Narrowing the Modality Gap: Dual-Encoder VLMs for Surveillance Video Anomaly Detection

Anonymous Authors[1]

Abstract

Vision-Language Models (VLMs) have recently been explored for Video Anomaly Detection (VAD) to provide natural-language explanations of anomalous events. In this work, we investigate a dual-encoder architecture that pairs a general-purpose CLIP vision encoder with a Vision Transformer(ViT) trained under Multiple Instance Learning (MIL) to inject anomaly-specific knowledge at the visual encoding stage, while keeping the Large Language Model (LLM) frozen and applying only LoRA-based fine-tuning to keep training costs low. A closed-set classification probe provides direct evidence that this design succeeds at the representation level: the dual-encoder variant surpasses the base VLM on Top-2 (0.534 vs. 0.525) and Top-3 (0.636 vs. 0.613) anomaly classification, showing that MIL-ViT injection contributes additional discriminative signal beyond CLIP alone. At Top-1 and at the caption level, the picture is more constrained: the model partially recovers the Sentence-BERT(SBERT) drop caused by LoRA fine-tuning yet still tends toward stereotyped templates, and trails the base model on Top-1 classification. We characterize this pattern as a *calibration gap*—the correct anomaly class is reliably available among the top candidates but is not consistently surfaced as the most probable token—rather than a loss of class information. These results identify the fusion-to-generation interface as the primary bottleneck, and we discuss two concrete paths forward (explicit alignment pretraining and late-fusion strategies) that may close this gap.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

In Video Anomaly Detection (VAD), Vision-Language Models (VLMs) have recently been adopted not only to detect anomalies but also to provide natural-language explanations of why a given event is anomalous (Lv & Sun, 2024; Zanella et al., 2024). Because VLMs typically pair a vision encoder with a Large Language Model (LLM) containing billions of parameters, fully fine-tuning both components is prohibitively expensive; recent research has therefore focused on prompt tuning as a parameter-efficient alternative to improve performance (Ye et al., 2025). However, prompt tuning operates entirely in the text space—it does not adapt the visual representations themselves for anomaly detection, but instead relies on general-purpose object recognition features and delegates anomaly reasoning to the language model.

To overcome this limitation, we propose a dual-encoder architecture that pairs a general-purpose VLM vision encoder with an anomaly-specialized Vision Transformer (ViT) in parallel, thereby injecting anomaly-specific knowledge at the visual encoding stage. Specifically, to keep training costs low, we freeze the LLM backbone and apply only parameter-efficient fine-tuning (PEFT) techniques—namely LoRA (Hu et al., 2022)—to integrate the anomaly encoder's outputs into the frozen LLM.

We empirically evaluate this design through a comparison of three configurations—the unmodified base VLM, a single-encoder LoRA baseline, and a dual-encoder variant with concatenation-based fusion—and report two complementary findings. First, a closed-set classification probe shows that anomaly knowledge injection is successful at the representation level: the dual-encoder model surpasses the base VLM on Top-2 and Top-3 anomaly classification, indicating that MIL-ViT injection contributes additional discriminative signal beyond CLIP alone. Second, this representation-level gain does not fully translate into open-ended captioning: the dual-encoder model trails the base at Top-1 classification and tends toward stereotyped, template-like captions. We interpret this gap as a calibration issue at the fusion-to-generation interface rather than a loss of class information—the correct class is reliably available among the top candidates but is not consistently surfaced as the most probable token. We discuss two concrete directions—dedicated

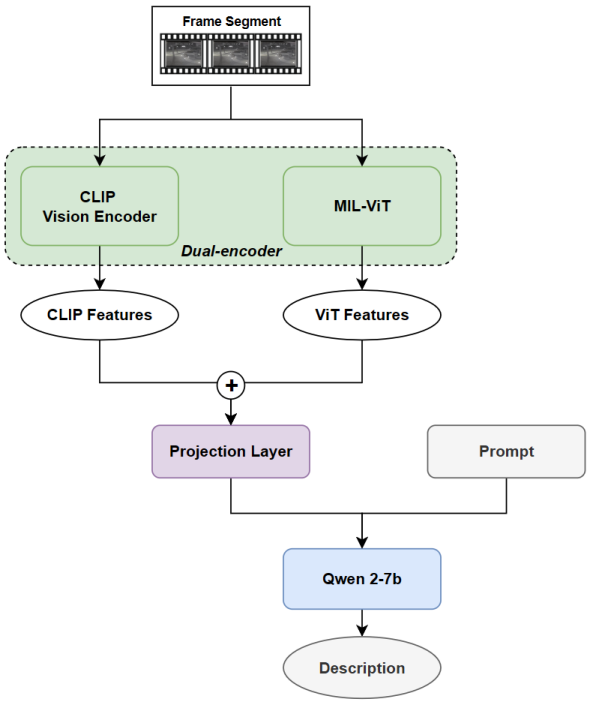

Figure 1. The overall architecture of the proposed method. Dual-encoder architecture combining frozen CLIP and MIL-ViT streams via concatenation, projected into a LoRA-tuned Qwen2-7B.

alignment pretraining and late-fusion alternatives—that may close this gap in future work.

