Do MLLMs Know When Visual Evidence Is Missing? Probing Visual-Support States in Unsupported VQA

Anonymous ACL submission

Anonymous code repository:
https://anonymous.4open.science/r/Do-MLLMs-Know-When-Visual-Evidence-Is-Missing--6CBD/

Abstract

Multimodal Large Language Models (MLLMs) frequently produce concrete answers to visual questions even when the paired image provides no supporting evidence. We formulate this as an evidence–behavior alignment problem admitting two explanations: the model may fail to represent the absence of evidence (*representational failure*), or it may represent this absence yet fail to act on it during answer selection (*coupling failure*). Using a matched-evidence protocol that holds the question and answer options fixed while varying only the visual context, we find evidence for the latter. Pre-answer hidden states encode a linearly readable **visual-support state** (VSS) that separates supported from unsupported answers across four open-weight MLLMs. Targeted controls indicate that this signal reflects image–question composition rather than image identity and captures graded evidence adequacy rather than binary correctness. Cross-dataset transfer shows that the signal generalizes beyond source-specific artifacts, though operating thresholds remain domain-sensitive. A diagnostic intervention routing VSS into the abstention decision yields positive but model-dependent leverage, reducing unsupported answering by 6.3–15.8 percentage points. These findings characterize unsupported visual answering as a gap between evidence encoding and answer selection, suggesting that improved abstention may be achievable by strengthening this link rather than retraining the visual encoder.

1 Introduction

A reliable Multimodal Large Language Model (MLLM) must not only answer when evidence is present but also abstain when it is missing (Geifman and El-Yaniv, 2017; Kamath et al., 2020). In document understanding, chart analysis, and safety-critical inspection, an unsupported concrete answer is not merely a wrong Visual Question Answering (VQA) prediction—it asserts a visual fact that the input image does not warrant (Liu et al., 2024a; Bai et al., 2024). Users naturally assume that a concrete answer reflects what the image shows; when that assumption fails, the cost is misplaced trust in a fabricated visual claim.

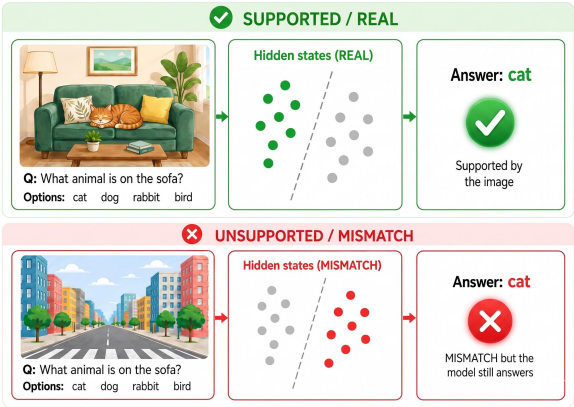

Figure 1: An MLLM may encode that the image does not support a candidate answer, yet still output a concrete option. We test this gap between the model's internal evidence representation and its abstention behavior.

When a model produces a concrete answer despite missing visual evidence, two fundamentally different failures may be responsible. **Representational failure** occurs when the model never forms an internal state distinguishing supported from unsupported evidence (Tong et al., 2024). **Coupling failure** occurs when the model forms such a state, but the answer selection process does not act on it (Burns et al., 2023; Azaria and Mitchell, 2023). The distinction carries practical implications: representational failure calls for improved visual grounding during pretraining or instruction tuning; coupling failure suggests that the requisite signal already exists and need only be connected to the abstention decision.

We address this question through a **matched-evidence protocol** that holds the question and

answer options fixed while varying only the visual context, isolating the effect of visual evidence on the model's internal state and behavior. Across four open-weight MLLMs, pre-answer hidden states contain a linearly readable **visual-support state** (VSS) that separates supported from unsupported concrete answers (Alain and Bengio, 2017; Belinkov, 2022). However, linear separability alone does not establish that VSS reflects genuine evidence adequacy—the classifier could succeed through shortcuts (Hewitt and Liang, 2019). We therefore organize the experiments to address alternative explanations before testing behavioral relevance.

A natural concern is that the VSS gate exploits shortcuts rather than genuine evidence signals. Same-image controls address this by holding the image fixed while varying the question: Layer 0 performs at chance (AUC = 0.5), while later layers achieve strong separation, indicating that the signal emerges from image–question composition rather than image identity. Quadrant analysis further reveals that gate scores form a graded evidence-adequacy continuum rather than functioning as a binary correctness detector. External compatibility baselines confirm that generic CLIP or SigLIP image–text matching does not account for the signal.

Beyond shortcut controls, the signal must generalize to be meaningful. Cross-dataset experiments show that a VSS gate trained on one source often preserves useful rankings on another, although operating thresholds remain domain-sensitive, suggesting that VSS captures properties of evidence adequacy beyond dataset-specific artifacts.

The critical test, however, is whether VSS can influence behavior. We introduce **Latent Evidence Alignment** (LEA), a diagnostic intervention that adjusts only the abstention option's score according to the gate score. LEA does not retrain the model or modify concrete-answer scores; it routes the existing VSS signal into the abstention decision. Across all four models, this intervention reduces unsupported answering by 6.3–15.8 percentage points, with **positive but model-dependent leverage**.

Together, these results support a specific characterization: unsupported visual answering frequently reflects **a gap between evidence encoding and answer selection**. The internal signal for missing evidence exists; the difficulty is that it does not reliably influence the final answer, identifying a concrete target for improving MLLM reliability—strengthening the evidence–behavior link—without retraining the visual encoder or language backbone.

Our contributions are:

- We formulate unsupported visual answering as an evidence–behavior alignment problem, distinguishing representational failure from coupling failure.
- We introduce a matched-evidence protocol that changes only the visual evidence while keeping the question and answer options fixed.
- We show that a pre-answer visual-support signal is linearly readable across four MLLMs and is not well explained by image identity, correctness detection, or generic image–text compatibility.
- We test whether this signal is behaviorally actionable via LEA, a diagnostic intervention that adjusts only the abstention logit, and find positive but model-dependent leverage.

2 Related Work

Multimodal language models. Recent MLLMs connect visual encoders with instruction-tuned language backbones to support image-conditioned dialogue and visual reasoning (Alayrac et al., 2022; Li et al., 2023a; Dai et al., 2023). LLaVA (Liu et al., 2023), Qwen-VL (Bai et al., 2023; Team, 2025), and InternVL (Chen et al., 2024) are representative open-weight systems. Their strong VQA performance makes unsupported answering especially consequential: when a capable model produces a concrete answer, users are likely to assume the image justified it.

