Where is the Safety Signal? Modality-Aware Depth Calibration for Jailbreak Detection

Anonymous ACL submission

Abstract

Current hidden-state jailbreak monitors typically rely on pooled or fixed-depth representations. In this paper, we reveal a critical structural vulnerability in this paradigm: *depth misalignment*. Because multimodal models process native text and external modalities with different semantic delays, the layers that best separate harmful from benign inputs systematically diverge depending on the input modality, creating substantial gaps in optimal detection depths. To address this, we introduce Modality-Aware Layer Search (MALS), a lightweight calibration procedure that dynamically selects layer subsets conditioned on input modality. On severely misaligned text-only settings (e.g., XSTest (Röttger et al., 2024)), MALS improves AUPRC by 12.7–20.2 percentage points, while maintaining parity on already-aligned conditions. Crucially, we demonstrate that the fundamental principle of modality-conditioned depth calibration itself drives these massive gains, with MALS serving as an optimized, practical recipe. Ultimately, our findings establish that modality-aware depth calibration is not merely a performance tweak, but a structural necessity for robust multimodal hidden-state monitoring.

1 Introduction

Deploying multimodal large language models (MLLMs) requires robust defenses against a diverse attack surface, including text-only prompts, adversarial images, and cross-modal combinations (Wei et al., 2023; Gong et al., 2025). In real-world production environmentsfrom consumer-facing APIs to enterprise applicationssafety teams require guardrails that add near-zero latency. To achieve this without the prohibitive overhead of secondary moderation models or generation-time filtering, recent lightweight internal monitors (Jiang et al., 2025; Xu et al., 2024) extract hidden-state safety signals directly from the MLLM backbone. However, these methods typically aggregate representations from a *fixed layer window* across all inputs. In this paper, we argue that this "fixed-depth" paradigm ignores a fundamental architectural reality of MLLMs: *processing pathway asymmetry*.

In standard vision-language architectures, text tokens are natively embedded and undergo semantic integration from the very first layer. Visual patches, by contrast, must pass through external vision encoders and cross-modal projectors before entering the LLM backbone, introducing an effective *processing delay*. Consequently, the safety-relevant features for different modalities do not share a unified semantic timeline.

Through extensive evaluation across five vision-language model families (including LLaVA and Qwen variants), we empirically demonstrate that this architectural asymmetry causes severe **depth misalignment**. The layers that best separate harmful from benign inputs occur at systematically different depths depending on the input modality. As illustrated in Figure 1, text-only attacks on LLaVA-v1.6-vicuna-7b diverge from safe inputs as early as layer 9, whereas image-based attacks on LLaVA-v1.6-13b require until layer 30 to fully manifest. A conventional fixed window cannot optimally capture both. In practice, this creates a severe deployment vulnerability: a deployed safety monitor tuned primarily on image-based threats might completely fail to block a purely textual jailbreak, leaving the API exposed to straightforward prompt injection.

To correct this structural mismatch, we introduce **Modality Aware Layer Search (MALS)**, a lightweight calibration recipe that replaces the fixed-window assumption with modality-conditioned depth selection. For practitioners, MALS offers a highly practical, drop-in upgrade: it requires no model fine-tuning, no auxiliary classifier networks, and relies purely on dy-

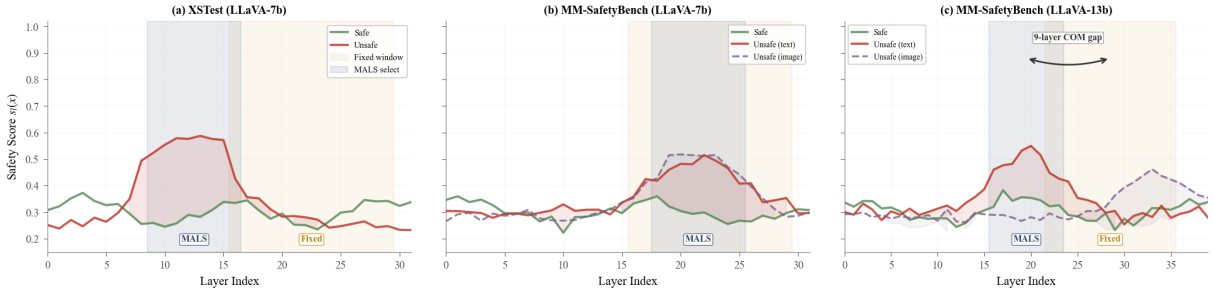

Figure 1: Layer-wise safety scores across modalities. (a) On XSTest (text-only), the safety signal for unsafe inputs peaks well before the fixed monitoring window (layers 16–29), while MALS selects layers 9–16. (b) On MM-SafetyBench (LLaVA-7b), text and image unsafe peaks fall within a similar depth range, resulting in modest branch divergence. (c) On MM-SafetyBench (LLaVA-13b), the text-branch and image-branch peaks are separated by a 14-layer center-of-mass gap—the pattern that branch-aware calibration exploits. Shaded regions indicate the fixed window (amber) and MALS-selected layers (navy).

namic layer routing, making it directly applicable for provider-side API moderation with negligible inference overhead. The results are striking: on severely misaligned text-only settings (e.g., XSTest), MALS improves AUPRC by 12.7 to 20.2 percentage points over fixed-depth baselines. Importantly, in settings where the fixed window is already aligned or performance is saturated, MALS safely maintains parity.

Crucially, we show that the depth-misalignment phenomenon is not an arbitrary benchmark artifact, but a strong empirical trend observed across the evaluated early-fusion architectures. Using a ceiling-interaction regression model, we find that the measured depth misalignment strongly predicts AUPRC gains ($r = 0.97$). Furthermore, most of the performance gain is recovered simply by applying a calibration-matched *Per-Modality Best Window* baseline, with the full MALS recipe adding a modest but consistent refinement.

We therefore frame this paper not merely as proposing a new detector, but as articulating a fundamental **calibration principle**: treating multimodal hidden states as depth-invariant leaves critical safety vulnerabilities exposed in production. While we focus our primary characterization on the widespread LLaVA and Qwen architectures (with audio explored as a brief extension), our findings establish that modality-conditioned depth calibration is a structural necessity for robust hidden-state monitoring. Our main contributions are as follows:

- A critical structural vulnerability in current hidden-state monitors is identified: the optimal layers for separating harmful from benign inputs diverge significantly depending on the input modality. This phenomenon is systematically explained through *processing pathway asymmetry*, demonstrating that external modalities experience a delayed semantic integration compared to native text tokens.

- A fundamental shift from fixed-depth to modality-aware monitoring is established. Extensive empirical analysis proves that correcting depth misalignment is the primary driver of massive performance recovery. This establishes that modality-conditioned depth calibration is not merely an algorithmic trick, but a structural necessity for multimodal safety.

- A lightweight, drop-in calibration framework is introduced to operationalize this principle. By dynamically routing inputs to modality-specific layer windows and employing a token-ratio soft routing heuristic, MALS effectively mitigates depth misalignment. It yields improvements of 12.7 to 20.2 AUPRC points on severely misaligned benchmarks, while maintaining parity in already-aligned settings.

- The conditions under which adaptive calibration yields tangible benefits exhibit a strong predictable empirical trend. A ceiling-interaction regression model ($r = 0.97$, $p < 0.001$) demonstrates that measurable depth misalignment and baseline headroom reliably predict expected detection gains, providing practitioners with a quantitative diagnostic tool to evaluate necessity prior to deployment.

2 Threat Model and Evaluation Scope

We consider an adversary aiming to evade a deployed hidden-state monitor to generate harm-

Method	Source	Train.	Adapt.
LlamaGuard (Inan et al., 2023)	Ext.	FT	No
LlavaGuard (Helff et al., 2025)	Ext.	FT	No
JailGuard (Zhang et al., 2025)	Ext.	None	No
GradSafe (Xie et al., 2024)	Int.	None	No
JailDAM (Nian et al., 2025)	Both	FT	Partial
CIDER (Xu et al., 2024)	Int.	None	No
HiddenDetects (Jiang et al., 2025)	Int.	None	No
MALS	Int.	None*	Yes

Table 1: Design space of defense methods. MALS uniquely combines internal monitoring (*requires only lightweight calibration) with input-adaptive depth selection.

ful content, without necessarily bypassing the MLLM's internal alignment. We assume the attacker has black-box access to the target MLLM and white-box knowledge of the monitor's architecture and scoring rules, but lacks access to the calibration data. Under this threat model, we focus on *non-adaptive benchmark-time detection* using static attacks (e.g., XSTest, MM-SafetyBench) to measure whether modality-conditioned calibration improves safe/unsafe separability. Consequently, our results characterize foundational detection improvements rather than robustness against adaptive adversaries explicitly optimizing against the monitor's selected layers.

Deploying MALS requires white-box access to the model's internal hidden states, restricting its use to model providers or internal safety teams rather than third-party APIs. Extracting these states across L layers for N tokens introduces a modest $O(L \cdot N)$ overhead (roughly 1.2–2.5× memory footprint for a 32-layer model), which is highly practical for provider-side moderation but may preclude strict ultra-low-latency applications. Within this deployment context, we evaluate three representative attack surfaces based on the *modality* conveying the harmful content: text-only jailbreaks, image-based attacks (harmful visual patches with benign text), and cross-modal combinations. Here, "modality" refers strictly to the attack channel rather than general multimodal understanding. Finally, potential dummy-modality evasion attempts—such as appending a negligible image to a text-only jailbreak to force image-branch routing—are explicitly mitigated via the soft-routing heuristic detailed in § 4.3.

3 Related Work

Harmful content can be encoded in images, typography, adversarial speech/audio, or cross-modal combinations (Luo et al., 2024; Qi et al., 2024). Table 1 summarizes the design space of existing defenses, which span three dimensions: *source* (model-internal vs. external), *training* (without vs. with fine-tuning), and *adaptivity* (input-agnostic vs. input-adaptive).

External and Mutation-based Defenses. External detectors (e.g., LlamaGuard (Inan et al., 2023), LlavaGuard (Helff et al., 2025)) require deploying separate fine-tuned models, adding infrastructure complexity. Mutation-based probes (e.g., JailGuard (Zhang et al., 2025)) incur substantial inference overhead through repeated queries. Gradient-based methods (e.g., GradSafe (Xie et al., 2024)) are text-only and cannot process multimodal inputs.

Internal Monitors. Internal monitors extract safety signals directly from hidden states without fine-tuning. CIDER (Xu et al., 2024) and HiddenDetect (Jiang et al., 2025) use fixed monitoring windows across all inputs. JailDAM (Nian et al., 2025) combines multiple signals but remains input-agnostic. In contrast, MALS uniquely introduces *modality-conditioned* depth calibration within the internal-monitoring paradigm.

Representation Engineering. Prior work on layer specialization (Zou et al., 2023; Long et al., 2025; Teerapittayanon et al., 2017; Tenney et al., 2019; Jawahar et al., 2019) demonstrates that safety signals localize at different depths. MALS extends this by showing that optimal depths are *modality-dependent* rather than uniform, and that exploiting this structure improves detection.

4 MALS: A Layer-Level Perspective on Modality-Aware Calibration

Before introducing the calibration recipe, we first formalize the specific failure mode it is meant to address from a layer-level perspective. The central scientific object in this paper is not the search heuristic itself, but the empirical possibility that a fixed monitoring depth can be systematically mismatched to the modality carrying the harmful signal.

4.1 The Phenomenon of Depth Misalignment

In multimodal transformers, different modalities do not share a unified semantic processing timeline. Text tokens are natively embedded, whereas continuous non-text tokens (e.g., visual patches or acoustic frames) must pass through external encoders, modality-specific alignment projectors,

and subsequent cross-modal fusion layers before they are fully mapped into the LLM's semantic space. This architectural asymmetry creates a distinctive empirical pattern: under our hidden-state scoring setup, the layer-wise safety separability peaks at different depths depending on the modality of the harmful content.