2. Proposed Method

Our initial design replaced the CLIP vision encoder of the base VLM entirely with the MIL-trained ViT. This single-encoder MIL-ViT configuration showed clear limitations in generation capability: the model produced incoherent text or regurgitated fragments of training captions. We attributed this at design time to a mismatch between the MIL-ViT's compressed, discriminative feature space and the CLIP-aligned semantic space the LLM expects, and accordingly moved to a dual-encoder design that preserves the original CLIP stream while injecting MIL-ViT features in parallel (Figure 1). We revisit the validity of this design-time hypothesis in light of our results in Section 4.2.

For the anomaly-specific encoder, we adopt a ViT trained under a Multiple Instance Learning (MIL) framework. Compared to a fully supervised ViT, a MIL-trained model produces contextual patch-level features rather than representations tied to pixel-level labels, which we hypothesized would align more naturally with CLIP's semantic space. Details of the MIL training setup are provided in Section 3.1.

2.1. Dual-encoder Architecture and Cross-Modal Fusion

The proposed dual-encoder architecture combines visual features from two complementary sources: the general-purpose vision encoder of the base VLM and the MIL-trained ViT described above. Given a sequence of video frames, each encoder processes the input independently to produce its own set of feature representations. These two feature sets are then merged through a fusion step into a single unified embedding before being passed to the language model.

We adopt a concatenation-based fusion strategy: the feature sequences from both encoders are directly concatenated along the token dimension and fed into the LLM. This preserves all information from both sources without any learned transformation or selection. The fused embedding is then mapped into the LLM's input space through a two-layer Multi-Layer Perceptron (MLP) projection layer (Liu et al., 2024).

3. Experiments

3.1. Experimental Setup

We use LLaVA-OneVision 7B (Li et al., 2024) as the base VLM, with its pretrained weights kept intact. For the anomaly-specific encoder, we train a DeiT-Small (Touvron et al., 2021) on UCF-Crime (Sultani et al., 2018) under a Multiple Instance Learning (MIL) framework with a pairwise ranking loss between anomalous and normal segments. The trained encoder achieves a frame-level Area Under the Curve (AUC) of 0.86 on the UCF-Crime test set.

All experiments are conducted on the UCF-Crime Annotation (UCA) dataset (Yuan et al., 2024), which provides event-level natural-language captions for UCF-Crime videos. Each annotation pairs a temporal segment with a caption describing the event, covering both anomalous scenes (e.g., arson, burglary, fighting) and normal activities. We construct video-frame-sequence–caption pairs from this dataset for training and evaluation. To assess the quality of generated captions, we report Sentence-BERT (SBERT) cosine similarity between model outputs and ground-truth annotations as a semantic similarity measure (Table 2). We compare three configurations: (A) the unmodified base VLM (Original LLaVA), (B) a single-encoder LoRA baseline (LLaVA + LoRA), and (C) the proposed dual-encoder variant with concatenation-based fusion (Dual-Encoder Concat). To complement this caption-level evaluation, we additionally probe each model as a closed-set classifier over UCF-Crime anomaly categories; the motivation, setup, and results of this probe are presented in Section 4.2.

Table 1. Qualitative comparison of generated captions. All models receive the same prompt: *"Describe what is happening in this surveillance video."* Sample 1 is an arson scene where all configurations fail to identify the anomaly. Sample 2 is a fighting scene where (B) captures the action while (C) tends toward more generic, template-like phrasing.

Configuration	Sample 1 (Arson036)	Sample 2 (Fighting044)
Ground Truth	The man put down the gasoline can and tore up the cash register machine on the table.	The two people above stood up after being beaten.
(A) Original LLaVA	A person in a white shirt working at a bar, organizing and cleaning the area.	A chaotic scene with a group of people engaged in a physical altercation.
(B) LLaVA + LoRA	The man in white walked to the counter and bent down to pick up the bottle on the ground.	A man in blue clothes walked to the man in white clothes and pushed him to the ground.
(C) Dual (Concat)	The man in white walked to the counter and took out a bag from the cabinet.	A man in a blue shirt and a man in a black shirt walked out of the door.

