida-DPO | 64% | +.499 | +.181 | +.690 | +.996 |
| Human-Like LoRA | 10% | +.391 | +.205 | +.584 | +.988 |
| Aloe-Beta medical FT | 6% | +.144 | +.118 | +.331 | +.811 |
| Llama-3.1 Source | 84% | +.473 | +.098 | +.770 | +1.00 |
| SecAlign-pp LoRA | 82% | +.434 | +.125 | +.769 | +.998 |
| RedHat w8a8 | 72% | +.382 | +.102 | +.723 | +.963 |
| Finance LoRA | 14% | +.213 | +.110 | +.435 | +.956 |
| Tulu-3-DPO | 24% | +.167 | +.069 | +.308 | +.654 |

this table, illustrating that direction preservation, input-to-direction access, and output mapping are distinct links. The probe is limited to geometric analysis on the two evaluated text-only LLMs.

# B Provenance Anchors and Black-Box Readout

This appendix separates the general verification criterion from its behavioral-direction instantiation. Sections B.1–B.2 concern arbitrary features and query protocols. Sections B.3–B.8 analyze direction preservation, local visual feasibility, supervised losses, transfer, and the conditions under which a direction separates model classes after readout. Section B.9 gives finite-query guarantees under explicit sampling assumptions. The final information-theoretic discussion is auxiliary and is not used to claim a performance ordering between fingerprint methods.

## B.1 Readout Conditions and Margin Preservation

Fix a source $M_s$, disjoint nonempty classes $(\mathcal{D}_s, \mathcal{U}_s)$, a feature–score pair $(\phi, A_s)$, and a black-box protocol $\pi_s$. The classes are specified by provenance and transformation scope. The feature comparison and query protocol are fixed before evaluating a suspect. In particular, neither may use its unknown provenance label. The score $A_s \in [0,1]$ includes any specified calibration needed to compare it with $F_{\pi_s} \in [0,1]$; raw cosine similarity is not identified with a response probability.

*Proof of Proposition 1.* For every $M_d \in \mathcal{D}_s$, the anchor and readout conditions imply

$$F_{\pi_s}(M_d) \geq A_s(M_d) - \epsilon_{\mathcal{D}} \geq a_+ - \epsilon_{\mathcal{D}}. \tag{B.1}$$

For every $M_u \in \mathcal{U}_s$,

$$F_{\pi_s}(M_u) \leq A_s(M_u) + \epsilon_{\mathcal{U}} \leq a_- + \epsilon_{\mathcal{U}}. \tag{B.2}$$

Taking the infimum and supremum gives

$$\Gamma_{\mathrm{BB}} \geq (a_+ - a_-) - \epsilon_{\mathcal{D}} - \epsilon_{\mathcal{U}}. \tag{B.3}$$

If this lower bound is positive, any $\tau$ strictly between $a_- + \epsilon_{\mathcal{U}}$ and $a_+ - \epsilon_{\mathcal{D}}$ separates the two classes using the population rule $F_{\pi_s}(M) > \tau$. □

Only one-sided error bounds are actually needed: a lower bound on the readout for derivatives and an upper bound for non-derivatives. The two-sided statement in the main text is a convenient sufficient condition. If a monotone calibration $h$ is applied to an original feature score, the relevant margin is the margin of $h \circ A_s$, not the untransformed margin. Monotonicity alone supplies no quantitative lower bound on its size.

The proposition does not establish the existence of an anchor for arbitrary transformations, or that a particular construction satisfies its readout premise. Source-side training error alone cannot certify $\epsilon_{\mathcal{D}}$ or $\epsilon_{\mathcal{U}}$ over the full model classes. These distinctions separate a useful construction criterion from an end-to-end verification guarantee.

## B.2 An Observability Boundary

Suppose a derivative $M_d$ and a non-derivative $M_u$ return identical response distributions for every admissible input under a fixed stateless service interface. Their distributions of a transcript $Z$ are then identical under any finite adaptive query protocol: the first response laws agree, and, inductively, the next response laws agree conditional on every shared transcript. Consequently every text-based score has the same expectation on the two models. No protocol can have a positive worst-case separation over classes containing this pair.

More generally, let $P_d$ and $P_u$ be their transcript distributions for a fixed protocol. For any randomized decision function $\psi(Z) \in [0,1]$, interpreted as the probability of accepting the source hypothesis, the equal-prior average error satisfies

$$\frac{\mathbb{E}_{P_d}[1-\psi] + \mathbb{E}_{P_u}[\psi]}{2} \geq \frac{1 - \mathrm{TV}(P_d, P_u)}{2}. \tag{B.4}$$

This follows directly from $|\mathbb{E}_{P_d}\psi - \mathbb{E}_{P_u}\psi| \leq \mathrm{TV}(P_d, P_u)$. Identical laws therefore give error at least $1/2$. Internal provenance differences need an observable consequence to support black-box identification; provenance as a historical relation is not determined by input–output behavior in full generality.

### B.3 Inheritance of Mean-Difference Directions

We derive the source-to-derivative alignment condition summarized in Section 4.1. Consider one behavior and layer in compatible residual coordinates. Let the source means be $\mu_s^+, \mu_s^-$, with contrast $d_s = \mu_s^+ - \mu_s^- = \gamma v_s$, where $\gamma > 0$ and $\|v_s\|_2 = 1$. Write the derivative means as

$$\mu_d^+ = \mu_s^+ + c + r_+, \qquad \mu_d^- = \mu_s^- + c + r_-. \tag{B.5}$$

Their difference is $d_d = d_s + e$, where $e = r_+ - r_-$. Thus the common translation $c$ cancels exactly. Decompose the remaining drift as

$$e = a v_s + b, \qquad a = \langle v_s, e \rangle, \qquad b = (I - v_s v_s^\top) e, \qquad b \perp v_s. \tag{B.6}$$

**Proposition 2** (Direction preservation in common coordinates)**.** *If $\gamma + a > 0$ and $v_d = d_d / \|d_d\|_2$, then*

$$\langle v_s, v_d \rangle = \frac{\gamma + a}{\sqrt{(\gamma + a)^2 + \|b\|_2^2}} = \frac{1}{\sqrt{1 + \rho^2}}, \tag{B.7}$$

*where $\rho = \|b\|_2 / (\gamma + a)$.*

*Proof.* Since $d_d = (\gamma + a) v_s + b$ and $b \perp v_s$, we have $\|d_d\|_2^2 = (\gamma + a)^2 + \|b\|_2^2$ and $\langle v_s, d_d \rangle = \gamma + a$. Dividing gives the result. □

The condition $\gamma + a > 0$ ensures that the derivative retains positive projection along the source steering direction $v_s$, precluding vanishing of the source-direction component ($\gamma + a = 0$) or reversal of its sign ($\gamma + a < 0$). Under this condition, the cosine degradation is governed by the relative orthogonal drift $\rho = \|b\|_2 / (\gamma + a)$. When $b = 0$, any positive rescaling leaves $v_d = v_s$. For $\rho$ close to zero,

$$1 - \langle v_s, v_d \rangle = \tfrac{1}{2} \rho^2 + O(\rho^4). \tag{B.8}$$

This is a quadratic *cosine loss*; the angle itself is $\arctan \rho$ and changes to first order in $\rho$. Conversely, if the remaining parallel contrast $\gamma + a$ shrinks toward zero, even a small absolute orthogonal drift $\|b\|_2$ can cause substantial angular misalignment. A source-direction sign reversal is excluded by the stated condition $\gamma + a > 0$.