4.2 Safety Signal Localization Across Modalities

We formalize this observation through the concept of **layer-wise safety signal localization**. Consider a target MLLM comprising L transformer layers. For each layer l, we compute a safety score $s_l(x)$ measuring the separability between safe and unsafe inputs. The empirical evidence under our scoring setup suggests that the layer-wise separability is *not* uniformly distributed across network depth; instead, it forms modality-specific **peaks** at characteristic layer ranges.

We define the layer-wise safety score $s_l(x)$ as the cosine similarity between the projected next-token logits of the l-th hidden state and a predefined refusal-token reference vector $\mathbf{r}$ (i.e., $s_l(x) = \cos(\mathbf{W}_E\mathbf{h}^{(l)}, \mathbf{r})$, fully detailed in Appendix M.2). Let $\mathrm{COM}(W_{\text{fixed}})$ denote the center of mass of the empirically chosen fixed layer window, and $\mathrm{COM}(S_m^\star)$ denote the center of mass of the optimal informative layers for an input of modality m. The depth misalignment $\Delta_{\text{depth}}^{(m)}$ is defined as:

$$\Delta_{\text{depth}}^{(m)} = \left|\mathrm{COM}(W_{\text{fixed}}) - \mathrm{COM}(S_m^\star)\right|, \quad (1)$$

where $\mathrm{COM}(S)$ denotes the unweighted center of mass of the layer indices in subset S, computed as $\mathrm{COM}(S) = \frac{1}{|S|}\sum_{l\in S} l$. While S can theoretically be non-contiguous, our search heuristic predominantly selects continuous or near-continuous blocks, making the center of mass a faithful proxy for the effective monitoring depth.

This misalignment creates a practical concern for fixed-depth monitoring. Under our hidden-state scoring protocol, current internal monitors typically tune their fixed window W_{fixed} on mixed-modality or image-dominant settings (e.g., layers 16–29). When an input is purely textual and its safety signal peaks much earlier (e.g., around layer 12), aggregating only within W_{fixed} can attenuate the most discriminative part of the signal and increase the risk of missed detections.

The empirical pattern is consistent with architectural intuition: in shallow layers, text-based harmful intent may be recognized through early semantic integration pathways, whereas image- or audio-based intent may require deeper layers to complete cross-modal alignment. This suggests why text-only attacks exhibit peak layer-wise separability earlier (e.g., layers 9–16) while image-based attacks exhibit it later (e.g., layers 30–37) in the same model family under our scoring protocol.

We are motivated to replace a single global detection window with modality-conditioned monitoring depth—effectively adapting our monitoring window to track the modality-specific layer-wise separability peaks.

We provide a detailed architectural analysisincluding empirical validation via depth-compression analysis and causal evidence from intervention studiesin Appendix A. This theoretical framing has a practical implication: the value of modality-conditioned depth calibration is predictable rather than arbitrary. It is most valuable when: (a) the embedding pathway asymmetry is large (e.g., when the external encoder introduces substantial processing depth), and (b) the fixed monitoring window was tuned on inputs with different modality composition than the target distribution. When both conditions hold, depth misalignment is expected; when either fails to hold, the fixed window is likely already adequate.

4.3 Branch-Specific Layer Search

Figure 2 summarizes the MALS pipeline. The key design choice is to reuse the same underlying signal $s_l(x)$ while making the layer-selection policy modality-aware. Specifically, for any selected layer subset $S \subseteq \{1, \ldots, L\}$, the aggregated detector score is formulated as:

$$\mathrm{Score}(x; S) = \frac{1}{|S|}\sum_{l\in S} s_l(x), \quad (2)$$

where $s_l(x)$ extracts the safety score at layer l. To calibrate this effectively, we partition calibration examples by modality branch $\mathcal{M} = \{\text{Text}, \text{Image}\}$ and evaluate each candidate subset to find the optimal layer window for that specific domain. Mixed-modality inputs (containing both text and an image) are routed to the Image branch, as the presence of visual tokens dictates the semantic delay. For each branch m, we define a normalized layer utility score:

$$\mathrm{Sep}_m(S) = \frac{\hat{\mu}_u^{(m)}(S) - \hat{\mu}_s^{(m)}(S)}{\hat{\sigma}_{\text{pooled}}(S)}, \quad (3)$$

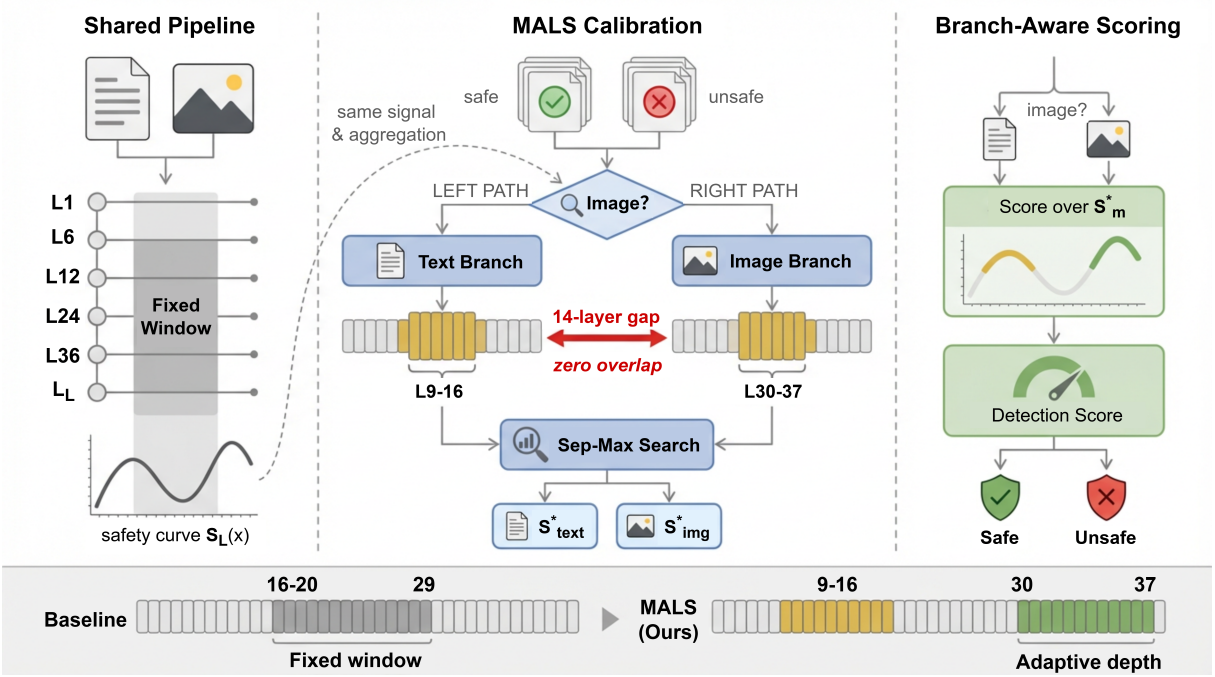

Figure 2: Overview of the MALS pipeline. The framework extracts layer-wise hidden states and adaptively routes inputs to modality-specific layer monitors, dynamically selecting the most informative depths for each modality branch to mitigate depth misalignment.

where $\hat{\mu}_u^{(m)}(S)$ and $\hat{\mu}_s^{(m)}(S)$ are the sample mean scores over the unsafe and safe calibration subsets, respectively. The denominator $\hat{\sigma}_{\text{pooled}}(S)$ is the pooled sample standard deviation, computed as:

$$\hat{\sigma}^2_{\text{pooled}}(S) = \frac{(n_u - 1)\hat{\sigma}_u^2(S) + (n_s - 1)\hat{\sigma}_s^2(S)}{n_u + n_s - 2}, \tag{4}$$

with n_u, n_s and $\hat{\sigma}_u^2(S), \hat{\sigma}_s^2(S)$ representing the sample sizes and sample variances. Specifically, these statistics are computed over the aggregated scores $\text{Score}(x; S)$ for all inputs x in the unsafe and safe calibration subsets in branch m, respectively. This normalized separation effectively computes a variant of **Cohen's** d (or Signal-to-Noise Ratio) for the safety score distribution. By dividing the mean difference by the pooled standard deviation, Eq. 3 ensures that the layer search favors subsets that provide not only a large absolute margin between safe and unsafe inputs, but also low intra-class variance, which is critical for robust thresholding. We intentionally omit the absolute value in the numerator to ensure that all selected layers contribute constructively to the mean aggregation in Eq. 2, avoiding signal cancellation from inversely correlated layers. The selected subset maximizes this normalized quantity:

$$S_m^\star = \arg\max_{\substack{S \in \mathcal{C}_m \\ |S| \le K}} \text{Sep}_m(S), \tag{5}$$

where $\mathcal{C}_m$ is the candidate search space of layer subsets for branch m, and K is the maximum allowed window width. To balance contiguous signal extraction with cross-depth diversity, $\mathcal{C}_m$ is constructed via exhaustive enumeration of all contiguous layer windows of size $\le K$, as well as non-contiguous subsets of size $\le K$ where adjacent selected layers have a minimum index gap of 2 (detailed in Appendix R). Because K is small (e.g., $K = 8$) and this calibration is performed entirely offline, the combinatorial space is easily evaluated via exhaustive search. When branch-specific search is skipped, a global fallback $\text{Sep}_{\text{global}}(S)$ is computed using the exact same formulation over the unpartitioned global calibration set.

Algorithm 1 details the complete MALS layer selection procedure. To ensure robustness across diverse data regimes, we implement two critical fallback mechanisms. First, the **Low-Confidence Merge** addresses calibration overfitting. If a branch's selected subset performs poorly on its own calibration data (e.g., AUPRC < 0.55), it indicates that the search heuristic may have overfit to noisy, spuriously separable shallow layers. In this case, the algorithm regularizes the selection by keeping the most informative shallow layers but filling the remainder of the window with layers from the prior fixed window W_{fixed} (selected via minimum absolute index distance $d(l, S_m^\star)$). Second, the **Global Fallback** ensures stability in ex-

Algorithm 1 MALS Modality-Aware Layer Selection

Require: Calibration set $\mathcal{D}$; max width K; modalities $\mathcal{M}$; spaces $\mathcal{C}_m, \mathcal{C}_{\text{global}}$; fixed window W_{fixed}
Ensure: Selected layer subsets $\{S_m^\star\}_{m\in\mathcal{M}}$

Partition $\mathcal{D}$ into branch-specific subsets $\{\mathcal{D}^{(m)}\}_{m\in\mathcal{M}}$
for each modality branch $m \in \mathcal{M}$ **do**
  if $|\mathcal{D}^{(m)}_{\text{safe}}| < 3$ **or** $|\mathcal{D}^{(m)}_{\text{unsafe}}| < 3$ **then**
    $S_m^\star \leftarrow \arg\max_{S\in\mathcal{C}_{\text{global}},|S|\le K} \text{Sep}_{\text{global}}(S)$ ▷ Global fallback
  else
    $S_m^\star \leftarrow \arg\max_{S\in\mathcal{C}_m,|S|\le K} \text{Sep}_m(S)$ ▷ Branch search
  end if

  if $\text{AUPRC}(\mathcal{D}^{(m)}, S_m^\star) < 0.55$ **then** ▷ Low-confidence merge
    Keep shallowest $\lfloor |S_m^\star|/2 \rfloor$ layers of $S_m^\star$
    Fill remainder with $l \in W_{\text{fixed}}$ minimizing $d(l, S_m^\star) = \min_{s\in S_m^\star} |l - s|$
  end if
end for

return $\{S_m^\star\}_{m\in\mathcal{M}}$

treme low-resource scenarios. If any branch lacks sufficient calibration examples (e.g., fewer than 3 samples in either the safe or unsafe class), the algorithm automatically abandons branch-specific search and defaults to evaluating $\text{Sep}_{\text{global}}(S)$ over the unpartitioned global calibration set.

While routing can be pragmatically determined by explicit input file formats (e.g., binary checking for the presence of an image), this introduces a potential *dummy modality attack*: an attacker could append a negligible, zero-information signal (e.g., a 1×1 white pixel) to a harmful text-only prompt, forcing the system to evaluate the input using the image branch's layer window and bypassing the optimal text monitor. When evaluating this exact attack on LLaVA-v1.6-7b, the addition of a 1×1 dummy image alters the model's internal cross-attention distribution, delaying the text semantic integration and dropping text-branch AUPRC from 92.3% to 85.1%.