3.2. Training Strategy

Both vision encoders—the CLIP encoder within LLaVA and the MIL-trained DeiT-Small—are fully frozen throughout training to preserve their respective pretrained representations. The original multi-modal projector of LLaVA is also frozen to prevent distortion of the base model's visual-semantic alignment and to reduce GPU memory consumption.

For the language model of LLaVA, namely Qwen2-7B, we apply LoRA to the query and value projection matrices (`q_proj`, `v_proj`) of the self-attention layers. All other LLM parameters remain frozen. Consequently, the only trainable components in the entire pipeline are: (1) the LoRA adapters within the LLM and (2) the MIL projection layer that maps the MIL encoder's features into the LLM embedding space.

The model is optimized with a standard causal language modeling loss, in which cross-entropy is computed only over the response tokens; prompt tokens are masked with a label of -100 and excluded from the loss calculation.

4. Results

4.1. Overview of Captioning Results

Table 2 reports SBERT cosine similarity for each configuration. The ordering—(A) $>$ (C) $>$ (B)—is informative. The base model (A) achieves the highest similarity, the dual-encoder variant (C) sits in the middle, and the LoRA-only baseline (B) is lowest. Notably, (C) $>$ (B): adding the MIL-ViT stream on top of LoRA partially *recovers* the SBERT drop induced by LoRA fine-tuning alone, indicating that the second visual stream does not corrupt caption-level grounding—if anything, it mitigates LoRA's degradation. The qualitative examples in Table 1 are consistent with this ordering: (A) produces verbose generic descriptions that miss the anomaly but share vocabulary with the descriptive UCA captions; (B) generates short, action-shaped sentences that occasionally capture the action (e.g., the fighting scene in Sample 2) but more often oversimplify; and (C) tends toward more stereotyped phrases such as "a man walked to the door/road." This mixed caption-level picture motivates a complementary probe of what visual information actually reaches the model's predictions, presented next.

Table 2. SBERT cosine similarity between generated and ground-truth captions on the UCA test set.

Configuration	SBERT Similarity
(A) Original LLaVA	0.388
(B) LLaVA + LoRA	0.315
(C) Dual-Encoder (Concat)	0.353

Table 3. Top-k closed-set classification accuracy on filtered UCF-Crime test segments. Each model is prompted to choose among 13 anomaly classes. Random baselines: $1/13 \approx 0.077$, $2/13 \approx 0.154$, $3/13 \approx 0.231$.

Configuration	Top-1	Top-2	Top-3
(A) Original LLaVA	0.434	0.525	0.613
(B) LLaVA + LoRA	0.328	0.463	0.581
(C) Dual-Encoder (Concat)	0.381	0.534	0.636

4.2. Closed-Set Classification: Evidence of Successful Knowledge Injection

Caption-level metrics conflate generation fluency with the underlying visual discrimination. To isolate the latter, we evaluate each configuration as a closed-set classifier. For each video in the UCF-Crime test set, we localize the anomaly segment via class-specific keyword filtering on UCA captions (e.g., *fire*, *smoke* for *Arson*), extract 16 uni-

formly sampled frames for the CLIP encoder and 32 sliding-window features for the MIL-ViT, and prompt the model with: *"What is the anomaly in this video? Choose one from [Abuse, Arson, . . . , Vandalism]."* We report Top-k accuracy ($k = 1, 2, 3$) to characterize both peak prediction and the breadth of class-relevant signal in the output distribution.

Table 3 reports the results, which provide the most direct evidence for the dual-encoder design.

(C) surpasses the base model at Top-2 and Top-3. The dual-encoder variant achieves Top-2 accuracy of 0.534 versus 0.525 for (A), and Top-3 of 0.636 versus 0.613. Although the absolute margins are modest, the direction is informative: MIL-ViT injection contributes *additional* anomaly-discriminative signal beyond what CLIP alone provides. This is direct empirical support that the dual-encoder architecture achieves its intended goal of injecting anomaly-specific knowledge at the representation level—a result not visible from caption-level metrics alone.

