# SEPARATE TO DELIBERATE, FUSE TO DECIDE: PERSPECTIVE-FACTORIZED LATENT REASONING FOR VIDEO ANOMALY UNDERSTANDING

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

## ABSTRACT

Video Anomaly Understanding (VAU) requires models not only to recognize unusual visual content, but also to understand how scene states change and how events evolve over time. These complementary perspectives provide distinct yet mutually informative cues for interpreting complex anomalies. However, existing LLM-based VAU methods typically encode heterogeneous visual information into a shared representation stream or perform explicit reasoning through a single textual chain of thought, which may prematurely entangle information from complementary perspectives. To address this limitation, we propose **BRAID**, a perspective-factorized latent reasoning framework based on an observe–deliberate–synthesize paradigm. Specifically, for perspective-aligned observation, BRAID captures object structures, spatial configurations, and state changes in the scene, and models temporal interactions and event evolution. It then performs perspective-specific latent deliberation over these observations, allowing each reasoning stream to develop from the visual evidence captured under its corresponding perspective while avoiding premature cross-perspective interference. Finally, since the relevance of different perspectives varies across contexts, BRAID employs a contextualized latent synthesis mechanism to dynamically assess and integrate perspective-specific information, enabling the model to jointly leverage complementary information from different perspectives for anomaly understanding and response generation. Extensive experiments on multiple widely used benchmarks demonstrate that our method outperforms existing MLLM-based VAU methods in both accuracy and computational efficiency.

## 1 INTRODUCTION

Video anomaly detection (VAD) aims to automatically identify anomalous events in videos and plays an important role in public safety and surveillance applications. Traditional VAD methods Tian et al. (2021); Wu et al. (2024); Sun & Gong (2023); Cao et al. (2023) primarily focus on predicting frame-level anomaly scores, but often lack the ability to explain why an event is considered anomalous. With the rapid development of multimodal large language models (MLLMs), video anomaly detection is evolving toward a higher-level task of video anomaly understanding (VAU). Beyond simply identifying anomalies, models are increasingly expected to provide meaningful explanations for the detected abnormal events.

Existing MLLM-based VAU methods typically improve anomaly interpretation through task-specific instruction tuning. Representative methods such as Holmes-VAU Zhang et al. (2025) constructs multi-granularity instructions for anomaly description and analysis, while HAWK Tang et al. (2024) develops an instruction-tuning dataset that equips the model with both anomaly understanding capability and the flexibility to address open-ended user queries. More recent methods further improve anomaly understanding through reinforcement learning while generating explicit intermediate reasoning text to enhance anomalous understanding ability. VAU-R1 Zhu et al. (2025) optimizes output format, answer correctness, and temporal grounding; Cue-R1 Yu et al. (2026) uses format and task-specific rewards for context-aware anomaly understanding; In contrast, VAD-R1 Huang et al. (2026) introduces a counterfactual reward strategy encourages the model to base its reasoning on genuine anomaly-related visual evidence rather than superficial correlations.



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.

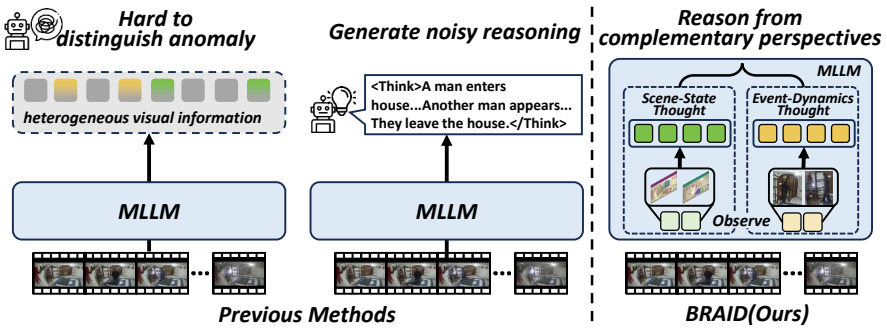


Figure 1: Compared with previous methods, **BRAID** captures and reasons over perspective-specific information within the MLLM, avoiding information entanglement and noise from explicit CoT.

Although existing methods have substantially improved video anomaly understanding, as shown in Figure 1, they often entangle complementary scene-state and event-dynamics cues within a shared representation or reasoning stream. Meanwhile, explicit CoT may introduce noisy intermediate reasoning, hindering effective anomaly understanding.

To address the above issues, we propose **BRAID**, a perspective-factorized latent reasoning framework based on an observe–deliberate–synthesize paradigm. BRAID forms visual observations from two complementary perspectives, capturing object structures, spatial configurations, and scene changes in scene-state latent tokens, while modeling motion states, action interactions, and event evolution in event-dynamics tokens. Furthermore, BRAID forms thought latent tokens based on the corresponding visual observations, enabling perspective-specific reasoning and more effective use of complementary information for anomaly understanding. Finally, since the importance of different perspectives may vary across task scenarios, we design a contextualized latent synthesis mechanism to adaptively integrate the two perspective-specific representations for final response generation. Extensive experiments demonstrate that BRAID improves anomalous event understanding while maintaining a favorable balance between accuracy and inference efficiency.

In summary, our contributions are as follows:

- We propose BRAID, a perspective-factorized latent reasoning framework for VAU that enables the model to reason over heterogeneous visual cues.
- We introduce a contextualized latent synthesis mechanism that dynamically evaluates the relevance of different perspectives allowing the model to comprehensively leverage complementary information for anomaly understanding.
- Extensive experiments on multiple VAU benchmarks demonstrate that BRAID consistently improves anomaly understanding while maintaining favorable inference efficiency, outperforming existing VAU methods in both accuracy and computational efficiency.

## 2 Related Works

**Video Anomaly Detection.** Traditional video anomaly detection methods typically learn normal behavior patterns through reconstruction-based or one-class classification objectives Cao et al. (2023); Yan et al. (2023); Yang et al. (2023) to distinguish anomalous events, or extract 3D video features and train models with a Multiple Instance Learning (MIL) objectives Zhang et al. (2023a); Lv et al. (2023); Majhi et al. (2025). Although these approaches perform well in anomaly localization, their practical applicability is limited by the lack of interpretable explanations for anomalous events.

