Text-guided Fine-Grained Video Anomaly Understanding

Jihao Gu[1] Kun Li[2] He Wang[1] Kaan Akşit[1]
[1]University College London
[2]CVLab, College of Information Technology, United Arab Emirates University
{jihao.gu.23, he_wang, k.aksit}@ucl.ac.uk kunli.hfut@gmail.com

Abstract

*Subtle abnormal events in videos often manifest as weak spatio-temporal cues that are easily overlooked by conventional anomaly detection systems. Existing video anomaly detection approaches typically provide coarse binary anomaly decisions without interpretable evidence, while large vision-language models (LVLMs) can produce textual judgments but lack precise localization of subtle visual signals. To address this gap, we propose Text-guided Fine-Grained Video Anomaly Understanding (**T-VAU**), a framework that grounds subtle anomaly evidence into multimodal reasoning. Specifically, we introduce an Anomaly Heatmap Decoder (**AHD**) that performs visual-textual feature alignment to extract pixel-level spatio-temporal anomaly heatmaps from intermediate visual representations. We further design a Region-aware Anomaly Encoder (**RAE**) that converts these heatmaps into structured prompt embeddings, enabling the LVLM to perform anomaly detection, localization, and semantic explanation in a unified reasoning pipeline. To support fine-grained supervision, we construct a target-level fine-grained video-text anomaly dataset derived from ShanghaiTech and UBnormal with detailed annotations of object appearance, localization, and motion trajectories. Extensive experiments demonstrate that **T-VAU** significantly improves anomaly localization and textual reasoning performance on both benchmarks, achieving strong results in BLEU-4 metrics and Yes/No decision accuracy while providing interpretable pixel-level spatio-temporal evidence for anomaly understanding. The code will be available at https://github.com/momiji-bit/T-VAU.*

1. Introduction

Video Anomaly Detection (VAD) [2, 17, 29, 38, 39, 47] is critical for safety-critical visual systems, *e.g.*, public security monitoring and industrial inspection. In real-world scenarios, anomalies often manifest as *subtle spatio-temporal cues*—such as slight motion deviations, short-lived interactions, or fine-grained appearance changes—that are easily obscured by background clutter, compression artifacts, occlusions, and scene dynamics. This low-SNR characteristic makes reliable detection highly dependent on *interpretable pixel-level evidence*, a key aspect of subtle visual computing [9, 10, 14, 15, 23, 34, 50].

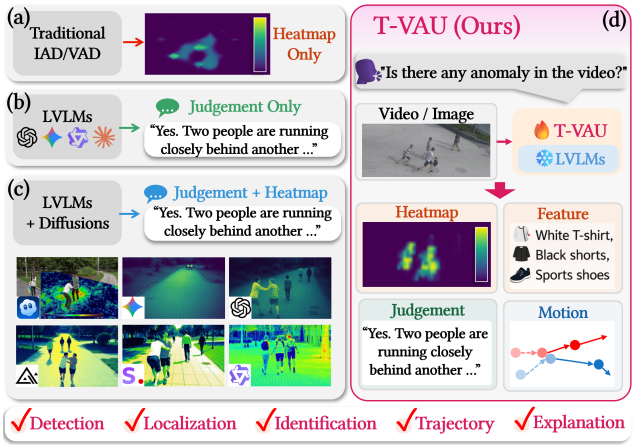

Figure 1. Comparison of anomaly detection paradigms and our **T-VAU**. (a) Traditional IAD/VAD outputs anomaly scores/heatmaps but lacks semantic explanation. (b) LVLMs provide textual judgments but lack pixel-level localization of subtle cues. (c) LVLM–diffusion hybrids combine visualization and text but may be unstable/inconsistent. (d) **T-VAU** couples a text-aligned **AHD** with a **RAE** to deliver pixel-level spatio-temporal localization and evidence-grounded reasoning (judgment, appearance, motion) for fine-grained anomaly understanding.

Mainstream VAD pipelines, including reconstruction-, prediction-, memory-based, and discriminative approaches, typically output video- or frame-level anomaly scores and rely on manual thresholds for decision-making [29, 39]. Although effective in many cases, such coarse outputs are insufficient for subtle and localized anomalies: fine-grained cues may be diluted by feature aggregation, and the lack of explicit evidence reduces robustness, generalization, and human trust. In practice, anomaly understanding often requires more than binary prediction, including *where* the

anomaly occurs, *which* target is responsible, and *how* it evolves over time. Beyond detection, anomaly understanding therefore aims to localize abnormal events, identify responsible targets, and provide interpretable explanations of their temporal evolution, aligning with localization-oriented evaluation protocols, such as region-based and track-based criteria [37].

Recently, vision–language pretraining [18, 25, 26, 36, 40] and large vision–language models (LVLMs) have significantly advanced open-vocabulary perception and multimodal reasoning [3, 5, 27, 55]. These models are appealing for anomaly understanding because language provides a flexible interface for *interactive querying* and *semantic explanation*. However, when directly applied to VAD, general-purpose LVLMs often struggle to *ground* subtle anomalies in pixel-level evidence, leading to unreliable localization and potentially unfaithful textual claims (Fig. 1). Meanwhile, text-guided anomaly localization has been explored in image anomaly understanding using VLP priors and fine-grained descriptions [11], but extending this capability to *video* remains challenging due to temporal dynamics and subtle motion cues.

To address these limitations, we propose **T-VAU**, an LVLM-based framework that closes the loop from *pixel-level evidence* to *language-based reasoning*. Our key idea is to (i) extract spatio-temporal anomaly evidence through visual–text alignment and (ii) inject this evidence into LVLMs as structured anomaly prompts for multi-task, multi-turn reasoning. Specifically, we introduce an Anomaly Heatmap Decoder (**AHD**) that aligns intermediate visual tokens with a *built-in normal–abnormal prompt pair* to generate *threshold-free* anomaly heatmaps with high spatial precision. We further design a Region-aware Anomaly Encoder (**RAE**) that converts heatmap evidence into region- and motion-aware prompt embeddings, bridging low-level anomaly evidence with high-level LVLM priors. These components enable unified anomaly judgment, localization, target identification, appearance description, and motion or trajectory reasoning through multi-turn interaction.

To facilitate fine-grained anomaly understanding, we construct a target-level video–text anomaly dataset based on ShanghaiTech and UBnormal, with explicit supervision on *appearance, localization, and motion trajectories*. This dataset allows **T-VAU** to learn consistent mappings from pixel-level evidence to faithful language descriptions, improving both interpretability and generalization.