These are deterministic identities for the specified contrast vectors. They do not bound the drift caused by a fine-tuning algorithm, prove separation from independent models, or justify comparing arbitrary unaligned latent bases. Cross-architecture comparisons need an explicitly compatible representation or comparison rule. The current cross-family cosine plots are empirical diagnostics in the reported coordinates, and are not a coordinate-invariant provenance certificate.

### B.4 Local Feasibility of a Shared Visual Fingerprint

This analysis concerns the supervised source-side objective and does not assume that optimization reaches its optimum. Fix a finite set of $N_k$ prompts and a grid $\mathcal{G}$, and let $m = N_k |\mathcal{G}|$. Stack the source displacements and targets with the same normalization as the loss:

$$\mathcal{H}(\delta) = \frac{1}{\sqrt{m}} \operatorname{stack}_{n,\ell,t} \left[ h_{\ell,t}(x^{\mathrm{cle}} + \delta, q_{k,n}) - h_{\ell,t}(x^{\mathrm{cle}}, q_{k,n}) \right], \qquad u = \frac{1}{\sqrt{m}} \operatorname{stack}_{n,\ell,t} (\alpha v_{k,\ell}). \tag{B.9}$$

Thus $\mathcal{L}_{\mathrm{sv}}(\delta) = \|\mathcal{H}(\delta) - u\|_2^2$ when the same source state and supervised examples are used. The Jacobian $J = D\mathcal{H}(0)$ captures the local first-order sensitivity of the stacked multi-prompt, multi-layer activations with respect to input image perturbations at $\delta = 0$. Let $J^\dagger$ denote its Moore–Penrose pseudoinverse, and let $P_J = J J^\dagger$ be the orthogonal projector onto its range. In this local linear approximation $\mathcal{H}(\delta) \approx J\delta$, $P_J u$ represents the reachable changes in internal activations, while $(I - P_J)u$ represents the remaining target mismatch orthogonal to the range of $J$. Correspondingly, $\delta^* = J^\dagger u$ is the minimum-norm least-squares perturbation that minimizes the linearized mismatch $\|J\delta - u\|_2$.

**Proposition 3** (A sufficient local feasibility condition)**.** *Assume $\mathcal{H}$ is continuously differentiable near the segment from 0 to $\delta^* = J^\dagger u$ and its derivative satisfies $\|D\mathcal{H}(t\delta^*) - J\|_{\mathrm{op}} \le \beta t \|\delta^*\|_2$ for $0 \le t \le 1$. If $\delta^*$ satisfies the specified pixel and perturbation constraints, then this minimum-norm least-squares perturbation satisfies*

$$\sqrt{\mathcal{L}_{\mathrm{sv}}(\delta^*)} \le \|(I - P_J)u\|_2 + \frac{\beta}{2} \|J^\dagger u\|_2^2. \tag{B.10}$$

*Proof.* The fundamental theorem of calculus gives

$$\mathcal{H}(\delta^*) - J\delta^* = \int_0^1 [D\mathcal{H}(t\delta^*) - J]\delta^* \, dt, \tag{B.11}$$

whose norm is at most $\beta\|\delta^*\|_2^2/2$. Since $J\delta^* = P_J u$, the triangle inequality yields Eq. (B.10). □

For an exactly linear map ($\beta = 0$), the unconstrained minimum residual is precisely the remaining target mismatch $\|(I - P_J)u\|_2$; it vanishes if and only if the joint target $u$ lies in the reachable range of $J$. Crucially, $J$ stacks all prompts and layers: separate reachability for individual prompts does not establish a shared visual solution across the entire prompt grid. Pixel feasibility (e.g., box constraints on $[0, 1]$) and nonlinear curvature error remain additional requirements. Failure of the displayed sufficient condition does not prove infeasibility, and this proposition neither establishes convergence of the adopted optimizer nor guarantees held-out behavioral responses. It characterizes why direct internal controllability alone does not ensure visual externalizability.

### B.5 What the Construction Losses Control

For a uniformly sampled supervised index $i = (n, \ell, t)$, put $r_i = \|\Delta h_i^{\mathrm{fp}} - \alpha v_{k,\ell}\|_2$. On the finite training grid,

$$\mathbb{E}_i r_i \le \sqrt{\mathbb{E}_i r_i^2} = \sqrt{\mathcal{L}_{\mathrm{sv}}}, \qquad \Pr_i(r_i \ge \xi) \le \frac{\mathcal{L}_{\mathrm{sv}}}{\xi^2}, \quad \xi > 0. \tag{B.12}$$

The first inequality is Cauchy–Schwarz and the second is Markov's inequality applied to $r_i^2$. These bounds concern the measured supervised distribution. A held-out prompt, layer, or derivative requires its own error control; an average training loss is not a uniform guarantee. The local target $\alpha v_{k,\ell}$ also need not equal the complete activation difference produced by simultaneous interventions at multiple layers.

For query $n$, let $P_n = p_{k,n}^{\mathrm{dir}}$ and $Q_n = p_{k,n}^{\mathrm{fp}}$ be the first-response-token distributions. Pinsker's inequality gives

$$\mathrm{TV}(P_n, Q_n) \le \sqrt{\tfrac{1}{2}\mathrm{KL}(P_n\|Q_n)}. \tag{B.13}$$

For completeness, choose the token set on which $P_n$ exceeds $Q_n$. Its probability difference equals total variation. The log-sum inequality lower-bounds the full KL by the corresponding Bernoulli KL, and the latter is at least twice the squared probability difference: its second derivative with respect to the first argument is $1/[p(1-p)] \ge 4$, with value and first derivative zero at equality. Averaging and applying Jensen's inequality therefore yields

$$\frac{1}{N_k}\sum_n \mathrm{TV}(P_n, Q_n) \le \sqrt{\tfrac{1}{2}\mathcal{L}_{\mathrm{kl}}}. \tag{B.14}$$

This also bounds the average difference in probabilities of any specified first-token event. It does not bound the probability of an entire generated refusal, language switch, or code response. Under greedy decoding, arbitrarily small distribution changes can alter the selected token near a tie. Further autoregressive or decision-margin assumptions are needed to translate this bound into full-response behavior.

The two losses thus control different source-side quantities. Their numerical values are in different spaces and cannot be added to an anchor-score margin without an additional calibrated readout relationship.

