REPRESENTATION–BEHAVIOR ALIGNMENT FOR EXPLAINABLE WEAKLY-SUPERVISED VIDEO ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) provide a natural way to make video anomaly detection more explainable. However, their final decisions do not always fully use the discriminative information contained in their hidden states, an issue we refer to as representation–behavior misalignment. We decompose this gap into a capacity component that measures discriminative information never aggregated into the readout position, and a directional component that measures the angular mismatch between the optimal and the native normal–abnormal axis at that position. Across multiple video anomaly detection benchmarks and MLLM backbones the directional component dominates, and residual-stream tracing shows that native-axis separability rises sharply in several mid-to-late attention layers. Because both components are governed by attention rather than MLP updates, we propose Representation–Behavior Alignment (RBA), a parameter-efficient method that adapts those layers using video-level labels alone while updating about 0.012% of the backbone parameters. Experiments on three benchmarks show that RBA improves native-readout performance and better aligns the model's decision direction with discriminative representations, and it produces anomaly decisions and explanations through a single generative process.

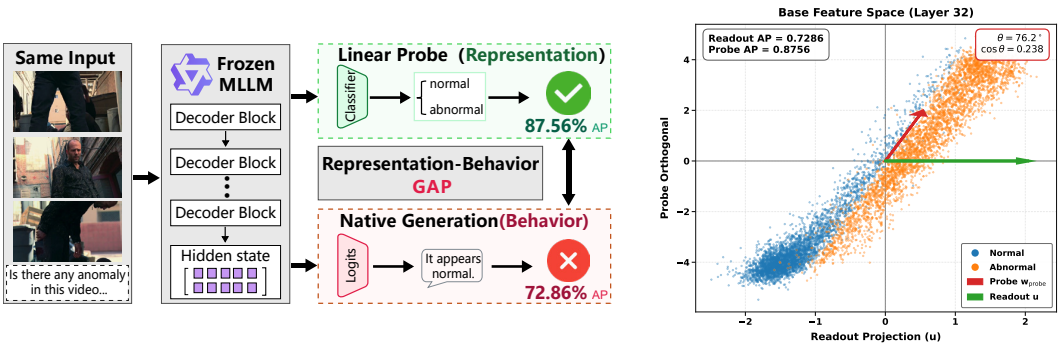
1 INTRODUCTION

Multimodal Large Language Models (MLLMs) unify visual understanding, decision making, and natural language explanation in one interface. Yet the decisions they generate do not necessarily reflect the task-relevant information their internal representations already hold. Work on linear probing and latent prediction has shown that frozen hidden states often encode task signals the output interface cannot reliably express [2, 3, 5, 4]. We examine this representation–behavior gap in weakly-supervised Video Anomaly Detection (VAD), asking how much of the performance loss is tied to the native readout direction, where native-readout alignment emerges across depth, and whether one targeted parameter-efficient intervention can mitigate it.

Why Video Anomaly Detection. VAD suits this question well. Judging a clip as anomalous depends on how a short span of frames evolves and on the surrounding scene, so no single frame settles it. Training data typically carry one label per video, yet a useful system must localize anomalies and explain its judgment in language. Existing methods cover only part of this need. Weakly-supervised detectors [25, 28] are accurate but sit outside the language-generation pathway, training-free MLLMs explain their outputs yet lack precision, and instruction-tuned MLLMs [33] lean on costly frame annotations or reasoning traces. We fix a four-frame clip window throughout. That caps the absolute accuracy any readout can reach, but it applies identically to every readout we compare, so the gaps we report are internal to the model rather than a property of the input window.

Core Diagnosis. Evaluating the frozen final-layer features under identical inputs reveals a clear divergence between decision pathways. The native readout reaches only 0.7286 Frame-AP on XD-Violence, whereas the best one-dimensional readout on the same state attains 0.9588 and a weakly supervised linear probe already recovers 0.8756, as depicted in Figure 1(a). Chaining two controlled comparisons attributes 97.5% of that gap to the readout direction rather than to information missing

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



(a) Representation–Behavior Gap (b) Directional Misalignment

Figure 1: Diagnosis of representation–behavior misalignment on XD-Violence. (a) Native readout versus one-dimensional readouts on frozen hidden states. (b) 2D projection showing directional misalignment between the Fisher-optimal axis $\mathbf{w}_{\mathrm{LDA}}$ and the native readout axis $\mathbf{u}$.

from the readout position. Null-space projections agree, since stripping out the frame-supervised direction $\mathbf{w}_{\mathrm{LDA}}$ degrades accuracy substantially. Figure 1(b) shows this optimal direction sitting 76.2° away from the native readout axis $\mathbf{u}$, a cosine similarity of only 0.238.

Tracking how well the residual stream separates normal from abnormal clips along the native direction locates where that alignment appears. It emerges abruptly at an attention sublayer in the middle-to-late network. Separability more than triples at Block 20 and then levels off, with attention updates improving the alignment and MLP updates weakening it. The probe, meanwhile, is already near its final performance before the transition. Useful anomaly features thus form early, but the model connects them to its language readout late and only partially.

Method and Contributions. We propose Representation–Behavior Alignment (RBA) in response. RBA adds only about 0.012% trainable parameters to the attention sublayers pinpointed by the analysis, and uses pseudo-labels from the probe to push representations toward the native readout direction. A single forward pass at test time then yields anomaly predictions together with explanations that match those decisions. Our contributions can be summarized as follows.

- **Geometric diagnosis for VAD.** Comparing readouts that differ in one constraint at a time shows that projection direction, not the information available at the readout position, accounts for nearly all of the observed performance gap. Separating the actionable directional term from the capacity term reveals that the missing performance reflects a misoriented native axis rather than uninformative features.
- **Layer-wise attribution.** We locate the sharp change in native-axis separability at specific attention sublayers and distinguish the roles of the attention and MLP modules. This attribution turns a diffuse "the model underperforms" observation into a data-driven criterion for which layers to adapt.
- **A diagnosis-guided, lightweight intervention.** With video-level labels alone and 0.012% trainable parameters, RBA rotates the discriminative direction toward the native readout axis at the identified layers. It delivers consistent gains across three VAD benchmarks and several MLLM backbones while preserving native decision explainability.

2 Related Work

Reading Internal Representations. Linear probing [2, 3], together with unembedding-projection tools such as Logit Lens [4, 21] and Representation Engineering [35], has shown that the internal representations of language models often encode more task-relevant information than their explicit textual outputs reveal. Most of this work, however, remains at the level of passive diagnostic observation. While these studies confirm that such a performance discrepancy exists, they pay less attention to geometrically decomposing the gap into direction and capacity components, to localizing where it emerges across network depth, or to deriving targeted parameter-efficient interventions

from those findings. In this work, we extend passive inspection into an actionable pipeline that covers diagnosis, layer-wise attribution, and a minimal intervention for MLLM-based decision tasks.

Video Anomaly Detection. Existing Video Anomaly Detection (VAD) paradigms fall broadly into two categories. Weakly-supervised specialized detectors [25, 28], typically built upon Multiple Instance Learning (MIL), achieve strong detection accuracy on standard benchmarks. Their scalar predictions, however, bypass the generative language pathway and decouple textual explanations from the detection mechanism. In contrast, MLLM-based VAD frameworks comprise training-free inference [31] and instruction tuning [33]. Training-free approaches unify judgment and explanation but suffer from lower precision due to readout misalignment. Instruction tuning improves accuracy yet demands dense frame annotations or reasoning chains that conflict with realistic weakly-supervised settings. Our proposed RBA reformulates this challenge as a directional alignment problem along the native readout axis. Supervised solely by video-level labels, it performs minimal low-rank adaptation to deliver accurate detection and consistent explanations with no external inference classifier.

3 Representation–Behavior Misalignment: Diagnosis and Attribution

Hidden states in MLLM-based VAD can carry discriminative information that the generated answer never uses, a shortfall we call *representation–behavior misalignment*. To measure it, we compare readouts that differ in exactly one constraint at a time, namely how much representational capacity a readout may use and which direction it must read along. This yields a two-term accounting of the gap, and residual-stream tracing then locates the depth at which each term is resolved.

3.1 Formalization and Two-Term Gap Decomposition

Task Setup and Readout Position. For a video $\mathcal{V}$, each clip i is placed in a fixed prompt template, so every clip yields a token sequence of the same length N built from prefix tokens, the visual tokens of four frames, and the question tokens. Let $\tilde{\mathbf{h}}_{L,i}^{(p)} \in \mathbb{R}^d$ denote the RMSNorm-normalized final-layer residual state at position p [32], the state produced after the last MLP update and read by the output head. We read at the final position $p = N$, a single prompt token, and abbreviate $\tilde{\mathbf{h}}_{L,i} := \tilde{\mathbf{h}}_{L,i}^{(N)}$. We write $\mathbf{a}_{l,i}$ for the attention-sublayer output of block l at that position, the only update in the block able to aggregate information from other token positions into it. The model's *native readout axis* $\mathbf{u}$ is the unit direction in the pre-trained output embedding matrix pointing from the "Normal" token to the "Abnormal" token,

$$\mathbf{u} = \frac{\mathbf{W}_U[\text{Abnormal}] - \mathbf{W}_U[\text{Normal}]}{\|\mathbf{W}_U[\text{Abnormal}] - \mathbf{W}_U[\text{Normal}]\|_2}, \tag{1}$$

and the projection $d_i = \mathbf{u}^\top \tilde{\mathbf{h}}_{L,i}$ is the native abnormal-versus-normal score of clip i. Every readout below shares the same prompt, clips, visual input, and network weights, so only the available positions and the reading direction vary. Properties of the input itself, including the temporal window length, are held fixed and cannot account for differences among them.

Three Readout Tiers. The *sequence-level reference* P_{full} fits linear discriminant analysis on the concatenation of final-layer states at *every* position,

$$\mathbf{g}_{L,i} = \left[\tilde{\mathbf{h}}_{L,i}^{(1)}; \tilde{\mathbf{h}}_{L,i}^{(2)}; \ldots; \tilde{\mathbf{h}}_{L,i}^{(N)}\right] \in \mathbb{R}^{Nd}, \tag{2}$$

exposing the visual and prompt tokens alongside the readout token. It bounds the linearly extractable signal in the whole sequence whether or not the model has routed that signal into the readout token, though not the nonlinear information the sequence may also hold. The *position-constrained reference* P_{last} fits Fisher LDA on the single last-position state with closed-form optimum $\mathbf{w}_{\text{LDA}} = \boldsymbol{\Sigma}_w^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$, the best one-dimensional readout once the signal has been compressed into the one token the output head consumes. The *native generation readout* P_{read} projects that same state onto the fixed direction $\mathbf{u}$. Both references use held-out frame annotations retrospectively, so they are diagnostic quantities rather than attainable operating points.

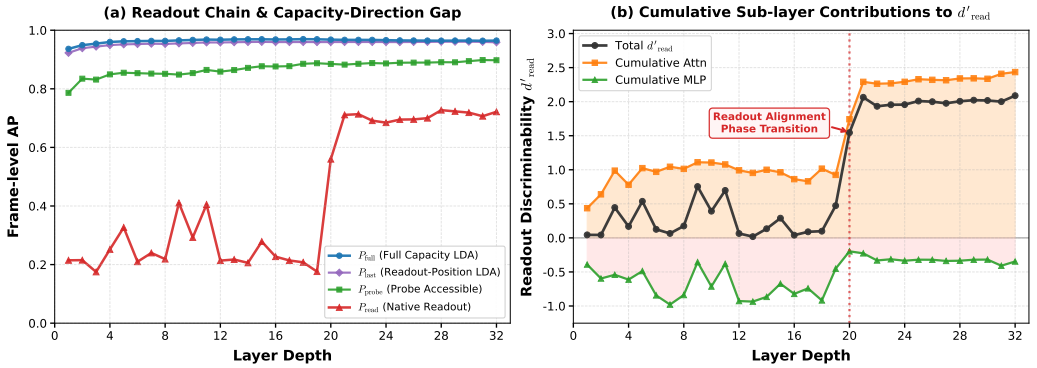


Figure 2: Layer-wise residual-stream diagnostic. The two Fisher references nearly coincide at every depth and saturate early, whereas native separability increases sharply at the Block 20 attention sublayer. Additive decomposition isolates positive gains from attention and negative interference from MLP sublayers.