In summary, our contributions are threefold:

(1) We present **T-VAU**, a fine-grained video anomaly understanding framework built upon large vision–language models. It integrates an Anomaly Heatmap Decoder (**AHD**) for precise pixel-level anomaly localization and a Region-aware Anomaly Encoder (**RAE**) for injecting region-aware and motion-aware evidence into language reasoning, enabling a unified solution for anomaly detection, localization, and multi-turn explanation.

(2) We construct a fine-grained anomaly understanding dataset based on two large-scale benchmarks, ShanghaiTech and UBnormal, supporting both target localization and language-based anomaly understanding. It provides frame-wise annotations for ShanghaiTech (4,108 training / 1,028 validation) and target-aligned descriptions for ShanghaiTech (5,136 samples) and UBnormal (7,912 samples), covering abnormal targets, appearance attributes, and motion trajectories.

(3) We achieve state-of-the-art performance across anomaly judgment, localization, and dialog-based explanation benchmarks, demonstrating the effectiveness of our framework for fine-grained and explainable video anomaly understanding.

2. Related Work

2.1. Video Anomaly Understanding

Video anomaly detection (VAD) has long been dominated by the paradigm of learning normality and identifying deviations at test time, with representative directions including weakly supervised multiple-instance learning, predictive modeling, memory-based modeling, and more recent discriminative or density-based formulations [29, 30, 33, 39, 54]. While these methods have substantially advanced benchmark performance, most of them primarily output frame- or clip-level anomaly scores, making it difficult to support fine-grained reasoning about *where* an anomaly occurs, *which* target is responsible, and *how* it evolves over time. This limitation has motivated a shift toward localization-oriented evaluation and finer anomaly localization, as reflected by region- and track-based protocols [37] and benchmarks with pixel-level annotations such as UBnormal [1]. In parallel, vision-language pretraining has introduced a new line of work that injects semantic priors into VAD, including prompt-based weakly supervised methods such as VadCLIP and STPrompt [45, 46], open-vocabulary formulations such as OVVAD [44], and training-free or reasoning-oriented frameworks built upon large language models, *e.g.*, LAVAD and AnomalyRuler [48, 52].

Recently, video anomaly understanding has emerged as a distinct setting, with datasets and systems such as UCA, CUVA, HAWK, VERA, Holmes-VAU, VANE-Bench, and VAU-R1 pushing the field from score prediction toward explanation, dialogue, and anomaly-aware reasoning [6, 7, 41, 49, 51, 53, 56]. Nevertheless, existing approaches still often rely on coarse global descriptions or weak spatial grounding, leaving a gap between pixel-level anomaly evidence and faithful language-based interpretation.

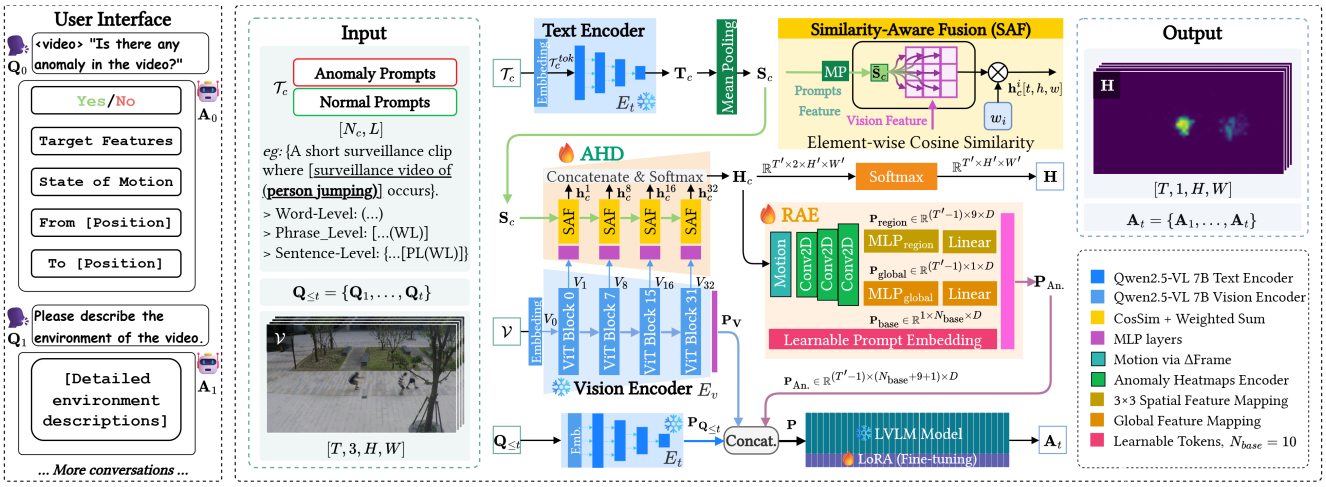

Figure 2. The proposed **T-VAU** model. The framework consists of three modules: a Text Encoder (E_t) that generates class-specific text embeddings $\mathbf{S}_c$ from binary prompts; an Anomaly Heatmap Decoder (**AHD**) that fuses $\mathbf{S}_c$ with visual features $\mathcal{V}$ to produce spatio-temporal pixel-level anomaly heatmaps $\mathbf{H}$; and a Region-aware Anomaly Encoder (**RAE**) that projects $\mathbf{H}_c$ into the LoRA-tuned LVLM semantic space and integrates it with a video $\mathcal{V}$ and a sequence of incrementally refined question $\mathbf{Q}_{\leq t}$ to yield the final anomaly understanding response $\mathbf{A}_t$.

2.2. Subtle Visual Cues & Fine-grained Supervision

A key bottleneck in fine-grained anomaly understanding lies in the mismatch between the desired level of reasoning and the granularity of available supervision. Classical VAD datasets, such as ShanghaiTech and UCF-Crime, mainly provide video- or frame-level labels, which are sufficient for anomaly scoring but inadequate for learning consistent mappings from localized evidence to target-aware semantic explanations [29, 39]. Although subsequent benchmarks have gradually enriched the supervision space, their focus varies substantially: Street Scene emphasizes localization-oriented evaluation [37], UBnormal provides pixel-level masks under an open-set setting [1], UCA extends surveillance anomaly understanding with sentence-level annotations [51], and CUVA further introduces causal supervision over the *what*, *why*, and *how* of anomalous events [6]. Recent large-scale resources, including Holmes-VAU, SurveillanceVQA-589K, and VANE-Bench, continue this trend by expanding anomaly understanding toward long-horizon, hierarchical, and conversational settings [7, 24, 53]. These developments are also closely aligned with the agenda of subtle visual computing (SVC) [12, 13, 19–23, 28, 35, 42, 43], which emphasizes low-SNR visual cues, robustness, generalization, interpretability, and unified modeling across tasks and modalities in complex real-world environments [31]. From this perspective, fine-grained video anomaly understanding can be viewed as an instance of SVC: subtle anomalies are often local, short-lived, and easily drowned by background clutter, requiring not only stronger spatio-temporal representations but also evidence-aware explanation.