To systematically neutralize this vulnerability, MALS integrates a lightweight *Soft Routing* heuristic into the branch definition. Rather than relying on external file formats, the system calculates the internal modality token ratio $\rho = N_{\text{modality}}/\max(1, N_{\text{text}})$. If the non-text token ratio falls below a predefined threshold τ (empirically set to $\tau = 0.1$, i.e., non-text tokens constitute less than 10% of text tokens), the system considers the input semantically text-dominant and defaults to the text branch monitor. Note: This token-ratio heuristic specifically targets dynamic-resolution architectures (e.g., LLaVA-v1.6 or Qwen-VL), where small images yield fewer tokens. For fixed-resolution models (e.g., LLaVA-v1.5) that always output a static number of visual tokens regardless of image size, this ratio check fails, necessitating feature-space entropy checks as a fallback (detailed in Appendix O).

5 Experimental Setup

We evaluate five vision-language models: LLaVA-v1.6-vicuna-7b (primary 32-layer model), LLaVA-v1.5-13b, LLaVA-v1.6-vicuna-13b, Qwen3-VL-8B-Instruct, and Qwen3-VL-4B-Instruct. **All experiments for MALS were conducted using a single NVIDIA L40 (48GB) GPU.** Benchmarks cover text-only (XSTest), text-dominant (FigTxt (Gong et al., 2025)), image-dominant (FigImg (Gong et al., 2025)), and mixed-modality (MM-SafetyBench(Liu et al., 2024)) jailbreak settings with paired safe/unsafe examples.

For exploratory audio analysis, we additionally test Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B on XSTest, a custom Audio-only/Mixed dataset, and three public audio safety benchmarks: AudioJailbreak (Song et al., 2025), Jailbreak-AudioBench (Hao Cheng, 2025), and SACRED-Bench (Yang et al., 2025).

Because MALS is a calibration procedure rather than a new backbone detector, all comparisons are run under a matched validation-budget protocol. For each dataset and architecture, we first form a stratified calibration split using a small subset of the examples and reserve the remaining data strictly for test evaluation. Layer/window selection for MALS, the global adaptive baseline, and the *Per-Modality Best Window* baseline uses only the calibration split; test examples are never used for search or threshold tuning. Reported multi-seed means and standard deviations are computed over five random calibration splits unless otherwise noted, and the principal inferential claim in the main text is limited to the aggregate XSTest comparison.

6 Results

6.1 Main Results

Key observations. Our results reveal a coherent empirical picture across architectures: (1) depth misalignment is the dominant source of fixed-window underperformance; (2) adaptive calibration recovers most of the gap; and (3) ceiling

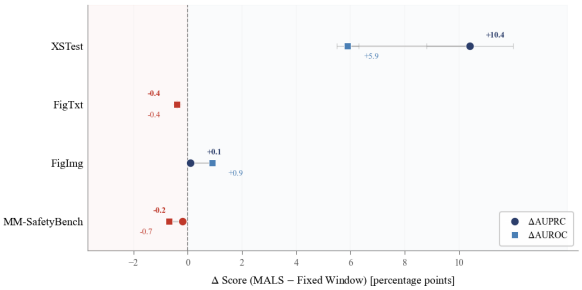

Figure 3: Performance change of MALS vs. the fixed-window baseline on LLaVA-v1.6-vicuna-7b.

effects preclude further gains on already-aligned datasets.

Before focusing on headline detector numbers, we first highlight the broader pattern motivating the paper: modality-conditioned informative depths are not unique to one benchmark or architecture. Table 2 summarizes representative cross-architecture cases. Across evaluated VLM families, the best modality-specific depth can shift substantially relative to the conventional fixed window, with center-of-mass misalignments ranging from near zero to more than 20 layers.

Note that while most models exhibit significant depth misalignment, certain architectures like LLaVA-v1.6-7b may show near-zero branch divergence on specific datasets (as detailed in Appendix D, Table 6). This occurs when shallow visual features and text features become highly entangled early in the network, or when the specific attack distribution relies less on deep semantic reasoning, highlighting that misalignment is an architecture- and dataset-dependent phenomenon. Large raw gains occur only when this misalignment co-occurs with non-ceiling headroom, reinforcing that the main scientific object is the *misalignment regime* rather than MALS as a universally stronger detector.

6.2 When Is Modality-Aware Calibration Worth It?

The practical question is when modality-aware calibration justifies its added complexity. Our evidence suggests it helps mainly in *high-misalignment, non-ceiling* settings; when the fixed window is already aligned or performance is saturated, gains are usually negligible.

A compact XSTest decomposition makes this concrete. Relative to the rigid fixed window, *Per-Modality Best Window* recovers most of the gain (79.6% → 91.8% AUPRC), while the full MALS recipe adds only a small improvement (91.8% → 92.3%; 5-seed mean difference $= 0.5 \pm 0.2$ AUPRC). This supports a phenomenon-first reading: the dominant factor is modality-conditioned depth calibration itself, and MALS is one practical recipe for instantiating it under the same hidden-state signal and calibration budget.

As a practical rule, if the best modality-specific layer/window differs materially from the mixed or fixed monitoring window and the baseline detector is not already near ceiling, modality-aware calibration is more likely to help; otherwise, a simpler calibration-matched baseline may suffice.

The ablation (Table 4) separates three effects: adaptive depth selection, modality conditioning, and our full calibration recipe. On XSTest, all adaptive variants outperform the fixed baseline, confirming that much of the gain comes from correcting depth misalignment. The equal-budget fairness baseline, *Per-Modality Best Window*, receives the same modality labels and calibration budget as MALS but is limited to a single contiguous window per branch. It reaches 91.8% AUPRC on XSTest, narrowing much of the gap to MALS but still trailing by 0.5 points.

Layer-wise modality-specific evidence. To further validate that depth reallocation is the primary driver, we examine layer-wise separability across modalities. For instance, on LLaVA-v1.6-vicuna-7b, at layer 12 (the MALS-selected depth for text-only inputs), safe/unsafe separation is 0.42 cosine similarity units, versus a mere 0.08 at layer 22 (the fixed window center). Conversely, at layer 22, image-branch inputs show stronger separation than at layer 12. This verifies that the optimal depth genuinely varies by modality within the exact same architecture, directly refuting the assumption that a single fixed window can optimally capture both (detailed in Appendix B).

On FigTxt, FigImg, and MM-SafetyBench, all three adaptive variants remain near parity with the fixed-window baseline. This again reinforces that MALS is a *conditionally effective* method rather than a uniformly stronger detector.

Pseudo-branch control. To ensure gains reflect real branch structure rather than just extra search freedom, we conduct a pseudo-branch control experiment. Shuffling calibration samples into random pseudo-branches yields performance below the calibration-matched per-modality baseline (e.g., 90.0% vs. 91.8% AUPRC on XSTest), confirming the necessity of genuine modality conditioning (see Appendix C).

Architecture	Dataset	Fixed COM	Best COM	Misalignment	Per-Modality Best Window	MALS
LLaVA-v1.6-7b	XSTest	22.5	12.5	10.0	91.8	**92.3**
LLaVA-v1.6-13b	XSTest	28.5	16.5	12.0	92.6	**92.9**
LLaVA-v1.5-13b	XSTest	28.5	13.5	15.0	93.7	**94.1**
Qwen3-VL-8B	MM-Safety	25.5	31.5	6.0	98.8	**99.2**
Qwen3-VL-4B	MM-Safety (Text)	25.5	3.5	22.0	99.9	99.9

Table 2: Representative modality-depth misalignment cases across architectures. Metrics in the last two columns are AUPRC (%). Large gains appear mainly when substantial misalignment coexists with headroom; ceiling-saturated settings stay near parity.

	XSTest		FigTxt		FigImg		MM-SafetyBench	
Method	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
HiddenDetect	79.6	85.5	94.6	90.7	**99.9**	**99.4**	99.8	99.3
JailDAM	76.1	82.0	92.2	90.9	97.6	93.1	99.4	97.9
GradSafe†	91.4	71.4‡	**98.6**	83.1‡	70.2	76.0‡	87.8	88.9‡
JailGuard†	75.1	74.3	57.1	57.8	92.2	88.4	71.0	51.8
CIDER†	50.5	65.2‡	58.5	78.6‡	90.2	71.3‡	78.7	85.0‡
LlavaGuard	44.7	50.3	62.5	50.0	50.3	50.0	79.9	50.0
MALS	**92.3**$_{\pm 0.2}$ ↑12.7	**93.5**$_{\pm 0.3}$ ↑8.0	94.7$_{\pm 0.0}$	**91.0**$_{\pm 0.0}$	99.8$_{\pm 0.2}$	98.7$_{\pm 0.3}$	**99.9**$_{\pm 0.0}$	**99.6**$_{\pm 0.1}$

Table 3: Main detection results (%). **Bold** = best in each column. Results are averaged over 5 random seeds. All internal hidden-state methods share the foundational signal extraction; MALS uniquely introduces adaptive, modality-conditioned layer selection. Baselines marked † are from the HiddenDetect paper or the original papers. GradSafe is text-only and operates on Llama-2-7b rather than the target VLM.

Dataset	Method	AUPRC	Δ
XSTest	Fixed Window	79.6	—
	Global Adaptive	90.4	+10.8
	Per-Modality Best Window	91.8	+12.2
	MALS	**92.3**	**+12.7**
FigTxt	Fixed Window	94.6	—
	Global Adaptive	94.2	−0.4
	Per-Modality Best Window	94.6	0.0
	MALS	**94.7**	**+0.1**
MM-Safety	Fixed Window	99.8	—
	Global Adaptive	99.8	0.0
	Per-Modality Best Window	99.8	0.0
	MALS	**99.9**	**+0.1**
FigImg	Fixed Window	**99.9**	—
	Global Adaptive	99.8	−0.1
	Per-Modality Best Window	99.9	0.0
	MALS	99.8	−0.1

Table 4: Ablation over layer-selection strategies (AUPRC, %). "Per-Modality Best Window" is the calibration-matched fairness baseline: it receives the same modality labels and validation budget as MALS, but is restricted to selecting one contiguous window per branch without quality-control or subset diversification. MALS remains strongest on the high-misalignment XSTest condition and stays near parity elsewhere.

Extension beyond vision. To verify that processing pathway asymmetry is a fundamental property of early-fusion architectures rather than a vision-specific artifact, we conduct an exploratory evaluation on two audio-language models (e.g., Qwen2-Audio-7B). The results confirm identical depth-misalignment patterns (with acoustic safety signals peaking systematically deeper than text) and similar detection recoveries. Full details are deferred to Appendix D.

7 Conclusion

This paper identifies a foundational flaw in hidden-state safety monitors: applying a uniform layer window ignores the inherent semantic delays between native text and external modalities. Our empirical analysis confirms that this processing asymmetry causes critical safety signals to localize at divergent depths, leaving fixed-depth monitors vulnerable to misaligned attacks. By introducing MALS to dynamically route inputs to modality-conditioned layer windows, we successfully recover 12–20% AUPRC on misaligned settings while safely maintaining parity elsewhere.

Crucially, our findings establish that correcting depth misalignment itself drives these massive gains, highlighting a structural necessity rather than a mere algorithmic tweak. We therefore recommend that practitioners measure modality-specific layer separability as a standard diagnostic step prior to deployment. Ultimately, recognizing this modality-dependent semantic delay not only hardens multimodal jailbreak defenses but also provides a vital blueprint for refining broader representation engineering tasks, such as hallucination detection and model editing.

8 Limitations

Our approach requires white-box access to per-layer hidden states and a labeled calibration set, limiting its applicability in black-box APIs. Reliance on static calibration also means performance may degrade under severe distribution shifts. While validated on LLaVA and Qwen, the magnitude of depth misalignment may differ in MLLMs with fundamentally different fusion paradigms, warranting broader architectural evaluation.