Decomposition. Comparing P_{full} with P_{last} varies the available positions while both sides use the Fisher-optimal direction, and comparing P_{last} with P_{read} fixes the position and varies only the direction. Chaining the two gives

$$\Delta_{\text{total}} = P_{\text{full}} - P_{\text{read}} = \underbrace{\left(P_{\text{full}} - P_{\text{last}}\right)}_{\Delta_{\text{cap}}\ \text{(Capacity / Aggregation Term)}} + \underbrace{\left(P_{\text{last}} - P_{\text{read}}\right)}_{\Delta_{\text{dir}}\ \text{(Directional Term)}}, \tag{3}$$

where Δ_{cap} measures discriminative information present somewhere in the sequence, notably in the visual tokens, yet unrecoverable from the readout position alone, and Δ_{dir} measures the cost of reading that position along $\mathbf{u}$ rather than $\mathbf{w}_{\text{LDA}}$. Equation 3 is exact by construction and claims no statistical independence between the terms. Its value lies in the controls it enforces, since each term isolates one varied constraint while the shared intermediate P_{last} holds the other fixed.

The Probe as an Estimator. Frame labels are unavailable at training time, so $\mathbf{w}_{\text{LDA}}$ cannot be computed in the deployed setting. We therefore fit a Top-K Multiple Instance Learning (MIL) linear probe at the readout position as its weakly supervised estimator using only the video-level label Y_j,

$$\min_{\mathbf{w},b} \sum_{j} \mathcal{L}_{\text{BCE}}\Big(Y_j,\ \operatorname*{pool}_{i\in\text{Top-}K}\ \sigma(\mathbf{w}^\top \tilde{\mathbf{h}}_{L,i} + b)\Big). \tag{4}$$

The probe is not a tier in Equation 3. We judge it by its fidelity to the Fisher optimum, reported as $\cos(\mathbf{w}_{\text{probe}}, \mathbf{w}_{\text{LDA}})$ and $\varepsilon_{\text{probe}} = P_{\text{last}} - P_{\text{probe}}$, and it only produces offline pseudo-labels rather than being retained at inference. Because $\mathbf{g}_L$ has N times the dimension of $\tilde{\mathbf{h}}_L$, we reduce it by PCA fitted on the training split, estimate $\mathbf{\Sigma}_w$ with Ledoit–Wolf shrinkage, and fit every discriminant on the training split alone.

3.2 Geometric Attribution and Residual-Stream Readout Transition

The Directional Term Dominates. On XD-Violence with Qwen3.5-9B, the sequence-level reference reaches 0.9646 Frame-AP and the position-constrained reference 0.9588, whereas the native readout obtains only 0.7286. Within the total gap of 0.2360, the directional term accounts for 0.2302, or 97.5%, leaving 0.0058 to the capacity term. Projecting out the normalized LDA direction via $\mathbf{h}^{\perp} = (\mathbf{I} - \hat{\mathbf{w}}_{\text{LDA}}\hat{\mathbf{w}}_{\text{LDA}}^\top)\tilde{\mathbf{h}}_L$ causes a sharp drop, confirming that the linearly accessible signal concentrates along this single vector. The Fisher-optimal axis deviates from the native axis by $76.2°$, a cosine similarity of 0.238, while the weakly supervised estimator tracks that optimum closely at $\cos(\mathbf{w}_{\text{probe}}, \mathbf{w}_{\text{LDA}}) = 0.907$. The readout position is thus not where the signal is lost. Recoverable performance is forfeited mainly to a misoriented axis.

A Localized Readout Transition and the Opposition of Sublayers. Sweeping Equation 3 over all 32 blocks shows Δ_{cap} staying uniformly small and shrinking with depth, from 1.35 AP points at Block 1 to 0.35 at Block 28, so the two references in Figure 2 are nearly indistinguishable and saturate within the first few layers. Separability along $\mathbf{u}$, measured at 64 sampling points in the residual

stream, follows a very different trajectory. The statistic $d'_{\mathrm{read}}(l)$ jumps from 0.474 to 1.545 at the attention sublayer of Block 20, a rise of 1.07 with no gradual buildup beforehand. The representation supports anomaly discrimination long before the model can express that distinction through its own output tokens, which rules out late feature formation as an explanation.

The right panel of Figure 2 accumulates each attention or MLP residual update's contribution to native-axis separability. The attention curve takes a large positive step at the transition while the MLP curve stays negative. This asymmetry matches what the two operators can do, since attention alone moves information across token positions and can therefore reduce Δ_{cap} or write context-dependent content onto the readout axis, whereas MLP updates act per position and only rotate what is already there. We read the transition as a routing event, where discriminative content present elsewhere in the sequence since the early layers is written into the readout position and partially projected onto $\mathbf{u}$. This account describes the configurations we measure rather than a verified causal mechanism, and Appendix C adds further depth-wise analysis.

3.3 Diagnostically Derived Design Principles

Three design choices follow. First, attention is the operator implicated in both terms, since only attention moves information into the readout position and the same residual updates write the direction there, which makes attention projections at the transition our intervention point and motivates freezing the MLPs rather than leaving that choice arbitrary. Second, the signal that must be redirected occupies the small subspace spanned by $\mathbf{w}_{\mathrm{LDA}}$ and $\mathbf{u}$, so a low-rank update at the identified blocks suffices without reshaping general representations. Third, Δ_{dir} is an angular quantity and anomaly detection is a ranking problem, so we optimize the margin along the fixed native direction instead of matching the probe's probability distribution.

4 Methodology: Representation–Behavior Alignment

The goal of RBA is to align discriminative features with the model's native generation direction under weak supervision. Figure 3 summarizes the whole procedure: we first train a probe and generate pseudo-labels, then add low-rank adapters at the transition layers, and finally optimize the native decision margin along the fixed readout axis. The diagnosis in Section 3.2 is run before adapter training and determines *where* RBA acts, so it adds no inference-time component. RBA needs neither an external teacher nor frame annotations, and no separate classifier at inference.

4.1 Overall Pipeline

We split each video into temporal clips and place them in a fixed prompt made up of a prefix, visual tokens, and a question. In every stage, we extract a segment representation at the prediction-token position, which keeps the representation space consistent between probe training and alignment. Guided by the residual-stream diagnosis in Section 3.2, the adapters are inserted into the identified transition layer and the block that immediately follows it. In this way, the probe supplies the training signal while the residual-stream diagnosis supplies the adapter location.

4.2 Probe-Derived Internal Supervision

Let video j denote an instance bag with video-level label $Y_j \in \{0, 1\}$. We train a linear probe on the normalized final hidden states using Top-K MIL pooling:

$$\min_{\mathbf{w},b} \sum_{j} \mathcal{L}_{\mathrm{BCE}}\Big(Y_j, \operatorname*{pool}_{i \in \text{Top-}K} \sigma(\mathbf{w}^\top \tilde{\mathbf{h}}_{L,i} + b)\Big). \tag{5}$$

To mitigate class imbalance, we perform stratified class-balanced sampling strictly at the video level, preventing leakage between temporal segments from the same video.

After training, the probe produces a posterior score s_i for each segment. We threshold this score at 0.5 to obtain the pseudo-label $\hat{y}_i$, where -1 denotes normal and $+1$ denotes abnormal. These labels remain fixed while training RBA. We use only their direction, rather than the probe's potentially uncalibrated probability values. Notably, the probe is not retained as an inference-time head; it provides offline supervision only.

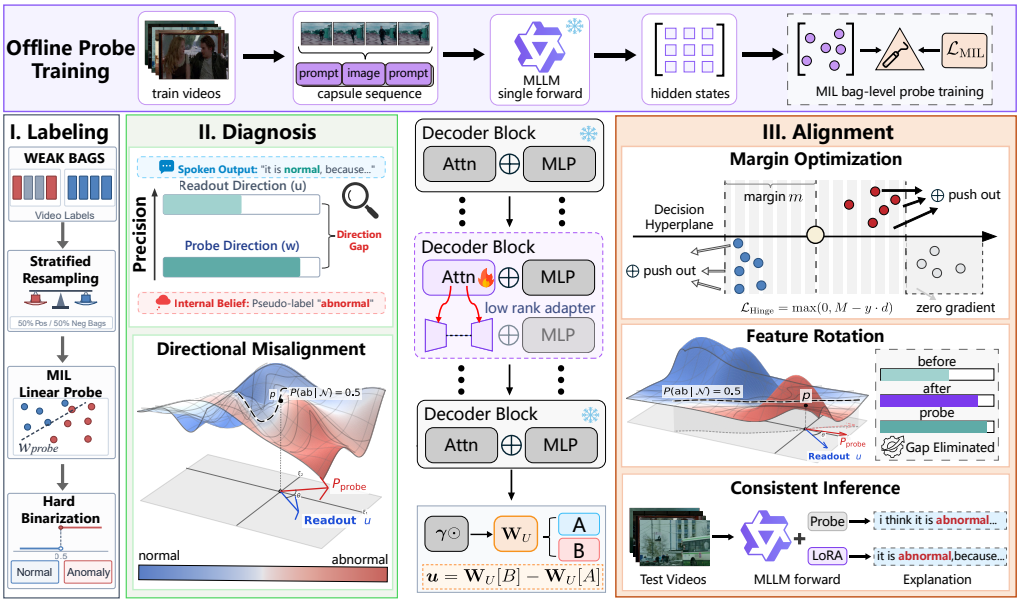


Figure 3: Overview of the RBA framework. Probe-derived pseudo-labels supervise low-rank adapters injected into the transition attention layers, optimized via a hinge-margin loss along the fixed native readout axis $\mathbf{u}$.

4.3 Targeted Low-Rank Adaptation

We add LoRA parameters [11] only to the attention projections in the transition layers. The following equation defines the adapted weight $\mathbf{W}'$ as the frozen base weight plus a low-rank update:

$$\mathbf{W}' = \mathbf{W}_{\text{base}} + \frac{\alpha}{r}\mathbf{B}\mathbf{A}, \qquad \mathbf{A} \in \mathbb{R}^{r \times d_{\text{in}}}, \quad \mathbf{B} \in \mathbb{R}^{d_{\text{out}} \times r}. \tag{6}$$

We use rank $r = 16$, scaling factor $\alpha = 32$, and apply LoRA to the attention projections in zero-indexed Blocks 20–21. The MLP sublayers and all remaining backbone parameters are frozen, restricting trainable parameters to roughly 0.012% of the backbone. The adapters can be merged into the base weights at inference without introducing any additional computational latency.

Low-Rank Geometric Motivation. Because our empirical diagnosis reveals that the critical discriminative signal occupies the low-dimensional subspace spanned by $\mathbf{w}_{\text{probe}}$ and $\mathbf{u}$, updating the entire backbone or reshaping general representations is unnecessary. Instead, lightweight low-rank modifications suffice to rotate internal features toward the native axis $\mathbf{u}$. We set the LoRA rank to $r = 16$ to accommodate cross-layer dynamics alongside inter-video variance. As reported in Table 2, this targeted intervention substantially increases alignment with the native readout direction.