**Video Anomaly Understanding.** Recent studies have leveraged MLLMs to enable interpretable video anomaly understanding. Training-free methods exploit the inherent reasoning capabilities of MLLMs without task-specific optimization. Representative approaches such as LAVAD Zanella et al. (2024), EventVAD Shao et al. (2025), VADTree Li et al. (2026), and PANDA Yang et al. (2025) explore frame-level, event-level, and agentic reasoning pipelines, but often involve multi-stage inference with relatively high computational overhead. Other works instead adapt MLLMs through task-

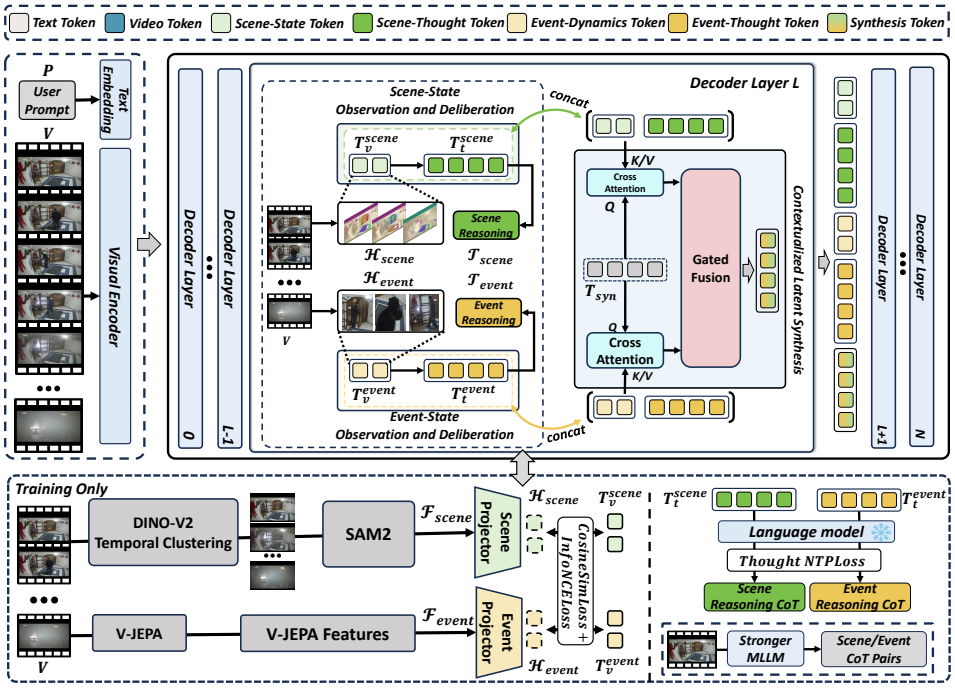


Figure 2: Overall architecture of BRAID. BRAID forms perspective-specific visual observations and thoughts within intermediate language-model layers, and integrates them through Contextualized Latent Synthesis for final reasoning. The modules in the lower training pipeline are used only during training and removed at inference.

specific training. Holmes-VAU Zhang et al. (2025) constructs a multi-granularity instruction-tuning dataset, enabling the model to better handle anomalous events across videos of varying lengths. HAWK Tang et al. (2024) constructs an anomaly-focused instruction dataset for open-world user interactions, improving the model's ability to handle open-ended queries. VADER Cheng et al. (2026) and TargetVAU Zhou et al. (2026) employ conventional VAD models as anomaly scorers to identify anomalous video frames, thereby enhancing the model's sensitivity to abnormal events. More recently, VAU-R1 Zhu et al. (2025), Cue-R1 Yu et al. (2026), and VAD-R1 Huang et al. (2026) introduce reinforcement learning with task-specific rewards and explicit CoT generation to strengthen anomaly reasoning. While these methods improve anomaly understanding, some rely on auxiliary VAD models, while others incur additional inference overhead from explicit CoT generation. More importantly, they still share a common limitation. VAU requires reasoning from complementary perspectives, particularly scene states and event dynamics, whereas existing methods typically process heterogeneous visual information within a shared representation stream or a single textual reasoning chain. This may prematurely entangle information with distinct semantic roles and limit the effective exploitation of their complementarity.

# 3 METHOD

## 3.1 FRAMEWORK OVERVIEW

To address the premature entanglement of heterogeneous visual information, we propose **BRAID**. As shown in Figure 2, BRAID follows a standard LLaVA-style MLLM architecture with a visual encoder $visual(\cdot)$, text embedding $embedding(\cdot)$, and LLM backbone $language(\cdot)$. Given a video $V$ and prompt $P$, the video is encoded and projected as $\mathbf{v} = visual(V)$, while $P$ is embedded into textual tokens $\mathbf{p}$. These tokens are concatenated and fed into the LLM to generate the response.

Most VAU methods fine-tune this architecture or optimize it with reinforcement learning. However, encoding heterogeneous visual information into a shared representation stream may obscure complementary cues from different perspectives, limiting comprehensive anomaly understanding. Video

anomalies can be characterized from two complementary perspectives: scene-state changes and the temporal evolution of events. Accordingly, BRAID performs perspective-specific latent reasoning within the LLM, using $n$ Visual Latent Tokens ($T_v$) to capture visual information and $m$ Thought Latent Tokens ($T_t$) to deliberate over the corresponding observations.

During inference, BRAID first generates scene-state and event-dynamics visual latent tokens, $T_v^{\text{scene}} = \{t_{v,i}^{\text{scene}}\}_{i=1}^{n}$ and $T_v^{\text{event}} = \{t_{v,i}^{\text{event}}\}_{i=1}^{n}$, to capture scene context and event evolution, respectively. It then produces thought latent tokens $T_t^C = \{t_{t,i}^C\}_{i=1}^{m}$ for perspective-specific deliberation, where $C \in \{\text{scene}, \text{event}\}$. Finally, the two perspectives are synthesized to generate the response.

### 3.2 Perspective-Specific Observation

**Scene-State Observation**. Scene context provides important cues for identifying anomalous events; for example, theft and robbery are more likely in shopping malls, banks, or secluded areas. Scene changes, such as fire progression or layout variations, further reflect the severity and evolution of anomalous events.