(C) trails at Top-1: a calibration gap, not a signal gap. At Top-1, (C) lags (A) (0.381 vs. 0.434). The Top-1-to-Top-3 expansions are revealing: $+0.179$ for (A), $+0.253$ for (B), and $+0.255$ for (C). (C)'s larger expansion indicates a flatter, less-peaked output distribution—the correct class is reliably present in the top candidates but does not always surface as the single most probable token. What MIL-ViT injection costs is therefore not class information but *calibration*: the fusion-to-generation interface fails to commit to one anomaly class with high confidence, even when it has identified the correct candidate set.

(B) as a control baseline. (B) underperforms (A) at every k but remains well above the random baselines, consistent with the modest SBERT drop observed earlier. LoRA fine-tuning alone shifts the output distribution without contributing new discriminative signal—confirming that the gains in (C) at Top-2/Top-3 originate from the MIL-ViT stream, not from LoRA itself.

4.3. Caption-Level Limitations and the Path Forward

The classification probe shows that (C)'s underlying visual discrimination is intact and, on Top-2/Top-3 criteria, improved over (A). Free-form captions, however, tell a more constrained story.

Stereotyped templates in (C). Despite receiving both CLIP and MIL-ViT tokens, configuration (C) often produces caption-level outputs of the form "a man walked to the door/road" that show limited variation across input scenes (Table 1). This is not a collapse of grounding—(C)'s SBERT score lies between (A) and (B), and the classification probe shows class-relevant information remains accessible—but the LLM, faced with two heterogeneous visual streams, tends to gravitate toward high-prior linguistic templates rather than committing to scene-specific descriptions. We read this as the open-ended counterpart of the calibration gap visible in the Top-1 results: the model has the right candidates but does not commit to scene-specific tokens.

Why early fusion makes confident generation hard. Our architecture performs early fusion—merging the two visual streams before the language model—which forces the LLM to consume an unaligned mixed signal. Successful multi-encoder VLMs such as BLIP-2 (Li et al., 2023) avoid this regime via a dedicated alignment stage (e.g., Q-Former pretraining on image-text pairs) before the LLM sees the fused features. Our results indicate that when this alignment stage is replaced by LoRA-based fine-tuning alone, anomaly-specific discriminative signal does pass through—enough to surpass the base model on Top-2/Top-3—but is not yet translated into confidently calibrated generation. Closing this gap, rather than re-engineering the visual encoders, is the natural next step.

5. Conclusion and Future Work

This paper investigated whether injecting anomaly-specific visual knowledge into a VLM through a dual-encoder architecture could improve Video Anomaly Detection. Our closed-set classification probe provides direct evidence that the approach succeeds at the representation level: the dual-encoder variant surpasses the base model on Top-2 and Top-3 anomaly classification (0.534 vs. 0.525 and 0.636 vs. 0.613), demonstrating that MIL-ViT injection contributes additional discriminative signal beyond what CLIP alone provides. At the caption level the picture is more nuanced: the dual-encoder model partially recovers the SBERT degradation induced by LoRA fine-tuning yet still tends toward stereotyped templates, and at Top-1 it trails the base model. The pattern of Top-k expansions identifies the bottleneck as a *calibration gap*—the fusion-to-generation interface fails to commit confidently to a single scene-specific token, even though the correct class is reliably available among the top candidates.

These findings suggest two concrete directions for future work. First, an explicit alignment pretraining stage—such as a lightweight bridging module (e.g., a Q-Former) trained on image-text pairs to project MIL-ViT features into the CLIP-aligned space—appears necessary *before* LoRA-based fine-tuning of the LLM, as our results indicate that this alignment cannot be deferred to the LoRA stage itself. Second, late-fusion strategies that preserve each encoder's representation until the final decoding step, rather than forcing early integration through token concatenation, could sidestep the unaligned mixed signal that produces the observed calibration gap.

Impact Statement

This paper aims to advance the general field of vision-language modeling for video understanding. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., and Li, C. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Lv, H. and Sun, Q. Video anomaly detection and explanation via large language models. *arXiv preprint arXiv:2401.05702*, 2024.

Sultani, W., Chen, C., and Shah, M. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6479–6488, 2018.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, volume 139, pp. 10347–10357, 2021.

Ye, M., Liu, W., and He, P. VERA: Explainable Video Anomaly Detection via Verbalized Learning of Vision-Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 8679–8688, 2025.

Yuan, T., Zhang, X., Liu, K., Liu, B., Chen, C., Jin, J., and Jiao, Z. Towards surveillance video-and-language understanding: New dataset baselines and challenges. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22052–22061, 2024.

Zanella, L., Menapace, W., Mancini, M., Wang, Y., and Ricci, E. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18527–18536, 2024.