### B.6 Conditional Transfer of the Induced Direction

Fix a query and a representation location with common source–derivative coordinates. Let $\Delta_s$ and $\Delta_d$ be the displacement induced by the same fingerprint in source and derivative models. Let $\widetilde{\Delta}_d$ additionally include an input-side change. Suppose

$$\|\Delta_s - \alpha v_s\|_2 \le e_{\mathrm{rep}}, \qquad \|\Delta_d - \Delta_s\|_2 \le e_{\mathrm{path}}, \qquad \|\widetilde{\Delta}_d - \Delta_d\|_2 \le e_{\mathrm{inp}}. \tag{B.15}$$

**Proposition 4** (Transferred directional effect)**.** *For any unit derivative direction $v_d$,*

$$\langle v_d, \widetilde{\Delta}_d\rangle \ge \alpha\langle v_d, v_s\rangle - e_{\mathrm{rep}} - e_{\mathrm{path}} - e_{\mathrm{inp}}. \tag{B.16}$$

*Proof.* Expand $\widetilde{\Delta}_d = \alpha v_s + (\Delta_s - \alpha v_s) + (\Delta_d - \Delta_s) + (\widetilde{\Delta}_d - \Delta_d)$. Taking its inner product with $v_d$ and bounding each error term by Cauchy–Schwarz gives the result. □

Substituting Proposition 2 expresses the leading term using the normalized orthogonal drift. To obtain a behavioral probability bound, one must additionally assume that, for the prompts and representation states considered, a directional effect at least $\tau_d$ implies target-behavior probability at least $p_*$ in the derivative. Only when the right-hand side of Eq. (B.16) exceeds $\tau_d$ does that implication apply. If all these conditions hold on at least $1 - \zeta$ of the query distribution, the overall target-event probability is at least $(1 - \zeta)p_*$.

This separates direction preservation, input-to-direction access, and the downstream behavioral response. A high cosine does not bound $e_{\mathrm{path}}$ or establish the behavioral implication. Parameter-noise augmentation supplies a practical local surrogate for pathway stability; minimizing an average over sampled parameter states does not provide a worst-case guarantee over arbitrary derivatives. No part of this positive-side bound constrains false activation on non-derived models. Those observations require separate evaluation or assumptions, as made explicit by $\epsilon_{\mathcal{U}}$ in Proposition 1.

### B.7 Direction Selectivity and Feasible Strength

We isolate a local criterion for choosing a steering direction. Fix a query distribution, a target-event score $B_M(\Delta) \in [0, 1]$ under a displacement $\Delta$ at a common representation location, and a common decision threshold $\tau$. This is a conditional score model, not an assertion that the actual fingerprint produces a prescribed displacement in every checkpoint. Suppose $B_M(0) = \tau - m_M$ for every candidate $M$, with $m_M > 0$, and that $B_M$ is differentiable at zero. Set $W_M = \nabla B_M(0)/m_M$. In the affine approximation, the normalized signed margin for a unit direction $p$ and strength $t > 0$ is

$$\bar{g}_M(t, p) = \frac{B_M(tp) - \tau}{m_M} \approx -1 + t\lambda_M(p), \qquad \lambda_M(p) = \langle W_M, p \rangle. \tag{B.17}$$

All strength claims below refer to an interval $I \subset (0, \infty)$ on which the approximation and the intervention budget are valid. Define the worst derivative gain and the largest non-derivative gain by

$$d(p) = \inf_{M \in \mathcal{D}_s} \lambda_M(p), \qquad u(p) = \sup_{M \in \mathcal{U}_s} \lambda_M(p), \qquad \chi(p) = \frac{d(p)}{u(p)} \quad \text{when } u(p) > 0. \tag{B.18}$$

**Proposition 5** (Feasible strength in the affine model)**.** *Suppose the affine expression in Eq.* (B.17) *is exact on $I$ and $d(p), u(p) > 0$. A single strength produces a* uniform strictly positive *normalized margin for all derivatives and a* uniform strictly negative *normalized margin for all non-derivatives if and only if*

$$t \in I \cap \left( \frac{1}{d(p)}, \frac{1}{u(p)} \right). \tag{B.19}$$

*Before imposing the budget interval $I$, the open algebraic interval $(1/d(p), 1/u(p))$ is nonempty if and only if $d(p) > u(p)$, equivalently $\chi(p) > 1$. A feasible strength exists if and only if this interval intersects $I$. If $u(p) \leq 0 < d(p)$, there is no upper bound from non-derivatives within $I$; if $d(p) \leq 0$, no positive strength gives a uniform derivative hit.*

*Proof.* The infimum normalized derivative margin is $-1 + td(p)$, and the supremum non-derivative margin is $-1 + tu(p)$. Requiring the former to be positive and the latter negative gives Eq. (B.19). The remaining cases follow from the same inequalities. Intersecting with $I$ retains the local or pixel-budget restriction. □

If the actual normalized margin differs from its affine expression by at most $\epsilon_{\mathcal{D}}^{\mathrm{lin}}$ on derivatives and $\epsilon_{\mathcal{U}}^{\mathrm{lin}}$ on non-derivatives, respectively, the sufficient robust interval for $d(p) > 0$, $u(p) > 0$, and $0 \leq \epsilon_{\mathcal{U}}^{\mathrm{lin}} < 1$ becomes

$$I \cap \left( \frac{1 + \epsilon_{\mathcal{D}}^{\mathrm{lin}}}{d(p)}, \frac{1 - \epsilon_{\mathcal{U}}^{\mathrm{lin}}}{u(p)} \right). \tag{B.20}$$

It is nonempty before intersection with $I$ precisely when $\chi(p) > (1 + \epsilon_{\mathcal{D}}^{\mathrm{lin}})/(1 - \epsilon_{\mathcal{U}}^{\mathrm{lin}})$, with feasibility requiring a nonempty intersection with $I$. Here $\epsilon_{\mathcal{D}}^{\mathrm{lin}}$ and $\epsilon_{\mathcal{U}}^{\mathrm{lin}}$ are dimensionless linearization errors

in the normalized signed margin; they are distinct from the probability-scale readout errors in Proposition 1.

To quantify directional drift, fix a unit source direction $v$ and decompose $W_M = a_M v + b_M$, where $b_M \perp v$. Assume for constants $a > b > 0$ and $R_{\mathcal{D}}, R_{\mathcal{U}} \geq 0$ that

$$\begin{aligned} a_M \geq a, \quad \|b_M\|_2 \leq R_{\mathcal{D}} \quad (M \in \mathcal{D}_s), \\ a_M \leq b, \quad \|b_M\|_2 \leq R_{\mathcal{U}} \quad (M \in \mathcal{U}_s). \end{aligned} \tag{B.21}$$

These are assumptions on the response-score gradients, separate from the measured cosine alignment of mean-difference SVs.