Visual hallucination and unsupported generation. Object hallucination and visually unsupported generation have been studied in image captioning and VQA (Rohrbach et al., 2018; Li et al., 2023c; Guan et al., 2024; Liu et al., 2024b; Tong et al., 2024). Benchmarks such as VQAv2 (Antol et al., 2015; Goyal et al., 2017), GQA (Hudson and Manning, 2019), POPE (Li et al., 2023c), and MME (Fu et al., 2024) expose failures in recognition, grounding, and visual reasoning. Mitigation methods—including visual contrastive decoding (Leng et al., 2024), post-hoc correction (Yin et al., 2023), hallucination-aware prompting (Zhou et al., 2024), and over-trust penalty (Huang et al., 2024)—reduce some unsupported outputs. Our focus is

complementary: we ask whether VSS is already present in the model before the answer is decoded.

Selective prediction and calibration. Selective prediction studies when a model should abstain rather than answer (Geifman and El-Yaniv, 2017; Kamath et al., 2020); calibration studies whether confidence reflects empirical reliability (Guo et al., 2017). Related work on uncertainty in language models shows that confidence and verbalized uncertainty can be informative but imperfect (Jiang et al., 2021; Kadavath et al., 2022; Lin et al., 2022; Kuhn et al., 2023). In our setting, confidence alone is insufficient: a model can be confident in a language-prior answer even when the image provides no support.

Positioning. Our goal is not to introduce another hallucination score but to test a diagnostic distinction: whether unsupported answers arise because the model lacks an evidence signal or because the answer selection process fails to act on one. Whereas hallucination benchmarks primarily measure unsupported outputs and selective-prediction methods often rely on confidence, we directly probe the pre-answer hidden state for VSS (Alain and Bengio, 2017; Conneau et al., 2018; Tenney et al., 2019). This framing extends beyond probing studies that stop at separability: we complement the analysis with targeted controls, cross-dataset transfer, and a behavioral intervention that tests whether VSS can be routed into abstention (Li et al., 2023b; Zou et al., 2023).

3 Task and Evidence Conditions

We introduce three key terms. **Visual-support state** (VSS) refers to the model's internal representation of whether the image supports the answer. **Gate score** is the classifier-estimated probability that the current hidden state corresponds to an unsupported answer. **Latent Evidence Alignment** (LEA) is a diagnostic intervention that adds this gate score to the abstention option's score.

Our diagnostic task is a multiple-choice VQA problem where the linguistic content (question and answer options) remains fixed while only the visual evidence changes. This design keeps the classifier interpretable: if a VSS gate trained on hidden states separates supported from unsupported cases, it is unlikely to rely on the question, answer options, or dataset format alone.

Task definition. An item consists of a question q, four concrete options $\mathcal{A} = \{A, B, C, D\}$, an explicit abstention option E, a gold answer $a^\star \in \mathcal{A}$, and an image x. The model outputs an option $y \in \mathcal{A} \cup \{E\}$. The explicit abstention option makes unsupported answering measurable without free-form normalization.

Matched evidence conditions. Each item is evaluated under four visual contexts (Figure 2). REAL pairs q with its original image x. TEXT-ONLY removes the image entirely, measuring reliance on language priors. BLANK supplies a blank image, controlling for the mere presence of visual tokens. MISMATCH pairs q with a different image x' sampled from the same dataset (e.g., another GQA image for a GQA question), making the mismatch visually plausible. Image-content hashing prevents leakage across evaluation boundaries.

Visual-support contrast sets. Let $h_i^c \in \mathbb{R}^d$ be the hidden state for item i under condition c, extracted at the final prompt position immediately before answer generation. We construct two contrast sets for training the VSS gate:

$$\begin{aligned} \mathcal{S} &= \{h_i^{\text{Real}} : y_i^{\text{Real}} = a_i^\star\}, \\ \mathcal{U} &= \{h_i^{\text{Mismatch}} : y_i^{\text{Mismatch}} \in \mathcal{A}\}. \end{aligned} \tag{1}$$

That is, $\mathcal{S}$ collects hidden states from supported cases (correct image, correct answer), while $\mathcal{U}$ collects states from unsupported cases (mismatched image, concrete answer). Native abstentions under MISMATCH are excluded because they already exhibit the desired behavior; incorrect REAL answers are excluded because they conflate visual-support failure with ordinary task errors. This construction yields a clean contrast for testing whether VSS is linearly readable before answer generation.

Behavioral metrics. The main behavioral quantity is the unsupported-answer rate (UAR):

$$\text{UAR} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(y_i^{\text{Mismatch}} \in \mathcal{A}). \tag{2}$$

This counts the fraction of mismatched-image cases where the model picks a concrete option instead of abstaining. We also report the real-answer drop (RAD), the decrease in supported-answer accuracy caused by an intervention. These two metrics form the main behavior tradeoff: a useful intervention should reduce unsupported answers without suppressing supported ones. Appendix C gives the full metric table.