Our work follows this direction by coupling pixel-level anomaly grounding with target-level appearance, localization, and trajectory supervision, thereby enabling anomaly judgment, localization, and language explanation within a unified framework.

3. Proposed Method

3.1. Framework Overview

We propose **T-VAU**, a unified framework for fine-grained video anomaly understanding built on a frozen Large Vision-Language Model (LVLM). Unlike conventional video anomaly detection methods that mainly output anomaly scores or coarse localization results, **T-VAU** jointly supports anomaly judgment, spatio-temporal localization, and interactive language-based analysis. Given an input video $\mathcal{V} \in \mathbb{R}^{T\times 3\times H\times W}$, a sequence of natural language queries $\mathbf{Q}_{\leq t} = \{\mathbf{Q}_1, \ldots, \mathbf{Q}_t\}$, and binary text prompts $\mathcal{T}_c \in \mathbb{R}^{N_c \times L}$ for each category $c \in \{\text{normal}, \text{abnormal}\}$, the model predicts pixel-level spatio-temporal anomaly heatmaps $\mathbf{H}$ and the response $\mathbf{A}_t$ at dialogue round t:

$$\mathbf{H}, \mathbf{A}_t = \mathcal{M}(\mathcal{V}, \mathbf{Q}_{\leq t}, \mathbf{A}_{\leq t-1}, \mathcal{T}_c; \Theta), \tag{1}$$

where $\mathcal{M}(\cdot;\Theta)$ denotes **T-VAU**, and T', H', and W' denote the temporal and spatial resolutions after the visual encoder.

As shown in Fig. 2, **T-VAU** consists of a frozen LVLM backbone, including a Vision Encoder E_v, a Text Encoder E_t, and a Large Language Model Decoder D_l, together with two lightweight trainable modules: the **AHD** and the **RAE**. The **AHD** aligns multiscale visual features with sentence-level text embeddings to produce frame-wise anomaly heatmaps. The **RAE** further converts these

heatmaps into structured prompt embeddings, which are fused with visual and dialogue context to enhance anomaly-aware reasoning. In this way, **T-VAU** couples explicit anomaly localization with language-guided reasoning in a unified framework.

3.2. Model Design

3.2.1. Anomaly Heatmap Decoder

AHD takes multiscale visual features from E_v and binary text prompts from E_t as input. For each scale, it computes visual-text cosine similarity and fuses the resulting similarity maps to generate anomaly heatmaps. During this stage, only **AHD** is optimized.

Textual Feature Extraction. We construct template-based prompts for two categories, *normal* and *abnormal*, and tokenize them as $\mathcal{T}_c^{\text{tok}} \in \mathbb{R}^{N_c \times L}$. The text encoder E_t extracts token-level embeddings, which are mean-pooled to obtain sentence-level features:

$$\mathbf{T}'_c = E_t(\mathcal{T}_c^{\text{tok}}) \in \mathbb{R}^{N_c \times L \times D}, \tag{2}$$

$$\mathbf{T}_c = \frac{1}{N_c L} \sum_{n=1}^{N_c} \sum_{\ell=1}^{L} \mathbf{T}'_c[n, \ell, :] \in \mathbb{R}^D, \tag{3}$$

where D is the text feature dimension.

Visual Feature Extraction. Given a video clip $\mathcal{V} \in \mathbb{R}^{T \times 3 \times H \times W}$, we first encode it using the visual encoder to obtain an initial feature tensor $\mathbf{V}_0 \in \mathbb{R}^{T' \times D_v \times H' \times W'}$. Let $\phi_i(\cdot)$ denote the output of the i-th transformer block in the visual encoder. We extract intermediate visual representations from multiple layers to capture multi-scale semantics:

$$\mathbf{V}_i = \phi_i(\mathbf{V}_0), \quad i \in \{1, 8, 16, 32\}, \tag{4}$$

where $\mathbf{V}_i \in \mathbb{R}^{T' \times D_v \times H' \times W'}$ denotes the feature map from the i-th transformer layer. These features encode visual information at different levels of abstraction and are later fused to construct anomaly heatmaps.

Anomaly Heatmaps Generation. We project each intermediate visual feature $\mathbf{V}_i$ to the shared text-visual space via an MLP, yielding $\mathbf{V}'_i \in \mathbb{R}^{T' \times D \times H' \times W'}$. We then compute cosine similarity between each spatio-temporal visual token and the sentence-level text feature:

$$\mathbf{h}_c^i[t, h, w] = \text{CosineSimilarity}(\mathbf{V}'_i[t, :, h, w], \mathbf{T}_c). \tag{5}$$

To aggregate information across layers, we introduce learnable weights w_i and fuse the similarity maps:

$$\mathbf{H}_c = \sum_i w_i \cdot \mathbf{h}_c^i \in \mathbb{R}^{T' \times 2 \times H' \times W'}. \tag{6}$$

Finally, we apply softmax along the category dimension and take the anomaly channel as the final heatmap:

$$\tilde{\mathbf{H}} = \text{Softmax}(\mathbf{H}_c) \in \mathbb{R}^{T' \times 2 \times H' \times W'}, \tag{7}$$

$$\mathbf{H} = \tilde{\mathbf{H}}[:, \text{abnormal}, :, :] \in \mathbb{R}^{T' \times 1 \times H' \times W'}. \tag{8}$$

3.2.2. Region-aware Anomaly Encoder

RAE transforms predicted anomaly heatmaps into structured prompt embeddings, which are injected into the LVLM decoder to improve anomaly reasoning.