4.4 Decision-Axis Margin Alignment Objective

During fine-tuning, the native decision score d_i is the abnormal-minus-normal token logit difference at the first answer position, equivalent to a projection onto the fixed native readout direction up to a positive constant scale. The unembedding matrix and the final RMSNorm remain frozen, ensuring that optimization changes the hidden representations rather than the classification axis itself.

We optimize the unweighted hinge-margin loss:

$$\mathcal{L}_{\text{align}} = \frac{1}{B}\sum_{i=1}^{B} \max\left(0, M - \hat{y}_i \cdot d_i\right), \tag{7}$$

where $M = 2.0$. Once a representation satisfies $\hat{y}_i d_i \geq M$, its gradient vanishes, concentrating the limited adapter capacity on samples that are incorrectly oriented or insufficiently separated.

We choose the hinge loss over cross-entropy or KL divergence because alignment calls for correct orientation and separation along the native axis, not replication of the probe's probability distribution. Compared with distribution-matching objectives, the hinge objective directly improves native-axis separation without assuming that the probe posterior is calibrated.

4.5 Capacity Isolation Principle

If we add an explanation loss $\mathcal{L}_{\text{explain}}$ to the same adapters, detection AP on XD-Violence drops from 0.8302 to 0.7279, completely erasing the alignment gain and falling slightly below the zero-shot baseline of 0.7286. This happens because the decision loss acts on a single direction, whereas the explanation loss affects many token positions and directions. When the adapter rank is only 16, the two objectives compete for capacity.

We call this the *capacity isolation principle*. During fine-tuning, $\mathcal{L}_{\text{align}}$ is evaluated only at the first answer position and updates only the selected attention adapters, and we do not add an explanation-generation loss. Explanations draw directly on the frozen generative ability of the pre-trained model. At inference, scores and explanations then come from the same aligned pathway.

5 Experiments

Three questions organize our study. Does the representation–behavior gap appear consistently across datasets and MLLM backbones? Can a diagnosis-guided low-rank update improve native anomaly scores from video labels alone? And do the gains stem from better directional alignment and layer placement rather than from added parameters?

5.1 Experimental Setup

Datasets and Evaluation Metrics. We evaluate on XD-Violence [27], UCF-Crime [23], and UB-normal [1], reporting Frame-AP and Frame-AUC for XD-Violence and Frame-AUC for the other two. Training uses binary video-level labels only, with no frame annotations or text descriptions. At test time we map each segment score to its covered frames and compare against the official frame labels. Probes, normalization statistics, and pseudo-labels are fitted on the training splits alone.

Implementation Details. Our primary backbone is Qwen3.5-9B [24], and we also apply RBA to InternVL3.5-8B [26] and Gemma-4-12B [9] to test cross-architecture generality. RBA uses LoRA [11] with rank 16 and scaling factor 32. Guided by the layer-wise diagnostics, we adapt the attention projections in zero-indexed Blocks 20–21 for Qwen3.5-9B, Blocks 25–26 for InternVL3.5-8B, and Blocks 33–34 for Gemma-4-12B. All other weights, including the output embedding and the final RMSNorm scale, stay frozen, leaving about 0.012% of the backbone trainable. We use AdamW [18] with learning rate 1×10^{-4}, batch size 32, and three epochs. Training takes under two hours on a single NVIDIA A100 GPU.

5.2 Main Benchmark Comparison

Table 1 compares RBA with recent VAD methods, which trade off detection accuracy, natural-language output, and weak supervision. Specialized detectors can be accurate yet cannot explain through their decision pathway, while interpretable MLLMs show weaker native readout accuracy or need extra fine-tuning. On the same backbone, RBA lifts the XD-Violence native-readout AP from 72.86% to 83.02% and the AUC from 90.37% to 94.30%. Among explanation-capable MLLMs it achieves the best result reported here on both XD-Violence and UCF-Crime, exceeding LRPO by 8.93 AP points on XD-Violence. On UBnormal it reaches 76.84% AUC, above LRPO and PANDA and close to Anom-π. Aligning the readout direction therefore improves native decisions under weak supervision while retaining language generation.

5.3 Generalization Across VAD Benchmarks and MLLM Backbones

Running the three-tier readout chain and the gap decomposition across backbone architectures and video domains yields a gap that is almost entirely directional in every configuration, with Δ_{dir}

Table 1: Quantitative comparison on VAD benchmarks (%). ✓ and ✗ denote native support for natural language explanations.

Method	Venue	Interpretable	UCF AUC	UB AUC	XD AUC	XD AP
Specialized Detectors (Non-interpretable)						
VadCLIP [28]	AAAI'24	✗	88.02	—	—	84.51
π-VAD [19]	CVPR'25	✗	**90.33**	—	—	85.37
RefineVAD [12]	AAAI'26	✗	88.92	—	—	88.66
LAVIDA [8]	CVPR'26	✗	82.18	76.45	—	**90.62**
HeadHunt-VAD [6]	AAAI'26	✗	87.03	—	—	82.63
PRISM [7]	ICML'26	✗	88.85	—	93.42	77.48
MLLM-based Methods (Interpretable)						
LLaVA-1.5 [16]	CVPR'24	✓	72.84	—	79.62	50.26
LAVAD [31]	CVPR'24	✓	80.28	51.06	85.36	62.01
VERA [30]	CVPR'25	✓	86.55	—	88.26	70.54
GLM-4.1V-9B [10]	-	✓	61.80	60.81	72.73	52.93
EventVAD [22]	ACM MM'25	✓	82.03	—	87.51	64.04
Unified-VAD [15]	NeurIPS'25	✓	84.28	69.02	91.34	68.07
VADTree [14]	NeurIPS'25	✓	84.74	65.80	90.55	68.85
PANDA [29]	NeurIPS'25	✓	84.89	75.78	—	70.16
Anom-π [20]	ICML'26	✓	84.46	**78.75**	—	72.29
LRPO [13]	ICML'26	✓	86.59	76.24	—	74.09
RBA (Ours)	-	✓	**87.01**	76.84	**94.30**	**83.02**

Table 2: Two-factor gap decomposition and alignment gains across benchmarks and backbones, reporting Frame-AP for XD and Frame-AUC for UCF/UB. P_{full} and P_{last} are retrospective Fisher ceilings used only for decomposition, whereas P_{probe} is the weakly supervised attainable ceiling against which the recovered fraction of Δ_{dir} is reported in the text. The final column gives the cosine between the Fisher-optimal axis and the native readout axis before and after adaptation.

Backbone / Dataset	P_{full}	P_{last}	P_{probe}	P_{read}	Δ_{cap}	Δ_{dir}	Dir. Ratio	Aligned	$\cos(\mathbf{w}_{\text{LDA}}, \mathbf{u})$
Qwen3.5-9B / XD	0.9646	0.9588	0.8756	0.7286	0.0058	0.2302	97.5%	**0.8302**	0.238 → **0.774**
Qwen3.5-9B / UCF	0.9168	0.9120	0.8847	0.7714	0.0048	0.1406	96.7%	**0.8701**	0.287 → **0.789**
Qwen3.5-9B / UB	0.8974	0.8910	0.8523	0.7290	0.0064	0.1620	96.2%	**0.7684**	0.263 → **0.746**
InternVL3.5-8B / XD	0.9251	0.9180	0.8734	0.6493	0.0071	0.2687	97.4%	**0.7842**	0.164 → **0.728**
Gemma-4-12B / XD	0.9386	0.9320	0.8938	0.6980	0.0066	0.2340	97.3%	**0.8058**	0.201 → **0.756**

Table 3: Adapter placement ablation on XD-Violence (Frame-AP).

Placement	Sublayer	Params	AP
Native Readout	—	0	0.7286
Blocks 5–6	Post-Attn	≈0.012%	0.7314
Blocks 17–18	Post-Attn	≈0.012%	0.7976
Blocks 20–21 (Ours)	**Post-Attn**	**≈0.012%**	**0.8302**
Blocks 30–31	Post-Attn	≈0.012%	0.7738
Blocks 20–21	Post-MLP	≈0.024%	0.7415

Table 4: Decision–explanation consistency under prefix-token intervention (%).

Method	Consist.	F^+	R^-	Quality
Zero-Shot Qwen	78.2%	94.5%	86.1%	3.63
	79.1%	93.8%	85.4%	3.57
Probe + Caption	43.2%	71.3%	94.4%	2.85
	44.0%	70.8%	93.9%	2.78
RBA (Ours)	**91.4%**	**96.3%**	**36.7%**	**4.12**
	92.0%	**95.9%**	**35.8%**	**4.09**

† First row: GPT-5.4 mini; second row: Gemini 3.6 Flash.

above 96% of Δ_{total} in all five settings and Δ_{cap} never exceeding 0.71 points. The cosine similarity between $\mathbf{w}_{\text{LDA}}$ and $\mathbf{u}$ starts below 0.29 everywhere and RBA reliably raises it past 0.72. Measured against the weakly supervised ceiling P_{probe}, the intervention closes 69.1% of the directional gap on XD-Violence, 87.1% on UCF-Crime, and 32.0% on UBnormal, alongside 60.2% and 55.1% reductions on InternVL3.5-8B and Gemma-4-12B. Accuracy and axis alignment improve together, which points to directional correction as the main driver of the recovery.

5.4 Ablation Studies and Mechanistic Verification

Diagnostic-Guided Adapter Placement. Table 3 shows that adapters in Blocks 20–21 recover most of the gain with 0.012% trainable parameters. Blocks 5–6 barely help at 0.7314 AP because the useful features have not yet formed. Blocks 17–18 reach 0.7976 and Blocks 30–31 reach 0.7738, a pattern favoring updates near the start of the transition. Adapting MLP sublayers gives only 0.7415 AP, consistent with attention carrying the positive alignment role.

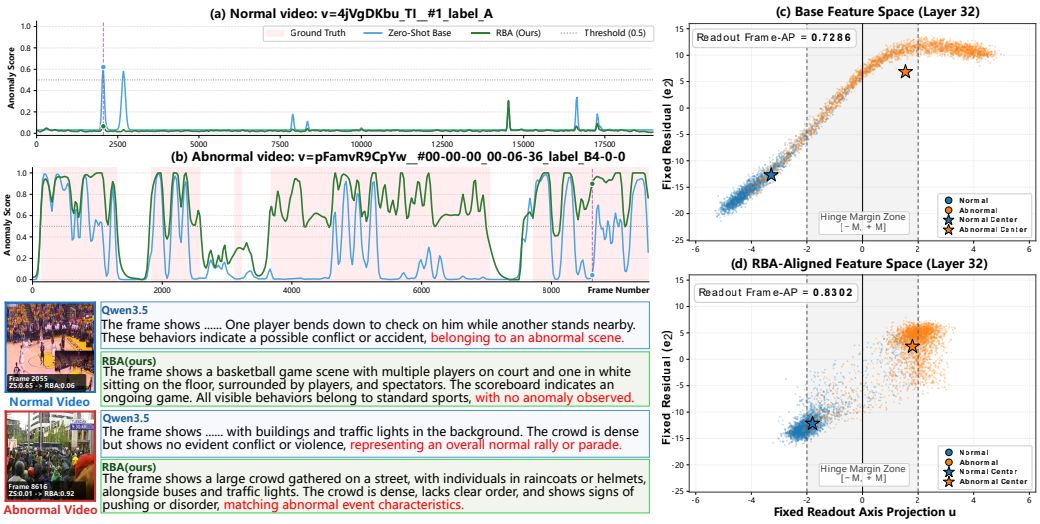


Figure 4: Qualitative and geometric analysis. (a)(b) Temporal anomaly scores and corresponding explanations before and after RBA. (c)(d) Final-layer feature distributions along the native readout axis $\mathbf{u}$, reflecting the downstream effect of adapting Blocks 20–21.