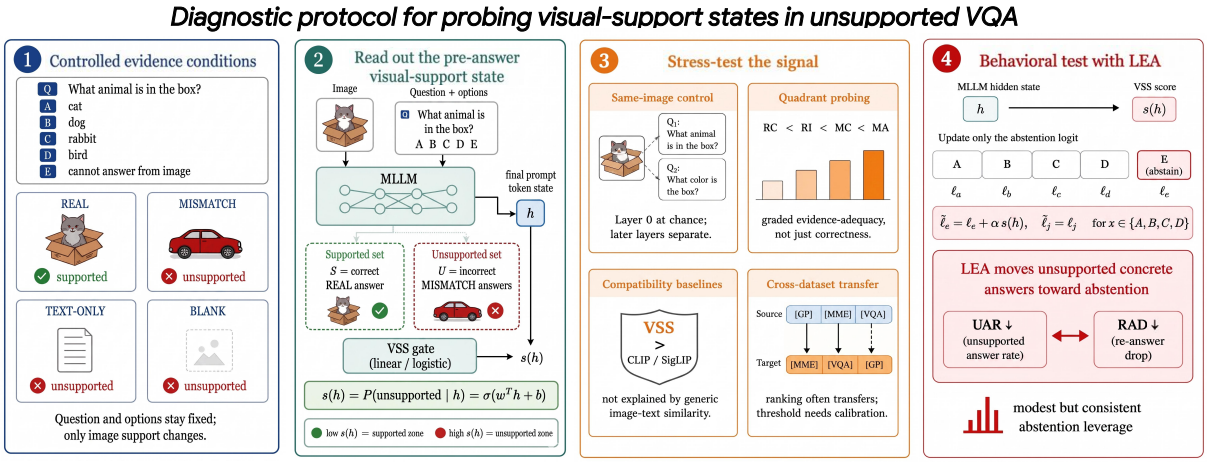

Figure 2: Diagnostic protocol. Each question is evaluated under matched evidence conditions (REAL, TEXT-ONLY, BLANK, MISMATCH) with fixed answer options. Pre-answer hidden states from supported and unsupported concrete answers are used to train a VSS gate; controls and cross-dataset transfer test whether VSS is robust; LEA tests whether it can steer abstention.

Metric families. Readout metrics (probe AUC, same-image AUC) measure whether VSS is separable in hidden states; AUC (area under the receiver operating characteristic curve) quantifies ranking quality, where 0.5 indicates chance and 1.0 indicates perfect separation. Native-behavior metrics (UAR) measure what the unmodified model does under mismatched evidence. Intervention metrics (UAR reduction, RAD) measure how much LEA or a baseline changes behavior. Higher AUC and UAR reduction indicate better separability or intervention effect; lower UAR and RAD indicate better behavior and lower cost.

Latent Evidence Alignment. To test whether VSS is behaviorally useful, we train a logistic VSS gate on standardized hidden states:

$$s(h) = P(z = 1 \mid h) = \sigma(w^\top h + b), \quad (3)$$

where $z = 1$ denotes the unsupported class and σ is the logistic sigmoid. The gate outputs a score between 0 and 1: higher scores indicate the model is likely answering without sufficient visual support. LEA uses this gate score to adjust only the abstention option's score:

$$\tilde{\ell}_E = \ell_E + \alpha s(h), \quad \tilde{\ell}_a = \ell_a \quad \forall a \in \mathcal{A}. \quad (4)$$

When the gate detects weak visual support (high $s(h)$), the abstention option receives a boost of $\alpha s(h)$, making it more likely that the model selects "cannot determine." Concrete-option scores remain unchanged, so any behavioral shift is attributable solely to the gate signal. The parameter α controls intervention strength. We measure the supported-answer cost with

$$\mathrm{RAD} = \mathrm{Acc}_{\mathrm{Real}}(\text{native}) - \mathrm{Acc}_{\mathrm{Real}}(\text{intervened}), \quad (5)$$

which captures how much correct answering is sacrificed. The minimal design of LEA keeps interpretation straightforward: if unsupported-answer reduction improves at limited supported-answer cost, the VSS gate carries useful information about missing support.

4 Experimental Setup

We evaluate whether VSS is readable, robust to shortcuts, transferable across data sources, and useful for changing behavior. Appendix F gives additional protocol details and hyperparameters.

Models. We evaluate four open-weight MLLMs: Qwen2.5-VL-7B (Team, 2025), LLaVA-1.5-7B (Liu et al., 2023), InternVL3-2B, and InternVL3-8B (Chen et al., 2024). These span three architectural families and two InternVL scales, providing diversity in vision–language integration while ensuring full reproducibility. All four models support interleaved image–text input and have been instruction-tuned for visual question answering. For all models, we score the next-token logits of the option letters A–E at the prompt end and select the highest-probability option; no stochastic sampling is used.

Prompt format. Each prompt presents the question, four concrete options, and an explicit abstention option:

```
Question: [question]
A. [option A]
B. [option B]
C. [option C]
D. [option D]
E. cannot determine from the provided
image
Answer with the option letter only.
```

The explicit option E lets us distinguish unsupported concrete answering from abstention without free-form text normalization.

Datasets and splits. The evaluation pool combines a cleaned GQA/POPE slice, MME, and a filtered VQAv2 multiple-choice slice. GQA/POPE contributes spatial and object-presence questions with scene-graph-derived distractors; MME contributes existence, count, and color questions designed for MLLM evaluation; VQAv2 contributes broader natural-image questions with human-annotated answer distributions. We construct a mixed split of 960 training and 240 test items, each manually verified for unambiguous gold answers and absence of incidental support in mismatched images; the dataset is compact by design because the matched-evidence protocol requires per-item human verification, prioritizing annotation quality over scale. All splitting and grouped validation use image-content hashes rather than filenames to prevent data leakage from repeated images.

Hidden-state extraction. We extract hidden states at the final prompt position—immediately before the model generates its answer—after the model has processed the image, question, and answer options. We retain the last few language-model layers (e.g., layers 18–22 for a 24-layer model) and use these states for probe training and LEA analyses.

Gate training. Unless otherwise stated, VSS gates use pre-answer hidden states from late language-model layers. Features are standardized inside an sklearn pipeline and classified with L2-regularized logistic regression using class-balanced weights. Probe AUC is estimated via five-fold cross-validation with GroupKFold grouped by image-content hash, ensuring that states derived from the same image cannot appear in both train and validation folds. The VSS gate used for LEA is fit on the training split; intervention strength α and baseline thresholds are selected on a validation slice and evaluated once on the held-out test split.

Baselines. We compare LEA with confidence thresholding, entropy thresholding, prompted abstention, random gates, and external compatibility baselines (CLIP ViT-L/14 (Radford et al., 2021) and SigLIP ViT-SO400M (Zhai et al., 2023)). Confidence and entropy baselines abstain when the model's output statistics cross a validation-selected cutoff; prompted abstention adds an instruction encouraging option E; compatibility baselines score image–question similarity outside the MLLM. All thresholds are validation-selected under the same protocol as LEA to ensure fair comparison.

5 Results

The results are organized around four questions that progressively build the evidence chain for the main claim: unsupported answering often reflects a gap between VSS and answer selection. Each subsection addresses one alternative explanation: absence of signal (§5.1), shortcut-based signal (§5.2), dataset-specific signal (§5.3), and non-actionable signal (§5.4). Throughout, AUC measures signal separability in hidden states, while UAR/RAD measure behavioral outcomes.

