# FROM DETECTION TO TRANSLATION: RETHINKING EVIDENCE SAMPLING FOR VIDEO ANOMALY UNDERSTANDING

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

## ABSTRACT

Video Anomaly Understanding (VAU) aims to identify abnormal events in videos and explain their visual evidence, event process, and semantic causes in natural language. Existing VAU pipelines commonly rely on uniform sampling or anomaly-score-based filtering before Multimodal Large Language Model (MLLM) reasoning, where subtle but critical abnormal frames may be removed before the reasoning stage. To address this challenge, we propose a translation-based sampling framework: instead of making coarse anomaly decisions during sampling, our method translates all video clip units into compact visual-semantic representations and performs text-driven redundancy merging to retain representative clips with their corresponding text priors. We further integrate scene context into motion-aware translation and accelerate evidence construction via dynamic token pruning to improve both accuracy and efficiency. Experiments on HIVAU-70k show that, compared with Holmes-VAU baseline, our method improves average BLEU and ROUGE performance by 13.4% and 20.4%. The codes will be public after acceptance.

## 1 INTRODUCTION

Video Anomaly Understanding (VAU) extends conventional detection to language-grounded event analysis, aiming to identify abnormal events from an input video and explain their anomaly type, event process, visual evidence, and semantic cause in natural language (Yuan et al., 2023; Lv & Sun, 2024; Zhang et al., 2025; Zhu et al., 2025; Cheng et al., 2026). This task holds profound practical value across safety-critical domains, including intelligent surveillance, traffic accident analysis, emergency response, and video forensics. However, VAU is inherently challenging due to the sparsity, brevity and high context-dependency of abnormal events.

Existing VAU methods typically adopt a two-stage pipeline. First, they sample a small set of frames from the entire input video, using the anomaly-score-based selection strategy, where a lightweight anomaly detector assigns preliminary anomaly scores to all candidate frames (Lv & Sun, 2024; Zhang et al., 2025; Cheng et al., 2026). Second, the selected visual content is fed into a Multimodal Large Language Model (MLLM) for anomaly analysis (Maaz et al., 2024; Zhang et al., 2023; Lin et al., 2023). Finally, the model generates anomaly judgments, event descriptions, or natural language answers based on the sampled visual input.

However, the main limitation of such a detection-based sampling is that it couples evidence selection with an early and coarse anomaly decision, as shown in Figure 1(a). Since the front-end detector usually has limited semantic understanding, it can be easily affected by low-resolution inputs, background distractions or contextual ambiguity. If a truly abnormal clip is incorrectly filtered out at this stage, downstream MLLM no longer receives the critical visual evidence and can only reason over incomplete cues. Thus, early false negatives are irreversibly propagated to the final prediction, often leading to unreliable anomaly understanding.

To address this problem, we propose a translation-based sampling method for VAU, as shown in Figure 1(b). Different from detection-based sampling, our method does not directly decide which part is abnormal during sampling. Instead, it converts video content into textual semantic representations faithfully. Specifically, we first divide the whole video into many short clips with equal duration,



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.

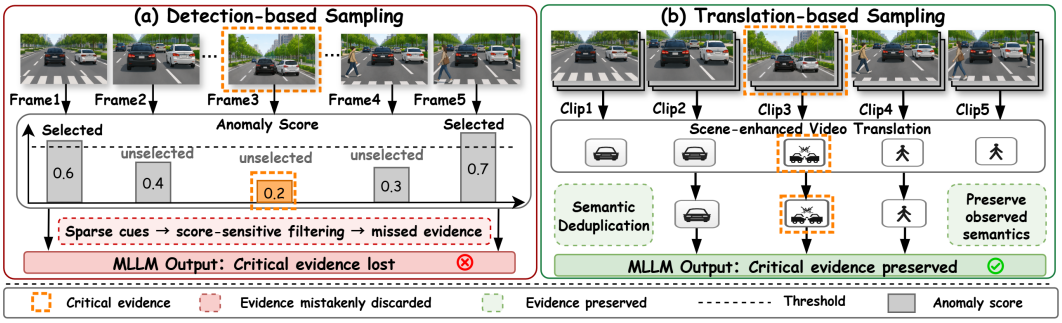


Figure 1: Illustration of different sampling pipelines. (a) Traditional detection-based sampling relies on front-end anomaly scoring that may miss brief abnormal cues. (b) Our translation-based sampling translates clips into semantic descriptions and merges redundancy, preserving critical evidence for downstream MLLM reasoning.

and then translate each clip into text that describes its visual content. We then merge repeated or highly similar clips according to textual semantics, removing only redundant information without actively filtering out clips that appear normal or abnormal. Finally, the retained clips and their textual descriptions are fed into the downstream multimodal large model, which performs anomaly judgment and explanation in a unified manner. This strategy changes the goal of sampling from early anomaly detection to video redundancy compression with semantic coverage, thereby reducing the risk that key abnormal clips are mistakenly discarded at the front-end stage. Since it does not rely on a specific MLLM, it integrates into existing VAU pipelines as a plug-and-play component.