Region-Aware Feature Extraction. Given category-wise anomaly heatmaps sequence $\mathbf{H}_c \in \mathbb{R}^{T' \times 2 \times H' \times W'}$, we first compute temporal differences between adjacent frames:

$$\mathbf{X}[t] = \mathbf{H}_c[t+1] - \mathbf{H}_c[t], \quad \{t = 1, \ldots, T' - 1\}, \tag{9}$$

where $\mathbf{X} \in \mathbb{R}^{(T'-1) \times 2 \times H' \times W'}$ represents the temporal differences between consecutive anomaly heatmaps. These motion-aware heatmaps are then processed by a lightweight convolutional backbone to extract region-aware features:

$$\mathbf{F} = \text{ConvBackbone}(\mathbf{X}) \in \mathbb{R}^{(T'-1) \times C_h \times H' \times W'}. \tag{10}$$

Prompt Embedding Generation. To encode both local and global anomaly cues, we partition each frame into a regular grid (e.g., 3×3) and apply adaptive average pooling to obtain regional features $\mathbf{F}_{\text{grid}} \in \mathbb{R}^{(T'-1) \times 9 \times C_h}$. These features are mapped into region-specific prompts:

$$\mathbf{P}_{\text{region}} = \text{MLP}_{\text{region}}(\mathbf{F}_{\text{grid}}) \in \mathbb{R}^{(T'-1) \times 9 \times D}. \tag{11}$$

A global prompt is generated by spatial mean pooling followed by an MLP:

$$\mathbf{p}_{\text{global}} = \text{MLP}_{\text{global}}(\text{MeanPool}(\mathbf{F})) \in \mathbb{R}^{(T'-1) \times 1 \times D}. \tag{12}$$

We also introduce a learnable base prompt $\mathbf{P}_{\text{base}} \in \mathbb{R}^{1 \times N_{\text{base}} \times D}$, shared across time. The final anomaly prompt sequence is constructed as

$$\mathbf{P}_{\text{An}} = \left[\mathbf{P}_{\text{base}}, \mathbf{P}_{\text{region}}, \mathbf{p}_{\text{global}}\right] \in \mathbb{R}^{(T'-1) \times (N_{\text{base}}+9+1) \times D}. \tag{13}$$

We then concatenate anomaly prompts with visual prompts and dialogue context:

$$\begin{aligned} \mathbf{P}_{\mathbf{V}} &= \text{MLP}(\mathbf{V}_{32}), \\ \mathbf{P}_{\mathbf{Q}_{\le t}} &= E_t(\mathbf{Q}_{\le t}, \mathbf{A}_{\le t-1}), \\ \mathbf{P} &= \left[\mathbf{P}_{\mathbf{V}}, \mathbf{P}_{\text{An}}, \mathbf{P}_{\mathbf{Q}_{\le t}}\right], \end{aligned} \tag{14}$$

where $\mathbf{P}_{\mathbf{V}}$ is the visual prompt, $\mathbf{P}_{\mathbf{Q}_{\le t}}$ encodes the query-response history, and $\mathbf{P}$ is the final prompt sequence fed into D_l for anomaly-aware reasoning.

3.3. Fine-grained Anomaly Understanding

To support fine-grained anomaly understanding, we construct a structured video–text supervision pipeline covering three aspects: appearance attributes, spatial localization, and motion trajectory, as illustrated in Fig. 3. Since

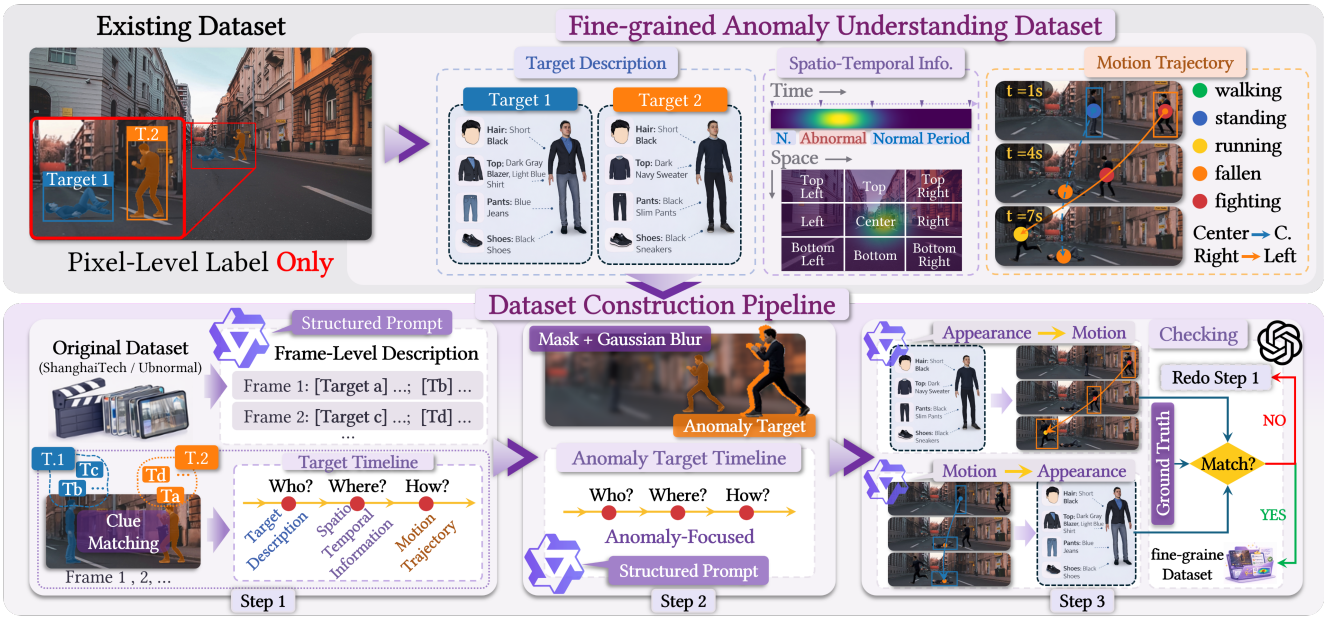

Figure 3. Fine-grained anomaly dataset construction pipeline. Starting from existing datasets with only pixel-level anomaly labels (e.g., ShanghaiTech and UBnormal), we build a structured video–text dataset through three stages: (1) frame-level structured prompting to extract target attributes and spatial information and aggregate them into target timelines; (2) anomaly-focused refinement using anomaly masks and background suppression to emphasize abnormal evidence; and (3) cross-modal consistency verification between appearance and motion cues. The resulting dataset provides aligned annotations of appearance, spatial localization, and motion trajectory for fine-grained anomaly understanding.

directly applying LVLMs to videos often loses fine-grained spatio-temporal details, we adopt a frame-wise extraction and temporal aggregation strategy to align visual evidence with textual descriptions.

Frame-level extraction and temporal aggregation. Each frame is processed with a unified structured prompt to extract target-level attributes and bounding boxes. The frame-wise outputs are then linked across time through identity association and spatial consistency, forming target-level timelines that describe who appears, where the event occurs, and how it evolves.

Anomaly-focused refinement. To emphasize anomalous evidence, we further refine the data by applying anomaly masks and suppressing background regions with Gaussian blurring. We then re-extract target-level descriptions from the refined frames and aggregate them into high-confidence anomaly timelines, providing stronger supervision for anomaly localization and motion reasoning.