To capture scene changes, we uniformly sample $M$ frames $F = \{I_i\}_{i=1}^{M}$ from the input video $V$ and extract DINO-v2 Oquab et al. (2023) features $f_i$. We perform temporally constrained clustering based on inter-frame similarities to select $N$ frames $F_{sel} = \{I_j\}_{j=1}^{S}$ that characterize scene evolution. The dense scene representations $\mathcal{F}_{\text{scene}}$ are then extracted using SAM2 Ravi et al. (2025):

$$\mathcal{F}_{\text{scene}} = \{f_i^{\text{scene}}\}_{i=1}^{S} = \text{SAM2}(F_{sel}), \quad f_i^{\text{scene}} \in \mathbf{R}^{H \times W \times C} \tag{1}$$

where $H$, $W$, and $C$ denote the height, width, and channel of the SAM2 feature map, respectively. The dense scene representations are projected into the LLM latent space by $\text{SceneProjector}$, producing $n$ supervisory targets for the scene-state tokens:

$$\mathcal{H}_{\text{scene}} = \{h_{v,i}^{\text{scene}}\}_{i=1}^{n} = \text{SceneProjector}(\mathcal{F}_{\text{scene}}), \quad h_{v,i}^{\text{scene}} \in \mathbf{R}^{D} \tag{2}$$

where $D$ denotes the hidden dimension of MLLM.

**Event-Dynamics Observation**. Event dynamics provide critical motion cues for understanding anomalous events. Modeling temporal dependencies in human motion, object interactions, and event evolution is essential for understanding the dynamics underlying anomalous events.

We employ V-JEPA Assran et al. (2025) to extract temporally structured representations as informative priors for event-dynamics modeling. Given the $M$ sampled frames, V-JEPA produces spatiotemporal features $\mathcal{F}_{\text{event}}$ encoding motion patterns, interaction dynamics, and event evolution:

$$\mathcal{F}_{\text{event}} = \text{VJEPA}(F), \mathcal{F}_{\text{event}} \in \mathbf{R}^{N_e \times D'} \tag{3}$$

where $N_e$ denotes the sequence length and $D'$ is the hidden dimension of V-JEPA. Similarly, we design an $\text{EventProjector}$ to obtain the supervisory targets for event-dynamic tokens $T_v^{\text{event}}$:

$$\mathcal{H}_{\text{event}} = \{h_{v,i}^{\text{event}}\}_{i=1}^{n} = \text{EventProjector}(\mathcal{F}_{\text{event}}), \quad h_{v,i}^{\text{event}} \in \mathbf{R}^{D} \tag{4}$$

### 3.3 Perspective-Specific Deliberation

Although perspective-aligned observations preserve distinct visual cues, observation alone is insufficient for perspective-specific reasoning. We therefore enable BRAID to perform latent deliberation over each perspective's visual observations, allowing the model to better exploit the captured perspective-specific information.

We place thought latent token immediately after the visual latent token, allowing them to reason over the corresponding visual information through causal attention. We further modify the attention mask to block cross-perspective interactions between latent tokens, enabling each reasoning stream to focus on its own perspective.

### 3.4 Contextualized Latent Synthesis

BRAID has completed perspective-specific observation and deliberation. Since different scenarios may emphasize different perspectives, we introduce a contextualized latent synthesis mechanism to

adaptively assess and integrate their contributions. Before response generation, we insert synthesis slots after the prompt, initialized with the fixed instruction "*Let's consider this comprehensively.*" After tokenization, they form the synthesis tokens $T_{syn}$. At the $L$-th language-model layer, we extract the visual and thought latent tokens from their corresponding positions:

$$\mathbf{H}^{(L)} = LM^{(L)}(\mathbf{H}^{(L-1)}), \quad \mathbf{H}^{(L)} \in \mathbf{R}^{N\times D} \tag{5}$$

$$T_v^{C,(L)} = \mathbf{H}^{(L)}[\mathcal{I}_v^C], \quad T_t^{C,(L)} = \mathbf{H}^{(L)}[\mathcal{I}_t^C], \quad C \in \{\text{scene}, \text{event}\}. \tag{6}$$

where $\mathcal{I}_v^C$ and $\mathcal{I}_t^C$ denote the position indices of visual and thought latent tokens, respectively.

We concatenate the visual latent tokens and thought latent tokens within each perspective and use the synthesis tokens as queries to perform cross-attention over the concatenated latent representations:

$$\mathbf{a}^C = \text{CrossAttention}(T_{syn}^{(L)}, [T_v^{C,(L)}, T_t^{C,(L)}]), \quad \mathbf{a}^C \in \mathbf{R}^{k\times D} \tag{7}$$

where $[\cdot]$ denotes concatenation, and $T_{syn}^{(L)}$ represents the hidden states of the $k$ synthesis tokens at layer $L$. A gating network then predicts perspective scores for weighted fusion of the latent tokens:

$$s^C = \text{Gate}([T_{syn}^{(L)}, \mathbf{a}^{scene}, \mathbf{a}^{event}]). \tag{8}$$

$$T'^{(L)}_{syn} = s^{scene}\mathbf{a}^{scene} + s^{event}\mathbf{a}^{event} + T_{syn}^{(L)} \tag{9}$$

$T'^{(L)}_{syn}$ then replaces the original synthesis tokens at their corresponding positions.

### 3.5 Perspective-Aligned Training Objective

We apply perspective-specific supervision at the $L$-th layer, where the visual and thought latent tokens have sufficiently interacted and formed a more complete understanding of the video and task. Applying supervision at $L$-th layer encourages subsequent layers to better exploit perspective-specific observations and reasoning.