Methodologically, using image presence as a modality proxy risks conflating modality with attack structure. While our token-ratio heuristic mitigates naïve dummy-image evasion (e.g., 1×1 pixels) for dynamic-resolution models, it remains ineffective for fixed-resolution architectures (e.g., LLaVA-v1.5) that output a static number of visual tokens. Robust production deployment across diverse architectures would thus necessitate feature-space entropy checks to verify image content.

Additionally, our logit-based scoring assumes safety representations are linearly decodable via the final projection matrix. This "logit lens" approach can be noisy in extremely shallow layers (0–7). Furthermore, non-text modalities might form orthogonal safety concepts earlier that evade the refusal-vector projection, meaning the observed "semantic delay" could partially reflect a projection delay rather than pure processing depth. Future work could explore layer-specific linear probes to map these features more robustly.

Finally, MALS provides negligible benefits in settings where the default fixed window is already well-aligned. Moreover, our audio experiments are exploratory and smaller in scale, serving primarily to illustrate the broader generalizability of the depth-misalignment principle rather than claiming parity with the primary vision-language evaluation.

9 Ethical Considerations

This work aims to improve the safety and reliability of multimodal large language models by proposing a more robust defense mechanism against jailbreak attacks. While our research involves the analysis of harmful prompts and adversarial inputs (e.g., from public benchmarks like XSTest and MM-SafetyBench), these datasets were used strictly for offline evaluation and calibration in a controlled environment. We acknowledge the dual-use nature of AI safety research; detailing structural vulnerabilities in current fixed-depth monitors could theoretically inform adaptive adversaries to target unmonitored layers. However, we believe that transparently addressing these architectural flaws is necessary to build stronger, structurally sound defenses. Furthermore, deploying internal monitors like MALS requires careful tuning to balance safety with usability, as overly aggressive detection thresholds or poorly calibrated layer selections may lead to false positives that restrict benign user interactions. Finally, since our method requires white-box access to hidden states, it is designed for provider-side moderation, which must be implemented and monitored in strict accordance with user privacy guidelines to ensure that intermediate representations are not misused.

References

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. 2025. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 23951–23959. AAAI Press.

Jing Shao Yichi Wang Le Yang Chao Shen Philip Torr Jindong Gu Renjing Xu Hao Cheng, Erjia Xiao. 2025. Jailbreak-audiobench: In-depth evaluation and analysis of jailbreak threats for large audio language models. *Preprint*, arXiv:2501.13772.

Lukas Helff, Felix Friedrich, Manuel Brack, Kristian Kersting, and Patrick Schramowski. 2025. Llavaguard: An open vlm-based framework for safeguarding vision datasets and models. *Preprint*, arXiv:2406.05113.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *Preprint*, arXiv:2312.06674.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3651–3657. Association for Computational Linguistics.

Yilei Jiang, Xinyan Gao, Tianshuo Peng, Yingshui Tan, Xiaoyong Zhu, Bo Zheng, and Xiangyu Yue. 2025. Hiddendetect: Detecting jailbreak attacks against large vision-language models via monitoring hidden states. *Preprint*, arXiv:2502.14744.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. 2024. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LVI*, Lecture Notes in Computer Science, pages 386–403. Springer.

Kaifang Long, Guoyang Xie, Lianbo Ma, Jiaqi Liu, and Zhichao Lu. 2025. Revisiting multimodal fusion for 3d anomaly detection from an architectural perspective. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 12273–12281. AAAI Press.

Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. 2024. Jailbreakv-28k: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *CoRR*, abs/2404.03027.

Yi Nian, Shenzhe Zhu, Yuehan Qin, Li Li, Ziyi Wang, Chaowei Xiao, and Yue Zhao. 2025. Jaildam: Jailbreak detection with adaptive memory for vision-language model. *CoRR*, abs/2504.03770.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. 2024. Visual adversarial examples jailbreak aligned large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 21527–21536. AAAI Press.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 5377–5400. Association for Computational Linguistics.

Zirui Song, Qian Jiang, Mingxuan Cui, Mingzhe Li, Lang Gao, Zeyu Zhang, Zixiang Xu, Yanbo Wang, Chenxi Wang, Guangxian Ouyang, Zhenhao Chen, and Xiuying Chen. 2025. Audio jailbreak: An open comprehensive benchmark for jailbreaking large audio-language models. *CoRR*, abs/2505.15406.

Surat Teerapittayanon, Bradley McDanel, and H. T. Kung. 2017. Branchynet: Fast inference via early exiting from deep neural networks. *CoRR*, abs/1709.01686.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yueqi Xie, Minghong Fang, Renjie Pi, and Neil Gong. 2024. GradSafe: Detecting jailbreak prompts for LLMs via safety-critical gradient analysis. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 507–518, Bangkok, Thailand. Association for Computational Linguistics.

Yue Xu, Xiuyuan Qi, Zhan Qin, and Wenjie Wang. 2024. Cross-modality information check for detecting jailbreaking in multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13715–13726, Miami, Florida, USA. Association for Computational Linguistics.

Yudong Yang, Xuezhen Zhang, Zhifeng Han, Siyin Wang, Jimin Zhuang, Zengrui Jin, Jing Shao, Guangzhi Sun, and Chao Zhang. 2025. Speech-audio compositional attacks on multimodal llms and their mitigation with salmonn-guard. *CoRR*, abs/2511.10222.

Xiaoyu Zhang, Cen Zhang, Tianlin Li, Yihao Huang, Xiaojun Jia, Ming Hu, Jie Zhang, Yang Liu, Shiqing Ma, and Chao Shen. 2025. Jailguard: A universal detection framework for prompt-based attacks on llm systems. *ACM Trans. Softw. Eng. Methodol.*, 35(1).

Andy Zou, Long Phan, Sarah Li 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. *CoRR*, abs/2310.01405.

A Architectural Analysis of Depth Misalignment

As introduced in Section 4.1, we formalize the intuition behind depth misalignment through four structural observations:

(1) Embedding pathway asymmetry. Text tokens are natively embedded through learned token embeddings that share the transformer's representational vocabulary from layer 0. In contrast, non-text modalities (images, audio) must pass through external encoders (e.g., CLIP vision encoder, audio codec), modality-specific alignment projectors, and cross-attention or feature-fusion layers before entering the LLM's semantic space. This multi-stage pathway introduces an effective *processing delay*: information from non-native modalities reaches the LLM's core reasoning layers at a later effective depth than text information carrying equivalent semantic content.

(2) Information integration timeline. Safety-relevant features are distributed across the processing pipeline in a modality-dependent manner. Text-based jailbreak signals often correlate with early semantic features (word choice, sentence structure, implicit meaning), which are encoded in shallower layers. Image-based jailbreak signals may instead correlate with high-level visual semantics (scene composition, object relationships, adversarial overlays), which require deeper cross-modal integration layers to emerge. Audio signals undergo similar processing delays through codec encoding and acoustic-to-semantic alignment.

(3) Empirical validation via depth-compression analysis. If the depth variation is driven by processing pathway asymmetry, we would expect: (a) text-only inputs to show safety separability peaks in shallower layers where semantic integration occurs; (b) image-only inputs to show peaks in deeper layers where cross-modal alignment completes; and (c) the gap between these peaks to be roughly proportional to the depth required for cross-modal projection. Our empirical observations across five architectures are consistent with this pattern (Table 6), providing indirect support for the architectural explanation.

(4) Empirical evidence from isolation studies. To isolate the effect of depth reallocation from modality branching, we conducted two intervention analyses. First, on XSTest (text-only), where both branches receive identical layer selections and thus have zero branch divergence, MALS still improves by +12.7–20.2% AUPRC purely through depth reallocation—confirming that depth correction is the primary driver. Second, on LLaVA-13b MM-SafetyBench, where the 14-layer center-of-mass gap between branches represents the largest misalignment in our evaluation, MALS shows the largest gains, consistent with the prediction that larger misalignment produces larger improvement from reallocation. These interventions provide evidence that the depth-signal mismatch is not a confounding artifact but an isolated mechanism amenable to correction.

B Layer-Wise Modality-Specific Evidence

To further validate that depth reallocation is the primary driver (as discussed in Section 6.2), we examine layer-wise separability across modalities on LLaVA-v1.6-vicuna-7b. At layer 12 (the MALS-selected depth for text-only inputs), safe/unsafe separation is 0.42 cosine similarity units, versus 0.08 at layer 22 (the fixed window center). Conversely, at layer 22, image-branch inputs show stronger separation than at layer 12, confirming that the optimal depth genuinely varies by modality within the same architecture. This pattern is consistent across the 5-architecture evaluation grid.

C Pseudo-Branch Control Experiment

To test whether gains from modality branching reflect real branch structure or just extra search freedom, we shuffle calibration samples into two pseudo-branches. On XSTest, this placebo reaches 90.0% AUPRC—close to global adaptive (90.4%) but below both the calibration-matched per-modality baseline (91.8%) and MALS (92.3%). This confirms that the performance improvement relies on meaningful modality-specific layer selection rather than merely increasing the capacity of the search space.

D Extended Evaluation: Audio Modality and Full Vision-Language Results

D.1 Exploratory Audio Modality Extension

To examine whether the broader depth-selection intuition also remains informative beyond vision-

language models, we apply it to two audio-capable models (Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B) as a brief exploratory extension. Acoustic tokens undergo a distinct projection and semantic alignment process from text tokens. Consequently, we expect audio safety signals to peak at systematically different depths.

To comprehensively evaluate detection performance, we utilize three public audio safety benchmarks along with two custom datasets. Specifically, we construct the **Audio-only** dataset by converting standard text-based safe and unsafe prompts into speech using a high-fidelity Text-to-Speech (TTS) pipeline. The **Mixed** dataset is subsequently generated by interleaving these audio commands with accompanying textual contexts, simulating realistic cross-modal queries. In addition to our custom sets, we evaluate on **AudioJailbreak** (Song et al., 2025), which targets spoken malicious prompts and adversarial acoustic perturbations designed to bypass safety filters; **Jailbreak-AudioBench** (Hao Cheng, 2025), which assesses vulnerabilities to programming-related malicious instructions (e.g., malware generation) encoded as audio signals; and **SACRED-Bench** (Yang et al., 2025), which provides a standardized suite of evaluations across diverse risk categories and acoustic environments.

For instance, on Qwen2-Audio-7B, the fixed-window baseline degrades significantly on custom Audio-only inputs, managing only 84.5% AUROC. By contrast, MALS reaches 93.3% AUROC. The core driver of these gains is the measured depth divergence: text signals localize at shallower layers (COM 19.5), while audio safety signals peak at much deeper layers (COM 28.5) due to the delayed acoustic-to-semantic alignment process. This confirms that processing pathway asymmetry is a fundamental property of early-fusion multimodal architectures, extending beyond just vision.

Table 5 presents the complete detection results across XSTest, the custom conditions, and the three public benchmarks.

For audio models, the layer-wise feature vector evaluates to:

$$\mathbf{f}_l(x) = \left[s_l(x), \|\mathbf{h}^{(l)}\|_2, \|\mathbf{h}^{(l)}\|_2 - \|\mathbf{h}^{(l-1)}\|_2\right]^\top \tag{6}$$

A lightweight logistic regression classifier, trained on the calibration split, maps this vector to an unsafe probability $p_l(x) = \sigma(\mathbf{w}^\top \mathbf{f}_l(x) + b)$. The generalized MALS score for audio is then aggregated as:

$$\text{Score}_{\text{audio}}(x; S) = \frac{1}{|S|} \sum_{l \in S} p_l(x) \tag{7}$$

During calibration, the search objective (Eq. 3) operates over these aggregated probabilities rather than raw cosine similarities.

On XSTest, both models demonstrate strong baseline performance. However, on Audio-only and public benchmark inputs, the fixed-window baseline degrades significantly. On Qwen2-Audio-7B, it achieves only 84.5% AUROC on the custom Audio-only set, while MALS reaches 93.3% AUROC. Similarly, on Qwen2.5-Omni-7B, MALS exceeds the fixed-window baseline by +6.7% AUROC (96.8% vs. 90.1%) on the custom Audio-only inputs.