5.1 Do MLLMs encode visual support before answering?

We first ask whether any pre-answer visual-support signal exists. Table 1 shows that a linear gate separates supported REAL answers from unsupported concrete MISMATCH answers across all four models. The probe AUC values (0.698–0.760) are moderate but consistent, establishing the existence of linearly readable VSS in the pre-answer hidden state. The key pattern is dissociation: native unsupported-answer rates can be high while pre-answer states still carry readable VSS. This is most striking for LLaVA-1.5-7B, whose native UAR is 0.892 while its probe AUC remains 0.745. High unsupported-answer rates therefore need not imply the absence of VSS; they may instead reflect a gap between the encoded signal and the behavior selected by answer generation. This dissociation is the defining signature of coupling failure: the model encodes evidence adequacy but does not route it to the final answer.

The visual-dependence gap (VCD gap, denoted d) provides a complementary output-space view, measuring how much the model's log probability for its chosen answer changes when the correct image is restored. Supported answers exhibit sub-

Table 1: Core evidence for the visual-support state (VSS). Probe AUC measures how well the VSS gate separates supported from unsupported hidden states; VCD gap (visual-dependence gap) measures how much the model's answer probability changes when the correct image is restored; same-image AUC tests whether the signal emerges from image–question composition rather than image identity alone; native UAR measures unmodified behavior (lower is better). The central pattern is dissociation: all models carry readable pre-answer support information while native unsupported-answer rates remain substantial.

Model	Probe AUC ↑	VCD gap d ↑	Native UAR ↓	Same-image AUC (L0 → best)
Qwen2.5-VL-7B	0.698	1.322	**0.500**	0.500 → 0.808
LLaVA-1.5-7B	0.745	0.889	0.892	0.500 → 0.813
InternVL3-2B	**0.760**	**1.384**	0.546	0.500 → **0.866**
InternVL3-8B	0.705	1.233	0.621	0.500 → 0.785

stantially larger visual dependence (d ranging from 0.889 to 1.384) than unsupported concrete answers, confirming that the two classes differ in their reliance on the correct image.

Figure 3 shows the same-image layer sweep: the signal is absent at Layer 0 and emerges only in later layers where the model integrates visual and linguistic information, suggesting that VSS arises from the composition of these two modalities rather than from either alone. Together with the VCD gap, these results support the first part of our claim: VSS is often linearly readable before answer generation even when the model still produces a concrete answer.

5.2 Could the VSS signal be explained by shortcuts?

A readable VSS gate could succeed for the wrong reasons. We evaluate three shortcut explanations and find evidence against each: (1) the gate recognizes the image rather than the support relation; (2) the gate detects answer correctness rather than evidence adequacy; (3) generic image–text similarity accounts for the signal. Each control targets one of these alternatives.

Same-image control. The same-image control fixes the image and varies the question. If the gate primarily recognized image identity, separability should already appear at the input layer. Instead, Figure 3 shows Layer 0 at chance (AUC = 0.5) for every model, with separation emerging only in later layers after image–question composition. This pattern makes a simple image-identity account unlikely. Appendix Table 6 provides full layer-wise values.

Quadrant analysis. The signal is also not well characterized as a binary correctness detector. Gate scores form a graded evidence-adequacy continuum: REAL correct (RC) receives the lowest scores,

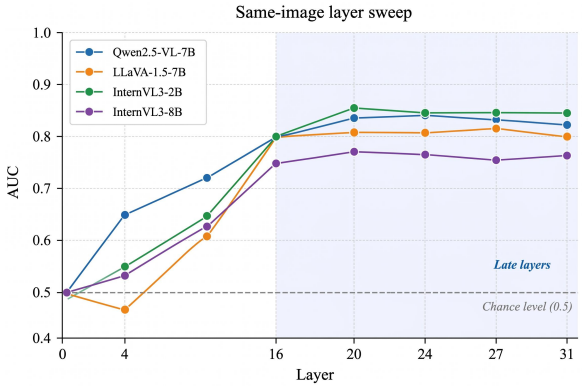

Figure 3: Same-image layer sweep. With the image fixed, the signal is absent at Layer 0 and emerges only in later layers, arguing against a simple image-identity shortcut.

Table 2: Quadrant probing. Scores form a graded evidence-adequacy continuum (RC < RI < MC < MA) rather than a binary correctness detector. RC/RI denote REAL correct/incorrect; MC/MA denote MISMATCH concrete/abstain.

Model	RC	RI	MC	MA
Qwen2.5-VL-7B	0.313	0.528	0.584	0.820
LLaVA-1.5-7B	0.409	0.634	0.732	0.962
InternVL3-2B	0.321	0.607	0.688	0.906
InternVL3-8B	0.393	0.591	0.668	0.891

REAL incorrect (RI) higher, concrete MISMATCH (MC) higher still, and MISMATCH abstentions (MA) highest (Table 2). For example, InternVL3-2B progresses from 0.321 (RC) to 0.607 (RI), 0.688 (MC), and 0.906 (MA). Native MISMATCH abstentions are not training positives, but their high gate scores are expected: they represent cases where the model correctly identifies missing support.

Compatibility baselines. The signal is not accounted for by generic image–text similarity. We compare VSS against CLIP-Q and SigLIP-Q, which score image–question compatibility using external encoders. Under the selective-prediction

protocol, the VSS gate achieves 0.875–0.902 AUROC across models, while CLIP-Q stays below 0.34 and SigLIP-Q below 0.51. Efficiency (UAR reduction per unit RAD) ranges from 9.63 to 12.65 for VSS compared with at most 1.63 for the compatibility baselines. The MLLM hidden state carries answer- and context-conditioned information that external encoders do not capture.

5.3 Does VSS survive dataset shift?

Dataset shift provides a stricter test. If VSS captured only source-specific quirks, a gate trained on one source should fail on another. We train on one source and test on another across GQA/POPE (GP), MME, and VQAv2 (VQA); the full transfer matrix appears in Appendix Figure 6.

The three datasets differ substantially in question types and image domains. GQA/POPE focuses on object presence and spatial relations in natural scenes; MME tests existence, counting, and color recognition; VQAv2 covers diverse open-domain questions with human-annotated answer distributions. If VSS primarily reflected dataset-specific artifacts, cross-dataset transfer should degrade to chance.