For visual latent tokens, we use cosine similarity loss and InfoNCE losses Oord et al. (2018). The cosine similarity loss aligns scene-state and event-dynamics tokens with their supervisory visual features, while InfoNCE strengthens their discriminative correspondence.

$$\mathcal{L}_{\text{rec}}^C = \text{CosSimLoss}\left(T_v^{C,(L)}, \mathcal{H}^C\right), \quad \mathcal{L}_{\text{InfoNCE}}^C = \text{InfoNCE}\left(T_v^{C,(L)}, \mathcal{H}^C\right). \tag{10}$$

$$\mathcal{L}_{\text{visual}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{nce}}\mathcal{L}_{\text{InfoNCE}} \tag{11}$$

For each training video $V$, we query a stronger teacher MLLM to generate perspective-specific reasoning texts. The stronger MLLM is instructed to keep the two perspectives disentangled: $\mathcal{T}_{\text{scene}}$ focuses on scene changes, while $\mathcal{T}_{\text{event}}$ emphasizes motion cues such as actions and object interactions. We generate multiple candidate pairs and select the one with the largest semantic discrepancy as supervision. A frozen copy of BRAID's language model is then used as an auxiliary decoder to reconstruct $\mathcal{T}_C = \{y_i^C\}_{i=1}^{N_y^C}$ solely from $T_t^{C,(L)}$ with a next-token prediction loss:

$$\mathcal{L}_{\text{ntp}}^C = -\frac{1}{N_y^C}\sum_{i=1}^{N_y^C} \log p_\theta\left(y_i^C \mid y_{<i}^C, T_t^{C,(L)}\right). \tag{12}$$

### 3.6 Training Strategy

We adopt a three-stage training strategy to progressively learn perspective-specific observation, latent deliberation, and contextualized synthesis. We first establish well-structured latent representations for each perspective, then learn how to adaptively synthesize them for response generation, and finally jointly refine the entire reasoning process.

**Perspective-Specific Representation Learning.** We first train BRAID to construct perspective-specific visual observations and derive corresponding latent thoughts. This stage encourages BRAID to acquire distinct visual information from different perspectives and perform deliberation conditioned on the corresponding observations. The training objective is:

$$\mathcal{L}_{\text{stage1}} = \lambda_{\text{visual}}\mathcal{L}_{\text{visual}} + \lambda_{\text{ntp}}\mathcal{L}_{\text{ntp}} \tag{13}$$

**Contextualized Synthesis Learning.** We then focus exclusively on the contextualized latent synthesis mechanism while keeping the previously learned perspective-specific representations fixed to train BRAID to integrate them into a contextualized representation for final response generation:

$$\mathcal{L}_{\text{stage2}} = \lambda_{\text{ans}} \mathcal{L}_{\text{ntp}}^{\text{ans}} \tag{14}$$

where $\mathcal{L}_{\text{ntp}}^{\text{ans}}$ is the next-token prediction loss of the final answer.

**Joint Refinement.** Finally, we jointly optimize the visual latent tokens, thought latent tokens, and contextualized latent synthesis mechanism. Allowing an end-to-end refinement of the overall latent reasoning process:

$$\mathcal{L}_{\text{stage3}} = \mathcal{L}_{\text{stage1}} + \mathcal{L}_{\text{stage2}} \tag{15}$$

# 4 Experiments

## 4.1 Experimental Setup

**Datasets and Evaluation Metrics.** We evaluate BRAID on four VAU benchmarks: VAU-Bench Zhu et al. (2025), HIVAU-70K Zhang et al. (2025), UCF-Crime Sultani et al. (2018), and XD-Violence Wu et al. (2020). **VAU-Bench** contains 2,939 training and 929 test videos from UCF-Crime, ECVA Du et al. (2026), and MSAD Zhu et al. (2024), covering QA, anomaly classification, and anomaly reasoning; we report Acc for the first two tasks and VAU-Eval for reasoning. For **HIVAU-70K**, we report BLEU Papineni et al. (2002), METEOR Banerjee & Lavie (2005), ROUGE-L Lin (2004), and CIDEr Vedantam et al. (2015). For anomaly localization, we report frame-level AUC on **UCF-Crime** and AUC/AP on **XD-Violence**.

**Implementations.** BRAID is built on Qwen3-VL-2B Bai et al. (2025a). We use at most $M = 32$ input frames with a maximum resolution of $256 \times 32 \times 32$, and fine-tune the model with LoRA ($r = 32$, $\alpha = 64$). The batch size is 32; the learning rates are $2\text{e}{-5}$ for LoRA and $1\text{e}{-4}$ for other modules. We use $n = 4$ visual latent tokens and $m = 10$ thought latent tokens, with supervision applied at layer $L = 13$. The loss weights are $\lambda_{visual} = 0.2$, $\lambda_{rec} = 1.0$, and $\lambda_{nce} = 0.5$ in Stages 1 and 3, and $\lambda_{ans} = 1.0$ in Stage 2. Training and testing are conducted on a single NVIDIA A100 and RTX 4090 GPU, respectively.

## 4.2 Baselines

We compare BRAID with a diverse set of state-of-the-art VAU methods spanning different learning and reasoning paradigms. For weakly supervised methods, we include VadCLIP Wu et al. (2024), Pi-VAD Majhi et al. (2025), LAS-VAD Wang & Zhao (2026), VERA Ye et al. (2025), and HeadHunt-VAD Cai et al. (2026). For training-free MLLM-based approaches, we compare with LAVAD Zanella et al. (2024), Monitor Feng et al. (2026), QVAD Bekit et al. (2026), and PANDA Yang et al. (2025).

We further compare with instruction-tuned VAU methods, including direct-answer approaches such as Holmes-VAU Zhang et al. (2025), VADER Cheng et al. (2026), ReactVAU Chen et al. (2026), VAD-DPO Zhang et al. (2026), VADor Lv & Sun (2024), and COPRA Jacob et al. (2026), as well as explicit chain-of-thought (CoT) reasoning methods such as VAU-R1 Zhu et al. (2025), Cue-R1 Yu et al. (2026), and VAD-R1 Huang et al. (2026). In addition, we evaluate several general-purpose MLLMs, including Qwen2.5-VL Bai et al. (2025b), Qwen3-VL Bai et al. (2025a), InternVL Chen et al. (2024), Video-ChatGPT Maaz et al. (2024), Video-LLaVA Lin et al. (2024), and Video-LLaMA Zhang et al. (2023b).