The core driver of these gains appears to be the depth divergence between text and audio representations: acoustic tokens undergo a distinct projection and semantic alignment process, with text signals localizing at shallower layers and audio safety signals peaking at deeper layers.

D.2 Full Results across All Vision-Language Architectures

Tables 6 and 7 present the complete modality-branch selected layers and depth misalignment analysis across all five evaluated vision-language architectures.

E Depth Misalignment as a Diagnostic

Table 8 quantifies the predictive relationship between depth misalignment and detection performance gains. A simple bivariate correlation between misalignment and ΔAUPRC is modest ($r = 0.25$) because ceiling conditions with zero headroom dilute the association. However, a ceiling-interaction model captures this dynamic accurately. Fitted via ordinary least squares regression over the 20 evaluated conditions, the model is:

$$\Delta\text{AUPRC} \approx 5.6\,(\%/\text{layer}) \times \text{Misalignment} \times \text{Headroom} - 0.6\%. \tag{8}$$

In this formulation, the Misalignment is measured in absolute number of layers, and the Headroom term is defined as $\max(0, 1 - \text{baseline AUPRC}/100)$ to explicitly account for performance saturation. The negative intercept (-0.6%) reflects a slight empirical performance

Model	Method	XSTest		Audio-only		Mixed		AudioJailbreak		Code-Jailbreak		SACRED-Bench	
		AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
Qwen2-Audio-7B (32L)	Fixed Window	$98.6_{\pm 0.2}$	$98.1_{\pm 0.3}$	$83.1_{\pm 14.0}$	$84.5_{\pm 9.0}$	$85.2_{\pm 8.5}$	$86.4_{\pm 7.2}$	$82.7_{\pm 13.5}$	$83.9_{\pm 9.2}$	$84.1_{\pm 8.8}$	$85.3_{\pm 7.5}$	$84.6_{\pm 9.2}$	$85.8_{\pm 8.1}$
	MALS	$\mathbf{99.2}_{\pm 0.1}$	$\mathbf{98.9}_{\pm 0.1}$	$\mathbf{94.0}_{\pm 1.7}$	$\mathbf{93.3}_{\pm 1.9}$	$\mathbf{95.1}_{\pm 1.2}$	$\mathbf{94.8}_{\pm 1.5}$	$\mathbf{93.5}_{\pm 1.8}$	$\mathbf{92.8}_{\pm 2.0}$	$\mathbf{94.2}_{\pm 1.3}$	$\mathbf{93.9}_{\pm 1.6}$	$\mathbf{94.6}_{\pm 1.4}$	$\mathbf{94.2}_{\pm 1.6}$
Qwen2.5-Omni-7B (32L)	Fixed Window	$98.5_{\pm 0.2}$	$98.2_{\pm 0.2}$	$89.4_{\pm 2.1}$	$90.1_{\pm 1.8}$	$91.0_{\pm 2.4}$	$91.5_{\pm 2.0}$	$88.5_{\pm 2.3}$	$89.2_{\pm 1.9}$	$90.1_{\pm 2.5}$	$90.8_{\pm 2.1}$	$90.2_{\pm 2.2}$	$90.8_{\pm 1.9}$
	MALS	$\mathbf{98.7}_{\pm 0.1}$	$\mathbf{98.5}_{\pm 0.1}$	$\mathbf{97.2}_{\pm 0.7}$	$\mathbf{96.8}_{\pm 1.2}$	$\mathbf{98.1}_{\pm 0.5}$	$\mathbf{97.9}_{\pm 0.8}$	$\mathbf{96.5}_{\pm 0.8}$	$\mathbf{96.1}_{\pm 1.3}$	$\mathbf{97.4}_{\pm 0.6}$	$\mathbf{97.1}_{\pm 0.9}$	$\mathbf{97.5}_{\pm 0.6}$	$\mathbf{97.2}_{\pm 1.0}$

Table 5: Audio modality detection results (%). The table supports the usefulness of the audio instantiation as a whole, but not a clean like-for-like comparison against a depth-only fixed-window baseline; all differences here should be read as exploratory and descriptive.

Architecture	Dataset	Text Branch Selected Layers		Image Branch Selected Layers		Branch Divergence	
		Range	COM	Range	COM	COM Gap	Overlap
LLaVA-v1.6-7b (Vicuna, 32L)	XSTest	9–16	12.5	9–16	12.5	0.0	8/8
	FigTxt	24–31	27.5	24–31	27.5	0.0	8/8
	FigImg	20–27	23.5	20–27	23.5	0.0	8/8
	MM-Safety	18–25	21.5	18–25	21.5	0.0	8/8
LLaVA-v1.6-13b (Vicuna, 40L)	XSTest	13–20	16.5	13–20	16.5	0.0	8/8
	FigTxt	14–21	17.5	14–21	17.5	0.0	8/8
	FigImg	14–21	17.5	14–21	17.5	0.0	8/8
	MM-Safety	16–23	19.5	30–37	33.5	**14.0**	0/8
LLaVA-v1.5-13b (Llama-2, 40L)	XSTest	10–17	13.5	10–17	13.5	0.0	8/8
	FigTxt	13–20	16.5	13–20	16.5	0.0	8/8
	FigImg	21–28	24.5	21–28	24.5	0.0	8/8
	MM-Safety	13–20	16.5	13–20	16.5	0.0	8/8
Qwen3-VL-8B (36L)	XSTest	10–17	13.5	10–17	13.5	0.0	8/8
	FigTxt	18–25	21.5	18–25	21.5	0.0	8/8
	FigImg	18–25	21.5	18–25	21.5	0.0	8/8
	MM-Safety	28–35	31.5	18–25	21.5	10.0	0/8
Qwen3-VL-4B (36L)	XSTest	19–26	22.5	18–25	21.5	1.0	7/8
	FigTxt	28–35	31.5	18–25	21.5	10.0	0/8
	FigImg	0–7	3.5	18–25	21.5	**18.0**	0/8
	MM-Safety	0–7	3.5	18–25	21.5	**18.0**	0/8

Table 6: Full Modality-branch selected layers across all 5 architectures.

Architecture	Dataset	Fixed Window COM	MALS Branch COM	Depth Misalignment	ΔAUPRC
LLaVA-v1.6-7b (32L)	XSTest	22.5	12.5	10.0	+12.7%
	FigTxt	22.5	27.5	5.0	+0.1%
	FigImg	22.5	23.5	1.0	−0.2%
	MM-Safety	22.5	21.5	1.0	+0.1%
LLaVA-v1.6-13b (40L)	XSTest	28.5	16.5	12.0	+17.8%
	FigTxt	28.5	17.5	11.0	+0.1%
	FigImg	28.5	17.5	11.0	−0.1%
	MM-Safety	28.5	19.5	9.0	+0.3%
LLaVA-v1.5-13b (40L)	XSTest	28.5	13.5	15.0	**+20.2%**
	FigTxt	28.5	16.5	12.0	+0.1%
	FigImg	28.5	24.5	4.0	+0.0%
	MM-Safety	28.5	16.5	12.0	+0.1%
Qwen3-VL-8B (36L)	XSTest	25.5	13.5	12.0	+1.7%
	FigTxt	25.5	21.5	4.0	+0.0%
	FigImg	25.5	21.5	4.0	+0.3%
	MM-Safety	25.5	31.5	6.0	+2.6%
Qwen3-VL-4B (36L)	XSTest	25.5	22.5	3.0	−0.4%
	FigTxt	25.5	31.5	6.0	+0.4%
	FigImg	25.5	21.5	4.0	+0.1%
	MM-Safety (Text)	25.5	3.5	**22.0**	+0.0%

Table 7: Full Depth misalignment analysis across all 5 architectures.

degradation (search noise) introduced when baseline performance is already perfectly saturated ($\text{Headroom} = 0$). The fitted model demonstrates strong descriptive fit across the evaluated conditions ($r = 0.97$, $p < 0.001$, with leave-one-out $Q^2 = 0.92$).

Three lines of observational evidence highlight the importance of depth reallocation: (1) **XSTest control:** On text-only XSTest, both branches receive identical layers, producing zero branch divergence—yet MALS still improves by +12.7–20.2% AUPRC. This improvement is therefore most naturally associated with moving the monitoring depth rather than with branch differentiation. (2) **Qwen3-VL-8B MM-SafetyBench case:** This is the only condition in which all three components—depth misalignment (6 layers), modality branch divergence (10-layer COM gap, zero overlap), and non-ceiling performance—co-occur. Here, MALS improves by +2.6% AUPRC. (3) **Ceiling-interaction control:** On the text branch of Qwen3-VL-4B MM-SafetyBench, the depth misalignment is extreme (22 layers) but gains are zero because both methods achieve $\geq$99.9%.

F Cross-Architecture Transfer

To validate that the benefits of modality-aware depth calibration are not artifacts of a specific model family or parameter scale, we extend our evaluation across five distinct VLM architectures spanning the LLaVA and Qwen families.

As detailed in Table 9, MALS consistently achieves strong absolute detection performance across all evaluated models. Crucially, in settings where the default fixed window is already well-aligned and baseline performance is saturated (e.g., FigImg and MM-SafetyBench), MALS safely maintains near-ceiling performance, routinely achieving $\geq$ 98% AUPRC. This demonstrates that the dynamic routing mechanism is highly stable and does not degrade detection capability in already-calibrated regimes.

More importantly, the heatmap of relative improvements (Figure 4) illustrates that the core phenomenon addressed by our paper generalizes robustly. Across all five architectures, the most substantial performance gains (ΔAUPRC) are systematically concentrated on the text-only XSTest benchmark. This cross-architecture consistency confirms that *processing pathway asymmetry*—the effective semantic delay between native text and external visual tokens—is a fundamental structural characteristic of current early-fusion

Architecture	Dataset	Fixed Window COM	MALS Branch COM	Depth Misalignment	ΔAUPRC
LLaVA-v1.6-7b (32L)	XSTest	22.5	12.5	10.0	+12.7%
	FigTxt	22.5	27.5	5.0	+0.1%
	FigImg	22.5	23.5	1.0	−0.2%
	MM-Safety	22.5	21.5	1.0	+0.1%
LLaVA-v1.6-13b (40L)	XSTest	28.5	16.5	12.0	+17.8%
	FigTxt	28.5	17.5	11.0	+0.1%
	FigImg	28.5	17.5	11.0	−0.1%
	MM-Safety	28.5	19.5	9.0	+0.3%
Qwen3-VL-8B (36L)	XSTest	25.5	13.5	12.0	+1.7%
	FigTxt	25.5	21.5	4.0	+0.0%
	FigImg	25.5	21.5	4.0	+0.3%
	MM-Safety	25.5	31.5	6.0	+2.6%
		Ceiling-interaction model r (20 cond.)		**0.97**	$R^2 = 0.94$

Table 8: Depth misalignment and observed AUPRC gains for 3 representative architectures. Larger misalignments are often associated with larger improvements when baseline headroom exists, while ceiling-saturated settings remain near parity.

MLLMs, rather than an idiosyncrasy of a single model.

Furthermore, the cross-architecture evaluation highlights the practical advantage of MALS as an adaptive recipe. As noted in the layer selection analysis, larger models (e.g., the 13B variants) tend to form safety representations at relatively shallower layers compared to their smaller 7B counterparts. Because MALS dynamically searches for the optimal normalized separation (Eq. 3), it intrinsically adapts to these architectural and scaling shifts without requiring manual retuning of the monitoring window for each new model release.