Cross-modal consistency verification. We introduce bidirectional verification between appearance and motion cues. In the appearance → motion path, motion trajectories are inferred from appearance-conditioned timelines; in the motion → appearance path, motion patterns are used to verify appearance-based descriptions. This process enforces consistency between localization, trajectory, and textual semantics, and aligns the supervision signals with pixel-level spatio-temporal anomaly heatmaps generated by **AHD**.

Following this pipeline, we obtain a structured anomaly supervision dataset with aligned annotations of appearance, localization, and temporal trajectory, which provides fine-grained supervision for evidence-grounded anomaly understanding.

4. Experiments

4.1. Experimental Setup

Datasets. **UBnormal** [1] is a synthetic dataset for supervised open-set anomaly detection. It contains 543 clips (236,902 frames) generated via Cinema4D across 29 virtual scenes, with pixel-level target annotations. The dataset covers 22 anomaly types (*e.g.*, fighting and fire), which are mutually exclusive across splits: training contains both normal and abnormal events, whereas testing includes disjoint abnormal categories. **ShanghaiTech** [29] is a widely used unsupervised benchmark collected in real campus environments. It consists of 330 normal training videos and 107 testing videos (including both normal and abnormal events) across 13 scenes. The test set contains 130 annotated anomalies, such as vehicles in pedestrian areas, fighting, and cycling. We use the augmented datasets derived from UBnormal and ShanghaiTech as training and validation data, as described in Section 3.3.

Method	Validation			
	AUC		RBDC ↑	TBDC ↑
	Micro ↑	Macro ↑		
Georgescu *et al*. [8]	58.5	94.4	18.580	48.213
Georgescu *et al*. [8] (FT)	68.2	**95.3**	28.654	58.097
Sultani *et al*. [39] (PT)	61.1	89.4	0.001	0.012
Sultani *et al*. [39] (FT)	51.8	88.0	0.001	0.001
Bertasius *et al*. [4] (FT, SR=1/32)	86.1	89.2	0.008	0.021
Bertasius *et al*. [4] (FT, SR=1/8)	83.4	90.6	0.009	0.023
Bertasius *et al*. [4] (FT, SR=1/4)	78.5	89.2	0.006	0.018
AHD (1S)	94.5	85.2	64.300	74.400
AHD (FT)	**94.8**	87.8	**67.800**	**76.700**

Table 1. Experimental results on UBnormal. We report micro-/macro-averaged frame-level AUC, RBDC, and TBDC (%) for the baselines [4, 8, 39]. PT, FT, and 1S denote pre-trained, fine-tuned, and one-shot; SR denotes frame sampling rate. Although only [8] supports anomaly localization, we report RBDC and TBDC for all baselines for completeness. Best results are highlighted in bold with a green background.

Evaluation Metrics for AHD. We use frame-level Area Under the Curve (AUC) [16], reporting both micro and macro versions following [8]. We also adopt the Region-Based (RBDC) and Track-Based (TBDC) Detection Criteria [37] to evaluate spatial localization.

Evaluation Metrics for RAE. We evaluate RAE on *anomaly explanation* and *discriminative ability*. For faithfulness, we compute BLEU-4 [32] ($\mathrm{BP} \cdot \exp(\frac{1}{4}\sum \log p_n)$) separately for *Target* and *Trajectory* descriptions, using the best-matching reference per video. For discriminative ability, we report binary accuracy (Acc) and balanced accuracy ($\mathrm{bAcc} = \frac{1}{2}(\mathrm{TPR} + \mathrm{TNR})$). We further evaluate threshold-free discrimination using frame-level micro- and macro-ROC–AUC. All experiments adopt a one-shot setting with minimal text normalization.

4.2. Main Comparison

Anomaly Detection Results. Table 1 summarizes the validation results on UBnormal. In the one-shot setting, AHD achieves 94.5% micro-AUC and 85.2% macro-AUC, together with 64.3% RBDC and 74.4% TBDC, indicating strong localization from a single exemplar. After fine-tuning, AHD further improves to 94.8% micro-AUC and 87.8% macro-AUC, with 67.8% RBDC and 76.7% TBDC, establishing SOTA performance on localization-oriented metrics. Compared with the strongest baseline (Georgescu *et al*. [8] + UBnormal anomalies), our fine-tuned model gains +26.6 points in micro-AUC (94.8 *vs*. 68.2), +39.1 in RBDC (67.8 *vs*. 28.7), and +18.6 in TBDC (76.7 *vs*. 58.1), while remaining competitive in macro-AUC (87.8 *vs*. 95.3). Their macro-AUC advantage relies on a multi-stream architecture, whereas our single-model approach achieves pixel-level anomaly localization while supporting multi-turn dialogue and explainable reasoning. Similar trends are observed against recent video transformers [4], where our method yields substantially higher RBDC/TBDC together with stronger micro-AUC.

Multi-turn Dialogue Results. To evaluate language understanding and interpretability in multi-turn interactions, we adopt a one-shot setting in which the dialogue follows a hierarchical flow: "anomaly judgment → target appearance attributes → motion trajectory/direction → temporal anchors and causal cues." Compared with representative LVLMs in Table 2, our RAE significantly improves textual faithfulness and discriminative power. On ShanghaiTech, BLEU-4 scores for *Target* and *Trajectory* reach **62.67** and **88.84**, improving by about 7 and 6 points over the strongest baseline (InternVL2-8B at 55.73/82.65), while Yes/No accuracy increases to **97.67%**. On UBnormal, RAE achieves **50.32/78.10** BLEU-4 for *Target*/*Trajectory* and **89.73%** Yes/No accuracy, again outperforming 7B- and 8B-scale baselines under one-shot prompting.

The performance gain stems not only from the serial coupling of AHD and RAE, but also from constructing a dedicated dataset enriched with target-level features, localization, and motion information. This structured dataset supports a curriculum-style SFT process: holistic training on appearance–motion narratives followed by anomaly-focused refinement. By aligning pixel-level heatmaps (AHD outputs) with structured target timelines and descriptive annotations, RAE learns to inject region- and time-sensitive prompts into the decoder's semantic space. As a result, the model improves entity disambiguation ("which person/vehicle"), temporal anchoring ("when did the anomaly occur"), and motion semantics ("entering, turning, accelerating"), while reducing hallucinations such as background drift and target switching. Evidence generation, dataset-driven supervision, and prompt injection together establish a closed-loop mechanism for high-fidelity, verifiable multi-turn dialogue.