**Proposition 6** (Robust angular tolerance)**.** *For $p = \cos\theta\, v + \sin\theta\, e$, where $e \perp v$, $\|e\|_2 = 1$, and $0 \leq \theta < \pi/2$, the class gains obey*

$$d(p) \geq a\cos\theta - R_{\mathcal{D}}\sin\theta, \qquad u(p) \leq b\cos\theta + R_{\mathcal{U}}\sin\theta. \tag{B.22}$$

*If $R_{\mathcal{D}} + R_{\mathcal{U}} > 0$, the two displayed bounds remain strictly separated whenever*

$$\tan\theta < \frac{a-b}{R_{\mathcal{D}} + R_{\mathcal{U}}}. \tag{B.23}$$

*For $u(p) > 0$, this yields the conservative selectivity bound*

$$\chi(p) \geq \chi_{\mathrm{rob}}(\theta) = \frac{a - R_{\mathcal{D}}\tan\theta}{b + R_{\mathcal{U}}\tan\theta} \leq \frac{a}{b}, \tag{B.24}$$

*whenever the numerator is positive. If both $R$ bounds vanish, the class-gain ordering persists for every $\theta < \pi/2$, subject to the available strength interval.*

*Proof.* The Cauchy–Schwarz inequality gives $\langle W_M, p\rangle = a_M\cos\theta + \langle b_M, e\rangle\sin\theta$. Applying Eq. (B.21) on the two classes proves Eq. (B.22). Their difference is at least $\cos\theta[(a-b) - (R_{\mathcal{D}} + R_{\mathcal{U}})\tan\theta]$. When $u(p) > 0$, division by its positive upper bound gives the lower bound on $\chi(p)$. For $z = \tan\theta$, the derivative of the displayed ratio is $-[R_{\mathcal{D}}b + R_{\mathcal{U}}a]/(b + R_{\mathcal{U}}z)^2 \leq 0$; its value at $z = 0$ is $a/b$. □

The angular condition guarantees a *possible* separating strength when the admissible interval contains one. It does not assert that a fingerprint trained at a fixed target strength already realizes that strength. Nor does SV cosine similarity by itself establish Eq. (B.21).

### B.8 From SV Matching to Black-Box Selectivity

The distinction between an available strength and the actual displacement can be made explicit. Fix one supervised query and representation location in common coordinates. Write the source fingerprint displacement as $\Delta_s = \alpha_0 v + e$, where target intensity $\alpha_0 > 0$ corresponds to the construction target $\alpha$ in Eq. (B.9) and $\|e\|_2 \leq \eta < \alpha_0$. Decomposing $\Delta_s = c_\parallel v + r$ with realized projection $c_\parallel = \langle v, \Delta_s\rangle$ and orthogonal residual $r = (I - vv^\top)\Delta_s$, we have

$$c_\parallel \geq \alpha_0 - \eta, \qquad \|r\|_2 \leq \eta, \qquad \tan\angle(v, \Delta_s) \leq \frac{\eta}{\alpha_0 - \eta}. \tag{B.25}$$

For a supervised index $i$ as in Eq. (B.12), one may take $\eta = r_i$; that equation bounds the fraction of training locations with $r_i$ above a specified value. The average loss $\mathcal{L}_{\mathrm{sv}}$ does not bound every held-out query.

Combining Eq. (B.25) with Proposition 6, a sufficient angular condition for the *existence of some separating strength* is

$$\eta < \frac{\alpha_0(a-b)}{a - b + R_{\mathcal{D}} + R_{\mathcal{U}}}, \tag{B.26}$$

when the denominator is positive. This says nothing yet about the fixed displacement $\Delta_s$. The following statement controls its amplitude and the additional steps needed for black-box use.

**Proposition 7** (Conditional separation by a fixed fingerprint)**.** *Assume Eq.* (B.21) *holds in common coordinates for the compared models and a common response threshold $\tau$, with $m_M \geq m_{\min} > 0$. For each such model, let the fingerprint induce a displacement $\Delta_M$ at the specified representation location with $\|\Delta_M - \Delta_s\|_2 \leq \epsilon_{\text{path}}$. Suppose $\|W_M\|_2 \leq W$, the normalized internal response-score map $\bar{B}_M(\Delta) = [B_M(\Delta) - \tau]/m_M$ has $L_B$-Lipschitz gradient on the segment from $0$ to $\Delta_M$, and*

$$\left| \frac{F_{\pi_s}(M) - \tau}{m_M} - \bar{B}_M(\Delta_M) \right| \leq \epsilon_{\text{out}}.$$

*Let $R_\Delta \geq \sup_M \|\Delta_M\|_2$ and $E = W\epsilon_{\text{path}} + \frac{1}{2}L_B R_\Delta^2 + \epsilon_{\text{out}}$. If*

$$G_{\mathcal{D}} := a\alpha_0 - 1 - (a + R_{\mathcal{D}})\eta - E > 0, \qquad G_{\mathcal{U}} := 1 - b\alpha_0 - (b + R_{\mathcal{U}})\eta - E > 0, \tag{B.27}$$

*then every derivative has $F_{\pi_s}(M) > \tau$, every non-derivative has $F_{\pi_s}(M) < \tau$, and*

$$\Gamma_{\text{BB}} \geq m_{\min}(G_{\mathcal{D}} + G_{\mathcal{U}}) > 0. \tag{B.28}$$

*Proof.* Taylor's theorem and the Lipschitz assumption give $\left|\bar{B}_M(\Delta_M) + 1 - \langle W_M, \Delta_M \rangle\right| \leq L_B R_\Delta^2/2$. Replacing $\Delta_M$ by $\Delta_s$ adds at most $W\epsilon_{\text{path}}$, and the response readout adds $\epsilon_{\text{out}}$. Thus the normalized observable margin differs from $-1 + \langle W_M, \Delta_s \rangle$ by at most $E$. Write $\Delta_s = c_\parallel v + r$, where $c_\parallel \in [\alpha_0 - \eta, \alpha_0 + \eta]$ is positive and $\|r\|_2 \leq \eta$. For a derivative, $\langle W_M, \Delta_s \rangle = a_M c_\parallel + \langle b_M, r \rangle \geq a(\alpha_0 - \eta) - R_{\mathcal{D}}\eta$; for a non-derivative the corresponding upper bound is $b(\alpha_0 + \eta) + R_{\mathcal{U}}\eta$. Equation (B.27) therefore puts their normalized observable margins above $G_{\mathcal{D}}$ and below $-G_{\mathcal{U}}$, respectively. Multiplying by $m_M \geq m_{\min}$ and taking the class infimum and supremum proves Eq. (B.28). □

This is a sufficient statement for a score model, a fixed query protocol, and checkpoints whose representation coordinates permit the assumed comparison. The requirement on $\epsilon_{\text{path}}$ over non-derived models and the full-response readout bound are independent assumptions; source-side loss and first-token KL alone do not supply them. For a multi-query detector, the proposition applies when its score and the stated bounds are defined for the corresponding query distribution. Across unmatched architectures, Proposition 1 remains the general criterion without a shared-coordinate gradient model.

### B.9 Finite-Query Verification

Condition on a fixed suspect model $M$, fingerprint protocol $\pi_s = (Q_s, g_s)$, and query distribution $Q_s$, all chosen independently of the test responses. For the $i$-th submitted query $z_i \sim Q_s$, let $Y_i \sim M(z_i)$ denote the returned text response, and let $S_i = g_s(z_i, Y_i) \in [0, 1]$ denote the resulting numerical score evaluated by the detector. The sample average score is $\widehat{F}_N(M) = \frac{1}{N}\sum_{i=1}^N S_i$. If $N$ queries and their response randomness yield independent bounded scores $S_i$ with average expectation $\frac{1}{N}\sum_{i=1}^N \mathbb{E}[S_i] = F_{\pi_s}(M)$, Hoeffding's inequality (Hoeffding, 1963) applied directly to the scores $S_i$ gives

$$\Pr\left(\left|\widehat{F}_N(M) - F_{\pi_s}(M)\right| \geq \eta\right) \leq 2\exp(-2N\eta^2). \tag{B.29}$$

Independent identically distributed queries are one sufficient sampling scheme. Equal-size direction strata also fit the independent bounded-score form when their average expectation defines $F_{\pi_s}$.