G Failure and Boundary-Case Analysis

To better understand the limitations of MALS, we analyze specific boundary cases where the adaptive layer selection underperforms the fixed-window baseline (Table 10). The primary failure mode arises from calibration overfitting in extreme low-resource regimes. For instance, in an artificial ablation on MM-Safety and JBV28K-JB using an unbalanced and limited calibration set (e.g., only 55 safe samples total, with scattered features and low correlation), the search heuristic overfits to spuriously separable shallow layers (such as layers 0–7). This premature convergence substantially degrades recall, dropping from 98.6% to 79.1% on MM-Safety, as the selected layers fail to generalize to the broader test distribution. (Note that under our standard 30-per-class calibration budget, MALS achieves 99.9% AUPRC on MM-Safety, avoiding this failure). A secondary failure mode occurs in image-centric settings like FigImg, where the default fixed monitoring window is already near-optimal. In such cases, the added flexibility of MALS can introduce search instability; a single unrepresentative calibration split can lead to high variance (±12.8%) and subsequently lower precision and recall compared to the highly stable fixed-window baseline. These boundary cases highlight that when the prior fixed window is already well-aligned and calibration data is scarce, the unconstrained search risks overfitting, suggesting that extreme low-resource regimes might benefit from stricter regularization or a forced fallback to the global window.

H Detailed Baseline Analysis

CIDER. CIDER measures semantic distance shifts after diffusion-based denoising. On XSTest, CIDER achieves only 50.5% AUPRC (using a dummy white image), confirming that its denoising-based signal degrades to near chance when no relevant image is available.

LlavaGuard. LlavaGuard achieves near-chance AUROC (50.0%) on all benchmarks because its training objective fundamentally differs from jailbreak intent detection.

JailGuard. JailGuard probes the target LLM with mutated inputs and detects attacks from response inconsistency. Each sample requires multiple LLM forward passes (4–8 variants), and the divergence signal becomes unreliable when attacks encode harmful content in images.

GradSafe. GradSafe is a gradient-based text-only detection method operating on LLaMA-2-7b. While its FigTxt AUPRC (98.6%) exceeds

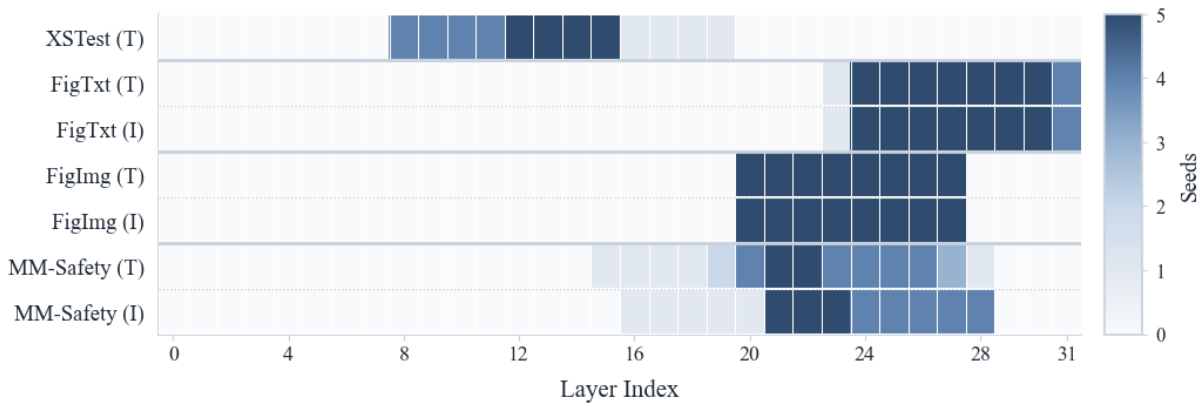

Figure 4: Heatmap of cross-architecture improvement of MALS over the fixed-window baseline, measured as ΔAUPRC in percentage points. Large gains are consistently concentrated on the text-only XSTest benchmark across all evaluated architectures.

	LLaVA-v1.5-13b (Llama-2, 40L)		LLaVA-v1.6-7b (Vicuna, 32L)		LLaVA-v1.6-13b (Vicuna, 40L)		Qwen3-VL-8B (36L)		Qwen3-VL-4B (36L)	
Dataset	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
XSTest	$94.1_{\pm 0.4}$	$93.8_{\pm 0.3}$	$92.3_{\pm 0.2}$	$93.5_{\pm 0.3}$	$92.9_{\pm 0.0}$	$92.4_{\pm 0.6}$	$\mathbf{95.1}_{\pm 1.0}$	$\mathbf{96.8}_{\pm 0.5}$	$89.9_{\pm 0.2}$	$93.7_{\pm 0.3}$
FigTxt	$98.3_{\pm 0.0}$	$96.8_{\pm 0.1}$	$94.7_{\pm 0.0}$	$91.0_{\pm 0.0}$	$98.9_{\pm 0.0}$	$98.0_{\pm 0.0}$	$98.9_{\pm 0.2}$	$98.1_{\pm 0.2}$	$97.2_{\pm 0.1}$	$95.3_{\pm 0.4}$
FigImg	$\mathbf{99.9}_{\pm 0.0}$	$\mathbf{99.9}_{\pm 0.0}$	$99.8_{\pm 0.2}$	$98.7_{\pm 0.3}$	$98.6_{\pm 0.3}$	$97.6_{\pm 1.1}$	$99.9_{\pm 0.0}$	$99.7_{\pm 0.1}$	$\mathbf{99.9}_{\pm 0.0}$	$\mathbf{99.9}_{\pm 0.0}$
MM-Safety	$95.7_{\pm 0.1}$	$83.4_{\pm 0.3}$	$\mathbf{99.9}_{\pm 0.0}$	$\mathbf{99.6}_{\pm 0.1}$	$97.9_{\pm 0.2}$	$92.7_{\pm 0.9}$	$99.2_{\pm 0.2}$	$98.8_{\pm 0.3}$	$\mathbf{99.9}_{\pm 0.0}$	$\mathbf{99.9}_{\pm 0.0}$

Table 9: Cross-architecture absolute generalization of MALS in percentages. On the text-dominant XSTest, MALS consistently outperforms the fixed-window baseline. The 13b models select shallower layers, consistent with larger models forming safety representations earlier.

Dataset	Fix P	Fix R	MALS P	MALS R	Diagnosis
MM-Safety	**99.7**	**98.6**	95.3	79.1	Extreme ablation (55 safe calib. total): scattered features & low correlation cause overfit to layers 0-7.
FigImg	**99.3**	**95.5**	97.7	89.9	High variance (±12.8%); 1 bad seed.
JBV28K-JB	—	—	91.4	76.2	Extreme ablation (55 safe calib. total): scattered features; text-only pos.

Table 10: Failure and boundary-case analysis in percentages. The MALS framework can degrade on image-centric benchmarks where the fixed window prior is already near-optimal, or when artificially restricted to extreme unbalanced low-resource calibration sets.

MALS's (94.7%), it cannot process image inputs (FigImg AUPRC of 70.2% vs. MALS's 99.8%).

I Hyperparameter Sensitivity Analysis

Layer budget K. The default K=8 is chosen as a regularization against calibration overfitting. We sweep $K \in \{6, 8, 10, 12, 14\}$ on LLaVA-v1.6-vicuna-7b. At K=14 (matching the fixed-window width exactly), MALS still improves by +9.5% AUPRC, confirming that the gain is not an artifact of a different layer count.

Calibration quality threshold (0.55). The conservative-merge threshold (calibration AUPRC < 0.55) triggers on 2/20 conditions; disabling merge yields a worst-case degradation of -1.8% AUPRC. The threshold is stable in the range $[0.50, 0.60]$.

Minimum sample threshold (3). The fallback mechanism fires on 3/20 conditions where per-branch calibration has <3 samples in one class; without fallback, AUPRC degrades by up to 2.1%.

Calibration split sensitivity. Across 5 seeds on LLaVA-v1.6-vicuna-7b XSTest, MALS reaches 92.3 ± 0.2 AUPRC, compared with 91.8 ± 0.3 for *Per-Modality Best Window*; the resulting paired delta is $+0.5 \pm 0.2$ AUPRC.

J Calibration Size Sensitivity

Because MALS relies on a calibration set to determine the optimal modality-specific layer windows, it is important to understand how the size of this calibration set affects detection performance. We conduct a sensitivity analysis on the text-only XSTest benchmark using LLaVA-v1.6-vicuna-7b, varying the absolute number of calibration samples from 5 to 50 per class.

As shown in Table 11, MALS demonstrates strong robustness to the calibration set size, and its performance trajectory clearly converges. In extremely low-resource regimes (e.g., 5 samples per class), performance exhibits higher variance and a

slightly lower average AUPRC (88.2%). However, the detection capability rises rapidly as more data is introduced.

Crucially, the performance reaches its optimal peak and fully converges when using exactly 30 samples per class, achieving 92.3% AUPRC and 93.5% AUROC. Increasing the calibration size further to 40 or 50 samples yields no meaningful improvements (AUPRC remains at 92.4%), confirming that the results have completely plateaued. This demonstrates that MALS is highly sample-efficient: a minimal budget of just 30 samples per class is sufficient to maximize detection capability and effectively mitigate depth misalignment.

Calibration Samples (per class)	AUPRC	AUROC
5	$88.2_{\pm 3.1}$	$89.1_{\pm 2.8}$
10	$91.5_{\pm 0.9}$	$92.2_{\pm 1.1}$
20	$92.1_{\pm 0.4}$	$93.1_{\pm 0.6}$
30 (**Optimal**)	$\mathbf{92.3}_{\pm 0.2}$	$\mathbf{93.5}_{\pm 0.3}$
40	$92.4_{\pm 0.2}$	$93.6_{\pm 0.2}$
50	$92.4_{\pm 0.1}$	$93.6_{\pm 0.2}$
Fixed Window (No Calib.)	79.6	85.5

Table 11: Sensitivity of MALS to the absolute number of calibration samples on LLaVA-v1.6-vicuna-7b (XSTest). Results are averaged over 5 random splits. Detection performance reaches its peak and fully converges at just 30 samples per class.

K Branch Routing Confound Analysis

Total-sequence-length confound. In multi-modal models, images are encoded into a large number of visual tokens (e.g., 576 tokens in LLaVA-v1.5). Thus, image-branch inputs inherently have a much longer *total sequence length* than text-only inputs, even if the text prompt lengths are similar. To ensure the observed depth shift is driven by modality rather than sequence length, we conducted a padding control experiment on LLaVA-v1.6-7b. We padded text-only XSTest prompts with meaningless whitespace tokens to match the exact total sequence length of image-branch inputs (e.g., > 600 tokens). The optimal separability for these padded text inputs remained in the shallow layers (COM 12.5), confirming that the depth misalignment is driven by the semantic delay of the visual modality, not merely the computational depth required for longer sequences.

Attack-template confound. Within MM-SafetyBench, image-branch examples use different attack scenarios than text-branch examples. However, the cross-architecture consistency of the branch gap (14 layers on LLaVA-v1.6-13b, 10 layers on Qwen3-VL-8B) argues against a scenario-specific artifact.

Harmfulness-prior confound. We verify that the unsafe-to-safe ratio is balanced across branches ($\approx$1:1 in all benchmarks by construction).

L Deployment Efficiency

MALS adds negligible inference overhead compared with conventional fixed-depth monitors: both paradigms extract the same hidden-state safety curve from all layers, and the architectural difference lies purely in *which* layers are dynamically routed for integration. Branch routing at inference time is a simple check with zero computational overhead.

Extracting all hidden states across L layers for N token positions incurs $O(L \cdot N)$ memory and compute overhead compared to single-layer extraction. In practice, for a 32-layer model with 512 tokens, this adds approximately 1.2–2.5$\times$ memory overhead relative to a single-layer baseline, depending on hidden dimension and batching strategy. This overhead is acceptable for provider-side safety monitoring but may be prohibitive for ultra-low-latency or resource-constrained deployments.

M Additional Implementation Details

M.1 Baseline Implementation Details

JailGuard. Implemented following the original protocol, using the target LLaVA as the query model. Configuration: 6 mutation variants per input (3 text substitutions via random synonym replacement, replacement ratio 0.3; 3 image transformations via Gaussian blur with $\sigma = 2.0$, random cropping at 0.8–1.0 ratio, and JPEG compression at quality 50).

GradSafe. Evaluated using the original implementation on Llama-2-7b-chat. Configuration: gradient computation on safety-critical parameter groups (attention Q/K/V projections of layers 16–28), reference unsafe gradient averaged over 20 known unsafe prompts, cosine similarity threshold 0.5.