## 4.3 Main Results

**Comparison with state-of-the-art methods across multiple tasks.** We compare BRAID with direct-answer and explicit CoT-based methods on VAU-Bench. As shown in Table 1, BRAID achieves the best QA accuracy (93.59%) and anomaly classification performance (81.77% B-Acc and 54.53% Acc), surpassing the previous best results by 1.25, 4.09, and 3.68 points, respectively, while achieving competitive results on anomaly reasoning with a VAU-Eval score of 28.43. Despite using only a 2B backbone, BRAID outperforms several 7B/8B models and explicit CoT-based methods. Moreover, enabling the thinking modes of Qwen2.5-VL and Qwen3-VL degrades performance

Table 1: Performance on VAU-Bench across anomaly QA, classification, and reasoning. B-Acc denotes binary normal/anomalous classification accuracy. $*$ indicates methods with explicit CoT reasoning.

| Methods | Param | QA | CLS | | VAU-Eval | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | B-Acc | Acc | CLS | KM | FLU | INF | FAC | Total |
| Qwen2.5VL | 7B | 86.11 | **77.68** | **50.85** | 4.64 | **4.37** | **8.89** | **6.93** | **4.82** | **29.78** |
| InternVL2.5 | 8B | 88.03 | 76.94 | 49.33 | 4.37 | 4.03 | 8.09 | 5.34 | 4.31 | 26.13 |
| Video-LLaVA | 7B | 52.42 | 28.00 | 28.82 | 1.73 | 1.27 | 3.76 | 0.72 | 1.23 | 8.71 |
| Video-LLaMA | 7B | **92.34** | 50.47 | 34.38 | 3.67 | 2.54 | 8.10 | 2.40 | 2.73 | 19.44 |
| Holmes-VAU | 2B | 86.45 | 57.69 | 47.41 | 3.05 | 1.97 | 6.30 | 3.08 | 2.39 | 16.79 |
| Qwen3VL | 2B | 81.67 | 69.49 | 41.17 | 4.03 | 2.61 | 8.34 | 3.61 | 3.40 | 21.99 |
| Qwen2.5VL$^*$ | 7B | 87.01 | 59.30 | 45.68 | 4.15 | **4.52** | 8.47 | **6.40** | **4.54** | 28.08 |
| Qwen2.5VL$^*$ | 3B | 77.68 | 52.45 | 39.64 | 3.20 | 2.71 | 7.73 | 5.39 | 3.04 | 22.07 |
| Qwen3VL$^*$ | 2B | 80.52 | 53.83 | 43.56 | 3.54 | 3.08 | 8.69 | 3.71 | 3.12 | 22.12 |
| VAU-R1$^*$ | 3B | 89.98 | 61.56 | 48.82 | 3.47 | 3.27 | 7.00 | 5.44 | 3.59 | 24.26 |
| Cue-R1$^*$ | 3B | 90.11 | 61.34 | 48.10 | 3.82 | 2.98 | 7.75 | 3.26 | 3.04 | 20.85 |
| VAD-R1$^*$ | 3B | 91.44 | 64.14 | 48.76 | **4.80** | 3.01 | 7.42 | 4.82 | 3.11 | 23.16 |
| **BRAID(Ours)** | 2B | **93.59** | **81.77** | **54.53** | **6.27** | 4.18 | **9.05** | 4.93 | 4.17 | **28.43** |

on multiple tasks, suggesting that explicit CoT may introduce noisy reasoning. In contrast, BRAID factorizes complementary visual perspectives and performs latent deliberation over them, enabling more effective anomaly understanding.

Table 2: Comparison of video anomaly understanding performance on HIVAU-70K across different temporal granularities. C, E, and V denote clip, event, and video levels, respectively. $*$ denotes methods that explicitly generate chain-of-thought (CoT) reasoning.

| Method | Param | BLEU | | | CIDEr | | |
|---|---|---|---|---|---|---|---|
| | | C | E | V | C | E | V |
| Video-LLAMA | 7B | 0.151 | 0.079 | 0.104 | 0.024 | 0.014 | 0.017 |
| Video-LLAVA | 7B | 0.164 | 0.046 | 0.055 | 0.032 | 0.009 | 0.013 |
| VADER | 8B | **1.266** | **1.246** | **1.268** | **1.040** | **1.763** | **1.812** |
| ReactVAU | 10B | **1.150** | **1.366** | **1.339** | **0.920** | **2.032** | **2.016** |
| Holmes-VAU | 2B | 0.913 | 0.804 | 0.566 | 0.467 | 1.519 | 1.437 |
| VAD-R1$^*$ | 3B | 0.102 | 0.928 | 0.813 | 0.471 | 1.533 | 1.269 |
| Cue-R1$^*$ | 3B | 1.020 | 1.149 | 1.048 | 0.384 | 1.572 | 1.021 |
| VADER | 2B | 1.035 | 1.163 | 1.068 | 0.612 | 1.650 | **1.403** |
| Ours | 2B | **1.103** | **1.212** | **1.092** | **0.787** | **1.718** | 1.381 |
| **Method** | **Param** | **METEOR** | | | **ROUGE** | | |
| | | C | E | V | C | E | V |
| Video-LLAMA | 7B | 0.112 | 0.076 | 0.057 | 0.156 | 0.067 | 0.090 |
| Video-LLAVA | 7B | 0.097 | 0.022 | 0.014 | 0.132 | 0.023 | 0.045 |
| VADER | 8B | **0.247** | **0.216** | **0.164** | **0.429** | **0.463** | **0.446** |
| ReactVAU | 10B | **0.222** | **0.222** | **0.179** | **0.401** | **0.488** | **0.465** |
| Holmes-VAU | 2B | 0.190 | 0.165 | 0.121 | 0.329 | 0.370 | 0.355 |
| VAD-R1$^*$ | 3B | 0.196 | 0.174 | 0.132 | 0.284 | 0.391 | 0.383 |
| Cue-R1$^*$ | 3B | 0.201 | 0.178 | 0.129 | 0.279 | 0.412 | 0.374 |
| VADER | 2B | 0.215 | 0.183 | 0.137 | 0.376 | 0.444 | **0.408** |
| **BRAID(Ours)** | 2B | **0.227** | **0.196** | **0.159** | **0.394** | **0.450** | 0.402 |

**Multi-granularity performance on HIVAU-70k.** As shown in Table 2, BRAID achieves strong performance across metrics and temporal granularities on HIVAU-70K. Among 2B and 3B models, it achieves the best results in 10 of 12 settings, consistently outperforming explicit CoT-based methods such as VAD-R1 and Cue-R1, as well as Holmes-VAU and VADER-2B. These results suggest that simply relying on direct instruction tuning or explicitly verbalized CoT does not necessarily lead to stronger anomaly understanding. In contrast, BRAID factorizes complementary visual information into different perspectives and performs perspective-specific deliberation in the latent space, enabling the model to more effectively exploit anomaly-relevant information before contextualized synthesis. Despite using only a 2B backbone, BRAID also remains competitive with substantially larger 8B/10B models on several metrics.