Instead, positive off-diagonal cells in the transfer matrix show that gates trained on one source often reduce unsupported answering on another. Some transfers are strong—for example, a LLaVA-1.5-7B gate trained on VQA substantially reduces unsupported MME answers—but operating costs vary across source–target pairs. We interpret this as evidence for a portable VSS ranking rather than a deployable fixed threshold: the ranking remains informative, but the score scale and operating point remain domain-sensitive.

5.4 Can VSS steer abstention behavior?

The preceding analyses establish that VSS exists and transfers; we now test whether it can affect behavior. Because LEA modifies only the abstention option's score, it provides a controlled test of whether the gate score can shift behavior toward abstention without altering concrete-answer rankings. At a shared operating point ($\alpha = 1.0$), LEA reduces unsupported concrete answers for all four models while incurring model-dependent costs to supported-answer accuracy (Table 3).

Positive reduction across all models provides behavioral evidence that VSS carries useful information. LLaVA-1.5-7B achieves the largest reduction with minimal RAD, suggesting many of its unsupported answers lie close to the decision boundary. InternVL3-2B exhibits a similarly favorable tradeoff. Qwen2.5-VL-7B and InternVL3-8B show smaller gains, indicating tighter coupling between the model's concrete-answer preference and its output scores. This variation illustrates that encoding the signal and using it for abstention are separate problems that different architectures couple to different degrees.

Table 3: Shared-α LEA at $\alpha = 1.0$. UAR reduction is the decrease in unsupported concrete answers (higher is better); RAD is the drop in supported REAL accuracy (lower is better).

Model	UAR red. ↑	RAD ↓
Qwen2.5-VL-7B	0.071	0.008
LLaVA-1.5-7B	0.146	**0.000**
InternVL3-2B	**0.158**	0.017
InternVL3-8B	0.063	0.013

Figure 4 places this operating point in the broader behavior tradeoff. Compared with confidence, entropy, prompting, compatibility scores, and random gates, LEA consistently offers a more favorable low-RAD curve because it leverages the hidden-state VSS rather than relying on generic output-distribution statistics or external compatibility scores.

Behavioral interpretation. The qualitative patterns behind the tradeoff curves are consistent with LEA using VSS rather than acting as a blanket refusal mechanism. Supported REAL examples with low gate scores retain their concrete answers, while MISMATCH, BLANK, and TEXT-ONLY cases with high gate scores often shift to abstention. Cases that do not flip are equally informative: some unsupported examples receive low gate scores, and others have high gate scores but require a stronger boost to overcome the model's concrete-answer preference.

Taken together, these results support the main diagnostic picture: VSS is readable, survives shortcut controls, often transfers across datasets, and can be used to reduce unsupported answering.

6 Additional Validation and Scope

6.1 Robustness across layers, probes, and sample budgets

We conduct focused InternVL3-8B checks to test whether VSS readability depends on a particular layer, probe type, or sample budget. A single Layer-20 probe reaches AUC 0.883; adding nearby late

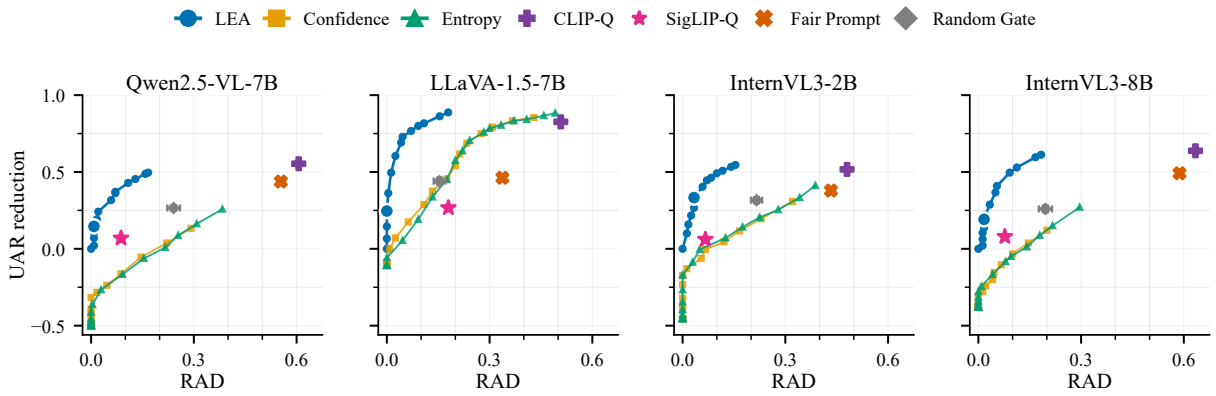

Figure 4: Tradeoff between real-answer drop (RAD) and unsupported-answer reduction for LEA and behavioral baselines. Curves closer to the upper-left are better: they reduce unsupported answers while preserving supported REAL accuracy. LEA uses the hidden-state VSS and often gives a more favorable low-cost abstention tradeoff than confidence, entropy, compatibility scores, prompting, or random gates.

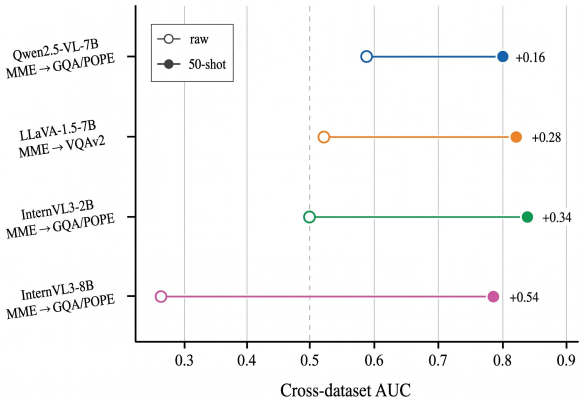

Figure 5: Few-shot calibration recovery. Weak raw transfers (left endpoints) often recover after adding 50 target-domain examples (right endpoints), suggesting score-scale mismatch rather than loss of the VSS signal.

layers does not improve performance; and training with only 50 or 200 examples reaches AUC 0.843 $\pm$ 0.043 and 0.889 $\pm$ 0.026, respectively. Logistic regression, RBF-SVM, and random forest probes all recover comparable signals, indicating that the readout is not a brittle artifact of one layer, classifier, or sample size. This consistency suggests that VSS occupies a linearly accessible subspace rather than requiring a specific extraction recipe. Full ablations appear in Appendix E.