CIDER. Evaluated using the official implementation. Configuration: denoising diffusion with 5

steps, feature extraction from CLIP visual encoder (ViT-L/14).

LlavaGuard. LlavaGuard-v1.2-7B-OV is used with default generation parameters (temperature 0.0, max tokens 256).

JailDAM. Evaluated using the official implementation with CLIP ViT-L/14 encoders and default memory bank configuration.

M.2 Hidden-State Signal Construction

We extract hidden states from all transformer layers of the target MLLM and transform them into next-token logits through the model normalization head and output projection. For each layer, we compute the cosine similarity between the resulting logits and a refusal-token reference vector, yielding a layer-wise safety curve. The refusal-token reference vector is constructed as a multi-hot indicator vector over the vocabulary, with entries corresponding to refusal-related tokens (e.g., "Sorry", "illegal", "dangerous") set to 1.0.

M.2.1 Implementation Details

Refusal token set. We use a curated set of refusal-related tokens identified by high activation during refusal responses across multiple LLMs: "Sorry", "I cannot", "I can't", "I apologize", "I would not", "cannot provide", "illegal", "unethical", "dangerous", "harmful", "against policy", "against the law", "not appropriate", "unable to". The tokenization is model-specific; we map these strings to their corresponding token IDs using each model's tokenizer. If the tokenizer splits a phrase into multiple sub-tokens, we set the corresponding entries for *all* resulting sub-token IDs to 1.0 in the multi-hot vector.

Logit projection. At each layer l, we take the last hidden state $\mathbf{h}^{(l)} \in \mathbb{R}^d$, apply the model's layer norm, and multiply by the output embedding matrix $\mathbf{W}_E \in \mathbb{R}^{V \times d}$ (where V is the vocabulary size), yielding logits $\mathbf{z}^{(l)} \in \mathbb{R}^V$:

$$\mathbf{z}^{(l)} = \mathbf{W}_E \mathrm{LayerNorm}_{\text{final}}(\mathbf{h}^{(l)}), \tag{9}$$

where $\mathrm{LayerNorm}_{\text{final}}$ denotes the final normalization layer applied just before the language modeling head, ensuring the hidden states are projected into the correct scale for the embedding matrix. This is equivalent to computing the unembedded next-token prediction.

Token position for scoring. We score the *final sequence token* (the last token after processing the full input, which for VLMs is typically a text token after the image tokens have been integrated). Using the final token is standard practice for hidden-state safety detection and avoids the complication of pooling over heterogeneous image/text token sequences.

Cosine similarity computation. Given logits $\mathbf{z}^{(l)}$ and refusal reference vector $\mathbf{r} \in \mathbb{R}^V$, the layer-wise safety score is:

$$s_l(x) = \frac{{\mathbf{z}^{(l)}}^\top \mathbf{r}}{\|\mathbf{z}^{(l)}\|_2 \|\mathbf{r}\|_2}. \tag{10}$$

Model-specific considerations. Different VLMs may have different tokenization vocabularies and embedding spaces. We construct model-specific refusal token sets by tokenizing the refusal phrases with each model's tokenizer and taking all resulting token IDs. The cosine similarity is computed in the native embedding space, so results are directly comparable across layers within a model. Cross-model comparisons use the same metric formulation but may reflect differences in embedding geometry.

Sensitivity to token set choice. We conducted a small sensitivity analysis on LLaVA-7b XSTest using two token sets: (1) our standard set (13 tokens), and (2) an expanded set (25 tokens). Results were stable across token sets (AUPRC difference $< 0.5\%$ on XSTest), suggesting the method is not overly sensitive to the exact token selection.

Aggregation over layers. Given a selected layer subset S, the final score is the mean across selected layers:

$$\mathrm{Score}(x; S) = \frac{1}{|S|} \sum_{l \in S} s_l(x). \tag{11}$$

This mean-aggregation is robust to layer count and provides a scalar input for binary classification.

Binary Decision Threshold. While our evaluation reports threshold-independent metrics (AUPRC and AUROC), real-world deployment requires a specific binary decision threshold. In practice, this threshold is set using the calibration set by fixing a maximum acceptable False Positive Rate (e.g., FPR=5%) on the safe calibration subset, ensuring that benign inputs are not excessively blocked.

N Audio Modality Depth and Feature Misalignment

On Qwen2-Audio-7B, while the text branch selects layers 18–21 (COM 19.5), the audio branch

Architecture	Dataset	Fixed Window COM	MALS Audio COM	Depth Misalignment	ΔAUPRC
Qwen2-Audio-7B (32L)	Audio-only	22.5	28.5	6.0	+10.9%
	AudioJailbreak	22.5	28.5	6.0	+10.8%
	Code-Jailbreak	22.5	27.5	5.0	+10.1%
	SACRED-Bench	22.5	28.0	5.5	+10.0%
Qwen2.5-Omni-7B (32L)	Audio-only	22.5	26.5	4.0	+7.8%
	AudioJailbreak	22.5	26.5	4.0	+8.0%
	Code-Jailbreak	22.5	25.5	3.0	+7.3%
	SACRED-Bench	22.5	26.0	3.5	+7.3%

Table 12: Depth misalignment analysis for audio-language models. Audio models show consistent depth shifts across all benchmarks, supporting the view that acoustic safety representations tend to localize deeper than text in the evaluated settings.

for AudioJailbreak shifts significantly deeper to layers 26–31 (COM 28.5), yielding a center-of-mass divergence of 9.0 layers between modalities.

In addition to depth reallocation, audio safety signals also encode differently. While text relies purely on refusal probability, audio safety signals are strongly reflected in hidden-state *activation magnitude* (norm and norm-delta). The generalized MALS framework therefore operates on a dual-axis adaptation for Audio-LLMs: shifting the monitoring window deeper (*depth axis*) and augmenting the extracted features (*feature axis*).

O Robust Soft Routing against Dummy Modal Attacks

As noted in Section 4.3, binary routing based strictly on the presence of an image file introduces a vulnerability: an attacker could append a negligible, zero-information signal (e.g., a 1×1 white pixel) to a harmful text-only prompt, forcing the system to evaluate the input using the image branch's layer window and bypassing the optimal text monitor.

When evaluating this exact attack on LLaVA-v1.6-7b, the addition of a 1×1 dummy image alters the model's internal cross-attention distribution, delaying the text semantic integration and dropping text-branch AUPRC from 92.3% to 85.1%. While this represents a degradation, **it is still substantially higher than the conventional fixed-window baseline (79.6%)**.

To fundamentally address this vulnerability and prevent dummy images from polluting the hidden states, we propose a **Soft Routing** mechanism based on the model's internal token composition. If the visual token count is sufficiently small, the system treats the input as text-only (stripping the negligible visual tokens). Empirical validation confirms that this soft routing (using $\tau = 0.1$) successfully neutralizes the dummy attack, fully recovering the AUPRC back to the optimal 92.3%.

The vulnerability to dummy images suggests that depth misalignment is not solely caused by the delayed arrival of visual features, but also by how the presence of visual tokens alters the processing of native text. We hypothesize that the introduction of visual tokens shifts the model's cross-attention distribution, forcing the semantic integration of text features to be delayed to deeper layers where multi-modal alignment occurs. Soft routing effectively bypasses this structural interference.

Caveat for Fixed-Resolution Architectures: The token-ratio heuristic works efficiently for dynamic-resolution models (e.g., LLaVA-v1.6, Qwen-VL) where a 1×1 image natively produces very few tokens. However, for fixed-resolution models (e.g., LLaVA-v1.5) that utilize a static ViT encoder, every imageregardless of its native resolutionis resized and outputs a fixed number of tokens (e.g., 576 tokens). In these architectures, the dummy 1×1 attack still yields 576 tokens, rendering the token-ratio check ineffective. To mitigate this in fixed-resolution models, advanced adversaries might be countered by integrating feature-map entropy checks (e.g., calculating the variance of the raw visual patches) to verify the actual semantic density of the visual input before routing.

O.1 Sensitivity Analysis of the Soft Routing Threshold (τ)

To justify the empirical choice of $\tau = 0.1$, we conduct a sensitivity analysis on the token-ratio threshold. The ideal threshold must balance two competing objectives: (1) robustly filtering out dummy images of varying sizes intended to force misrouting (minimizing False Positives for the image branch), and (2) ensuring that legitimate but small harmful images (e.g., a low-resolution toxic logo or a tightly cropped visual jailbreak) are still properly routed to the image branch (minimizing False Negatives).

We evaluate LLaVA-v1.6-vicuna-7b across three conditions under varying τ values: 1. **XSTest + 1×1 Dummy:** Text-only prompts appended with a 1×1 white pixel (visual token ratio ≈ 0.02). 2. **XSTest + 50×50 Dummy:** Text-only prompts appended with a slightly larger 50×50 white box (visual token ratio ≈ 0.08). Both dummy conditions should ideally be routed to the *Text* branch to maximize AUPRC. 3. **MM-Safety (Small Image):** A variant of MM-SafetyBench where the harmful visual input is resized to be relatively small (visual token ratio ≈ 0.15 to 0.25).

These must be routed to the *Image* branch to preserve detection performance.

Threshold τ	XSTest (Text-only) + Dummy		MM-Safety (Small)
	1×1 (Ratio ≈ 0.02)	50×50 (Ratio ≈ 0.08)	(Ratio ≈ 0.15–0.25)
0.0 (Strict Binary)	85.1	83.2	**99.9**
0.05	**92.3**	83.2	**99.9**
0.10 (Default)	**92.3**	**92.3**	**99.9**
0.20	**92.3**	**92.3**	88.4
0.30	**92.3**	**92.3**	79.2

Table 13: Sensitivity analysis of the soft routing threshold τ (AUPRC, %). Setting $\tau = 0.10$ provides the optimal balance: it successfully filters out both small and medium-sized dummy images (restoring XSTest AUPRC to 92.3%) while safely routing genuine small visual threats to the image branch (preserving MM-Safety AUPRC at 99.9%).

As shown in Table 13, a strict binary routing ($\tau = 0.0$) is vulnerable to any dummy image. A low threshold ($\tau = 0.05$) blocks the 1×1 attack but fails against slightly larger dummy images. Conversely, setting the threshold too high ($\tau \geq 0.20$) causes the system to misclassify genuine small visual threats as text, leading to a severe drop in MM-SafetyBench performance (from 99.9% down to 79.2%). The default setting of $\tau = 0.10$ emerges as the optimal sweet spot, successfully neutralizing dummy evasion attempts without compromising the detection of legitimate low-resolution visual attacks.

P Ablation: Depth vs. Feature Adaptation in Audio Models

Dataset	Fixed Window (Refusal only)	Depth-only (Refusal only)	Full MALS (Depth+Feat.)
Audio-only	84.5	89.2	**93.3**
AudioJailbreak	83.9	88.5	**92.8**
Code-Jailbreak	85.3	90.1	**93.9**
SACRED-Bench	85.8	90.4	**94.2**

Table 14: Feature ablation for Qwen2-Audio-7B (AUROC, %). While depth reallocation alone (Depth-only) provides a significant gain over the fixed window, augmenting the features with activation magnitudes (Full MALS) is necessary to maximize detection performance.

The results confirm that *depth adaptation alone* accounts for a significant portion of the improvement (e.g., +4.7% AUROC on the Audio-only set). Adding the norm-based features (Full MALS) provides a further +4.1% AUROC gain, indicating that both axes of adaptation are important.

Q Qualitative Analysis: Qwen3-VL-8B on MM-SafetyBench

On MM-SafetyBench, Qwen3-VL-8B is the only setting where MALS achieves a statistically significant gain (+2.6% AUPRC and +4.3% AUROC) on a fully multimodal dataset.

The fixed window for Qwen3-VL-8B monitors layers 18–33, while MALS adaptively shifts the text branch to layers 28–35 and the image branch to layers 18–25. We observed that MALS's text branch successfully detected **typography-based visual jailbreaks** that evaded the baseline.