**Evaluate anomaly localization performance on UCF-Crime and XD-Violence. Anomaly Localization on UCF-Crime and XD-Violence.** We evaluate BRAID in an online setting using only

Table 3: Anomaly localization performance on UCF-Crime and XD-Violence under different supervision settings. † denotes methods operating in the online setting without access to future frames.

| Annotation | Method | Size | UCF-Crime | XD-Violence | |
|---|---|---|---|---|---|
| | | | AUC (%) | AUC (%) | AP (%) |
| Weakly | VadCLIP | – | 88.04 | – | 84.51 |
| | VERA | 8B | 86.55 | **88.26** | 70.54 |
| | HeadHunt-VAD | 8B | 87.03 | – | 82.63 |
| | Pi-VAD | – | 90.33 | – | 85.37 |
| | LAS-VAD | – | **90.86** | – | **87.92** |
| Frame-level | Holmes-VAU | 2B | **88.96** | – | 87.68 |
| | StreamForest† | 7B | 85.26 | 92.82 | 75.92 |
| | ReactVAU† | 10B | 88.44 | **95.25** | **88.50** |
| Training-Free | LAVAD | 65B | 80.28 | **85.36** | 62.01 |
| | Monitor | 82B | 82.75 | 79.11 | 55.01 |
| | QVAD | 8B | 84.28 | – | 68.53 |
| | PANDA | – | **84.89** | – | **70.16** |
| Fine-Tuning | VADor | 7B | 85.90 | – | – |
| | VAD-DPO | 2B | 86.20 | 88.50 | – |
| | COPRA | 8B | 87.14 | – | 76.52 |
| | **BRAID(Ours)**† | 2B | **87.89** | **92.51** | **77.94** |

observed frames sampled at 1 FPS, and we categorize existing methods according to the supervision used during training. Specifically, the Fine-Tuning category includes methods trained only on instruction-tuning QA data without frame-level anomaly annotations. As shown in Table 3, BRAID achieves 87.89% AUC on UCF-Crime and 92.51% AUC / 77.94% AP on XD-Violence, outperforming all Fine-Tuning methods despite using only a 2B backbone. It also surpasses training-free multi-stage approaches and even the frame-level supervised StreamForest Zeng et al. (2026), whose result is reported in ReactVAU Chen et al. (2026). These results show that BRAID can effectively capture anomaly-related information and generalize to online localization without future frames or frame-level anomaly annotations.

Table 4: Ablation study on VAU-Bench QA and classification, and UCF-Crime anomaly localization. SS-O/ED-O denote scene-state/event-dynamics visual latent tokens, while SS-D/ED-D denote the corresponding thought latent tokens. Syn denotes Contextualized Latent Synthesis.

| SS-O | SS-D | ED-O | ED-D | Syn | QA (%) | Acc (%) | AUC (%) |
|---|---|---|---|---|---|---|---|
| × | × | × | × | × | 81.67 | 41.17 | 72.75 |
| × | × | ✓ | ✓ | × | 84.48 | 48.22 | 84.84 |
| ✓ | ✓ | × | × | × | 85.96 | 51.03 | 84.91 |
| ✓ | × | ✓ | × | ✓ | 87.13 | 50.48 | 83.59 |
| × | ✓ | × | ✓ | ✓ | 86.35 | 46.02 | 81.72 |
| ✓ | ✓ | ✓ | ✓ | × | 91.77 | 53.81 | 85.56 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **93.59** | **54.53** | **87.89** |

Table 5: Effect of the number of visual latent tokens $n$ and thought latent tokens $m$.

| $n$ | 2 | 4 | 8 | 16 |
|---|---|---|---|---|
| $m$ | 5 | 10 | 20 | 40 |
| Acc | 52.24 | 54.53 | 56.87 | 55.96 |

Table 6: Inference latency with different numbers of input frames.

| Frames | 8 | 16 | 32 | 64 |
|---|---|---|---|---|
| Latency (ms) | 137 | 194 | 362 | 669 |
| FPS | 58.4 | 82.5 | 88.39 | 95.6 |

### 4.4 Ablation Studies

We conduct ablation studies on different components of BRAID, including Perspective-Specific Observation (SS-O, ED-O), Perspective-Specific Deliberation (SS-D, ED-D), and Contextualized Latent Synthesis (Syn).

**Importance of Visual Observation.** As shown in Table 4, the baseline achieves 81.67% QA accuracy, 41.17% classification accuracy, and 72.75% AUC. Notably, using visual observations from both perspectives alone (Row 4) yields an average gain of approximately 8.5 points, demonstrating that perspective-specific visual observations help uncover richer anomaly-relevant details.

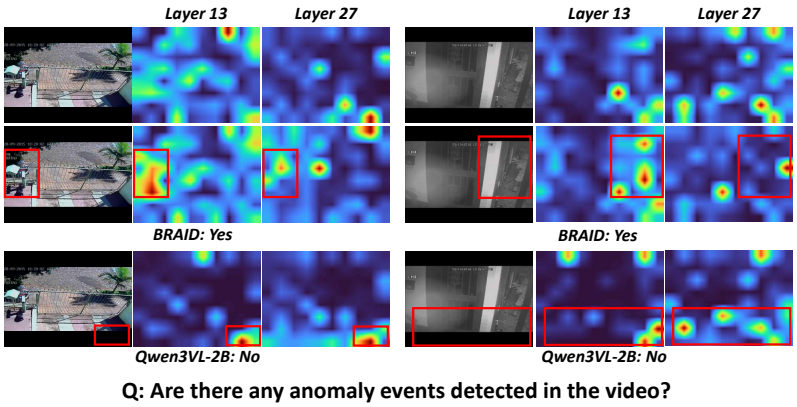


Figure 3: Visualization of attention maps from latent tokens to visual tokens.