To further improve the sampling quality of translation-based sampling, we incorporate scene context into the clip translation process. The reason is that the same action may be normal or abnormal under different scenes. For example, physical contact can be routine in a sports field but suspicious in a public street, and taking an object into pocket can be ordinary at home but abnormal in a store. Therefore, unlike prior sampling methods that mainly rely on motion or action cues, we design the gated motion-scene translation module to explicitly incorporate scene context. With this design, the translated clip representations better reflect both what happens and where it happens, providing the downstream MLLM with more complete and context-sensitive evidence. Our contributions are summarized as follows:

- We propose a translation-based sampling framework for the VAU task, which translates candidate video clips into compact visual-semantic evidence, removes redundancy at the text level, and provides both visual evidence and selected text priors to the MLLM. By avoiding premature anomaly judgment before MLLM reasoning, our framework better preserves sparse, short-lived, but critical anomaly evidence for downstream understanding.
- We further design an ultra-efficient scene-enhanced translator as the core of the proposed framework. The gated motion-scene translation module integrates motion-oriented dynamics with scene-oriented context, while the translator acceleration module reduces redundant computation and preserves semantically informative regions.
- Experiments on HIVAU-70k demonstrate consistent gains across three temporal granularities, improving average BLEU and ROUGE by 13.4% and 20.4% over the Holmes-VAU 2B baseline, with further generalization to other MLLM backbones.

## 2 Related Work

### 2.1 Anomaly Detection and Understanding

Video anomaly detection has evolved from reconstruction, prediction, and temporal modeling (Hasan et al., 2016; Liu et al., 2018) to weak supervision (Sultani et al., 2018; Wu et al., 2020; Tian et al., 2021) and vision-language alignment (Chen et al., 2023; Zhou et al., 2023; Wu et al., 2024). However, most methods output scores, boundaries, or categories rather than explanatory evidence. Recent VAU works introduce language annotations, multimodal explanations, and advanced

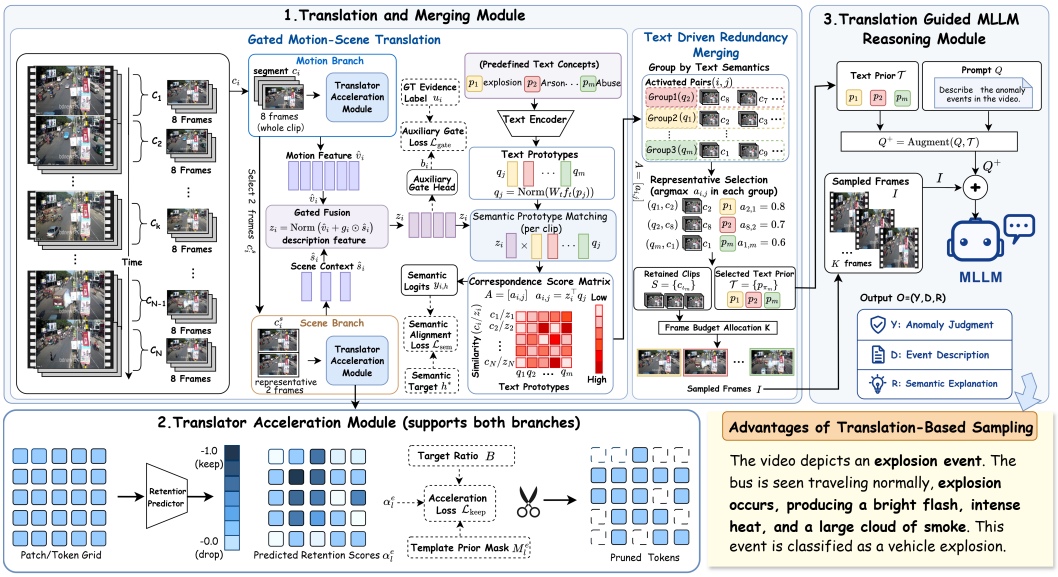


Figure 2: Illustration of the translation-based sampling framework. (a) Translation and Merging: the video is split into clips; each clip is fed into a gated motion-scene translator to obtain description features, which are matched with text prototypes to form a correspondence score matrix; semantically similar clips are grouped, and the most representative clip-text pair per group is retained. (b) Acceleration: branch-aware dynamic token pruning reduces redundant visual tokens for both branches. (c) MLLM Reasoning: sampled frames and selected text priors are augmented into the prompt for structured anomaly understanding.

reasoning (Yuan et al., 2023; Lv & Sun, 2024; Zhang et al., 2025; Zhu et al., 2025; Cheng et al., 2026), but still rely on sampling or filtering, making pre-reasoning evidence preservation difficult.

### 2.2 Evidence Sampling for Video LLMs

Video LLMs support video-language reasoning (Maaz et al., 2024; Zhang et al., 2023; Lin et al., 2023), while foundation models and benchmarks emphasize temporal and long-context understanding (Wang et al., 2024b; Fu et al., 2025). Yet VAU is constrained by frame and token budgets, where key evidence may be brief or scene-dependent. Long-video methods use memory, compact tokens, streaming, or adaptive compression (Li et al., 2023; Song et al., 2024; Qian et al., 2024; Shen et al., 2025; Shu et al., 2025), but sampling or compression may remove explanation-critical evidence. Scene context further shapes event semantics (Carreira & Zisserman, 2017; Wang & Gupta, 2018; Feichtenhofer et al., 2019).