For unequal $N_k$ in Eq. (8), each observation score $S_{k,n} = g_k(M(x_k^{\text{fp}}, q_{k,n})) \in \{0, 1\}$ has weight $w_{k,n} = 1/(K_b N_k)$, so that $\widehat{F}(M) = \sum_{k=1}^{K_b}\sum_{n=1}^{N_k} w_{k,n} S_{k,n}$. The weighted Hoeffding bound replaces $N$ by

$$N_{\text{w}} = \frac{1}{\sum_{k,n} w_{k,n}^2} = \frac{K_b^2}{\sum_k 1/N_k}. \tag{B.30}$$

This expression follows from the specified weights and independence, rather than an assumed effective sample size. A fixed shared fingerprint does not by itself invalidate conditional independence; adaptive detector selection or response-dependent query selection requires different analysis.

If the population scores satisfy $F_{\pi_s}(M_d) \geq p_+$ and $F_{\pi_s}(M_u) \leq p_-$ with $\Delta = p_+ - p_- > 0$, the midpoint threshold $\tau = (p_+ + p_-)/2$ yields, for each fixed model in either class,

$$\Pr(\text{classification error} \mid M) \leq \exp(-N\Delta^2/2). \tag{B.31}$$

For a pre-specified finite pool of $M_{\text{pool}}$ suspect models, a union bound gives simultaneous score deviations at most $\sqrt{\log(2M_{\text{pool}}/\delta)/(2N)}$ with probability at least $1-\delta$. Dependence across models due to using the same queries does not invalidate that union bound. It still controls only the specified models: finite-pool extrema do not certify the population infimum and supremum over unseen models.

These statements require known population separation or valid independent calibration, together with a suitable query-sampling scheme; scores evaluated on reused queries or under post-hoc detector selection do not constitute such confirmation. Appendix G provides descriptive pool-level summaries; deployment decision rules and finite-query certificates are governed by the calibration protocol (Appendix F) and the concentration bounds in this section.

### B.10 Auxiliary Information-Theoretic Perspective

The following identities provide additional motivation, rather than the main justification for the construction. Let $R$ be a random source feature across a population of sources, and $Z$ an observation. For a joint law $P_1$ and the auxiliary independence null $P_0 = P_1(R)P_1(Z)$,

$$\mathrm{KL}(P_1 \| P_0) = \mathbb{E}_{P_1} \log \frac{p_1(R,Z)}{p_1(R)p_1(Z)} = I_{P_1}(R;Z). \tag{B.32}$$

The randomness of the source feature is essential to this interpretation. For one fixed source, a deterministic $R$ carries no mutual information. Moreover, independently trained model weights do not imply independent query responses: a query selected from the source can itself convey source-dependent information. The product-of-marginals null is therefore a statistical modeling choice, not a proved property of the non-derived pool.

For a scalar parameter $\vartheta$, suppose an internal statistic $S \sim p_\vartheta$ produces observation $Y$ through a parameter-independent channel $K(y \mid s)$. Under differentiation and integrability conditions, let $u_\vartheta(S) = \partial_\vartheta \log p_\vartheta(S)$. Differentiating $p_\vartheta(y) = \int K(y \mid s) p_\vartheta(s)\, ds$ gives

$$\partial_\vartheta \log p_\vartheta(Y) = \mathbb{E}[u_\vartheta(S) \mid Y]. \tag{B.33}$$

The conditional variance identity then yields

$$\mathcal{I}_S(\vartheta) - \mathcal{I}_Y(\vartheta) = \mathbb{E}[\mathrm{Var}(u_\vartheta(S) \mid Y)] \geq 0. \tag{B.34}$$

Strict loss requires positive conditional variance of the score; a many-to-one observation alone is insufficient. A behavioral projection need not be sufficient to generate the model's text, and derivation can change the downstream channel. Consequently this identity does not establish that SV supervision always outperforms output-target optimization. Both methods ultimately face the same black-box observation constraint.

## C Notation Reference

We collect here the LVLM inference notation and terminology used throughout the paper. An LVLM is denoted $f_\theta$ and takes as input an image $x$ together with a textual query $q$, producing a textual response $y = f_\theta(x, q)$. The vision encoder maps $x$ to a sequence of visual tokens that are projected and concatenated with the tokenized query in the residual stream of the language backbone. Let $L$ be the number of transformer layers in the language backbone and $d$ the residual-stream width. For an input pair $(x, q)$ and parameters $\theta$, we write $h_\ell(x, q; \theta) \in \mathbb{R}^d$ for the residual activation at layer $\ell \in \{1, \ldots, L\}$ taken at the last prompt-token position, i.e. the position from which the first response token is decoded.

We fix the following core terminology and symbols:

- **Target behavior ($b_k$):** The high-level semantic behavioral axis indexed by $k \in \{1, \ldots, K_b\}$, such as code rewriting (Py→JS), language switching (En→Zh), or benign refusal (Ans→Ref).
- **Steering vector / SV ($v_{k,\ell} \in \mathbb{R}^d$):** The unit-normalized mean-activation difference direction extracted at layer $\ell$ of the source model that isolates behavior $b_k$ in the residual stream; adding $\alpha v_{k,\ell}$ during decoding shifts $f_\theta$ along the target behavioral axis. $\mathcal{V}(\theta_0)$ denotes the source-model steering-vector bank across the retained layer band $\mathcal{L}^\star \subset \{1, \ldots, L\}$.

- **Carrier image** ($x^{\text{cle}}$)**:** The clean base image (e.g., a neutral gray image or natural image at native model resolution) upon which perturbations are added.
- **Image perturbation** ($\delta_k$)**:** The learned additive pixel perturbation optimized to externalize the steering vector $v_k$ into the visual input.
- **Visual fingerprint** ($x_k^{\text{fp}}$)**:** The complete optimized image $x_k^{\text{fp}} = \text{clip}_{[0,1]}(x^{\text{cle}} + \delta_k)$ associated with direction $b_k$. When written alone, "fingerprint" refers to $x_k^{\text{fp}}$.
- **Textual query / Prompt** ($q_{k,n}$)**:** A natural language input drawn from a query pool $\mathcal{Q}_k$ that does not explicitly command the target behavior.
- **Fingerprint query** ($z_{k,n}$)**:** The joint multimodal input pair $(x_k^{\text{fp}}, q_{k,n})$ submitted by the verifier to the suspect model.

We denote the source-model parameters by $\theta_0$ and an arbitrary suspect model's parameters by $\theta'$.