4.3. Ablation Studies

We present a systematic ablation of the two key modules in T-VAU: AHD and RAE. These modules are serially coupled: AHD first generates pixel-level anomaly heatmaps, which are then transformed by RAE into learnable region-aware prompts injected into the language decoder. Removing both reduces the system to the frozen LVLM backbone, without additional anomaly localization or interpretability.

As shown in Table 3, the full T-VAU achieves the best performance across all metrics: anomaly localization (RBDC 67.8, TBDC 76.7), text generation quality (BLEU-4 of 62.67 and 88.84), and binary anomaly judgment accuracy (97.67%), all exceeding the ablated variants. When AHD is removed but RAE is retained, the model loses pixel-level anomaly evidence, making RBDC and TBDC inapplicable. In this case, language outputs rely mainly on the generalization ability of the backbone LVLM, and anomaly descriptions lack spatial interpretability. Con-

Method	Size ↓	ShanghaiTech			UBnormal		
		BLEU-4 ± Std ↑		Acc. ± Std ↑	BLEU-4 ± Std ↑		Acc. ± Std ↑
		Target	Trajectory	Yes/No	Target	Trajectory	Yes/No
Qwen2.5-VL (zero-shot)	7B	18.74 ± 0.82	27.33 ± 1.05	61.03 ± 0.38%	16.20 ± 0.85	24.18 ± 1.08	65.62 ± 0.41%
Qwen2.5-VL (one-shot)	7B	50.42 ± 0.58	78.91 ± 0.72	92.36 ± 0.24%	44.35 ± 0.62	70.82 ± 0.75	87.24 ± 0.27%
LLaVA-1.6 (one-shot)	7B	47.68 ± 0.60	75.42 ± 0.74	91.07 ± 0.26%	42.11 ± 0.64	68.07 ± 0.78	85.91 ± 0.28%
MiniCPM-V 2.6 (one-shot)	7B	52.34 ± 0.54	80.41 ± 0.69	93.11 ± 0.23%	46.70 ± 0.59	72.88 ± 0.73	86.94 ± 0.26%
Idefics2 (one-shot)	8B	44.29 ± 0.62	73.84 ± 0.76	90.12 ± 0.27%	39.51 ± 0.66	65.92 ± 0.80	84.03 ± 0.29%
InternVL (one-shot)	8B	55.73 ± 0.50	82.65 ± 0.66	94.28 ± 0.22%	49.84 ± 0.55	71.63 ± 0.70	88.65 ± 0.24%
RAE (Ours) (one-shot)	7B	**62.67 ± 0.45**	**88.84 ± 0.53**	**97.67 ± 0.12%**	**50.32 ± 0.49**	**78.10 ± 0.58**	**89.73 ± 0.18%**

Table 2. One-shot evaluation results of representative LVLMs on our constructed dataset. "Size" denotes the number of model parameters in billions. BLEU-4 is reported for Target and Trajectory, and Accuracy for Yes/No. All metrics are in percentages. Our RAE performs best across tasks.

Method	Size ↓	Heatmap (UBnormal) ↑		BLEU-4 (ShanghaiTech) ↑		Acc. (S.T.) ↑
	Parameters	RBDC ± Std	TBDC ± Std	Target ± Std	Trajectory ± Std	Yes/No ± Std
T-VAU w/o AHD	8299.71M	–	–	61.82 ± 0.42	85.47 ± 0.51	95.38 ± 0.18%
T-VAU w/o RAE	8317.13M	67.8 ± 0.36	76.7 ± 0.41	–	–	–
T-VAU w/o AHD & RAE	8274.74M	–	–	61.82 ± 0.42	85.47 ± 0.51	95.38 ± 0.18%
T-VAU	8324.67M	67.8 ± 0.36	76.7 ± 0.41	62.67 ± 0.45	88.84 ± 0.53	97.67 ± 0.12%

Table 3. Results of different T-VAU variants. "Parameters/Size" denotes the number of model parameters in millions. Heatmap metrics include RBDC and TBDC. BLEU-4 is reported for both Target and Trajectory. Accuracy is evaluated on Yes/No classification.

versely, removing RAE while keeping AHD still produces high-quality heatmaps (with RBDC/TBDC close to the full model), but without region-aware prompt injection, the language decoder cannot fully exploit the visual evidence, leading to lower BLEU-4 scores and anomaly judgment accuracy. When both AHD and RAE are removed, the system degenerates to the baseline LVLM. Although it can occasionally generate descriptions from large-scale pretraining, overall performance drops sharply: it neither provides reliable anomaly localization nor supports structured anomaly reasoning. This confirms the complementarity and necessity of both modules.

4.4. Qualitative Analysis

We qualitatively analyze how T-VAU uses heatmaps and region-aware prompts to ensure consistency between pixel-level evidence and language-based reasoning.

Fig. 4 compares the heatmaps produced by AHD with ground-truth (GT) pixel masks. In pedestrian scenes (*e.g.*, jaywalking and sudden running), AHD concentrates activation on the human silhouette with sharp boundaries while suppressing background structures. In traffic scenes (*e.g.*, abrupt vehicle entry and near-collision), activations adhere to target extents and propagate along the motion direction, producing elongated responses that mirror the trajectory. Decoder-driven heatmaps ("Video/Image Decoder" in Fig. 4) tend to blur across large regions or drift toward irrelevant textures. In contrast, AHD yields crisper spatial support and tighter overlap with GT masks, especially in crowded scenes.

Fig. 4 illustrates T-VAU in multi-turn interaction. With region-aware prompts from AHD, the decoder grounds descriptions on the correct spatial regions and time spans. Appearance attributes (*e.g.*, white top, grey shorts, green backpack), motion phrases (*e.g.*, enters from the right, turns upward-right, exits), and state changes (*e.g.*, walking → running at frame 10) align with the heatmap. Even with multiple candidates, RAE helps the decoder prioritize the instance supported by the heatmap. When the anomalous target is partially occluded, AHD responses attenuate but remain anchored and re-amplify upon reappearance. Peak activation follows the target across frames and supports reasoning about entry/exit events, turning points, and acceleration. Failure cases include micro-actions with minimal displacement, highly nonrigid motion that scatters activation, and scene-dependent appearance shifts (*e.g.*, specularities and fog). For cross-scene changes and unseen anomaly categories, AHD preserves localization fidelity when target scale is reasonable. For unseen categories sharing motion primitives, T-VAU often highlights the correct target and articulates motion phrases consistent with the evidence. Heatmaps provide a spatial witness for each claim, enabling justifications such as the silver SUV appearing abruptly from the left and accelerating. Multi-turn follow-ups (which target, "where", "when") reuse the same evidence chain: AHD, RAE, and the decoder. This yields concise, self-consistent narratives and enables error analysis through spatial and temporal evidence.