## 3 Methodology

### 3.1 Overview

Given an input video $V = \{x_t\}_{t=1}^{T}$ and a task instruction $Q$, where $x_t$ denotes the $t$-th video frame, $t$ is the temporal frame index, and $T$ is the total number of frames, video anomaly understanding aims to generate a structured explanatory output $\mathcal{O} = (Y, D, R)$. Here, $Y$ denotes the decision-level anomaly judgment, $D$ denotes the event-level description, and $R$ denotes the reasoning-level semantic cause. Our framework performs translation-based sampling before MLLM reasoning, as shown in Figure 2. First, the video $V$ is partitioned into consecutive candidate clips $\{c_i\}_{i=1}^{N}$, where $c_i$ denotes the $i$-th clip and $N$ is the total number of candidate clips. Second, a lightweight translator converts each clip $c_i$ into a description feature $z_i$ by combining motion-oriented and scene-oriented cues. Third, each description feature $z_i$ is matched with the predefined text set $\mathcal{P}$, where $\mathcal{P}$ contains textual descriptions used to describe event-related semantics. This matching process produces text-level correspondences and semantic relevance scores. Fourth, semantically similar text correspondences are merged, and the most relevant clip in each group is sampled under a fixed frame

budget to form $\mathcal{I}$, with the matched texts forming $\mathcal{T}$. Finally, the sampled video frames $\mathcal{I}$, the selected text prior $\mathcal{T}$, and the task instruction $Q$ are packed into a structured input for an MLLM $G_\theta$, where $\theta$ denotes model parameters. For simplicity, this reasoning process is written as follows:

$$\mathcal{O} = (Y, D, R) = G_\theta(\mathcal{I}, \mathcal{T}, Q). \tag{1}$$

This formulation shows that the translator constructs compact visual-semantic evidence, while the MLLM produces the final structured explanation across decision, event, and reasoning levels.

### 3.2 Translation and Merging Module

#### 3.2.1 Gated Motion-Scene Translation.

The proposed method first organizes the video stream into local consecutive clips and then performs semantic translation and scoring on these clip units. This avoids premature frame-level or score-level filtering before local events are represented in the text-related semantic space. The candidate clip is defined as:

$$c_i = \{x_{\eta_i}, x_{\eta_i+1}, \ldots, x_{\eta_i+L-1}\}, \quad i = 1, \ldots, N, \tag{2}$$

where $L$ denotes the clip length, and $\eta_i = (i-1)L + 1$ denotes the starting frame index of $c_i$. This clip-level formulation provides the basic units for subsequent semantic translation and text-level correspondence matching.

For each candidate clip, the translator uses two complementary inputs to capture diverse clues. The motion branch takes the entire clip $c_i$ as input and encodes temporal changes within the clip, producing the motion-oriented feature $\hat{v}_i$. The scene branch takes two representative frames from the same clip as input, denoted as $c_i^{\text{sce}} = \{x_{\rho_{i,1}}, x_{\rho_{i,2}}\} \subset c_i$, where $\rho_{i,1}$ and $\rho_{i,2}$ are the temporal indices of the selected frames and satisfy $\rho_{i,1}, \rho_{i,2} \in [\eta_i, \eta_i + L - 1]$. This component encodes relatively stable scene cues and produces the scene-oriented feature $\hat{s}_i$. Both features are projected and normalized inside the translator before fusion. To avoid injecting scene information unconditionally, a learnable gate is used to adaptively control the contribution of the scene-oriented feature. The final description feature is computed as:

$$g_i = \sigma\left(W_g[\hat{v}_i; \hat{s}_i]\right), \quad z_i = \mathrm{Norm}\left(\hat{v}_i + g_i \odot \hat{s}_i\right), \tag{3}$$

where $g_i$ denotes the learnable gate, $z_i$ denotes the final description feature of clip $c_i$, $W_g$ denotes a trainable gating projection, $\sigma(\cdot)$ denotes the sigmoid function, $[\cdot;\cdot]$ denotes feature concatenation, $\odot$ denotes element-wise multiplication, and $\mathrm{Norm}(\cdot)$ denotes feature normalization. This gated fusion allows the translator to preserve motion-sensitive evidence while adaptively incorporating scene-level context. In addition, the auxiliary gate head predicts a binary evidence logit $b_i = \phi_b(z_i)$ from the description feature, where $\phi_b(\cdot)$ denotes a trainable auxiliary gate head.

To compare the translated clip representation with textual semantics, we construct a predefined text set $\mathcal{P} = \{p_j\}_{j=1}^{P}$ that describes event-related visual concepts. Each text is encoded into a text prototype in the same latent space as the clip description feature, and the clip-text correspondence is computed by the normalized inner product:

$$a_{i,j} = z_i^\top q_j, \quad q_j = \mathrm{Norm}(W^{\text{text}} f^{\text{text}}(p_j)), \tag{4}$$

where $p_j$ denotes the $j$-th text in $\mathcal{P}$, $P$ denotes the total number of texts, $q_j$ denotes the text prototype of $p_j$, $f^{\text{text}}(\cdot)$ denotes the text encoding function, $W^{\text{text}}$ denotes the text projection layer, and $a_{i,j}$ denotes the correspondence score between clip $c_i$ and text $p_j$. This text-level correspondence provides the basis for subsequent semantic grouping and redundancy removal.