Decision–Explanation Consistency. We test whether explanations agree with decisions using an LLM-based protocol [17] on diagnostic sets of 10,000 clips each. We measure free-generation consistency, compliant faithfulness (F^+) on correct decisions, and adversarial rationalization (R^-) when a prefix forces a counterfactual decision. Table 4 reports 91.4% consistency for RBA, 96.3% F^+, and 36.7% R^-. Probe + Captioning attains only 43.2% consistency and rationalizes forced counterfactual decisions 94.4% of the time. A shared generative pathway keeps decisions and explanations consistent, and both judges agree on this ordering.

5.5 Qualitative Analysis and Geometric Evolution

Figure 4 shows scores, descriptions, and geometry improving together. In panels (a)(b), the base readout mistakes a player sitting on the floor for an injury and peaks at 0.65 on a normal basketball clip, whereas RBA drops the score to 0.06 and describes standard gameplay. On an abnormal crowd sequence the base model outputs only 0.01, while RBA raises the score to 0.92 and highlights dense pushing and protective helmets. Panels (c)(d) explain this geometrically by projecting Layer 32 representations along the fixed axis $\mathbf{u}$. The two classes heavily overlap inside the hinge band $[-M, +M]$ before alignment, and training the transition-layer adapters pushes representations past the margin boundaries and separates the class centroids cleanly. This redistribution recovers 69.1% of the directional gap relative to the weakly supervised ceiling P_{probe} and 44.1% relative to the retrospective ceiling P_{last}, with the remainder reflecting pseudo-label noise alongside the rank-16 and single-direction constraints. Updating attention projections at transition depth thus redirects features toward the unembedding head without any auxiliary inference classifier.

6 Conclusion

We studied the gap between what an MLLM represents and what it says in weakly supervised video anomaly detection. Across three VAD benchmarks and several MLLM backbones, linear probes on hidden states outperform the native normal–abnormal readout. Our single-axis analysis shows that most of the usable gap comes from a mismatch in readout direction, and residual-stream analysis locates the sharp rise in native separation at specific attention sublayers. Based on this result, RBA updates about 0.012% of the backbone parameters using video labels alone. It improves native anomaly detection while producing explanations through the same generative pathway. We believe representation–behavior alignment may prove equally useful in other multimodal decision tasks.

AI USE STATEMENT

In this work, we used generative AI tools for proofreading and improving text readability. We have not used generative AI tools for scientific idea generation or data simulation. All AI-assisted text was manually reviewed and verified by the authors. The authors take full responsibility for the content of this work.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. The datasets utilized are public benchmark datasets, and this work does not involve direct human subject experiments or known negative social/privacy impacts.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, all dataset preprocessing procedures, evaluation protocols, prompt templates, and full hyperparameter configurations are detailed in Appendix A. The complete source code, trained adapter checkpoints, and offline probing scripts are provided in the supplementary material and will be made publicly available upon release.

REFERENCES

[1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20111–20121. IEEE, 2022.

[2] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

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

[4] Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

[5] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

[6] Zhaolin Cai, Fan Li, Ziwei Zheng, Haixia Bi, and Lijun He. Headhunt-vad: Hunting robust anomaly-sensitive heads in mllm for tuning-free video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 19835–19843, 2026.

[7] Yuantong Chen, Zhengyan Ding, and YanFeng Shang. Prism: Training-free video anomaly detection via intrinsic statistical modeling. In *Forty-third International Conference on Machine Learning*, 2026.

[8] Zunkai Dai, Ke Li, Jiajia Liu, Jie Yang, and Yuanyuan Qiao. No need for real anomaly: Mllm empowered zero-shot video anomaly detection. *arXiv preprint arXiv:2602.19248*, 2026.

[9] Gemma Team. Gemma 4 technical report, 2026. URL `https://arxiv.org/abs/2607.02770`.

[10] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[12] Junhee Lee, ChaeBeen Bang, MyoungChul Kim, and MyeongAh Cho. Refinevad: Semantic-guided feature recalibration for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 5899–5907, 2026.

[13] Jiaxu Leng, Jiankang Zheng, Mengjingcheng Mo, Zhanjie Wu, Haosheng Chen, Ji Gan, and Xinbo Gao. Linguistic relative policy optimization for video anomaly reasoning. *arXiv preprint arXiv:2607.00654*, 2026.

[14] Wenlong Li, Yifei Xu, Yuan Rao, Zhenhua Wang, and Shuiguang Deng. Vadtree: Explainable training-free video anomaly detection via hierarchical granularity-aware tree. *Advances in Neural Information Processing Systems*, 38:148372–148404, 2026.

[15] Dongheng Lin, Mengxue Qu, Kunyang Han, Jianbo Jiao, Xiaojie Jin, and Yunchao Wei. A unified reasoning framework for holistic zero-shot video anomaly analysis. *Advances in Neural Information Processing Systems*, 38:29713–29744, 2026.

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

[17] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522, 2023.

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[19] Snehashis Majhi, Giacomo D'amicantonio, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, Egor Bondarev, and Francois Bremond. Just dance with π! a poly-modal inductor for weakly-supervised video anomaly detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24265–24274. IEEE, 2025.

[20] Mengjingcheng Mo, Jiaxu Leng, and Xinbo Gao. Learning to watch: Active video anomaly understanding via interleaved policy optimization. *arXiv preprint arXiv:2607.00622*, 2026.

[21] nostalgebraist. Interpreting GPT: the logit lens. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`, 2020. LessWrong.

[22] Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, Yuxuan Fan, Muyang Zhang, Ziyang Yan, Ao Ma, et al. Eventvad: Training-free event-aware video anomaly detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2586–2595, 2025.

[23] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488. IEEE, 2018.

[24] Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

[25] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4955–4966. IEEE, 2021.

[26] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[27] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision*, pages 322–339. Springer, 2020.

[28] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 6074–6082, 2024.

[29] Zhiwei Yang, Chen Gao, and Mike Zheng Shou. Panda: Towards generalist video anomaly detection via agentic ai engineer. *Advances in Neural Information Processing Systems*, 38: 83182–83211, 2026.

[30] Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8679–8688. IEEE, 2025.

[31] Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18527–18536. IEEE, 2024.

[32] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

[33] Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13843–13853. IEEE, 2025.

[34] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 3769–3777, 2023.

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

A Implementation and Experimental Details

A.1 Backbone Configurations and LoRA Setup

Our primary backbone is Qwen3.5-9B, whose language decoder contains 32 blocks indexed from Block 0 to Block 31. Based on the residual-stream diagnostic in Section 3, LoRA modules are inserted into the attention projections of Blocks 20 and 21. Unless otherwise specified, we use LoRA rank $r = 16$, scaling factor $\alpha = 32$, and dropout $p_{\text{drop}} = 0.05$. The default configuration adapts the query, key, value, output, and architecture-specific auxiliary attention projections. All MLP sublayers, normalization layers, vision encoders, and language unembedding parameters remain frozen. For a base projection matrix $W_{\text{base}} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, the adapted matrix is

$$W' = W_{\text{base}} + \frac{\alpha}{r} BA, \qquad A \in \mathbb{R}^{r \times d_{\text{in}}}, \quad B \in \mathbb{R}^{d_{\text{out}} \times r}. \tag{8}$$

For cross-backbone diagnostics and evaluations, Gemma 4 and InternVL3.5-8B use the same semantic readout tokens and categorical decision protocol as the primary Qwen3.5-9B experiment. We independently trace the residual stream for each architecture and identify the localized readout transition according to the largest characteristic increase in native-axis separability. The evaluated Gemma 4 and InternVL3.5-8B configurations exhibit their most prominent transitions around Layer 33 and Layer 25, respectively. For RBA alignment training on InternVL3.5-8B, we follow the same diagnostic-guided placement criterion: adapters are inserted into the identified transition attention block and the immediately subsequent attention block. The exact target projection module names are architecture-specific and are specified in the released configuration files.

A.2 Probe Training and Weakly-Supervised Pseudo-Labeling

The linear probe $\mathbf{w}_{\text{probe}} \in \mathbb{R}^d$ is trained on normalized final-layer hidden states $\tilde{\mathbf{h}}_L$. We use AdamW with an initial learning rate of 1×10^{-3} and train for 10 epochs. For video j containing T_j temporal clips, Top-K MIL pooling uses

$$K_j = \left\lfloor \frac{T_j}{16} \right\rfloor + 1. \tag{9}$$

Let

$$s_i = \sigma\left(\mathbf{w}_{\text{probe}}^{\top} \tilde{\mathbf{h}}_{L,i} + b\right) \tag{10}$$

denote the clip-level posterior. The video-level prediction is the mean probability of the K_j highest-scoring clips:

$$\hat{Y}_j = \frac{1}{K_j} \sum_{i \in \text{TopK}(j)} s_i. \tag{11}$$

The probe is trained with video-level binary supervision:

$$\mathcal{L}_{\text{MIL}} = \frac{1}{|\mathcal{V}|} \sum_{j \in \mathcal{V}} \text{BCE}\left(Y_j, \hat{Y}_j\right). \tag{12}$$

Class balancing is performed strictly at the video-bag level. Clips from the same video are never distributed across different data splits. After probe training, pseudo-labels are computed only on the training split:

$$\hat{y}_i = \begin{cases} +1, & s_i \geq 0.5, \\ -1, & s_i < 0.5. \end{cases} \tag{13}$$

Frame-level annotations are strictly held out and utilized solely for standard test-set benchmark evaluation and the oracle retrospective analysis presented in Section 3.2. Neither probe training, transition-layer localization, nor RBA adapter optimization utilizes ground-truth frame annotations at any stage.

Leakage-Free Transition Localization via MIL Pseudo-Labels. A central operational question is whether the critical intervention depth identified by the retrospective diagnostic in Section 3.2 can be reliably recovered without any ground-truth temporal supervision. We demonstrate here that evaluating the residual stream exclusively under training-set MIL pseudo-labels $\hat{y}_i$ consistently pinpoints the exact same transition layer (Blocks 20–21) across all three benchmarks.

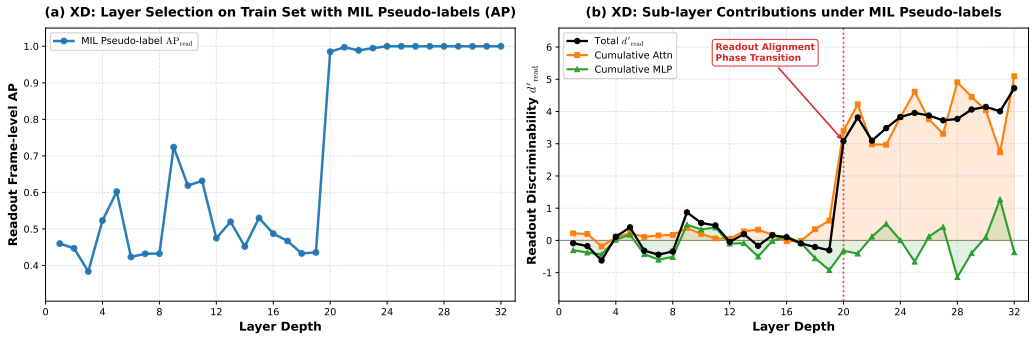


Figure 5: Pseudo-label layer diagnosis on the XD-Violence training set (Qwen3.5-9B). (a) Native Frame-AP against pseudo-labels across depth. (b) Pseudo-label discriminability d'_{read} and cumulative sublayer updates, showing a sharp phase transition at the Block 20 attention sublayer.