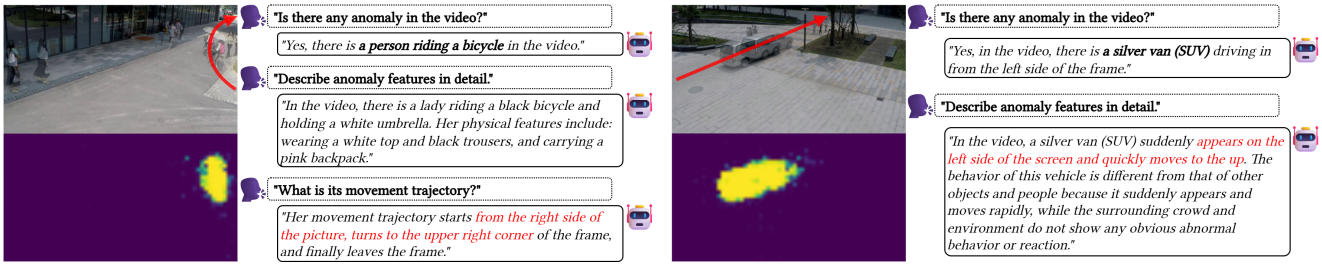

Figure 4. **Examples of interpretable anomaly detection and multi-turn QA across scenes.** Each group shows the raw frame, pixel-level anomaly heatmaps produced by **AHD**, and **T-VAU**'s dialogue outputs (anomaly yes/no, appearance/action details, and motion trajectory). **Left:** a cyclist (with umbrella and backpack) is localized as the anomalous target, with the trajectory "enter from right → turn toward the upper-right corner → exit." **Right:** a silver SUV suddenly appears from the left and moves rapidly; **AHD** consistently highlights the vehicle, and the QA module explains the abrupt appearance and fast motion. **T-VAU** first detects the anomaly, then describes the appearance (white top, grey shorts, green schoolbag) and the change from walking to running. Red arrows indicate the main motion directions, and heatmap intensity reflects anomaly confidence. **RAE** encodes the heatmaps into region-aware text prompts that guide the LVLM to produce consistent decisions and descriptions, closing the loop from pixel-level evidence to readable narratives.

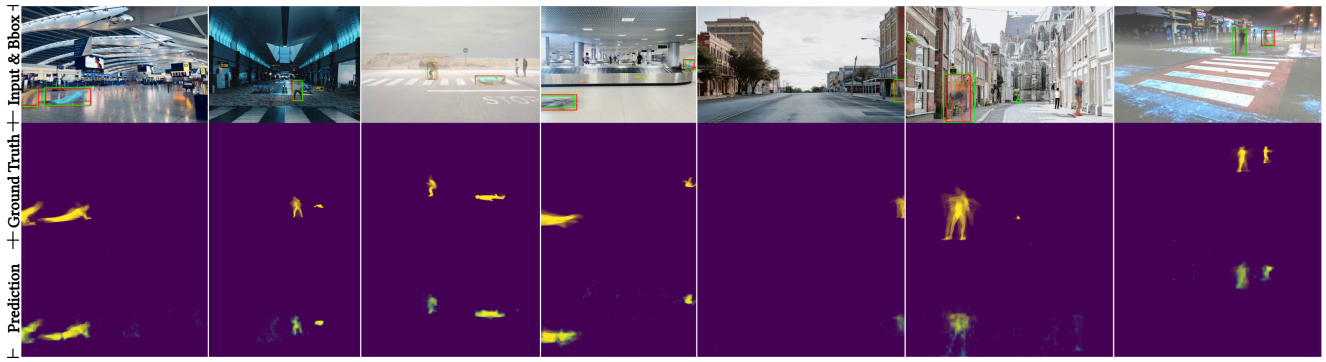

Figure 5. Trajectory visualization by accumulating frame-level outputs. The first row shows multi-frame overlays of the original video with green bounding boxes for GT and red bounding boxes for predictions. The second row overlays GT pixel-level masks to form fine-grained trajectories, while the third row overlays predicted pixel-level masks. Both bounding-box and pixel-level trajectories show strong spatial alignment with GT, indicating that our model accurately captures motion paths over time.

We further visualize trajectory consistency by overlaying outputs across frames (Fig. 5). The first row stacks raw frames with bounding boxes (green: GT, red: predictions) for coarse comparison. The second accumulates GT pixel-level masks to form fine-grained paths, while the third shows accumulated predicted masks. Both box- and pixel-level overlays closely match GT, indicating the method not only localizes anomalies per frame but also preserves temporal coherence, yielding trajectories consistent with the ground truth over time.

5. Conclusions

We present **T-VAU**, a closed-loop framework that unifies pixel-level anomaly grounding and high-level semantic reasoning by coupling an Anomaly Heatmap Decoder (**AHD**) with a Region-aware Anomaly Encoder (**RAE**). By aligning visual features with textual prompts, **T-VAU** achieves precise, threshold-free spatio-temporal anomaly localization, while its region- and motion-aware prompt design enables LVLMs to perform faithful, structured, and multi-turn anomaly reasoning. This unified formulation goes beyond conventional score-based paradigms, jointly supporting detection, localization, target identification, and explanation within a single framework. Extensive experiments on UBnormal and ShanghaiTech demonstrate consistent improvements over prior methods across localization accuracy, reasoning quality, and dialogue-based evaluation, while ablations confirm the strong complementarity between **AHD** and **RAE**.

Acknowledgments

We would like to thank Alex Chapiro for insightful discussions and constructive feedback on earlier versions of this manuscript. We also acknowledge the HPC system at the United Arab Emirates University for providing the computational resources.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20143–20153, 2022. 2, 3, 5

[2] Sunghyun Ahn, Youngwan Jo, Kijung Lee, Sein Kwon, Inpyo Hong, and Sanghyun Park. Anyanomaly: Zero-shot customizable video anomaly detection with lvlm. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3026–3035, 2026. 1

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 6

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 2

[6] Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, et al. Uncovering what why and how: A comprehensive benchmark for causation understanding of video anomaly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18793–18803, 2024. 2, 3

[7] Hanan Gani, Rohit Bharadwaj, Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. Vane-bench: Video anomaly evaluation benchmark for conversational lmms. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3123–3140, 2025. 2, 3

[8] Mariana Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4505–4523, 2021. 6

[9] Jihao Gu, Kun Li, Fei Wang, Yanyan Wei, Zhiliang Wu, Hehe Fan, and Meng Wang. Motion matters: Motion-guided modulation network for skeleton-based micro-action recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5461–5470, 2025. 1