#### 3.2.2 Text Driven Redundancy Merging.

To merge the textual redundancy, we first select a set of clips that represent distinct semantic types in the video. Specifically, we construct a semantic group set $\{\mathcal{G}_m\}_{m=1}^{N^{\text{ret}}}$, where each $G_m$ contains multiple clip-text correspondences of the same semantic category. Note that the number of retained groups, $N^{\text{ret}}$, is specified to preserve sufficient semantic diversity. Within each group, the clip with the highest correspondence score is selected as its semantic-peak clip, and all semantic-peak clips form the semantic-peak clip set $\mathcal{S}$, as follows:

$$\mathcal{S} = \{c_{\iota_m}\}_{m=1}^{N^{\text{ret}}}, \quad (\iota_m, \pi_m) = \underset{(i,j)\in\mathcal{G}_m}{\mathrm{argmax}}\, a_{i,j}, \tag{5}$$

where $\iota_m$ and $\pi_m$ denote the selected clip index and text index in $\mathcal{G}_m$. Each clip represents the strongest visual evidence for one semantic type, while their matched texts collectively form the corresponding semantic text set $\mathcal{T} = \{p_{\pi_m}\}_{m=1}^{N^{\text{ret}}}$.

Then, we adjust the frame numbers of clips within $\mathcal{S}$ to construct the final sampled frame set $\mathcal{I}$ under a total frame budget $K$. Each semantic-peak clip is assigned a target number of $k_i$ frames. The total budget $K$ is first evenly divided among all semantic-peak clips, and any remaining frames are assigned to clips with higher semantic relevance scores. Given the allocated $k_i$, we uniformly sample $k_i$ frames from each clip and arrange all sampled frames according to their original temporal order, calculated as follows:

$$\mathcal{I} = \text{Sort}\left(\bigcup_{c_i \in \mathcal{S}} \text{LinSample}(c_i, k_i)\right), \quad \sum_{c_i \in \mathcal{S}} k_i = K, \tag{6}$$

where $\text{LinSample}(c_i, k_i)$ uniformly extracts $k_i$ frames from clip $c_i$, and $Sort$ indicates the sorting process. For example, given three semantic groups with three 8-frame semantic-peak clips and a total budget of 12 frames, each clip is reduced to four uniformly sampled frames. This allocation preserves temporal order and assigns slightly more capacity to clips with stronger semantic evidence.

#### 3.2.3 Dual Objective Joint Optimization.

The translation module is optimized with two objectives: a semantic alignment objective and an auxiliary evidence-discrimination objective. The semantic alignment loss encourages each clip's representation to match its anomaly category. Each category $h \in \{1, \ldots, C\}$ corresponds to a set of anomaly texts $\mathcal{P}_h \subseteq \mathcal{P}$, where $C$ is the number of anomaly categories. For clip $c_i$, $a_{i,j}$ denotes its similarity to the anomaly text $p_j$. The category-level logit is obtained by: $y_{i,h} = \max_{p_j \in \mathcal{P}_h} a_{i,j}$, where the maximum selects the strongest prototype in category $h$. Given the ground-truth category $h^*$, the semantic alignment loss is:

$$\mathcal{L}_{\text{sem}} = -\log \frac{\exp(y_{i,h^*})}{\sum_{h'=1}^{C} \exp(y_{i,h'})}. \tag{7}$$

It enhances the GT category response and suppresses others.

The auxiliary gate loss improves the sensitivity of translated representations to informative evidence. Given $z_i$, the auxiliary gate head $\phi_b(\cdot)$ predicts $b_i = \phi_b(z_i)$. With binary label $u_i \in \{0, 1\}$, where 1 denotes informative evidence and 0 denotes normal or weak content, the gate loss is as follows:

$$\mathcal{L}_{\text{gate}} = \text{BCEWithLogits}(b_i, u_i). \tag{8}$$

This loss encourages evidence-aware clip representations. Jointly optimizing $\mathcal{L}_{\text{sem}}$ and $\mathcal{L}_{\text{gate}}$ produces semantically discriminative and evidence-sensitive representations for representative clip selection before MLLM reasoning.

### 3.3 Translator Acceleration Module

To reduce computation, we apply dynamic token pruning to both the motion ($e = \text{mot}$) and scene ($e = \text{sce}$) branches. At pruning layer $l$, a lightweight predictor $\phi_l^e(\cdot)$ assigns one importance logit to each patch token in $H_l^e$. A sigmoid converts these logits into soft retention scores $\alpha_l^e = \sigma(\phi_l^e(H_l^e))$, where each score indicates its token's retention priority.