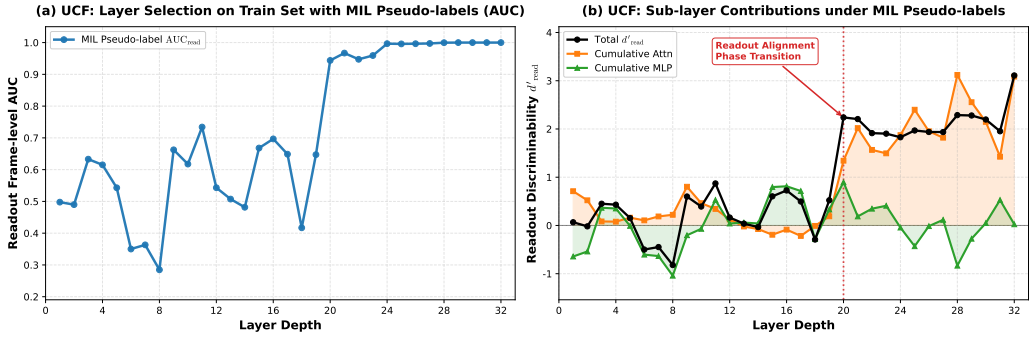


Figure 6: Pseudo-label layer diagnosis on the UCF-Crime training set (Qwen3.5-9B). (a) Native Frame-AUC against pseudo-labels across depth. (b) Pseudo-label discriminability d'_{read}, consistently locating the transition at Block 20.

Specifically, for each residual-stream sampling point l, we evaluate two diagnostic quantities driven solely by pseudo-labels. First, we compute the native abnormal-minus-normal readout score $d_i(l) = \mathbf{u}^\top \tilde{\mathbf{h}}_{l,i}$ and calculate the pseudo-label discriminability:

$$d'_{\text{read}}(l) = \frac{|\hat{\mu}_+(l) - \hat{\mu}_-(l)|}{\sqrt{\frac{1}{2}\left(\hat{\sigma}^2_+(l) + \hat{\sigma}^2_-(l)\right)}}, \tag{14}$$

where $\hat{\mu}_\pm$ and $\hat{\sigma}^2_\pm$ are the empirical mean and variance of $d_i(l)$ across pseudo-positive and pseudo-negative training clips. Decomposing $d'_{\text{read}}(l)$ into cumulative attention and MLP sublayer contributions precisely identifies the attention update producing the largest discrete increase in native alignment. Second, we record the overall frame-level readout metric (AP or AUC) against $\hat{y}_i$ across depth, which characterizes how native readout quality evolves with depth on the training split alone.

Consistent Transition Identification vs. Superficial Metric Fluctuations. Figures 5, 6, and 7 illustrate the training-set layer-selection trajectories on XD-Violence, UCF-Crime, and UBnormal, respectively. Two decisive observations emerge from these visualizations:

1. **Universal Phase Transition at Blocks 20–21:** As shown in Panel (b) of all three figures, d'_{read} consistently remains close to zero throughout early-to-mid layers, followed by a dramatic, discontinuous leap at the **Block 20 attention sublayer** before plateauing. Furthermore, the cumulative decomposition reveals that this sharp emergence is entirely propelled by attention residual updates, while MLP updates produce negative or negligible contributions across all datasets. This proves that the underlying alignment transition is an intrinsic structural landmark of the backbone architecture and can be discovered with 100% fidelity using only MIL pseudo-labels. Consequently, our policy of injecting adapters into Blocks 20–21 is entirely leakage-free and robustly reproducible across diverse video anomaly domains.

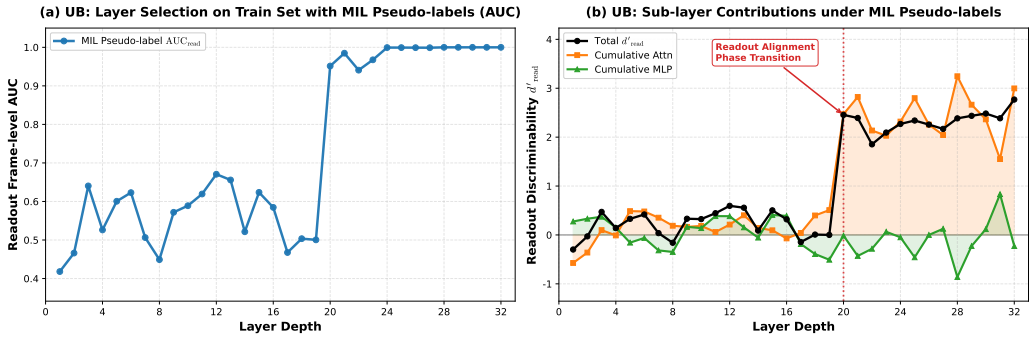


Figure 7: Pseudo-label layer diagnosis on the UBnormal training set (Qwen3.5-9B). (a) Native Frame-AUC against pseudo-labels across depth. (b) Pseudo-label discriminability d'_{read}, verifying the identical Block 20 attention transition under synthetic anomaly scenarios.

2. **Post-Transition Saturation Plateau:** In Panel (a), the readout metric enters an extended, near-flat saturation plateau once the representation passes the transition layer, and the residual variations within this plateau are marginal fluctuations caused by pseudo-label noise and depth-wise representation smoothing. Adapter placement is therefore not driven by where the curve attains its numerical maximum, since deep layers inside the plateau have already undergone heavy representation mixing and offer little headroom for redirecting the decision axis. RBA instead intervenes at the inception of the alignment transition (Blocks 20–21), where attention is actively rotating the decision axis.

In summary, weak-supervision MIL pseudo-labels faithfully preserve both the geometric phase transition and the sublayer opposition discovered in the oracle analysis. The automated detection of Blocks 20–21 across all benchmarks guarantees that RBA operates as a strictly weakly-supervised, leakage-free framework.

A.3 Margin Alignment Training Protocol

During alignment fine-tuning, we optimize the hinge-margin objective with margin $M = 2.0$:

$$\mathcal{L}_{\text{align}} = \frac{1}{B} \sum_{i=1}^{B} \max\left(0, M - \hat{y}_i d_i\right), \tag{15}$$

where d_i is the abnormal-minus-normal native logit difference. Optimization uses AdamW with an initial learning rate of 1×10^{-4}, cosine learning-rate scheduling, weight decay 0.01, and a batch size of 32 clips. Training is conducted for three epochs and takes approximately 1.8 hours on one NVIDIA A100 80GB GPU. All model-selection decisions, including the LoRA rank, dropout, target modules, target blocks, and margin, are made on the validation split. Benchmark test annotations are not used for hyperparameter selection.

A.4 Standardized Prompt and Native Readout Extraction

The following prompt is used consistently for hidden-state extraction, probe training, alignment training, categorical prediction, and open-ended generation:

```
<|im_start|>system
You are an expert video anomaly detection and reasoning
assistant.  <|im_end|>
<|im_start|>user
The following four consecutive video frames record the
temporal evolution of a scene.  Carefully observe the visual
content and determine whether an anomaly or abnormal event
occurs:
[VIDEO_TOKENS]
Select the most appropriate option:
- Normal
```

Table 5: Ablation of MIL pooling strategies on downstream RBA performance (%). Pseudo-label precision and recall are diagnostic quantities computed against training-set temporal annotations; these annotations are not used for optimization.

Pooling strategy	Pseudo-P	Pseudo-R	XD AP	XD AUC	UCF AUC	UB AUC
Mean Pooling	52.41	89.67	75.14	91.23	81.38	69.82
Max Pooling	84.16	41.29	78.43	92.47	83.94	72.18
Random-K Pooling	56.83	74.38	76.07	91.62	82.16	70.49
Top-K MIL	**81.79**	**79.34**	**83.02**	**94.30**	**87.01**	**76.84**
Dual-Memory Units	83.27	81.08	83.46	94.51	87.23	77.12

```
- Abnormal
Formatting requirement:  Output only ``Normal'' or
``Abnormal'' on the first line as the final judgment.  From
the second line onward, briefly describe the visual evidence
supporting the judgment.  <|im_end|>
<|im_start|>assistant
Answer:
```

The native decision score is the abnormal-minus-normal logit difference at the first answer position:

$$d_i = \mathrm{Logit}_i(\mathrm{Abnormal}) - \mathrm{Logit}_i(\mathrm{Normal}). \tag{16}$$

When either label is tokenized into multiple tokens, we use the teacher-forced sum of conditional token log-probabilities:

$$d_i = \log p(\mathrm{Abnormal} \mid V_i, P) - \log p(\mathrm{Normal} \mid V_i, P), \tag{17}$$

where V_i denotes the visual input and P denotes the shared prompt.

B Alignment Design Ablations

This section evaluates the principal design choices of RBA. We first study how video-level supervision is converted into clip-level instances, then examine adapter placement, rank, target modules, dropout, and finally the alignment objective and its margin.

B.1 MIL Pooling and Instance Mining

We compare five strategies for converting video-level supervision into clip-level probe scores:

1. **Mean Pooling:**
$$\hat{Y}_j = \frac{1}{T_j} \sum_{t=1}^{T_j} p_{j,t}. \tag{18}$$
2. **Max Pooling:**
$$\hat{Y}_j = \max_t p_{j,t}. \tag{19}$$
3. **Random-K Pooling:** randomly selects K_j clips and averages their probabilities.
4. **Top-K MIL:** averages the $K_j = \lfloor T_j/16 \rfloor + 1$ highest-scoring clips.
5. **Dual-Memory Units:** uses normal and abnormal memory banks $M_n, M_a \in \mathbb{R}^{60\times d}$ following UR-DMU [34], before applying Top-K pooling.

Mean pooling yields high recall but low precision because anomalous segments are diluted by normal background clips. Max pooling has the opposite behavior and is sensitive to isolated high-score clips. Top-K MIL provides a more balanced precision–recall trade-off and substantially improves downstream alignment. Dual-memory mining provides a marginal performance gain (+0.44% AP) but requires maintaining external bank matrices ($120 \times d$) and prototype distance updates. Top-K MIL introduces no auxiliary parameters or inference overhead, and we adopt it as our default design.

Table 6: Adapter placement ablation on XD-Violence. All attention placements use approximately the same trainable parameter budget.

Placement	Sublayer	Params	Frame-AP
Native Readout	–	0	0.7286
Blocks 5–6	Post-Attention	$\approx 0.012\%$	0.7314
Blocks 17–18	Post-Attention	$\approx 0.012\%$	0.7976
Blocks 20–21	**Post-Attention**	$\approx \mathbf{0.012\%}$	**0.8302**
Blocks 30–31	Post-Attention	$\approx 0.012\%$	0.7738
Blocks 20–21	Post-MLP	$\approx 0.024\%$	0.7415

Table 7: LoRA rank ablation in Blocks 20–21 with $\alpha/r = 2.0$.

Rank r	Params	XD AP	XD AUC	UCF AUC	UB AUC
1	$\approx 0.001\%$	77.83	92.14	82.47	73.19
4	$\approx 0.003\%$	80.46	93.18	84.82	74.93
8	$\approx 0.006\%$	82.17	93.84	86.29	76.15
16	$\approx \mathbf{0.012\%}$	**83.02**	**94.30**	**87.01**	**76.84**
32	$\approx 0.024\%$	83.09	94.36	87.04	76.81
64	$\approx 0.048\%$	82.93	94.25	86.97	76.68

B.2 Diagnostic-Guided Adapter Placement

We evaluate adapter placement at different depths while keeping the attention parameter budget strictly controlled.

The strong dependence on placement indicates that the improvement is not explained by parameter count alone. Adapters at Blocks 20–21 are located at the empirically identified readout transition and achieve the largest gain. Notably, placing adapters on the post-MLP sublayers yields an inferior 0.7415 AP despite consuming twice the parameter budget ($\approx 0.024\%$). This confirms that the positive native-axis alignment observed in the residual stream is driven specifically by attention sublayers.

B.3 LoRA Rank and Target Module Ablations

Rank-one adaptation already improves XD-Violence Frame-AP from 72.86% to 77.83%, showing that a low-dimensional intervention can recover part of the native-readout gap. Performance largely saturates at $r = 16$; larger ranks provide negligible gains.