[10] Jihao Gu, Fei Wang, Kun Li, Yanyan Wei, Zhiliang Wu, and Dan Guo. Mm-gesture: towards precise micro-gesture recognition through multimodal fusion. *arXiv preprint arXiv:2507.08344*, 2025. 1

[11] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Hao Li, Ming Tang, and Jinqiao Wang. Filo: Zero-shot anomaly detection by fine-grained description and high-quality localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2041–2049, 2024. 2

[12] Dan Guo, Kun Li, Bin Hu, Yan Zhang, and Meng Wang. Benchmarking micro-action recognition: Dataset, methods, and applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6238–6252, 2024. 3

[13] Dan Guo, Xiaobai Li, Kun Li, Haoyu Chen, Jingjing Hu, Guoying Zhao, Yi Yang, and Meng Wang. Mac 2024: Micro-action analysis grand challenge. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11304–11305, 2024. 3

[14] Xiaobao Guo, Nithish Muthuchamy Selvaraj, Zitong Yu, Adams Wai-Kin Kong, Bingquan Shen, and Alex Kot. Audio-visual deception detection: Dolos dataset and parameter-efficient crossmodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22135–22145, 2023. 1

[15] Xiaobao Guo, Zitong Yu, Nithish Muthuchamy Selvaraj, Bingquan Shen, Adams Wai-Kin Kong, and Alex C Kot. Benchmarking cross-domain audio-visual deception detection. *arXiv preprint arXiv:2405.06995*, 2024. 1

[16] Jin Huang and Charles X Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310, 2005. 6

[17] Fei Li, Wenxuan Liu, Jingjing Chen, Ruixu Zhang, Yuran Wang, Xian Zhong, and Zheng Wang. Anomize: Better open vocabulary video anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29203–29212, 2025. 1

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[19] Kun Li, Dan Guo, Guoliang Chen, Xinge Peng, and Meng Wang. Joint skeletal and semantic embedding loss for micro-gesture classification. *arXiv preprint arXiv:2307.10624*, 2023. 3

[20] Kun Li, Dan Guo, Guoliang Chen, Chunxiao Fan, Jingyuan Xu, Zhiliang Wu, Hehe Fan, and Meng Wang. Prototypical calibrating ambiguous samples for micro-action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4815–4823, 2025.

[21] Kun Li, Pengyu Liu, Dan Guo, Fei Wang, Zhiliang Wu, Hehe Fan, and Meng Wang. Mmad: Multi-label micro-action detection in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13225–13236, 2025.

[22] Kun Li, Jihao Gu, Fei Wang, Zhiliang Wu, Hehe Fan, and Dan Guo. Ma-bench: Towards fine-grained micro-action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[23] Xun Lin, Xiaobao Guo, Taorui Wang, Yingjie Ma, Jiajian Huang, Jiayu Zhang, Junzhe Cao, and Zitong Yu. Svc 2025: the first multimodal deception detection challenge. In *Proceedings of the 1st International Workshop & Challenge on Subtle Visual Computing*, pages 59–64, 2025. 1, 3

[24] Bo Liu, Pengfei Qiao, Minhan Ma, Xuange Zhang, Yinan Tang, Peng Xu, Kun Liu, and Tongtong Yuan. Surveillancevqa-589k: A benchmark for comprehensive surveillance video-language understanding with large models. *arXiv preprint arXiv:2505.12589*, 2025. 3

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 2

[27] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2

[28] Pengyu Liu, Fei Wang, Kun Li, Guoliang Chen, Yanyan Wei, Shengeng Tang, Zhiliang Wu, and Dan Guo. Micro-gesture online recognition using learnable query points. *arXiv preprint arXiv:2407.04490*, 2024. 3

[29] 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*, pages 6536–6545, 2018. 1, 2, 3, 5

[30] Jakub Micorek, Horst Possegger, Dominik Narnhofer, Horst Bischof, and Mateusz Kozinski. Mulde: Multiscale log-density estimation via denoising score matching for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18868–18877, 2024. 2

[31] SVC Organizers. Subtle visual computing workshop. CVPR Workshop, 2026. 3

[32] 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*, pages 311–318, 2002. 6

[33] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14372–14381, 2020. 2

[34] Wei Qian, Dan Guo, Kun Li, Xiaowei Zhang, Xilan Tian, Xun Yang, and Meng Wang. Dual-path tokenlearner for remote photoplethysmography-based physiological measurement with facial videos. *IEEE Transactions on Computational Social Systems*, 11(3):4465–4477, 2024. 1

[35] Wei Qian, Kun Li, Dan Guo, Bin Hu, and Meng Wang. Cluster-phys: Facial clues clustering towards efficient remote physiological measurement. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 330–339, 2024. 3

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[37] Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2569–2578, 2020. 2, 3, 6

[38] 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. 1

[39] 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*, pages 6479–6488, 2018. 1, 2, 3, 6

[40] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 2

[41] Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Yingcong Chen. Hawk: Learning to understand open-world video anomalies. *Advances in Neural Information Processing Systems*, 37:139751–139785, 2024. 2

[42] Fei Wang, Dan Guo, Kun Li, and Meng Wang. Eulermormer: Robust eulerian motion magnification via dynamic filtering within transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5345–5353, 2024. 3

[43] Fei Wang, Dan Guo, Kun Li, Zhun Zhong, and Meng Wang. Frequency decoupling for motion magnification via multi-level isomorphic architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18984–18994, 2024. 3

[44] Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18297–18307, 2024. 2

[45] Peng Wu, Xuerong Zhou, Guansong Pang, Zhiwei Yang, Qingsen Yan, Peng Wang, and Yanning Zhang. Weakly supervised video anomaly detection and localization with spatio-temporal prompts. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9301–9310, 2024. 2

[46] 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*, pages 6074–6082, 2024. 2

[47] Peng Wu, Chengyu Pan, Yuting Yan, Guansong Pang, Qingsen Yan, Peng Wang, and Yanning Zhang. Deep learning for video anomaly detection: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 2026. 1

[48] Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the rules: reasoning for video anomaly detection with large language models. In *European Conference on Computer Vision*, pages 304–322. Springer, 2024. 2

[49] Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8679–8688, 2025. 2

[50] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4186–4196, 2022. 1

[51] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22052–22061, 2024. 2, 3

[52] Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18527–18536, 2024. 2

[53] 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 *Proceedings of the computer vision and pattern recognition conference*, pages 13843–13853, 2025. 2, 3

[54] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1237–1246, 2019. 2

[55] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2

[56] Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning. *arXiv preprint arXiv:2505.23504*, 2025. 2