6.2 Few-shot calibration recovers weak transfer

The transfer results should be read as ranking evidence, not as a universal-threshold claim. Figure 5 shows that weak raw transfers improve after adding 50 target-domain examples, suggesting score-scale mismatch rather than loss of the underlying VSS signal. This pattern holds across models: even when raw transfer AUC drops, the relative ordering by gate score remains informative. VSS therefore provides a reusable ordering of examples by missing-support risk, not a plug-and-play decision rule. Practitioners deploying VSS-based abstention should expect to calibrate thresholds on a small target-domain sample.

6.3 Scope of the controlled setting

The multiple-choice format with an explicit abstention option bounds the claim but also enables it: it fixes the answer space, makes abstention directly measurable, and avoids free-form normalization. Open-ended generation introduces partial support and ambiguity that remain outside this controlled study.

7 Conclusion

This work formulated unsupported visual answering as an evidence–behavior alignment problem. A matched-evidence protocol reveals that pre-answer hidden states encode a linearly readable visual-support state (VSS) reflecting image–question composition and graded evidence adequacy across four open-weight MLLMs. Targeted controls confirm that this signal is not explained by image identity, correctness detection, or generic compatibility, and cross-dataset transfer shows it generalizes beyond source-specific artifacts. Routing VSS into the abstention decision yields positive but model-dependent leverage, reducing unsupported answering by 6.3–15.8 percentage points. These findings characterize unsupported visual answering as a gap between evidence encoding and answer selection—a gap addressable by strengthening this link rather than retraining the visual encoder or language backbone.

Limitations

Controlled answer space. The multiple-choice format provides a controlled foundation but does not cover open-ended generation, where partial support and ambiguity introduce additional challenges. Extending the protocol to free-form settings would test whether VSS generalizes beyond fixed answer spaces.

Linear gate. A linear gate establishes that VSS is accessible but does not explain how the model computes it or where in the forward pass it originates. Nonlinear probes or causal interventions could reveal finer-grained structure and clarify the mechanistic basis of VSS.

Diagnostic intervention. LEA demonstrates that VSS can influence behavior but requires a held-out calibration set. Training-time objectives that directly optimize the evidence–behavior link represent a natural next step for closing the gap more robustly.

Ethical Considerations

This work aims to improve MLLM reliability by diagnosing when models answer without sufficient visual evidence. The datasets used are publicly available VQA benchmarks containing no personally identifiable information. Understanding when models lack visual grounding may help prevent misinformation in downstream applications.

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. *International Journal of Computer Vision*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *ICLR*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CVPR*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. *ACL*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *NeurIPS*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *ICML*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *EMNLP-IJCNLP*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *ACL*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *ICLR*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. *NeurIPS*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *NeurIPS*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024b. Mmbench: Is your multi-modal model an all-around player? In *ECCV*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

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

Qwen Team. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *ACL*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *CVPR*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *ICCV*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *ICLR*.

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, and 2 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Cross-Dataset Transfer

The appendix expands the main-paper evidence. Unless stated otherwise, reported numbers use the same mixed train/test split and supported-versus-unsupported contrast defined in Section 3. When an ablation uses a different diagnostic setting, we describe it before interpreting the result.

Anonymous code repository.
The anonymized repository for reviewing the implementation and released experiment materials is available at
anonymous.4open.science/r/Do-MLLMs-Know-When-Visual-Evidence-Is-Missing-6CBD/.

The transfer experiment tests whether VSS is merely a dataset-specific pattern. For each source–target pair, a gate trained on one source is evaluated on another. Positive off-diagonal reductions indicate that the VSS ranking often remains informative beyond the training source.

The result should not be read as evidence for a universal abstention threshold. Several transfers reduce unsupported answering but incur different supported-answer costs, especially when the target has a different question distribution. This is why the main text describes VSS as a portable ranking signal whose operating point remains domain-sensitive.

B Baseline Comparison Details

Table 4 expands the compatibility-baseline comparison. This table uses a selective-prediction protocol: each method provides a score for deciding when to abstain, and the operating point is selected on validation data. The VSS row is a selective-prediction ranking score, not the mixed-split probe AUC in Table 1. Table 1 asks whether supported and unsupported states are linearly separable; Table 4 asks whether a score yields a useful abstention tradeoff.

The behavioral baselines play a different role from CLIP-Q and SigLIP-Q. Confidence and entropy abstain based on the model's own answer distribution rather than external compatibility. Prompted abstention changes the instruction to encourage option E. Their AUROC and AUPRC entries are left undefined, while RAD, UAR reduction, and efficiency report behavior at the selected operating point. The selected confidence thresholds are 0.95, 0.75, 0.95, and 0.95 for Qwen2.5-VL-7B, LLaVA-1.5-7B, InternVL3-2B, and InternVL3-8B; the selected entropy thresholds are 0.10, 0.80, 0.10, and 0.10.

The comparison supports the shortcut analysis in Section 5.2. CLIP-Q and SigLIP-Q access image–question compatibility but not the model's internal answer-conditioned state; their weak AUROC values indicate that generic external compatibility does not recover the VSS ranking. Confidence, entropy, and prompting can reduce some unsupported answers, but their tradeoffs are less consistently favorable than LEA, especially at low RAD.

C Extended Evidence and Diagnostics

Table 5 provides the full version of the core evidence table. It separates the held-out mixed-split probe AUC from the cross-validated probe AUC so that the two estimates are not conflated. The held-out AUC corresponds to the Probe AUC column in Table 1; the cross-validated AUC is a grouped estimate on the training contrast states.

Visual-dependence gap. The visual-dependence gap measures whether restoring the correct image increases the model's support for the selected option:

$$\mathrm{VCD}_i(y) = \log p_i^{\mathrm{Real}}(y) - \log p_i^{\mathrm{Mismatch}}(y), \quad (6)$$

where $p_i^c(y)$ is the A–E option probability assigned to option y under condition c. A positive value means that the chosen option receives more probability when the correct image is present. We use this as an output-space diagnostic of visual dependence, not as a correctness metric.

D Claim-Evidence Map

Table 7 maps main claims to supporting experiments, making the evidential role of each result explicit.