**Importance of Deliberation.** Building deliberation upon perspective-specific visual observations further improves performance. Rows 2 and 3 show that jointly introducing observation and deliberation from either the event-dynamics or scene-state perspective brings average gains of 7.3 and 8.8 percentage points, respectively. In contrast, using deliberation without the corresponding perspective-aligned visual observations (Row 5) yields a smaller gain of about 6.2 points, indicating that effective reasoning benefits from sufficiently grounded visual evidence.

**Importance of comprehensive reasoning.** Combining observation and deliberation from both perspectives without synthesis (Row 6) already achieves an average improvement of approximately 11.8 percentage points, confirming the complementarity of scene-state and event-dynamics information. Further introducing Contextualized Latent Synthesis (Row 7) improves all three metrics and achieves the best overall performance, showing that context-aware integration can further exploit the complementary information captured by the two perspectives.

**Effect of different numbers of latent tokens.** As shown in Fig. 5, increasing the number of latent tokens generally improves anomaly classification performance on VAU-Bench, but the gains become marginal at $n = 16$ and $m = 40$. Considering the additional computational overhead, we adopt $n = 4$ and $m = 10$ as the default setting throughout our experiments.

**Inference latency of BRAID.** As shown in Table 6, we evaluate the inference latency of BRAID under different numbers of input frames. Compared with the baseline, BRAID introduces an overhead of only 14 latent tokens, resulting in low inference latency.

### 4.5 Qualitative Results

As shown in Fig. 3, we visualize the attention from event-dynamics and scene-state tokens to video tokens (Rows 1–2), together with the baseline attention map (Row 3). Supervised at the 13th layer, our visual latent tokens consistently attend to key anomaly-relevant information through the final layer, while the baseline fails to capture them effectively. Moreover, the two perspectives focus on complementary visual information, enabling more comprehensive anomaly understanding.

## 5 Conclusion

In this paper, we propose BRAID, a perspective-factorized latent reasoning framework. BRAID captures useful visual information from two complementary perspectives: scene-state information, including objects, spatial layouts, and scene changes, and event-dynamics information, including temporal interactions and event evolution. These information is organized into latent tokens, further used to form perspective-specific deliberation, and finally integrated for final reasoning. Extensive experiments show that BRAID outperforms existing video anomaly understanding methods while maintaining efficient inference. These results demonstrate that perspective-factorized reasoning is more effective for anomaly understanding than conventional approaches that encode heterogeneous visual information into a shared representation stream.

## Reproducibility Statement

We provide detailed descriptions of the model architecture, training procedure, implementation settings, and evaluation protocols in the main paper and supplementary material. Additional details on data processing, hyperparameters, and supervision construction are also provided to facilitate reproduction of our results.

## Ethics Statement

This work is conducted using publicly available benchmark datasets and does not involve the collection of new personal data or human-subject experiments. We have considered the potential ethical implications associated with the use of the proposed method and are not aware of any additional ethical concerns beyond those commonly associated with research on publicly available visual data.

## AI Use Statement

Generative AI tools were used solely for language polishing and error checking of the manuscript. The authors reviewed all AI-assisted revisions and take full responsibility for the final content of this work.

## References

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

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, 2005.

Lokman Bekit, Hamza Karim, Nghia T Nguyen, and Yasin Yilmaz. Qvad: A question-centric agentic framework for efficient and training-free video anomaly detection. *arXiv preprint arXiv:2604.03040*, 2026.

Z Cai, F Li, Z Zheng, H Bi, and L He. Headhunt-vad: Hunting robust anomaly-sensitive heads in mllm for tuning-free video anomaly detection (2025), 2026.

Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20392–20401, 2023.

Chia-Hui Chen, Shih-Ying Yeh, Fu-En Yang, Min-Hung Chen, and Shang-Hong Lai. Reactvau: A slow-fast decoupled framework for streaming video anomaly understanding. *arXiv preprint arXiv:2609.07941*, 2026.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 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 *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7301–7311. IEEE, 2026.

Hang Du, Guoshun Nan, Jiawen Qian, Wangchenhui Wu, Wendi Deng, Hanqing Mu, Zhenyan Chen, Ji Zhang, Min Lei, Pengxuan Mao, et al. Exploring what why and how: A multifaceted benchmark for causation understanding of video anomaly. *International Journal of Computer Vision*, 134(8):391, 2026.

Yue Feng, Yingshi Liu, Jingrou Zhang, Jie Qin, et al. Monitor: Exploiting large language models with instruction for online video anomaly detection. *Advances in Neural Information Processing Systems*, 38:35851–35871, 2026.

Chao Huang, Benfeng Wang, Wei Wang, Jie Wen, Chengliang Liu, Li Shen, and Xiaochun Cao. Vad-r1: Towards video anomaly reasoning via perception-to-cognition chain-of-thought. *Advances in neural information processing systems*, 38:118486–118518, 2026.

Darryl Cherian Jacob, Xinyu Liu, Kai Wang, and Pan He. Copra: Conditional parameter adaptation with reinforcement learning for video anomaly detection. *arXiv preprint arXiv:2605.15325*, 2026.

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

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. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

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

Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8022–8031, 2023.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad 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 (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Snehashis Majhi, Giacomo D'Amicantonio, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, Egor Bondarev, and François Brémond. Just dance with pi! a poly-modal inductor for weakly-supervised video anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24265–24274, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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, 2002.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

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*, pp. 2586–2595, 2025.

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.

Shengyang Sun and Xiaojin Gong. Hierarchical semantic contrast for scene-aware video anomaly detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22846–22856. IEEE, 2023.

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.

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 (ICCV)*, pp. 4975–4986, October 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.

Yu Wang and Shengjie Zhao. Weakly supervised video anomaly detection with anomaly-connected components and intention reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28545–28556, 2026.

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

Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature prediction diffusion model for video anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5527–5537, 2023.

Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14592–14601, 2023.

Zhiwei Yang, Chen Gao, and Mike Zheng Shou. Panda: Towards generalist video anomaly detection via agentic ai engineer. *arXiv preprint arXiv:2509.26386*, 2025.