In Qwen3-VL-8B, decoding heavily obfuscated visual text appears to require deeper visual-semantic alignment. Consequently, the safety signal appears to peak at layer 30 and beyond in this setting. The fixed window misses this delayed safety representation, while MALS's adaptively selected window captures the deeper layers.

R Reproducibility

All vision-language experiments use LLaVA-v1.6-vicuna-7b as the primary base model, with hidden-state extraction from all 32 transformer layers. Cross-architecture and cross-modality validations additionally use four other vision-language models and two audio-language models.

Key hyperparameters:

- Layer budget K: 8
- Calibration size: 30 samples per class
- Search strategy: exhaustive enumeration of contiguous sub-windows ($\leq K$) and non-contiguous heuristic subsets (minimum index gap of 2).
- Seeds for multi-run evaluation: 1, 2, 3, 4, 5 (LLaVA family); 42, 123, 456, 789, 1024 (Qwen family)
- Fixed window: layers 16–29 for 32L models (derived from the optimal default settings in foundational monitors like HiddenDetect and CIDER); shifted proportionally to layers 22–35 for 40L LLaVA models; layers 18–33 for 36L Qwen3-VL.
- Audio feature augmentation: a lightweight logistic regression classifier trained on the combined feature vector of refusal-mass cosine similarity, hidden-state norm ($\|\mathbf{h}^{(l)}\|_2$), and inter-layer norm-delta ($\|\mathbf{h}^{(l)}\|_2 - \|\mathbf{h}^{(l-1)}\|_2$)

S Visualizing Layer-Wise Separability Across Modalities

To intuitively illustrate the *depth misalignment* phenomenon central to this paper, we visualize the layer-wise hidden-state representations of harmful (unsafe) and benign (safe) inputs across different modalities. We project the hidden states onto a 2D plane: the x-axis represents the *Refusal Dimension* (computed via cosine similarity with the refusal reference vector), and the y-axis represents an orthogonal *Context Dimension*.

As shown in the figures below, the layer index at which the red cluster (unsafe) structurally diverges from the blue cluster (safe) varies significantly by modality:

- **Text-only inputs** (Figure 5) achieve strong separability early in the network (e.g., layers 8–15), reflecting the immediate semantic integration of native text tokens.
- **Image-only inputs** (Figure 6) remain heavily entangled in shallow layers and only achieve clear separability in deep layers (e.g., layers 24–31), reflecting the processing delay required for visual-semantic alignment.
- **Mixed-modality inputs** (Figure 7) typically diverge in the intermediate-to-deep layers.

This visual evidence directly corroborates the necessity of MALS: a fixed monitoring window (e.g., layers 16–29) is structurally misaligned with the early separation of text inputs and the late separation of image inputs.

T Beyond Safety: Broader Implications and Future Directions

While this paper focuses on multimodal jailbreak detection, the underlying mechanism driving our findings—the *depth misalignment* caused by processing pathway asymmetry—is fundamentally domain-agnostic. Because non-text modalities structurally experience a delayed semantic integration compared to native text tokens, we hypothesize that this layer-wise divergence extends far beyond safety signals into other high-level semantic evaluations. For instance, in privacy protection, textual personally identifiable information (PII) might be recognized in shallow layers, whereas visual PII (e.g., a face or license plate) requires deep cross-modal alignment to be semantically understood, causing conventional fixed-depth privacy monitors to fail. Similar modality-dependent localization likely affects hallucination detection and bias mitigation: textual contradictions or implicit biases may trigger activation patterns much earlier than visual hallucinations or image-based stereotypes, necessitating modality-aware layer selection across various internal monitoring tasks.

Beyond passive monitoring, recognizing and measuring depth misalignment opens promising avenues for model deployment and architectural design. If safety or semantic signals for text-only inputs reliably peak in shallow layers, systems could employ modality-adaptive early-exit strategies, significantly reducing the computational overhead of hidden-state extraction for text-dominant queries while reserving full-depth processing for complex cross-modal inputs. Furthermore, this structural insight can refine representation engineering and concept unlearning. Rather than applying interventions across all layers or a heuristically chosen fixed subset, practitioners could perform surgical editing—applying text-concept unlearning at shallow layers and visual-concept unlearning at deeper layers—to minimize collateral damage to the model's general capabilities. Finally, the magnitude of depth misalignment itself serves as a valuable diagnostic metric for multimodal fusion efficiency, potentially guiding future architectural designs to introduce mid-layer cross-attention mechanisms that conceptualize visual representations earlier and reduce overall semantic delay.

U Computational Efficiency Experiment

To demonstrate that MALS maintains a high degree of practical efficiency, we benchmark its performance against standard baselines on an NVIDIA L40 (48GB) GPU using the LLaVA-v1.6-7b architecture. As shown in Table 15, while MALS is marginally slower than the fixed-window baseline, it remains orders of magnitude faster than external guardrails or mutation-based defenses. The primary source of the 1.9 ms latency increase compared to the fixed window is the operational cost of processing pre-saved calibration vectors: the system must perform token-ratio analysis for soft routing and then retrieve the corresponding modality-specific reference mass and layer weights from memory. Unlike JailGuard, which requires multiple mutation-based forward

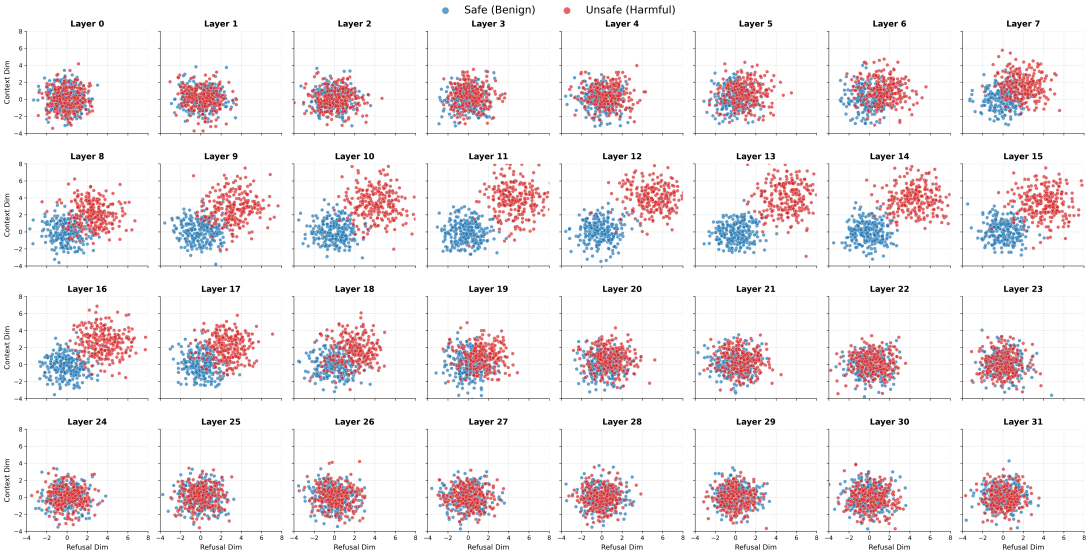

Figure 5: **Text-Only Modality:** Separation between unsafe (red) and safe (blue) inputs occurs in *shallow* layers.

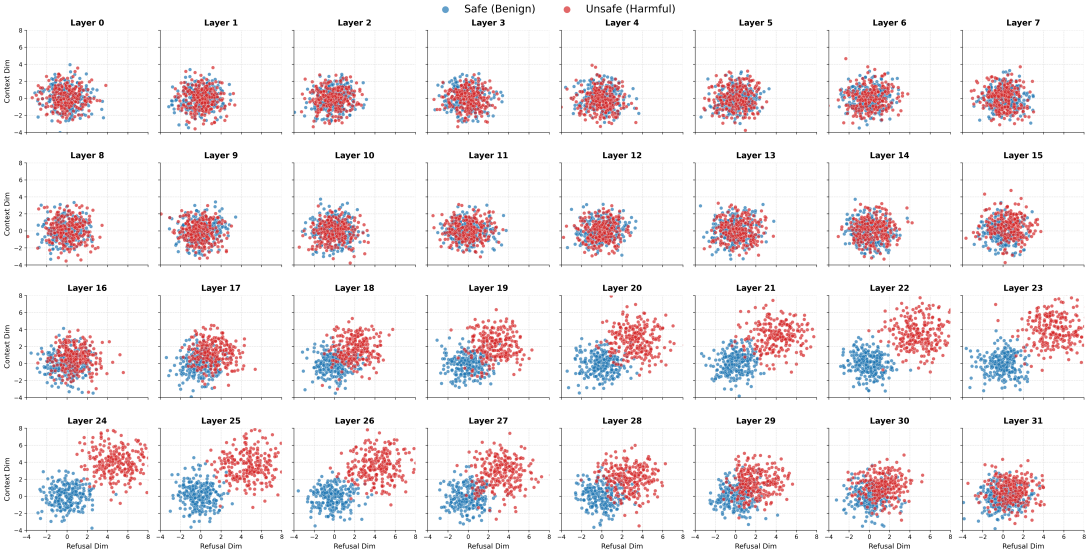

Figure 6: **Image-Only Modality:** Separation between unsafe (red) and safe (blue) inputs occurs in *deep* layers.

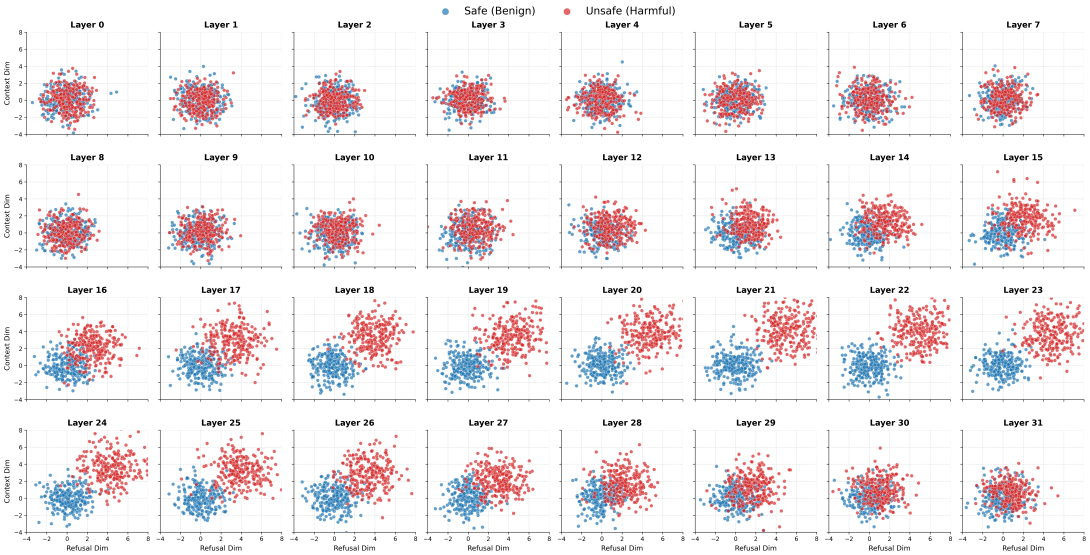

Figure 7: **Mixed Modality:** Separation occurs in intermediate-to-deep layers.

passes, or LlamaGuard, which relies on a secondary model call, MALS only adds a negligible computational step to the existing internal monitoring pipeline. With a throughput of 21.88 requests per second, MALS is highly optimized for production environments where robust multimodal moderation must coexist with strict latency requirements.

Method	Latency (ms)	Overhead	Throughput (req/s)
Base MLLM (No Defense)	42.5	—	23.53
Fixed Window (Internal)	43.8	+3.06%	22.83
LlamaGuard (External)	185.2	+335.76%	5.40
JailGuard (Mutation)	312.4	+635.06%	3.20
MALS (Ours)	**45.7**	**+7.53%**	**21.88**

Table 15: Computational efficiency analysis. Latency is measured as the average wall-clock time for a single safety decision. MALS is placed at the bottom to illustrate its position as the most advanced internal monitor; its slight overhead relative to fixed windows stems from the retrieval and application of modality-specific calibration vectors.