We use a predefined global retention-ratio hyperparameter $B \in (0, 1]$ to control the token budget. The retention scores $\alpha_l^e$ determine which tokens to preserve, while $B$ determines the overall proportion of retained tokens. During training, the average retention score is regularized toward $B$, and $\alpha_l^e$ acts as a differentiable soft mask without changing the sequence length. During inference, $\text{TopPrune}_B(\cdot)$ ranks tokens by $\alpha_l^e$, retains the highest-scoring fraction $B$ of tokens, and preserves their original order, calculated as follows:

$$H_l^{e,\text{train}} = \alpha_l^e \odot H_l^e, \quad H_l^{e,\text{infer}} = \text{TopPrune}_B(H_l^e, \alpha_l^e), \tag{9}$$

where $\odot$ denotes token-wise multiplication, and $H_l^{e,\text{train}}$ and $H_l^{e,\text{infer}}$ denote the training-time soft-masked representation and inference-time pruned representation, respectively.

Table 1: Comparison of VAU performance with SOTA methods at clip-level (C), event-level (E), and video-level (V). Time denotes the average sampling time for each 8-frame unit. * indicates reproduced results. The highest results are marked in **bold**.

| Method | Params | BLEU ↑ | | | CIDEr ↑ | | | METEOR ↑ | | | ROUGE ↑ | | | Time (ms) ↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | C | E | V | C | E | V | C | E | V | C | E | V | |
| Video-ChatGPT | 7B | 0.152 | 0.068 | 0.066 | 0.033 | 0.011 | 0.013 | 0.102 | 0.069 | 0.044 | 0.153 | 0.048 | 0.079 | – |
| Video-LLaMA | 7B | 0.151 | 0.079 | 0.104 | 0.024 | 0.014 | 0.017 | 0.112 | 0.076 | 0.057 | 0.156 | 0.067 | 0.090 | – |
| Video-LLaVA | 7B | 0.164 | 0.046 | 0.055 | 0.032 | 0.009 | 0.013 | 0.097 | 0.022 | 0.014 | 0.132 | 0.023 | 0.045 | – |
| LLaVA-Next-Video | 7B | 0.435 | 0.091 | 0.120 | 0.102 | 0.015 | 0.031 | 0.117 | 0.085 | 0.096 | 0.198 | 0.080 | 0.106 | – |
| QwenVL2 | 7B | 0.312 | 0.082 | 0.155 | 0.044 | 0.020 | 0.044 | 0.133 | 0.092 | 0.112 | 0.163 | 0.081 | 0.137 | – |
| InternVL2 | 8B | 0.331 | 0.101 | 0.145 | 0.052 | 0.022 | 0.035 | 0.141 | 0.095 | 0.101 | 0.182 | 0.102 | 0.122 | – |
| Qwen2.5-VL | 3B | 0.329 | 0.165 | 0.152 | 0.094 | 0.434 | 0.093 | 0.130 | 0.113 | 0.096 | 0.205 | 0.166 | 0.156 | – |
| **Qwen2.5-VL + Ours** | 3B | **0.406** | **0.215** | **0.212** | **0.144** | **0.494** | **0.157** | **0.185** | **0.191** | **0.173** | **0.251** | **0.214** | **0.208** | – |
| Holmes-VAU* | 2B | 0.439 | 0.311 | 0.301 | 0.109 | 0.964 | 0.119 | 0.185 | 0.097 | 0.095 | 0.193 | 0.133 | 0.117 | 177.92 |
| **Holmes-VAU + Ours** | 2B | **0.527** | **0.513** | **0.413** | **0.281** | **1.307** | **0.437** | **0.241** | **0.193** | **0.198** | **0.322** | **0.367** | **0.367** | **49.74** |




Figure 3: Qualitative comparison between Holmes-VAU and our method on two representative cases. Green words indicate descriptions consistent with the ground truth, while red words indicate hallucinated, incorrect, or missing evidence. (a) shows a case where the baseline introduces unsupported visual details. (b) shows a case where the baseline provides an incomplete anomaly description. In both cases, our method generates more faithful responses to sampled visual evidence.

The acceleration module uses a pruning-ratio loss and a template-prior loss. The pruning-ratio loss, defined as $\ell_l^{e,\mathrm{pr}} = (\mathrm{Mean}(\alpha_l^e) - B)^2$, constrains the average retention score to follow the predefined token budget. The template-prior loss guides the retention scores toward category-relevant spatial positions. During training, a predefined class-specific binary template $M_l^e$ is selected according to the ground-truth category and aligned with the patch-token sequence. For example, the template for Fighting contains multiple spatial regions to encourage the preservation of tokens covering multi-person interactions. The template is used only during training, and the corresponding loss is defined as $\ell_l^{e,\mathrm{tpl}} = \mathrm{MSE}(\alpha_l^e, M_l^e)$. The overall acceleration loss is:

$$\mathcal{L}_{\mathrm{keep}} = \sum_e \sum_l \left( \ell_l^{e,\mathrm{pr}} + \lambda_{\mathrm{tpl}} \ell_l^{e,\mathrm{tpl}} \right), \tag{10}$$

where the sums cover all branches and pruning layers, and $\lambda_{\mathrm{tpl}}$ controls the contribution of the template-prior loss. Together, the two losses control the token budget and guide the model to preserve informative spatial regions, producing compact representations for MLLM reasoning.