## D Implementation Details

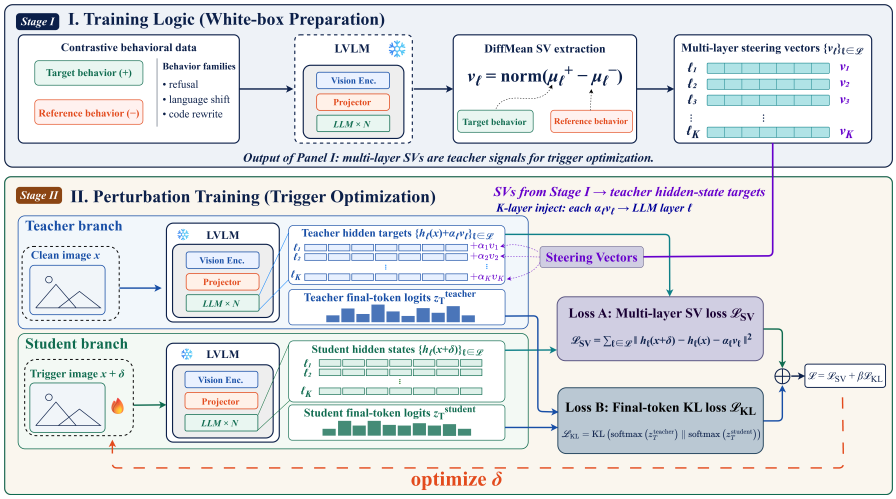



Figure 11: Overview of the white-box construction pipeline (corresponding to Stage I of Figure 1). Phase A extracts behavioral steering vectors from contrastive prompt pairs in the residual stream, while Phase B optimizes the visual fingerprint $\delta_k$ on carrier $x^{\text{cle}}$ via joint hidden-state displacement matching ($\mathcal{L}_{\text{sv}}$) and first-response-token distillation ($\mathcal{L}_{\text{kl}}$).

**Datasets.** **V7W**, **PaintingForm**, and **MathV360K** are subsampled to 28K, 20K, and 50K training examples; **ST-VQA** and **TextVQA** use the full training splits. All five datasets follow the original train/test partitioning of the respective benchmarks. The checkpoint-specificity study evaluates a panel of 27 non-derived models across diverse architectures (Section A.2).

**Post-release transformations.** Cross-architecture transformation experiments evaluate Qwen2-VL-2B (Wang et al., 2024) and LLaVA-1.5-7B (Liu et al., 2024). Quantization uses `bitsandbytes` at 8 and 4 bits (Dettmers et al., 2022; 2023). LoRA merging on Qwen2-VL-2B pairs V7W with other task adapters at equal ratios under Linear (Ilharco et al., 2023) and TIES (Yadav et al., 2023). Weight pruning uses Wanda (Sun et al., 2024) and SparseGPT (Frantar & Alistarh, 2023) at 10%, 20%, and 30% sparsity. Extended transformation studies on Qwen2.5-VL-7B are described in Appendix A.5.5.

**Visual-fingerprint construction.** Visual fingerprints are RGB/PNG images at the native input resolution of each source LVLM, standardized to an $\ell_\infty$ perturbation budget of $\epsilon = 16/255$ around the clean carrier image ($x_k^{\text{fp}} = \text{clip}_{[0,1]}(x^{\text{cle}} + \delta_k)$ with $\|\delta_k\|_\infty \leq 16/255$). During construction, source-model parameters are frozen and only the visual perturbation pixels $\delta_k$ are updated within this bounded range. Section A.7 reports the perturbation budget analysis. Optimization uses Adam at learning rate $5{\times}10^{-4}$, gradient clipping at $5{\times}10^{-4}$, and $\lambda_{\text{sv}}{=}1$. The SV injection strength, supervised prompt-end token count $K$, and layer window are selected during development (Section A.8).

**Parameter-noise augmentation.** To promote robust generalization across downstream weight modifications, SteerPrint incorporates parameter-noise augmentation during visual fingerprint optimization. Specifically, during each fingerprint forward pass, zero-mean Gaussian noise $\xi \sim \mathcal{N}(0, \sigma^2 I)$ with $\sigma = 0.01$ is added to the linear projection weights of the multimodal projector, refreshed independently at every optimization iteration. The teacher forward pass remains unperturbed to provide a stable distillation reference. This stochastic perturbation regularizes the input-side visual optimization, encouraging the learned fingerprint $x_k^{\mathrm{fp}}$ to reliably induce the source steering vector even under downstream model modifications.

**Baseline protocol.** For Ordinary, RNA, PLA, and AGDI, we follow each method's standard optimization recipe and evaluation protocol as established in prior literature (e.g., projected gradient descent (PGD) with $K{=}1000$ steps, step size $\eta{=}1/255$ for Ordinary; CLIP-aligned dual-injection for AGDI (Xia et al., 2026)). For bounded-perturbation baselines we sweep the $\ell_\infty$ perturbation budget $\epsilon \in \{4/255, 8/255, 16/255, 32/255\}$ and report each baseline at its best budget. Following standard evaluation protocols in AGDI (Xia et al., 2026), each baseline uses 5 predefined trigger pairs $\times$ 200 samples $= 1000$ queries per method, and verification evaluates the elicitation of target response phrases and semantic markers in the generated text. Unless otherwise stated, all evaluations use greedy decoding; Section A.8 reports SteerPrint stochastic-sampling checks over five runs with $T{=}0.7$, $p{=}0.9$.

Construction and verification cost, including runtime, peak memory, and query-count comparisons with the baselines, is reported in Appendix A.9.

# E SV-Extraction and Black-Box Query Data

This appendix specifies the per-direction corpora across the SteerPrint pipeline. For each target behavior $b_k$ (e.g., Python→JavaScript code rewriting, English→Chinese translation, or benign request refusal), data partition cleanly into three explicit stages:

1. **Stage 1: Steering-vector extraction pairs ($\mathcal{P}_k = \{(p_i^+, p_i^-)\}$).** Contrastive text pairs used exclusively offline on the protected source model to compute the internal representation direction $v_{k,\ell}$ as the normalized mean-activation difference. These contrastive pairs explicitly demonstrate opposite poles of the behavior (e.g., harmful vs. benign instructions, or language-specific completion prefixes) to isolate the target direction in the residual stream.
2. **Stage 2: Visual fingerprint training queries ($\mathcal{Q}_k^{\mathrm{tr}} = \{q_{k,n}^{\mathrm{tr}}\}_{n=1}^{N_k}$).** Prompts used during the optimization of the visual perturbation $\delta_k$ under the joint objective in Eq. (7), pairing the visual input with training prompts to align internal activations and output distributions.
3. **Stage 3: Black-box verification queries ($\mathcal{Q}_k^{\mathrm{te}} = \{q_{k,n}^{\mathrm{te}}\}_{n=1}^{N_k^{\mathrm{te}}}$).** Strictly held-out natural prompts paired with the frozen visual fingerprint $x_k^{\mathrm{fp}}$ and submitted to the suspect model during black-box auditing. Crucially, verification queries do not explicitly instruct the model to produce the target behavior (and in the refusal direction, they consist strictly of benign requests), serving as neutral conversational contexts to evaluate whether the visual fingerprint autonomously triggers the target behavior in generated text.

Across all three behavioral directions, these stages are strictly disjoint in samples and splits, ensuring that reported FTR is never measured on data seen during SV extraction or fingerprint optimization.