Adapting both the input and output sides of attention performs better than adapting either side alone. Including all projection matrices within the attention block (covering standard q, k, v, o along with the projection heads) provides a further gain of 0.88% AP, confirming that comprehensive attention transformation best guides directional rotation.

B.4 Dropout and Training Stability

We evaluate $p_{\text{drop}} \in \{0, 0.05, 0.1, 0.2\}$ across five random seeds. The default setting $p_{\text{drop}} = 0.05$ obtains 83.02 ± 0.17 AP on XD-Violence, compared with 82.46 ± 0.24 for $p_{\text{drop}} = 0$ and 82.13 ± 0.29 for $p_{\text{drop}} = 0.2$. These results suggest that mild dropout improves robustness to noisy pseudo-labels without materially reducing native-axis alignment.

B.5 Margin Sensitivity

We evaluate the hinge margin M while retaining the default LoRA rank, target layers, optimizer, and pseudo-labeling procedure.

Performance is stable for $M \in [1.5, 2.5]$ and peaks at $M = 2.0$. Small margins provide insufficient separation, whereas overly large margins continue to optimize hard or noisy pseudo-labeled clips after the useful directional correction has largely been achieved.

Table 8: Ablation of target attention projections in Blocks 20–21 using $r = 16$ and $\alpha = 32$.

Target projections	**Params**	**XD AP**	**UCF AUC**	**UB AUC**
Query, key, and value (q, k, v)	$\approx 0.006\%$	80.67	85.14	75.32
Attention output (o)	$\approx 0.002\%$	79.83	84.67	74.89
Standard attention (q, k, v, o)	$\approx 0.008\%$	82.14	86.38	76.23
All attention projections	$\approx \mathbf{0.012}\%$	**83.02**	**87.01**	**76.84**

Table 9: Sensitivity of the hinge margin M on XD-Violence.

Margin M	0.5	1.0	1.5	**2.0**	2.5	3.0
Frame-AP	78.94	81.23	82.68	**83.02**	82.89	81.95
Frame-AUC	92.81	93.74	94.18	**94.30**	94.22	93.80

B.6 Comparison of Alignment Objectives

To distinguish directional alignment from probability calibration, we compare the proposed hinge loss with BCE, KL, MSE, and joint decision–explanation training. Let

$$z_i = \frac{\hat{y}_i + 1}{2} \in \{0, 1\}. \tag{20}$$

The compared objectives are

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{B} \sum_{i=1}^{B} \left[z_i \log \sigma(d_i) + (1 - z_i) \log(1 - \sigma(d_i)) \right], \tag{21}$$

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{B} \sum_{i=1}^{B} D_{\mathrm{KL}} \left(\mathrm{Bern}(s_i) \parallel \mathrm{Bern}(\sigma(d_i)) \right), \tag{22}$$

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{B} \sum_{i=1}^{B} \left(d_i - \hat{y}_i M \right)^2, \tag{23}$$

$$\mathcal{L}_{\mathrm{joint}} = \mathcal{L}_{\mathrm{align}} + \lambda \mathcal{L}_{\mathrm{explain}}. \tag{24}$$

The hinge objective achieves the highest native detection accuracy along with the largest improvement in geometric alignment. Objectives like BCE and KL continue enforcing probability calibration even after representations cross the decision boundary, which misallocates the limited rank-16 capacity. Such probability fitting becomes counterproductive when dealing with potentially noisy pseudo-labels.

Introducing an auxiliary explanation loss alongside the margin term degrades both AP and AUC back to baseline levels, keeping the cosine similarity at a low value of 0.298. Autoregressive language modeling gradients diffuse the low-rank updates across numerous token positions, thereby diluting the targeted geometric correction. These comparisons demonstrate that placing adapters at the transition layers works best under a focused native-axis margin loss, which directly enhances boundary separation without requiring probability calibration.

C Mechanistic Interpretation of the Readout Transition

C.1 Potential Sources of Readout Misalignment

Several properties of general-purpose MLLMs may contribute to the representation–behavior gap:

1. **Multi-task pretraining.** Hidden states must support diverse visual and linguistic tasks rather than a single VAD decision boundary.
2. **Instruction post-training.** Instruction tuning optimizes broadly useful language behavior and does not directly optimize temporal anomaly ranking.
3. **Task-independent lexical readout.** The native direction

$$\mathbf{u} = \frac{\mathbf{W}_U[\mathrm{Abnormal}] - \mathbf{W}_U[\mathrm{Normal}]}{\left\| \mathbf{W}_U[\mathrm{Abnormal}] - \mathbf{W}_U[\mathrm{Normal}] \right\|_2} \tag{25}$$

Table 10: Comparison of alignment objectives on XD-Violence. The metric $\cos(\mathbf{w}'_{\text{probe}}, \mathbf{u})$ evaluates directional alignment between the adapted probe direction and the fixed native axis.

Training objective	Frame-AP	Frame-AUC	$\cos(\mathbf{w}'_{\text{probe}}, \mathbf{u})$
Zero-Shot Base Readout	72.86	90.37	0.238
$\mathcal{L}_{\text{BCE}}$ (hard labels)	79.82	93.12	0.684
$\mathcal{L}_{\text{KL}}$ (soft distribution)	78.45	92.68	0.641
$\mathcal{L}_{\text{MSE}}$ (logit regression)	77.19	91.85	0.612
$\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{explain}}$	72.79	90.41	0.298
$\mathcal{L}_{\text{align}}$	**83.02**	**94.30**	**0.781**

is inherited from pretraining rather than fitted to VAD data.

These factors provide plausible explanations for the measured angular difference between the weakly supervised probe and the native readout. They are not independently verified causal mechanisms.

C.2 Empirical Progression Across Depth

Probe performance rises rapidly and reaches approximately 0.85 AP around Block 15, indicating that linearly accessible anomaly information is already present before the final decoder blocks. On Qwen3.5-9B with XD-Violence, native-axis separability increases sharply at the attention sublayer of Block 20:

$$d'_{\text{read}} : 0.474 \longrightarrow 1.545. \tag{26}$$

We refer to this empirically observed region as the *readout transition region*. The term describes the measured behavior of this model–dataset configuration and does not imply a universal mathematical phase transition. After the transition, native-axis separability increases more slowly and eventually plateaus. Together with the adapter placement and objective ablations, this suggests that anomaly-discriminative features form early, while their projection into the language generation space is refined later.

C.3 Robustness Across Readout Tokens

The primary diagnostic in Figure 2 employs the default single-token verbalizers (`Normal` and `Abnormal`). To examine whether the observed depth-wise dynamics are sensitive to lexical realization, we keep the primary Qwen3.5-9B backbone, video inputs, prompt structure, extraction position, and evaluation protocol strictly fixed, and evaluate two alternative readout formulations. For any pair of single-token verbalizers $(\tau_{\text{abn}}, \tau_{\text{nor}})$, the native decision axis is defined as

$$\mathbf{u}_\tau = \frac{\mathbf{W}_U[\tau_{\text{abn}}] - \mathbf{W}_U[\tau_{\text{nor}}]}{\|\mathbf{W}_U[\tau_{\text{abn}}] - \mathbf{W}_U[\tau_{\text{nor}}]\|_2}. \tag{27}$$

When verbalizers comprise multiple sub-tokens, we use the teacher-forced log-probability difference in Equation 17. As shown in Figures 8 and 9, changing the readout tokens shifts the absolute numerical baseline because different token embeddings occupy distinct coordinates in the unembedding space. Crucially, both token variants preserve the central diagnostic traits:

1. **Early representation maturity:** The weakly-supervised linear probe (P_{probe}) ascends early in the network, confirming that anomaly features are formed well before native readout emergence.
2. **Localized transition:** Native readout discriminability undergoes a sharp, localized increase in the mid-to-late attention layers rather than accumulating linearly from the input.
3. **Sublayer dynamics:** Positive alignment gains remain driven predominantly by attention residual updates, while MLP sublayers provide smaller or opposing contributions.

These findings demonstrate that the representation–behavior gap is an intrinsic property of the model's depth-wise processing rather than an artifact of a specific lexical choice.

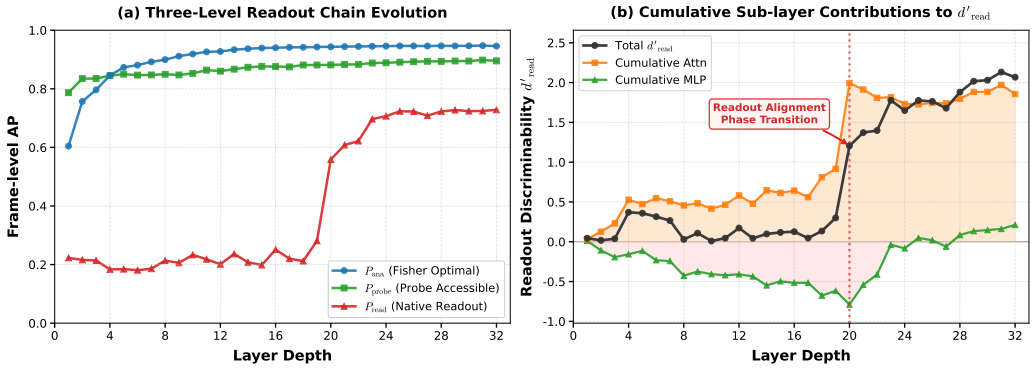


Figure 8: Layer-wise residual-stream diagnostic on Qwen3.5-9B using alternative word-level readout tokens (`Yes`/`No`) on XD-Violence. Left: three-tier readout trajectories across network depth. Right: native discriminability d'_{read} and cumulative sublayer contributions.

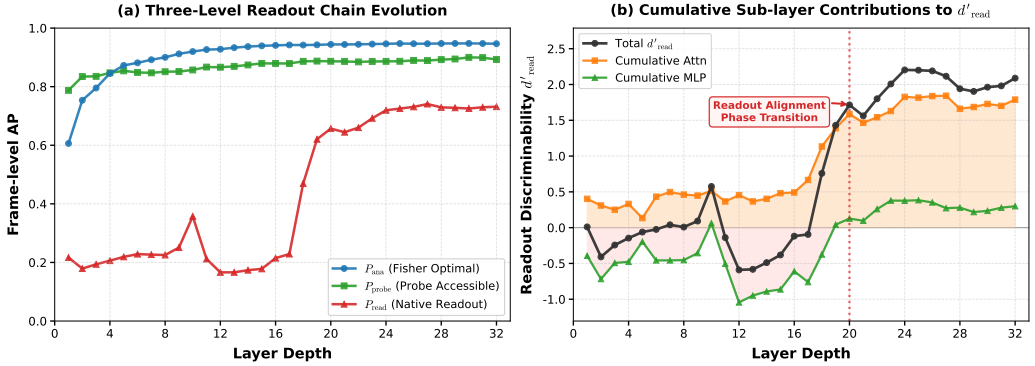


Figure 9: Layer-wise residual-stream diagnostic on Qwen3.5-9B using `Safe`/`Risky` readout tokens on XD-Violence. Left: three-tier readout trajectories across network depth. Right: native discriminability d'_{read} and cumulative sublayer contributions.