### 3.4 Translation-guided MLLM Reasoning Module

A conventional MLLM inference pipeline usually receives a small set of sampled frames and a task instruction, and then directly generates the anomaly understanding result. In our framework, the translator provides two forms of evidence: the sampled frame set $\mathcal{I}$ and the selected text prior $\mathcal{T}$. The selected text prior does not determine the final answer directly. Instead, it is injected into the original task instruction $Q$ as a compact visual-semantic hypothesis. Let $Q^+$ denote the augmented instruction, $\mathrm{Augment}(\cdot)$ denote the operation that inserts $\mathcal{T}$ into $Q$, $\mathcal{X}$ denote the structured MLLM

Table 2: Semantic composition of scene branch activations.

| Group | Examples | Share |
|---|---|---|
| Static Scene | *RoadAcc, Burglary,Shoplift Abuse, Fighting,* | **66.43%** |
| Interaction | *Shooting* | 20.00% |
| Auxiliary | *Explosion* | 13.57% |

Table 3: Ablation of model components.

| Components | | | BLEU ↑ | | | METEOR ↑ | | |
|---|---|---|---|---|---|---|---|---|
| Motion | Scene | Prior | C | E | V | C | E | V |
| | ✓ | ✓ | 0.513 | 0.367 | 0.317 | 0.142 | 0.111 | 0.096 |
| ✓ | | ✓ | **0.534** | 0.438 | 0.324 | 0.181 | 0.166 | 0.173 |
| ✓ | ✓ | | 0.524 | 0.504 | 0.401 | 0.235 | 0.191 | **0.199** |
| ✓ | ✓ | ✓ | 0.527 | **0.513** | **0.413** | **0.241** | **0.193** | 0.198 |

input, and $\mathrm{Pack}(\cdot)$ denote the operation that organizes the sampled frames and the augmented instruction into the model input:

$$\mathcal{X} = \mathrm{Pack}(\mathcal{I}, Q^+), \quad Q^+ = \mathrm{Augment}(Q, \mathcal{T}). \tag{11}$$

With this structured input, the MLLM $G_\theta$ verifies the visual-semantic hypothesis against the sampled frames and generates the structured explanatory output:

$$\mathcal{O} = (Y, D, R) = G_\theta(\mathcal{X}) = G_\theta\left(\mathrm{Pack}(\mathcal{I}, Q^+)\right). \tag{12}$$

This design feeds the MLLM both visually informative clips and a structured semantic hint, reducing repetitive-content interference and preserving critical short events.

The final objective combines semantic alignment, auxiliary gate, and acceleration losses, as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{sem}} + \lambda_{\mathrm{gate}}\mathcal{L}_{\mathrm{gate}} + \lambda_{\mathrm{keep}}\mathcal{L}_{\mathrm{keep}}. \tag{13}$$

Here, $\lambda_{\mathrm{gate}}$ and $\lambda_{\mathrm{keep}}$ are weighting coefficients. This objective makes the translator semantically aligned, evidence-sensitive, and computationally efficient before the final MLLM reasoning stage.

# 4 Experiments

## 4.1 Experiment Setup

**Dataset.** We conduct experiments on HIVAU-70k (Zhang et al., 2025), a large-scale video anomaly understanding benchmark built upon real-world anomalous video such as UCF-Crime (Sultani et al., 2018) and XD-Violence (Wu et al., 2020). HIVAU-70k further provides instruction-style annotations for text-grounded anomaly reasoning, enabling evaluation at multiple granularities including clip-level, event-level, and video-level understanding.

**Metric.** We evaluate from two aspects: anomaly prior prediction and anomaly reasoning. Classification accuracy measures clip alignment with anomaly semantics. BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), METEOR (Banerjee & Lavie, 2005), and ROUGE (Lin, 2004) compare responses with ground-truth answers, reflecting lexical overlap, semantic relevance, and sentence-level consistency.

### 4.1.1 Implementation Details.

We use XCLIP-base-patch16 (Ma et al., 2022) and CLIP ViT-B/32 (Radford et al., 2021) for motion and scene encoding, respectively. Each candidate clip contains 8 frames, with 2 representative frames used for scene encoding. The total frame budget $K$ is 12, and the number of retained semantic groups $N^{\mathrm{ret}}$ is 4. Input frames are resized to $224 \times 224$, and the maximum text prompt length is 16. The base translator is trained for 30 epochs using AdamW with a batch size of 2, learning rate $1 \times 10^{-4}$, and weight decay $1 \times 10^{-4}$. The accelerated translator is trained for 15 epochs with the same optimizer settings and token pruning at layers 10 and 11. The global token retention ratio $B$ is 0.7. The auxiliary gate, acceleration, and template-prior loss weights are 0.1, 0.5, and 0.05, respectively. All experiments use PyTorch (Paszke et al., 2019) and 2 NVIDIA RTX 4090 GPUs.