## E.1 Steering-Vector Extraction Pairs

For each direction $b_k$ we build a contrastive pair set $\mathcal{P}_k = \{(p_i^+, p_i^-)\}_i$. The steering vector at layer $\ell$ is the unit-normalized mean-activation difference at the last prompt token,

$$v_{k,\ell} = \mathrm{normalize}(\mu_{k,\ell}^+ - \mu_{k,\ell}^-),$$
$$\mu_{k,\ell}^{\pm} = \mathbf{E}_{p^{\pm}}\left[h_\ell(p^{\pm}; \theta)\right].$$

computed independently on the source base and on each variant that is later compared by cosine similarity. All prompts use the wrapper `USER: {prompt}\nASSISTANT:`.

**Py→JS (Python→JavaScript).** Source: items 1–200 of `greengerong/leetcode` (train split). Each item is parsed into a Python block `PY` and a JavaScript block `JS`; pairs use the half-prefix to force a non-trivial completion target:

- $p^+$: Python→JavaScript; source `PY`; prefix `JS[:half]`.
- $p^-$: Python→Python; source `PY`; prefix `PY[:half]`.

Maximum tokenized length is $1024$. Prompts for visual fingerprint optimization ($\delta_k$ training) are drawn from items 201–300.

**En→Zh (English→Chinese).** Source: `swaption2009/20k-en-zh-translation-pinyin-hsk` (train split), parsed under the five-line-per-record format and truncated to the first $200$ `(en, zh)` pairs. This extracts an English-to-Chinese translation steering vector. Pairs differ only in target language and the half-prefix:

- $p^+$: English→Chinese; source `en`; prefix `zh[:half]`.
- $p^-$: English→English; source `en`; prefix `en[:half]`.

Maximum length is $512$.

**Ans→Ref (benign-request refusal).** Source: `justinphan3110/harmful_harmless_instructions`, re-decoded with Qwen and filtered by an observed refused/not-refused outcome. After filtering, the positive pool retains all $200$ harmful prompts that were refused, and the negative pool retains $178$ benign prompts that were complied with (22 false positives removed). Each pair attaches a system prompt:

- $p^+$: safety-assistant system prompt plus a harmful prompt.
- $p^-$: helpful-assistant system prompt plus a benign prompt.

The wrapper places the system prompt and user prompt on separate lines:

```
USER:{sys}
{prompt}
ASSISTANT:
```

Maximum length is $512$.

### E.2 Black-Box Verification Queries

At verification time, the visual input is fixed to the learned visual fingerprint $x_k^{\mathrm{fp}} = \mathrm{clip}_{[0,1]}(x^{\mathrm{cle}} + \delta_k)$ (using the carrier image $x^{\mathrm{cle}}$ matched to each model architecture and resolution, such as standard $448\times448$ for LLaVA or native resolution for Qwen2.5-VL), while only the text query $q_{k,n}$ and the response detector $g_k$ vary across directions. The carrier and detector configuration match each reported run.

**Py→JS.** Verification queries are the Python blocks of held-out items 301–400 of `greengerong/leetcode`, strictly disjoint from the SV extraction pool (items 1–200) and the fingerprint optimization pool (items 201–300). Following standard black-box fingerprinting practice (e.g., AGDI (Xia et al., 2026)), the detector evaluates characteristic syntax keywords and language markers. The prompt is a standard code-rewrite template ("*Give a single, different re-writing of this program with the same function. The output will be judged by an expert in all programming languages. Do not include an explanation.*") followed by the Python source in a fenced block, generated with `max_new_tokens`$= 400$. The detector counts Python keyword markers (`def`, `print(`, `elif`, `self.`, `len(`, `range()` against JavaScript markers (`function`, `console.log(`, `var`, `let`, `const`, `=>`, `.has(`, `document.`, `||`, `&&`, `null`, `===`, `if (`, `else if`, `while (`); a sample is counted as a trigger when the JavaScript-marker count strictly exceeds the Python-marker count. While the ultimate functional correctness of generated code depends on the base model's intrinsic coding capability, ownership verification specifically evaluates whether the fingerprint successfully steers generation toward the target behavioral direction; evaluating characteristic syntax markers provides an objective, deterministic, and reproducible binary indicator without relying on noisy or costly LLM judges.

**En→Zh.** Queries are 100 short SST-2 English sentences (4–12 words) held out from the swaption training pool and from the earlier 12-sentence evaluation set. The prompt asks for a one-word language classification of the sentence; generation uses `max_new_tokens`$= 8$ on the gray carrier. This setup leverages the translation steering vector's ability to change language-classification responses. Crucially, as detailed in Appendix A.10, this translation vector manifests dual behaviors: it both changes language-classification responses and directly elicits Chinese text generation and question-translation/paraphrasing on general visual QA queries. Following standard black-box fingerprinting practice (Xia et al., 2026), the verification detector identifies characteristic language markers: the default CJK detector returns *chinese* if the first 300 characters of the output contain at least one CJK codepoint (U+4E00–U+9FFF), or if the first 80 characters contain the literal word *chinese* without *english*; otherwise it falls back to *english* or *mixed*. A strict variant additionally requires at least 3 CJK characters and a 3:1 ratio over Latin characters. A sample triggers when the detector returns *chinese*. Evaluating these objective language markers directly captures the induced behavioral steering with high sensitivity across diverse architectures.

**Ans→Ref.** Queries are 100 benign prompts drawn from the `test` split of `justinphan3110/harmful_harmless_instructions`, using a different split and a different sampling seed from the prompt pairs used at SV extraction. Verification is benign-only ($n_{\text{benign}} = 100$); no harmful prompts are queried at verification time. The query has no system preamble; optionally, the same safety-assistant system prompt used at SV extraction can be prepended, for a controlled overlap with the SV-extraction prompt format. Generation uses `max_new_tokens`$= 80$. The refusal detector is the union of an open-source refusal-template matcher and an extra regex set covering, among others, the following families of refusal templates:

- "unethical/illegal";
- "not allowed/should not/must not";
- "do not" followed by a harmful verb (*hijack*, *steal*, *attack*, *hack*, *exploit*, *create*, *provide*, *share*, *use*, *engage*);
- "please don't / please do not";
- "not acceptable / against the law";
- "the answer is no / don't / unethical / illegal";
- "this is / would be wrong / illegal / unethical / harmful / bad / inappropriate";
- "we should not / must not / cannot / refuse to / won't (do, help, answer, provide)".

We report two quantities, both computed on the same 100 benign prompts: the *benign refusal rate* (the fingerprint trigger signal reported as $\text{FTR}_{\text{refusal}}$ in the main text) and the fraction of benign outputs shorter than 10 words (a decoding-collapse diagnostic, not used in FTR aggregation). A sample is counted as a trigger when the refusal detector returns true, i.e. the benign prompt elicits a refusal-template response.

### E.3 Data Separation Across Pipeline Stages

Each direction enforces rigorous separation across all three data stages rather than relying on a single random split:

- **Py→JS:** Uses partitioned index splits from `greengerong/leetcode`, with items 1–200 reserved for SV extraction ($\mathcal{P}_k$), items 201–300 for visual fingerprint optimization ($\mathcal{Q}_k^{\text{tr}}$), and items 301–400 for black-box verification ($\mathcal{Q}_k^{\text{te}}$). The verification queries strictly exclude both prior index ranges, guaranteeing zero sample overlap with either SV extraction or fingerprint training.
- **En→Zh:** Employs cross-corpus dataset switching, extracting the translation steering vector from `swaption2009/20k-en-zh-translation-pinyin-hsk` and evaluating black-box verification on held-out English sentences from SST-2.
- **Ans→Ref:** Enforces domain and split separation: SV extraction utilizes contrastive harmful/harmless instruction pairs from the training partition of `justinphan3110/harmful_harmless_instructions`, while verification queries consist exclusively of 100 benign prompts drawn from the disjoint test split.