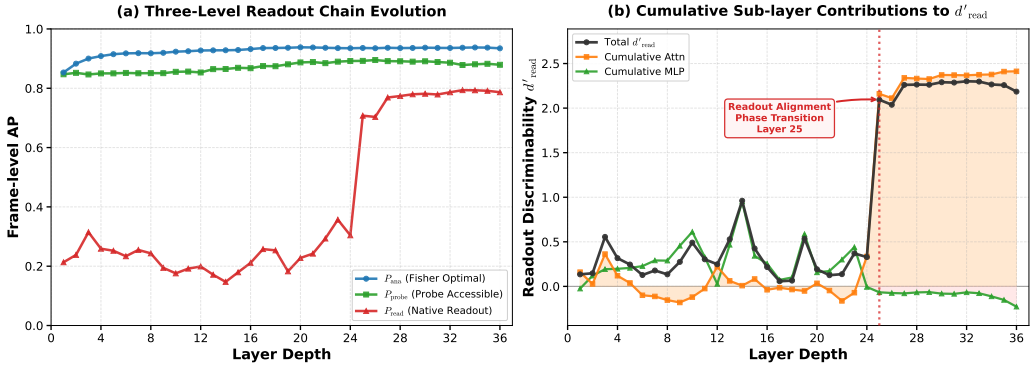


Figure 10: Layer-wise residual-stream diagnostic on InternVL3.5-8B on XD-Violence using the main-experiment readout tokens. Left: three-tier readout trajectories across network depth. Right: native discriminability d'_{read} and cumulative sublayer contributions.

C.4 Robustness Across MLLM Backbones

To assess cross-architecture generality, we evaluate the diagnostic pipeline on Gemma 4 and InternVL3.5-8B while strictly maintaining the main-experiment readout tokens, categorical protocol, and XD-Violence dataset split. Figures 11 and 10 confirm that the structural separation between representation accessibility and native readout alignment generalizes across diverse MLLM architectures. In both models, the probe reaches near-peak discriminative performance early, whereas the native readout remains unaligned until a localized transition in the deeper layers. Due to architectural variations in depth, attention topology, and pretraining distributions, the transition occurs at

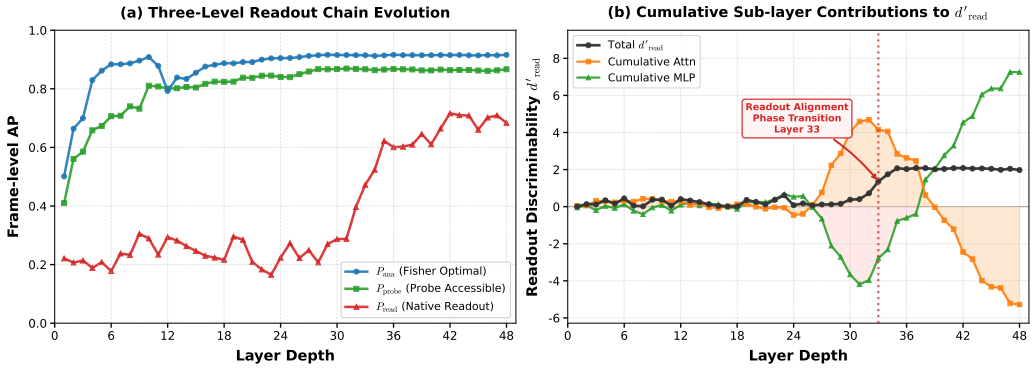


Figure 11: Layer-wise residual-stream diagnostic on Gemma 4 on XD-Violence using the main-experiment readout tokens. Left: three-tier readout trajectories across network depth. Right: native discriminability d'_{read} and cumulative sublayer contributions.

backbone-dependent depths: around Layer 33 in Gemma 4 and Layer 25 in InternVL3.5-8B. Across both architectures, attention residual updates remain the primary source of positive alignment gains. This validates our methodological premise: rather than hardcoding a fixed layer index, RBA applies a diagnostic-guided principle that localizes the critical transition layers specific to each backbone prior to adapter placement.

C.5 Objective-Based Evidence for Transition-Layer Specialization

The loss comparison in Table 10 provides functional evidence regarding transition-layer specialization. If these layers merely enhanced generic anomaly separability without directional preference, distinct objectives would produce comparable improvements under the same parameter constraints. Instead, the native-axis hinge loss yields the strongest detection metrics and maximizes the alignment value $\cos(\mathbf{w}'_{\text{probe}}, \mathbf{u})$. The inferior outcomes of BCE and KL suggest that matching probe probabilities is unnecessary for decision boundary correction. Furthermore, joint explanation training damages detection accuracy because dispersing the low-rank capacity over multi-token sequences compromises the primary rotation objective. These results indicate that the identified transition layers are well suited for an efficient representation-to-readout alignment rather than broad semantic fine-tuning. Residual-update decomposition shows that attention sublayers contribute positive gains to native-axis separability, whereas MLP sublayers introduce negative interference in the examined configuration. The post-attention placement result in Table 6, together with the inferior post-MLP result, provides a parameter-controlled empirical check of this observation. This pattern motivates freezing MLP sublayers during RBA training. The purpose is not to assume that MLPs are uninformative, but to avoid using the limited low-rank adaptation budget to modify components whose measured net contribution is not aligned with the desired native-axis correction.

C.6 Null-Space Diagnostic

We remove the normalized LDA direction from the final hidden states:

$$\mathbf{h}^{\perp} = \left(\mathbf{I} - \hat{\mathbf{w}}_{\text{LDA}}\hat{\mathbf{w}}_{\text{LDA}}^{\top}\right)\tilde{\mathbf{h}}_L. \tag{28}$$

A new probe trained and evaluated in the projected space obtains 0.5237 AP, compared with 0.8756 before projection. This result shows that the removed direction contains a large fraction of the linearly accessible discriminative signal. It does not rule out additional nonlinear or distributed anomaly information in the remaining subspace.

C.7 Implications for RBA

The diagnostics motivate three design choices:

- **Targeted placement.** Adapters are placed near the empirically identified readout transition.

- **Attention-only adaptation.** Attention projections are adapted while MLPs and the unembedding head remain frozen.
- **Native-axis optimization.** The hinge objective directly increases class separation along the fixed native decision axis without requiring probability matching.

D Decision–Explanation Evaluation Protocol

D.1 Diagnostic Sets

We evaluate decision–explanation behavior on fixed standardized sets constructed from the official test splits of XD-Violence, UCF-Crime, and UBnormal. The compliant faithfulness set $\mathcal{D}^+$ contains $N^+ = 10{,}000$ clips, balanced between normal and abnormal examples, for which the model's native decision is correct. The adversarial rationalization set $\mathcal{D}^-$ contains $N^- = 10{,}000$ clips subjected to counterfactual decision forcing. Each clip contains four temporally ordered frames, a binary ground-truth label y_i, and a human-verified reference description $\mathcal{C}_i^{\mathrm{ref}}$. Sampling is stratified across datasets and classes using seed 42.

D.2 Compared Generation Pipelines

We compare:

- **Zero-Shot Qwen:** the frozen backbone directly generates the categorical decision and explanation.
- **Probe + Caption:** the external probe supplies the categorical decision, while the frozen backbone independently generates the explanation.
- **RBA:** the aligned model generates both the categorical decision and explanation through one autoregressive pathway.

Greedy decoding is used for all methods. The categorical decision is parsed only from the first generated line. Invalid or unparsable first lines are counted as incorrect decisions and inconsistent outputs. For the adversarial benchmark, the first-line decision is replaced by the opposite label while keeping the visual input fixed. The model then generates the explanation conditioned on the forced prefix.

D.3 LLM-as-a-Judge Evaluation

We use two independent judges with temperature zero. Each judge returns a consistency label $c_i^m \in \{0, 1\}$ and an evidence-faithfulness grade $g_i^m \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$. The judge receives the ground-truth label, reference description, model decision, and model explanation, but not the method name.

```
You are an impartial evaluator of video anomaly decisions
and explanations.  Evaluate the model output using the
supplied reference.
Ground-truth label:  {ground_truth_label}
Reference description:  {reference_description}
Model decision:  {model_decision}
Model explanation:  {model_explanation}
Return exactly two fields:
Consistency:  YES or NO
Faithfulness:  A, B, C, or D
```

Consistency evaluates whether the explanation supports the model's own decision, independently of whether that decision is correct. Faithfulness evaluates the explanation against the reference evidence.

Table 11: Decision–explanation evaluation on fixed diagnostic sets. Consistency and F^+ are higher-is-better, while R^- is lower-is-better.

Method	Judge	Consist.	$\mathbf{F}^+$	$\mathbf{R}^-$	Quality
Zero-Shot Qwen	GPT-5.4 mini	78.2%	94.5%	86.1%	3.63
Zero-Shot Qwen	Gemini 3.6 Flash	79.1%	93.8%	85.4%	3.57
Probe + Caption	GPT-5.4 mini	43.2%	71.3%	94.4%	2.85
Probe + Caption	Gemini 3.6 Flash	44.0%	70.8%	93.9%	2.78
RBA	**GPT-5.4 mini**	**91.4%**	**96.3%**	**36.7%**	**4.12**
RBA	**Gemini 3.6 Flash**	**92.0%**	**95.9%**	**35.8%**	**4.09**

D.4 METRICS

We map the faithfulness grade to

$$\mathcal{Q}(g) = \begin{cases} 5, & g = \mathrm{A}, \\ 4, & g = \mathrm{B}, \\ 2, & g = \mathrm{C}, \\ 0, & g = \mathrm{D}. \end{cases} \tag{29}$$

Decision–explanation consistency is

$$\mathrm{Consist.}^m = \frac{1}{N} \sum_{i=1}^{N} c_i^m \times 100\%. \tag{30}$$

Compliant faithfulness is

$$F_m^+ = \frac{1}{N^+} \sum_{i \in \mathcal{D}^+} \frac{\mathcal{Q}(g_i^m)}{5} \times 100\%. \tag{31}$$

The adversarial rationalization rate is

$$R_m^- = \frac{1}{N^-} \sum_{i \in \mathcal{D}^-} \mathbb{I}\left[c_i^m = 1 \wedge g_i^m \in \{\mathrm{C}, \mathrm{D}\}\right] \times 100\%. \tag{32}$$

Lower R^- indicates greater sensitivity to the visual evidence under counterfactual decision forcing. Overall explanation quality is

$$\mathrm{Quality}^m = \frac{1}{N} \sum_{i=1}^{N} \mathcal{Q}(g_i^m). \tag{33}$$

These results provide evidence of improved decision–explanation coupling. They do not establish that the generated explanation is a complete causal account of the model's internal computation.

E STATISTICAL SIGNIFICANCE AND MULTI-SEED ROBUSTNESS

We conduct five independent runs with random seeds

$$S \in \{42, 100, 2024, 2026, 3407\}.$$

Across runs, we vary the LoRA initialization, mini-batch ordering, and dropout masks while keeping the dataset splits, pseudo-labeling procedure, training hyperparameters, prompt template, and evaluation protocol fixed. The standard deviations remain within 0.22 percentage points for all reported metrics. Two-sided one-sample t-tests on the per-seed improvements over the deterministic base readout yield $p < 10^{-5}$ for all evaluated dataset–metric pairs. These results indicate that the observed improvements are substantially larger than the variation caused by LoRA initialization, mini-batch ordering, and stochastic dropout. The multi-seed results also support the stability of the transition-layer placement and the default rank-16, margin-2.0 configuration.

Table 12: Multi-seed evaluation over five independent runs. Metrics are reported in percent. The base readout is deterministic under the fixed prompt and greedy decoding protocol.