## 4.2 Comparison with SOTA Methods

### 4.2.1 Quantitative Comparison.

Table 1 compares our method with general-purpose video MLLMs, including Video-ChatGPT (Maaz et al., 2024), Video-LLaMA (Zhang et al., 2023), Video-LLaVA (Lin et al., 2023),

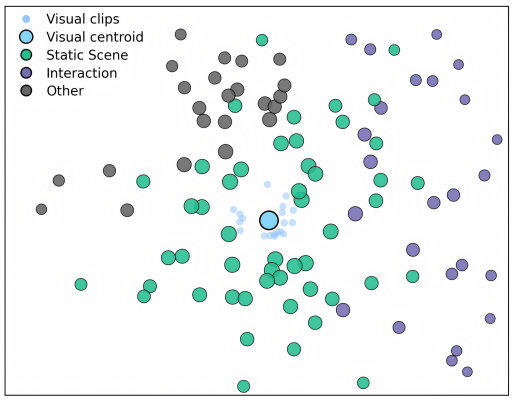


Figure 4: Probe-clip similarity layout for Burglary category. Blue points denote visual clip features extracted by the scene branch; colored points represent text-encoded probe features. Static scene probes (green) cluster near blue clips and the centroid, while interaction (purple) and auxiliary (gray) probes are more dispersed, confirming contextual activation.

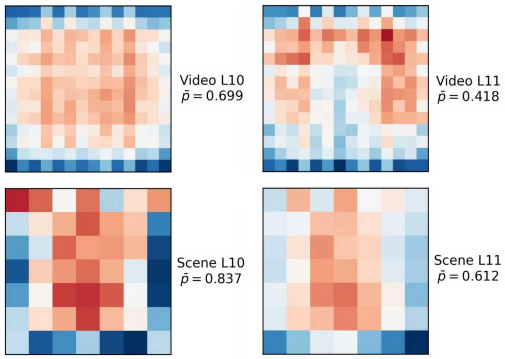


Figure 5: Spatial preference of the translator acceleration module relative to a uniform budget. Each heatmap shows the deviation of average keep probability from the branch/layer mean, where red indicates higher retention and blue indicates more aggressive pruning. The results demonstrate structured spatial preferences rather than uniform token removal.

LLaVA-Next-Video (Zhang et al., 2024), QwenVL2 (Wang et al., 2024a), InternVL2 (Chen et al., 2024), and the anomaly-specific Holmes-VAU baseline (Zhang et al., 2025). The gains over Holmes-VAU are metric-dependent: the baseline already excels at event-level CIDEr, while our method improves ROUGE more evidently at higher temporal granularities, yielding more complete and structurally aligned answers. METEOR gains are moderate at the clip level, as local descriptions are partially captured by the baseline, whereas our method better supports longer-range evidence aggregation. Meanwhile, the accelerated translator substantially reduces the average sampling time, demonstrating a favorable efficiency–performance trade-off. Consistent gains on Qwen2.5-VL-3B across all levels further confirm the module is backbone-agnostic and plug-and-play.

#### 4.2.2 Qualitative Comparison.

We present two qualitative examples in Figure 3 to verify whether explanations are grounded in sampled evidence rather than generic scene matching. Holmes-VAU either hallucinates unsupported details (a) or omits key anomaly evidence while describing generic scenes (b). In contrast, our method yields more faithful and evidence-aligned responses. It aligns with the quantitative gains in Table 1, confirming that improvements stem from reliable evidence selection and stronger visual-semantic grounding, not merely superficial lexical overlap.

#### 4.2.3 Scene Branch Interpretability.

We verify whether the scene branch captures genuine scene-oriented evidence. Using a manually grouped textual probe bank, we compute similarities between scene-branch features and text probe embeddings, categorizing top probes into static scene, interaction, and auxiliary cues. Static scene information dominates at 66.43% (Table 2), confirming the branch primarily responds to environment and entity-level cues. Figure 4 shows static scene probes cluster around visual clips and centroids, while others are more dispersed, supporting contextual and entity-level semantic activation. Probe details are in the appendix.

### 4.3 Ablation Study

#### 4.3.1 Main Component Contribution.

This ablation evaluates the contributions of motion evidence, scene branch and text prior $\mathcal{T}$, as shown in Table 3. Removing motion evidence causes the most severe degradation, confirming clips as the core source. Without the scene branch, higher-level stability drops, indicating scene cues pro-

Table 4: Sensitivity analysis of retained semantic groups. $N^{\text{ret}}$ denotes retained clips after text-driven redundancy merging.

| $N^{\text{ret}}$ | BLEU ↑ | | | METEOR ↑ | | |
|---|---|---|---|---|---|---|
| | C | E | V | C | E | V |
| 1 | 0.405 | 0.358 | 0.305 | 0.181 | 0.112 | 0.095 |
| 2 | 0.468 | 0.464 | 0.321 | 0.183 | 0.123 | 0.146 |
| 3 | 0.508 | 0.495 | 0.394 | **0.242** | 0.184 | 0.166 |
| 4 | **0.527** | **0.513** | **0.413** | 0.241 | **0.193** | **0.198** |

Table 5: Ablation of pruning strategies. "Time" and Spd. denote the average sampling time and "Spd" improvement over no pruning, respectively.