Consequently, no black-box verification query in $\mathcal{Q}_k^{\text{te}}$ overlaps with either the contrastive pairs $\mathcal{P}_k$ used to estimate the steering vector or the queries $\mathcal{Q}_k^{\text{tr}}$ used to train the visual fingerprint.

# F VERIFICATION PROTOCOL

At verification time the verifier holds the externalized fingerprint set $\{x_k^{\text{fp}}\}_{k=1}^{K_b}$ produced by the joint fingerprint objective in Eq. (7) together with the parameter-noise augmentation in Section 4.2, but has only black-box query access to the suspect model $f_{\theta'}$. For each behavioral direction $b_k$ we equip the fingerprint with the task-relevant query set $\mathcal{Q}_k = \{q_{k,n}\}_{n=1}^{N_k}$ and a target-behavior detector $g_k(\cdot) \in \{0, 1\}$. Each $g_k$ is a binary text detector over response surface forms—for example, the fraction of CJK characters in the response exceeding a threshold for a language-switch direction, the presence of JavaScript-specific syntax markers for a code-rewrite direction, or a refusal-template match on benign inputs for a refusal direction. A single fixed-string realization (e.g., the literal word `chinese`) is at most one instance that satisfies the indicator, not the indicator itself, so any response carrying the target behavior at the same surface level counts toward Eq. (10). The per-direction detectors used in our experiments are specified in Appendix E.2; in all cases $g_k$ is computed on the suspect's text output and does not require access to its activations.

**Per-direction and aggregate trigger rate.** For each direction $b_k$, the verifier submits each fingerprint–query pair $(x_k^{\text{fp}}, q_{k,n})$ with $n \in \{1, \ldots, N_k\}$ to $f_{\theta'}$ and records the binary outcome $g_k(f_{\theta'}(x_k^{\text{fp}}, q_{k,n}))$. The per-direction fingerprint trigger rate is the empirical mean of these outcomes,

$$\text{FTR}_k(\theta') \;=\; \tfrac{1}{N_k} \sum_{n=1}^{N_k} g_k\Big(f_{\theta'}(x_k^{\text{fp}}, q_{k,n})\Big), \tag{10}$$

and the aggregate trigger rate used for the ownership decision is its average over the $K_b$ behavioral directions,

$$\text{FTR}(\theta') \;=\; \tfrac{1}{K_b} \sum_{k=1}^{K_b} \text{FTR}_k(\theta'). \tag{11}$$

Aggregation by mean (rather than max) preserves evidence when one direction is degraded by a downstream modification, while direction-level reporting exposes that failure.

**Deployment decision protocol.** A deployed verifier should calibrate its threshold on held-out independently trained checkpoints $\mathcal{M}_{\text{cal}}$ that are disjoint from the reported hard-negative evaluation set. When evaluating $N$ fingerprint queries across $n_- = |\mathcal{M}_{\text{cal}}|$ non-owner checkpoints ($n_{\text{tot}} = n_- N$ total trials) under a binomial rate model where zero triggers are observed, the one-sided Clopper–Pearson upper confidence bound at level $\delta_{\text{conf}} = 0.05$ is

$$\widehat{p}_-^{\,U} = 1 - \delta_{\text{conf}}^{1/(n_- N)}. \tag{12}$$

One conservative protocol is

$$\tau \;=\; \max\!\left(\max_{M \in \mathcal{M}_{\text{cal}}} \text{FTR}(M),\;\; \widehat{p}_-^{\,U}\right), \tag{13}$$

and declares ownership only when $\text{FTR}(\theta') > \tau$. The strict inequality prevents the calibration model attaining the empirical maximum from being classified as owned. Our experiments report FTR and $\Delta$FTR to characterize separation independently of a deployment-specific calibration pool; a numerical ownership threshold must be calibrated for the intended model population. Formal pool-level descriptive statistics are given in Appendix G.

# G EVALUATION METRIC DEFINITIONS

This appendix gives the formal definitions of the FTR-based metrics summarized in Sec. 5.1. Throughout, $\mathcal{Q} = \{q_i\}_{i=1}^{N}$ is the fingerprint query set, $g(\cdot) \in \{0, 1\}$ is the target-behavior detector, $\mathcal{M}_{rel}$ is the related-model set (source model and its derivatives), and $\mathcal{M}_{unrel}$ is the unrelated-model set. The per-model FTR is given by Eq. (8).

**Pool-level statistics.** The related and unrelated pools are summarized by their means and worst cases:

$$\text{Rel-FTR} = \frac{1}{|\mathcal{M}_{rel}|} \sum_{M \in \mathcal{M}_{rel}} \text{FTR}(M), \tag{14}$$

$$\text{Min-Rel-FTR} = \min_{M \in \mathcal{M}_{rel}} \text{FTR}(M), \tag{15}$$

$$\text{Unrel-FTR} = \frac{1}{|\mathcal{M}_{unrel}|} \sum_{M \in \mathcal{M}_{unrel}} \text{FTR}(M), \tag{16}$$

$$\text{Max-Unrel-FTR} = \max_{M \in \mathcal{M}_{unrel}} \text{FTR}(M). \tag{17}$$

**Conservative margin.** The trigger margin is the worst-case gap between the two pools,

$$\text{Margin} = \text{Min-Rel-FTR} - \text{Max-Unrel-FTR}. \tag{18}$$

$\text{Margin} > 0$ is equivalent to the existence of a threshold $\tau$ with $\text{Max-Unrel-FTR} \leq \tau < \text{Min-Rel-FTR}$ under the strict decision rule $\text{FTR} > \tau$. When $\text{Max-Unrel-FTR} = 0$, the margin reduces to $\text{Min-Rel-FTR}$.

**Pass rate at threshold.** For a fixed threshold $\tau \in [0, 1]$, a model passes verification if $\text{FTR}(M) > \tau$. The pass rate on the related pool is

$$\text{Pass@}\tau = \frac{1}{|\mathcal{M}_{rel}|} \sum_{M \in \mathcal{M}_{rel}} \mathbf{1}[\text{FTR}(M) > \tau]\,. \tag{19}$$

With $N = 100$ queries and a requirement of at least 10 successes, the strict rule uses a threshold immediately below $0.10$; the count-based rule is not equivalent to the $\text{FTR} > 0.10$ decision.

**Remark.** The margin and pass-rate definitions above are descriptive summaries of the observed pool separation. Converting them to formal error guarantees (e.g., bounding the probability that a non-derived model exceeds $\tau$) requires specifying the model population, verifying query independence, and calibrating the threshold on a held-out pool. We report FTR and $\Delta$FTR to characterize empirical separation; deployment-level calibration and decision rules are detailed in Section 4.3 and Appendix F, while finite-query confidence guarantees under explicit sampling models are established in Appendix B.9.