Dataset	Metric	Base Readout	RBA	Absolute Gain
XD-Violence	Frame-AP	72.86	**83.02** $\pm$ **0.17**	+10.16 $\pm$ 0.17
	Frame-AUC	90.37	**94.30** $\pm$ **0.11**	+3.93 $\pm$ 0.11
UCF-Crime	Frame-AUC	77.14	**87.01** $\pm$ **0.19**	+9.87 $\pm$ 0.19
UBnormal	Frame-AUC	72.90	**76.84** $\pm$ **0.22**	+3.94 $\pm$ 0.22

F Quantitative Case Studies of Prediction Correction

This section provides two representative case groups to quantify how RBA corrects the native readout of the frozen Qwen3.5 backbone. Each group contains one normal video and one abnormal video. We compare the anomaly scores produced by the zero-shot native readout and RBA at representative frames, using 0.5 as the decision threshold. For a clip with ground-truth label $y \in \{0, 1\}$ and anomaly score s, we define the signed decision margin as

$$m(s, y) = (2y - 1)(s - 0.5). \tag{34}$$

A positive margin denotes a correct prediction, while a negative margin denotes an incorrect prediction. The margin improvement introduced by RBA is

$$\Delta m = m(s_{\mathrm{RBA}}, y) - m(s_{\mathrm{ZS}}, y), \tag{35}$$

where s_{ZS} and s_{RBA} are the zero-shot and RBA anomaly scores, respectively.

F.1 Case Group I: Traffic and Outdoor Crowd Scenes

Figure 12(a) presents a normal nighttime traffic video. At Frame 3711, the zero-shot readout assigns an anomaly score of 0.62, incorrectly exceeding the threshold. The prediction is associated with visually salient but non-anomalous cues, including vehicle lights, reflections on a wet road, and an orange-red glow. RBA reduces the score to 0.05, producing the correct normal prediction. For this normal example, the signed margin changes from

$$m(0.62, 0) = -0.12 \quad \text{to} \quad m(0.05, 0) = 0.45, \tag{36}$$

corresponding to a margin improvement of $\Delta m = 0.57$. Figure 12(b) shows an abnormal outdoor crowd scene. The zero-shot model assigns only 0.09 at Frame 2136 and therefore misses the anomaly. RBA increases the score to 0.96, correctly identifying the event as abnormal. The corresponding signed margin changes from

$$m(0.09, 1) = -0.41 \quad \text{to} \quad m(0.96, 1) = 0.46, \tag{37}$$

yielding $\Delta m = 0.87$. The generated explanations change consistently with the corrected decisions. In the normal traffic example, RBA recognizes that the observed lights and road reflections do not indicate collision, injury, or violence. In the abnormal crowd example, RBA focuses on the dense and disorderly crowd behavior rather than interpreting the scene as an ordinary gathering.

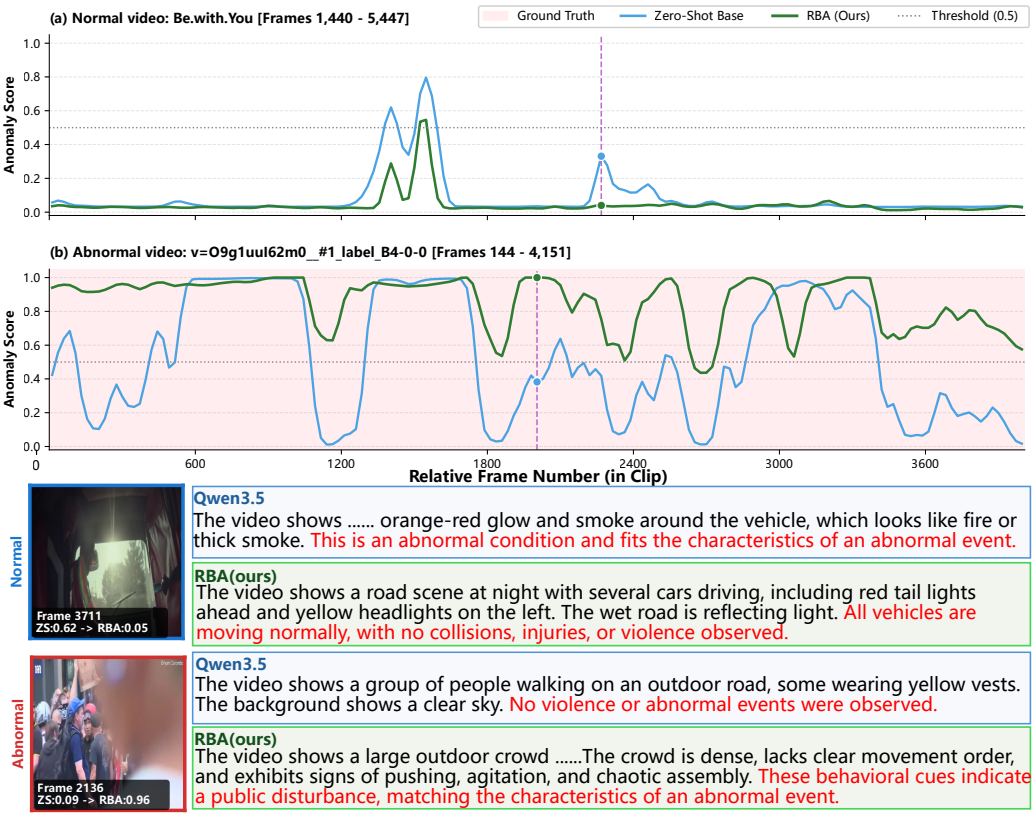


Figure 12: Quantitative and qualitative comparison for Case Group I. Left: temporal anomaly-score trajectories produced by the zero-shot Qwen3.5 readout and RBA, together with frame-level ground truth and the decision threshold. Right: explanations generated at representative frames. RBA suppresses a false alarm in the normal traffic video ($0.62 \rightarrow 0.05$) and recovers a missed anomaly in the abnormal crowd video ($0.09 \rightarrow 0.96$).

F.2 Case Group II: Sports and Dense Crowd Scenes

Figure 13(a) presents a normal sports video. At Frame 2919, the zero-shot model produces an anomaly score of 0.82, apparently treating a player bending over and nearby people as evidence of a possible conflict or accident. RBA lowers the score to 0.04 and correctly recognizes that the players, referees, spectators, and ongoing scoreboard belong to a normal sports context. The signed margin for this example changes from

$$m(0.82, 0) = -0.32 \quad \text{to} \quad m(0.04, 0) = 0.46, \tag{38}$$

which gives a margin improvement of $\Delta m = 0.78$. Figure 13(b) presents an abnormal dense-crowd scene. The zero-shot model assigns a score of 0.03 at Frame 6360 and describes the event as a normal gathering or parade. In contrast, RBA increases the anomaly score to 0.96. Its explanation identifies the dense and disorderly crowd, protective clothing, and signs of pushing or chaotic movement as evidence of an abnormal event. The signed margin changes from

$$m(0.03, 1) = -0.47 \quad \text{to} \quad m(0.96, 1) = 0.46, \tag{39}$$

resulting in the largest observed improvement among these examples, $\Delta m = 0.93$.

F.3 Quantitative Summary

Table 13 summarizes the representative-frame results. All four zero-shot decisions are initially located on the incorrect side of the 0.5 threshold. RBA moves every example to the correct side, with an average signed-margin improvement of

$$\overline{\Delta m} = \frac{0.57 + 0.87 + 0.78 + 0.93}{4} = 0.7875. \tag{40}$$

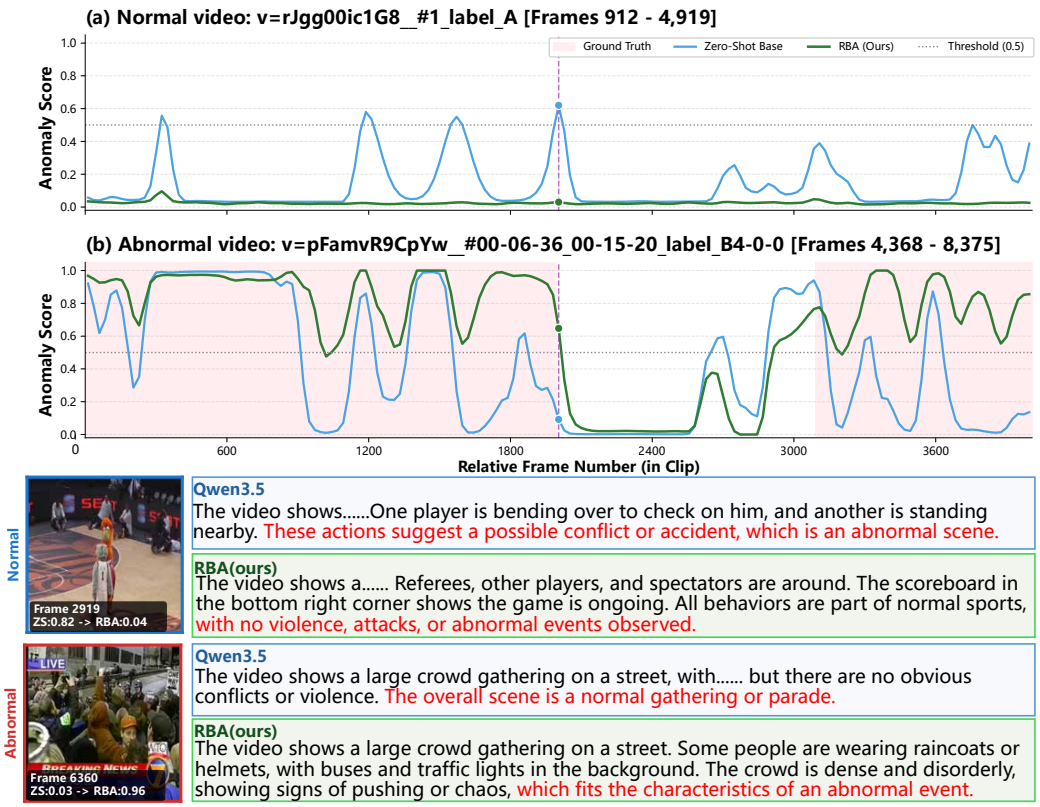


Figure 13: Quantitative and qualitative comparison for Case Group II. RBA corrects a false positive in a normal sports video ($0.82 \rightarrow 0.04$) and a false negative in an abnormal dense-crowd video ($0.03 \rightarrow 0.96$). The generated explanations are shown alongside the temporal score trajectories.

Table 13: Quantitative summary of the representative cases. ZS denotes the zero-shot native readout. A positive signed margin indicates a correct prediction under the 0.5 threshold.

Group	Scene	GT	Frame	ZS Score	RBA Score	ZS Margin	RBA Margin	Δm
I	Nighttime traffic	Normal	3711	0.62	**0.05**	-0.12	$+\mathbf{0.45}$	$+\mathbf{0.57}$
I	Outdoor crowd	Abnormal	2136	0.09	**0.96**	-0.41	$+\mathbf{0.46}$	$+\mathbf{0.87}$
II	Sports activity	Normal	2919	0.82	**0.04**	-0.32	$+\mathbf{0.46}$	$+\mathbf{0.78}$
II	Dense crowd	Abnormal	6360	0.03	**0.96**	-0.47	$+\mathbf{0.46}$	$+\mathbf{0.93}$
					Mean	-0.33	$+\mathbf{0.46}$	$+\mathbf{0.79}$

Across the four representative frames, the zero-shot readout obtains $0/4$ correct decisions, whereas RBA obtains $4/4$. For the two normal examples, the mean anomaly score decreases from

$$\frac{0.62 + 0.82}{2} = 0.72 \quad \text{to} \quad \frac{0.05 + 0.04}{2} = 0.045. \tag{41}$$

For the two abnormal examples, the mean anomaly score increases from

$$\frac{0.09 + 0.03}{2} = 0.06 \quad \text{to} \quad \frac{0.96 + 0.96}{2} = 0.96. \tag{42}$$

These cases illustrate two complementary effects of representation–behavior alignment. First, RBA suppresses false positives caused by visually salient but contextually normal patterns, such as road reflections and ordinary sports interactions. Second, it recovers false negatives in crowded scenes whose abnormality depends on contextual and behavioral evidence rather than isolated objects. The simultaneous correction of anomaly scores and generated explanations is consistent with the intended alignment of task-discriminative representations with the model's native generative readout.