| Strategy | Time (ms)↓ | Spd. (%)↑ | BLEU | METEOR |
|---|---|---|---|---|
| No Prune | 3281 | 0.0 | 0.484 | 0.211 |
| Random | 3055 | 7.4 | 0.451 | 0.206 |
| Uniform | 3110 | 5.5 | 0.443 | 0.197 |
| TokenNorm | 3169 | 3.5 | 0.448 | 0.209 |
| **Ours** | **3045** | **7.8** | 0.467 | 0.209 |

Table 6: Ablation of training objectives. Sem. CE, Gate BCE, and Contrastive denote semantic alignment CE, auxiliary gate BCE, and prompt contrastive loss, respectively.

| Objective | BLEU ↑ | | | METEOR ↑ | | |
|---|---|---|---|---|---|---|
| | C | E | V | C | E | V |
| Sem. CE | 0.508 | 0.481 | 0.369 | 0.211 | 0.163 | 0.145 |
| Gate BCE | 0.500 | 0.458 | 0.358 | 0.140 | 0.142 | 0.097 |
| Contrastive | 0.510 | 0.459 | 0.380 | 0.191 | 0.163 | 0.149 |
| **Ours** | **0.527** | **0.513** | **0.413** | **0.241** | **0.193** | **0.198** |

vide contextual compensation. Removing text prior $\mathcal{T}$ weakens cross-granularity balance, guiding MLLM hypothesis verification. The full model achieves the most balanced performance, demonstrating that the three components are complementary.

**Sensitivity to Retained Semantic Groups.** We further study the effect of the number of retained semantic groups in text-driven redundancy merging, as shown in Table 4. Here, $N^{\text{ret}}$ denotes the number of retained clips after semantic grouping, where each retained clip corresponds to an independent semantic group. A smaller $N^{\text{ret}}$ reduces the diversity of the selected text priors and may limit the MLLM's ability to reason over diverse anomaly-related evidence. Increasing $N^{\text{ret}}$ provides broader semantic coverage and improves reasoning stability across different temporal granularities. Under the fixed 12-frame budget, we restrict $N^{\text{ret}}$ to at most 4, since retaining more groups would leave only two to three frames for each clip and weaken its temporal evidence. The default setting $N^{\text{ret}} = 4$ achieves the best overall balance, especially for event-level and video-level reasoning.

**Spatial Preference of Dynamic Token Pruning.** We analyze whether the translator acceleration learns structured spatial preferences. By mapping retained tokens to the patch grid and computing keep-probability deviations (Figure 5), we observe that the motion branch retains central and vertically structured regions more frequently while suppressing boundaries, indicating event-aware selection. The scene branch shows smoother preferences, consistent with its global contextual role. Table 5 confirms that the learned strategy achieves a better efficiency-quality trade-off, supporting semantic-aware token selection over uniform compression.

### 4.3.2 Exploration of Training Objectives.

The effectiveness comparison of different training objectives is shown in Table 6. Semantic alignment CE provides competitive performance, indicating that category-level semantic supervision is effective for learning anomaly-related representations. Gate BCE alone performs less stably, suggesting that binary evidence supervision is insufficient to preserve diverse visual-semantic cues. Contrastive supervision improves semantic alignment in several settings, but the full CE+BCE objective (Ours) achieves the best overall results across all temporal granularities. This indicates that semantic classification and evidence-aware supervision are complementary for constructing reliable visual-semantic evidence.

## 5 Conclusion

We present a translation-based sampling framework for VAU that preserves critical evidence via scene-enhanced translation and dynamic pruning. Future work may extend it to longer videos and finer temporal reasoning.

### AI USE STATEMENT

Generative AI tools were used solely for language editing to improve the grammar, clarity, conciseness, and readability of the manuscript. They were not used to develop the proposed method, design experiments, analyze data, interpret results, or generate scientific claims. All AI-assisted edits were reviewed and verified by the authors.

## REFERENCES

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724–4733, 2017.

Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 387–395, 2023.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

Ying Cheng, Yu-Ho Lin, Min-Hung Chen, Fu-En Yang, and Shang-Hong Lai. Vader: Towards causal video anomaly understanding with relation-aware large language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6202–6211, 2019.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 733–742, 2016.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models, 2023.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection, 2023.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics.

Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection: A new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6536–6545, 2018.

Hui Lv and Qianru Sun. Video anomaly detection and explanation via large language models, 2024.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 638–647. Association for Computing Machinery, 2022. doi: 10.1145/3503161.3547910.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 12585–12602, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8024–8035, 2019.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. In *Advances in Neural Information Processing Systems*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54582–54599. PMLR, 2025.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, 2018.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4975–4986, 2021.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, 2015. doi: 10.1109/CVPR.2015.7299087.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a. doi: 10.48550/arXiv.2409.12191.

Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision*, pp. 399–417, 2018.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, and Yali Wang. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, 2024b.

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 *Proceedings of the European Conference on Computer Vision*, pp. 322–339, 2020.

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, pp. 6074–6082, 2024.

Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, and Zhenzhen Jiao. Towards surveillance video-and-language understanding: New dataset, baselines, and challenges, 2023.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding, 2023.

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, 2025.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: A strong zero-shot video understanding model, April 2024.

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, pp. 3769–3777, 2023.

Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning, 2025.