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*, pp. 8679–8688, 2025.

Yating Yu, Congqi Cao, Zhaoying Wang, Weihua Meng, Jie Li, Yuxin Li, Zihao Wei, Zhongpei Shen, and Jiajun Zhang. Cuebench: Advancing unified understanding of context-aware video anomalies in real-world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12187–12195, 2026.

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*, pp. 18527–18536, 2024.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, et al. Streamforest: Efficient online video understanding with persistent event memory. *Advances in Neural Information Processing Systems*, 38:75804–75835, 2026.

Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16271–16280, 2023a.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pp. 543–553, 2023b.

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*, pp. 13843–13853, 2025.

Menghao Zhang, Huazheng Wang, Pengfei Ren, Kangheng Lin, Qi Qi, Haifeng Sun, Zirui Zhuang, Lei Zhang, Jianxin Liao, and Jingyu Wang. Do lvlms truly understand video anomalies? revealing hallucination via co-occurrence patterns. *Advances in Neural Information Processing Systems*, 38: 106738–106765, 2026.

Lingru Zhou, Peng Wu, Manqing Zhang, Qingsheng Wang, Guansong Pang, and Peng Wang. Targetvau: Multimodal anomaly-aware reasoning for target behavior understanding in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13710–13718, 2026.

Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. *Advances in Neural Information Processing Systems*, 37:89943–89977, 2024.

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.

**Algorithm 1:** Temporal Clustering for Scene Representative Selection

**Input:** Similarity matrix $\mathbf{S}$, minimum frame number $K_{\min}$
**Output:** Segments $\mathcal{C}$ and representative frames $\mathcal{R}$
Compute adjacent-frame change scores:

$$d_i = 1 - S_{i,i+1}$$

Detect transition boundaries from peaks in $\{d_i\}$;
Split the video into temporally ordered segments $\mathcal{C} = \{C_m\}_{m=1}^{M}$;
**for** *each segment* $C_m$ **do**
  Select the frame with the highest average intra-segment similarity;
$K \leftarrow \min(\max(K_{\min}, M), N)$;
**while** $|\mathcal{R}| < K$ **do**
  **for** *each* $C_m$ *in temporal order* **do**
    Select the unchosen frame that is temporally farthest from the selected frames in $C_m$;
    Add it to $\mathcal{R}$;
    **if** $|\mathcal{R}| = K$ **then**
      **break**;
Sort $\mathcal{R}$ in temporal order;
**return** $\mathcal{C}, \mathcal{R}$

# A More Details on BRAID

**Architectures of the SceneProjector and EventProjector.** $\mathrm{SceneProjector}$ and $\mathrm{EventProjector}$ share a similar architecture. They are mainly composed of linear projections, multi-head attention layers, and MLP blocks. Specifically, the input features are first transformed by a LayerNorm followed by a two-layer MLP with GELU activation. The projected features are then processed by a lightweight MLP to produce attention weights for feature aggregation. The core latent resampler stacks multiple blocks, each consisting of a LayerNorm-based cross-attention layer, a self-attention layer, and a feed-forward MLP with residual connections. Finally, a LayerNorm and linear projection map the latent representations into the LLM embedding space.

**More training details.** During training, we use dinov2-vitb14-reg as the DINO-v2 encoder, sam2-hiera-base-plus for SAM2, and vjepa2-vitl-fpc64-256 for V-JEPA. For CoT supervision generation, we employ Qwen-3.7-Plus as the teacher MLLM, with the prompting template shown in the Figure 4 and Figure 5. We use Qwen-Embedding to measure the semantic similarity between generated texts.

As described in the Scene-State Observation paragraph of Perspective-Specific Observation section, after computing the DINO-v2 feature similarities between video frames, we employ a temporally constrained clustering algorithm to select $S$ representative frames, as shown in Algorithm 1.

**Prompt for Scene Reasoning CoT Generation**

**You are a video scene understanding assistant specialized in analyzing scene progression and environmental changes in videos.**
**You will be given a video. Your task is to extract Scene Chain-of-Thought (Scene CoT) information that focuses ONLY on the scene and environment.**

**Requirements:**
**1. Analyze only the scene characteristics and scene transitions.**
**2. If the video contains a fixed/static scene:**
**- Describe the scene layout, environment, background, lighting, weather, location type, and overall visual characteristics.**
**3. If the scene changes over time:**
**- Describe the scene evolution and transitions in chronological order.**
**- Include changes in environment, camera viewpoint, location, lighting, weather, traffic conditions, crowd density, or background elements.**
**4. DO NOT analyze:**
**- human intentions**
**- abnormality**
**- event reasoning**
**- interactions**
**- safety analysis**
**- emotions**
**- causality**
**5. Focus purely on visual scene progression and environmental context.**

**Provide:**
**- A detailed description of the scene characteristics or scene transition process.**
**Output format:**
**{**
**"Scene-CoT": <CoT>**
**}**

Figure 4: Prompts for Generating Scene Reasoning CoT.

**Prompt for Event Reasoning CoT Generation**

**You are a video event reasoning assistant specialized in extracting chronological event progression from detailed video analyses. Focus only on observable event flow and temporal development, without analysis or interpretation.**
**You are given a detailed analysis of a video event, including descriptions, explanations, and reasoning about the incident.**
**Your task is to extract ONLY the event development process (Chain-of-Thought for event progression) from the provided analysis.**

**Requirements:**
**1. Focus ONLY on the chronological sequence of what happened in the video.**
**2. Describe how the event developed step-by-step over time.**
**3. Include actions, interactions, movements, and visible event transitions.**
**4. DO NOT include:**
**- anomaly analysis**
**- safety analysis**
**- emotional interpretation**
**- legal judgment**
**- social commentary**
**- cause speculation unless explicitly visible in the event flow**
**- conclusions or summaries about abnormality**
**5. The output should strictly describe the observable event progression.**

**Output:**
**- A clear and complete event progression description.**
**- Preserve important temporal details and interactions.**
**Output format:**
**{**
**"Event-CoT": <CoT>**
**}**

Figure 5: Prompts for Generating Event Reasoning CoT.