E Robustness and Ablation Studies

These ablations test whether the InternVL3-8B result depends on a brittle probe choice. Table 1 reports a held-out AUC of 0.705 for InternVL3-8B on the mixed train/test protocol. The Layer-20 ablation below uses a source-specific GQA/POPE diagnostic contrast with grouped cross-validation, serving as a within-source robustness check rather than a second estimate of the main mixed-split AUC.

Single layers and layer combinations. In the source-specific diagnostic setting, a single Layer-20 probe reaches AUC 0.883. Combining late layers does not improve the result: Layers 20 and 24

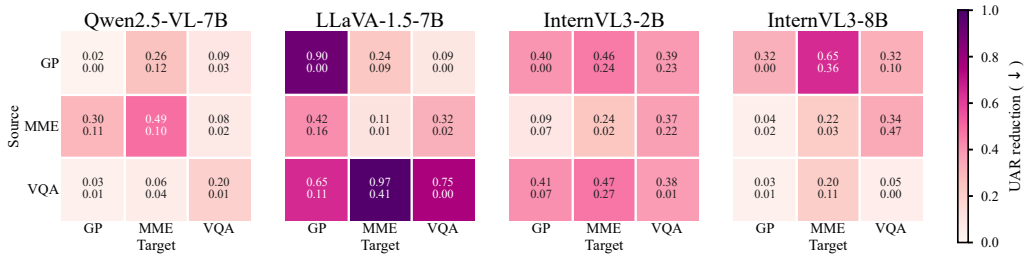

Figure 6: Cross-dataset transfer heatmaps (GP=GQA/POPE, VQA=VQAv2). Color denotes UAR reduction, where higher is better; cell annotations report UAR reduction and RAD. Positive off-diagonal cells show that source-trained gates often preserve useful rankings on other sources, while variation in RAD shows that fixed thresholds are domain-sensitive.

Table 4: VSS gate and baselines under the selective-prediction protocol. Efficiency is UAR reduction divided by RAD at the reported operating point; it should be read together with RAD, especially when RAD is near zero. Cov.-acc. is coverage-controlled accuracy.

Model	Method	AUROC	AUPRC	RAD	UAR red.	Eff.	Cov.-acc.
Qwen2.5-VL-7B	VSS	**0.886**	**0.886**	**0.038**	0.363	**9.63**	**0.827**
	CLIP-Q	0.328	0.380	0.606	0.554	0.91	0.163
	SigLIP-Q	0.492	0.474	0.087	0.069	0.80	0.600
	Confidence	—	—	0.292	0.133	0.46	—
	Entropy	—	—	0.383	0.258	0.67	—
	Prompted Abstention	—	—	0.554	0.438	0.79	—
LLaVA-1.5-7B	VSS	**0.902**	**0.930**	**0.054**	0.678	**12.65**	**0.856**
	CLIP-Q	0.300	0.498	0.508	0.826	1.63	0.512
	SigLIP-Q	0.469	0.598	0.180	0.267	1.49	0.521
	Confidence	—	—	0.233	0.688	2.95	—
	Entropy	—	—	0.242	0.704	2.91	—
	Prompted Abstention	—	—	0.338	0.463	1.37	—
InternVL3-2B	VSS	**0.899**	**0.914**	**0.032**	0.371	**11.69**	**0.790**
	CLIP-Q	0.290	0.404	0.480	0.517	1.08	0.093
	SigLIP-Q	0.487	0.513	0.067	0.062	0.93	0.477
	Confidence	—	—	0.321	0.308	0.96	—
	Entropy	—	—	0.388	0.413	1.06	—
	Prompted Abstention	—	—	0.433	0.379	0.88	—
InternVL3-8B	VSS	**0.875**	**0.869**	**0.033**	0.404	**12.33**	**0.820**
	CLIP-Q	0.331	0.402	0.634	0.638	1.01	0.326
	SigLIP-Q	0.504	0.509	0.077	0.079	1.03	0.651
	Confidence	—	—	0.200	0.121	0.60	—
	Entropy	—	—	0.296	0.271	0.92	—
	Prompted Abstention	—	—	0.588	0.492	0.84	—

Table 5: Full evidence summary. Gate n is the number of contrast states used for the held-out gate. Held-out AUC corresponds to the Probe AUC column in Table 1. CV AUC is five-fold grouped cross-validation on the training contrast states. Same-image columns test whether separability appears only after image–question composition rather than at Layer 0.

	Main mixed-split evidence					Same-image control			
Model	Gate n	Held-out AUC	CV AUC	VCD d	UAR	Pairs	L0 AUC	Best late	Multi
Qwen2.5-VL-7B	949	0.698	0.699	1.322	**0.500**	60	0.500	0.808	0.796
LLaVA-1.5-7B	1092	0.745	0.696	0.889	0.892	106	0.500	0.813	0.806
InternVL3-2B	788	**0.760**	**0.720**	**1.384**	0.546	66	0.500	**0.866**	**0.854**
InternVL3-8B	1030	0.705	0.696	1.233	0.621	87	0.500	0.785	0.784

reach 0.879, and Layers 20, 24, and 27 reach 0.874. This suggests that VSS is already readable from a single late representation.

Sample scaling. With only 10 training examples, the source-specific probe is noisy but above chance (mean AUC 0.728, std 0.250). Performance stabilizes at 50 examples (0.843 $\pm$ 0.043) and remains strong at 200 examples (0.889 $\pm$ 0.026). This argues against the ablation being driven by a large memorized training set, while showing that very small samples have high variance.

Classifier families. Logistic regression, RBF-SVM, and random-forest probes give similar performance on the same diagnostic states: RBF-SVM

Table 6: Same-image control AUC. The image is fixed while the question changes. Chance performance at Layer 0 and stronger late-layer performance argue against a simple image-identity explanation.

Model	L0	L16	L20	L24	Best
Qwen2.5-VL-7B	0.500	0.794	0.804	0.808	0.808
LLaVA-1.5-7B	0.500	0.808	0.809	0.813	0.813
InternVL3-2B	0.500	0.798	**0.866**	0.857	**0.866**
InternVL3-8B	0.500	0.753	0.774	0.772	0.785

Table 7: Claim-evidence alignment map. Each row states a main-text claim, the experiments that support it, and whether the evidence is sufficient.

Claim	Supporting empirical evidence	Status
VSS is linearly readable	Held-out AUC, grouped CV AUC, and VCD separation across all four models (Table 1; Appendix Table 5).	Supported
The signal is not a simple image-identity shortcut	Layer 0 stays at chance level (AUC = 0.5) while late layers become separable when the image is fixed and the question changes (Figure 3; Appendix Table 6).	Supported
The signal is not merely correctness or answer concreteness	Quadrant scores form a graded evidence-adequacy ordering across REAL correct, REAL incorrect, MISMATCH concrete, and MISMATCH abstention cases (Table 2).	Supported
The signal often transfers, but thresholds are domain-sensitive	Off-diagonal transfer cells are often positive, while RAD varies across source–target pairs (Appendix Figure 6; Figure 5).	Supported
The signal is behaviorally actionable but not a complete mitigation	LEA reduces unsupported answering for all four models at a shared operating point, with model-dependent RAD (Table 3; Figure 4).	Supported

reaches 0.895 $\pm$ 0.024, random forests 0.875 $\pm$ 0.025, and logistic regression 0.880 $\pm$ 0.019. The result is not specific to the linear classifier used by LEA, though the main paper keeps the linear gate for interpretability.

Leave-one-dataset-out generalization. A separate source-held-out ablation trains on two sources and tests on the third. AUCs are 0.613 (holding out GQA/POPE), 0.599 (holding out MME), and 0.834 (holding out VQAv2). The lower held-out AUCs motivate the calibration analysis: the representation can preserve useful rankings across sources, but thresholds are not invariant.

F Experimental Design Details

This section records methodological details important for interpreting the experiments.

Data construction. The paper-grade data pool combines a cleaned GQA/POPE slice (300 items), MME (400 items), and a filtered VQAv2 multiple-choice slice (500 items). The mixed split contains 960 training and 240 test items. Splits use image-content hashes rather than filenames, so repeated image content cannot appear on both sides.

Evidence conditions. Each question is evaluated under the four matched conditions defined in Section 3: original image, text-only, blank image, and same-source mismatch. Same-source mismatch keeps the visual domain plausible while changing whether the image supports the question. Because question and answer options are fixed across conditions, differences in behavior and hidden states can be attributed to visual evidence rather than linguistic changes.

Model scoring and hidden states. All models are evaluated deterministically by scoring next-token logits for options A–E at the final prompt position. The reported answer is the highest-probability option. Hidden states are extracted at the same position immediately before answer generation. The main gate uses late language-model layers; the same-image control additionally includes early layers to test whether the signal is present before image–question composition.

Gate estimation. Gate features are standardized using statistics fit on training data. The main gate is an L2-regularized logistic classifier with class-balanced weights. Probe AUC uses five-fold grouped cross-validation with grouping by image-content hash. The LEA gate is trained on the training split; intervention strengths and baseline thresholds are selected on a validation slice before evaluation on the held-out test split.

LEA and behavioral baselines. LEA changes only the abstention option's score according to the gate score, leaving concrete-answer scores unchanged. Shared operating-point results use the same intervention strength across models. Tuned LEA selects the strength from a fixed grid on validation data. Confidence, entropy, and prompted abstention use the same validation/evaluation separation, ensuring matched selection rules.

Calibration and transfer. Cross-dataset transfer trains a VSS gate on one source and evaluates on another. Calibration analyses reserve up to 50 target-domain examples for score-scale adjustment and evaluate on the remainder. These analyses separate two possibilities: whether the source-trained ranking disappears under dataset shift, or whether the ranking remains useful but the threshold needs target-domain calibration.

Sanity checks. Before using results for paper tables, we checked for leakage and shortcut failure modes with text-only probes, source-held-out transfer, image-held-out grouping, label-shuffle controls, and bootstrap uncertainty estimates. These checks support the validity of the reported contrast by testing whether it can be explained by repeated image content, accidental label structure, or unstable sampling.

G Representative Behavioral Checks

Table 8 provides representative held-out cases illustrating LEA behavior. These examples show how the gate score maps onto concrete decisions rather than serving as quantitative evidence. The desired pattern is selective: supported answers with low gate scores should be preserved, while unsupported conditions with high gate scores should shift toward abstention.

H Supplementary LEA Operating Points

The main text reports $\alpha = 1.0$ because it is a shared, moderate operating point that does not rely on per-model tuning. Table 9 shows nearby intervention strengths on the same held-out test protocol. All four models show positive UAR reduction at $\alpha = 1.0$, but the cost profile differs: LLaVA-1.5-7B has the strongest low-cost reduction, InternVL3-2B also shows a favorable tradeoff, and Qwen2.5-VL-7B and InternVL3-8B remain more conservative at mild intervention strengths.

Table 8: Representative held-out LEA checks. Gate scores range from 0.0 (more supported) to 1.0 (more unsupported).

Model	Condition	Native → LEA	Gate	Interpretation
InternVL3-8B	REAL	A → A	0.000	A supported answer is preserved when the gate assigns a very low gate score.
Qwen2.5-VL-7B	MISMATCH	E → E	0.992	A native abstention under mismatch remains abstained, as expected.
InternVL3-2B	MISMATCH	C → E	0.523	A concrete unsupported answer flips to abstention when the gate assigns a moderate gate score.
Qwen2.5-VL-7B	MISMATCH	C → C	0.001	A failure case: the unsupported answer receives a low gate score and is not corrected.
LLaVA-1.5-7B	TEXT-ONLY	A → E	1.000	A prior-driven text-only answer is suppressed when the gate assigns a high gate score.

Table 9: LEA α sweep on the shared 240-example test protocol. Each cell reports UAR reduction / RAD / efficiency. Efficiency follows the same convention as Table 4 and is descriptive when RAD is near zero.

Model	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 1.5$	$\alpha = 2.0$
Qwen2.5-VL-7B	0.021 / 0.008 / 2.50	0.071 / 0.008 / 8.50	0.104 / 0.008 / 12.50	0.146 / 0.008 / 17.50
LLaVA-1.5-7B	0.067 / 0.000 / 66.67	0.146 / 0.000 / 145.83	0.246 / 0.000 / 245.83	0.363 / 0.004 / 87.00
InternVL3-2B	0.100 / 0.013 / 8.00	0.158 / 0.017 / 9.50	0.217 / 0.025 / 8.67	0.267 / 0.033 / 8.00
InternVL3-8B	0.021 / 0.013 / 1.67	0.063 / 0.013 / 5.00	0.104 / 0.013 / 8.33	0.138 / 0.017 / 8.25