🦅 HAWK:
Learning to Understand Open-World Video Anomalies

Jiaqi Tang[1,2,3*]   **Hao Lu**[1,2*]   **Ruizheng Wu**[4]   **Xiaogang Xu**[5,6]   **Ke Ma**[7]
Cheng Fang[7]   **Bin Guo**[7]   **Jiangbo Lu**[3,4]   **Qifeng Chen**[2]   **Ying-Cong Chen**[1,2,3†]

[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology   [3]HKUST(GZ) – SmartMore Joint Lab
[4]SmartMore Corporation   [5]The Chinese University of Hong Kong   [6]Zhejiang University
[7]Northwestern Polytechnical University
{jtang092, hlu585}@connect.hkust-gz.edu.cn
{ruizheng.wu, jiangbo}@smartmore.com   xiaogangxu00@gmail.com
{2544552413, sura}@mail.nwpu.edu.cn   guob@nwpu.edu.cn
cqf@ust.hk   yingcongchen@hkust-gz.edu.cn

Abstract

Video Anomaly Detection (VAD) systems can autonomously monitor and identify disturbances, reducing the need for manual labor and associated costs. However, current VAD systems are often limited by their superficial semantic understanding of scenes and minimal user interaction. Additionally, the prevalent data scarcity in existing datasets restricts their applicability in open-world scenarios. In this paper, we introduce **HAWK**, a novel framework that leverages interactive large Visual Language Models (VLM) to interpret video anomalies precisely. Recognizing the difference in motion information between abnormal and normal videos, **HAWK** explicitly integrates motion modality to enhance anomaly identification. To reinforce motion attention, we construct an auxiliary consistency loss within the motion and video space, guiding the video branch to focus on the motion modality. Moreover, to improve the interpretation of motion-to-language, we establish a clear supervisory relationship between motion and its linguistic representation. Furthermore, we have annotated over 8,000 anomaly videos with language descriptions, enabling effective training across diverse open-world scenarios, and also created 8,000 question-answering pairs for users' open-world questions. The final results demonstrate that **HAWK** achieves SOTA performance, surpassing existing baselines in both video description generation and question-answering. Our codes/dataset/demo will be released at https://github.com/jqtangust/hawk.

1   Introduction

*"Have eyes like a **HAWK**!" – Longman Dictionary*

In recent years, the deployment of Video Anomaly Detection (VAD) systems has seen a significant uptick across a diverse array of domains, including but not limited to, autonomous driving [45, 25], surveillance [6, 23], and crime scene analysis [33]. The inherent capability of these systems to autonomously monitor and identify disturbances within a scene has markedly diminished the reliance on manual labor, thereby streamlining operational efficiency and reducing associated costs.
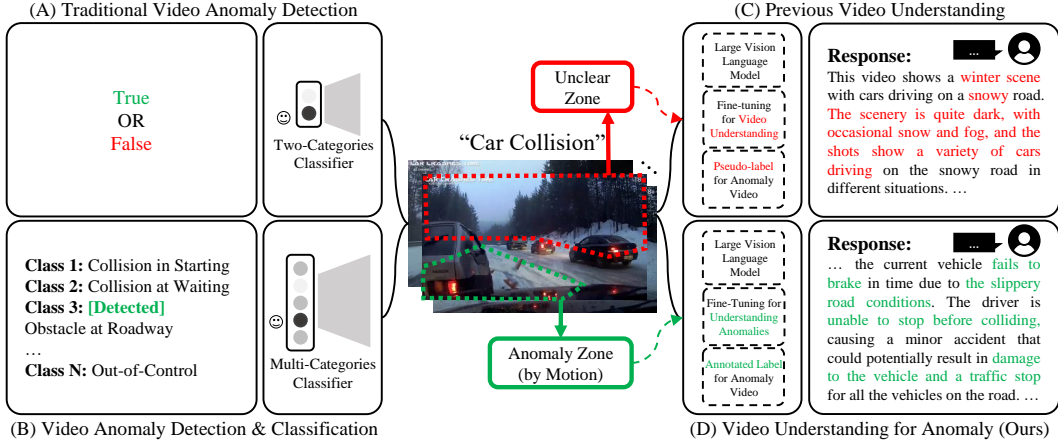
---

[*]Equal contribution.
[†]Corresponding author.

Figure 1: Different framework in video anomaly detection. (A) shows traditional video anomaly detection methods, which use binary classifiers to detect anomalies. (B), following (A), introduces a multi-class classifier for integrating semantic information, allowing users to obtain different types of anomaly information. Neither (A) nor (B) can interact with users. (C) is a previous video understanding framework that can interactively provide richer semantic information for users, but cannot specifically locate video anomalies. Our framework (D) enhances the anomaly understanding capability and provides annotated labels with rich semantic information.

Despite the extensive focus on anomaly detection in most existing VAD systems [23, 44, 33, 31, 9, 13, 19, 34, 40, 48, 52] (as shown in Fig. 1 (A)), there is often a lack of deeper semantic understanding of the scenes and insufficient interaction with users. While Pu et al. [31] and Wu et al. [42] incorporated semantic information for video anomaly detection, their frameworks are limited as multiple-class classifiers (as displayed in Fig. 1 (B)). Consequently, the functionality of these systems is confined to the detection of anomalous frames, necessitating further manual analysis by users to analyze the detected anomalies comprehensively. Although Lv et al. [27] has pioneered the development of a large language model for the video anomaly explanation, their approach primarily relies on *pseudo labels* for training. The lack of robust training data severely constrains its practical applicability. Besides, such a method focuses more on acquiring long-range context information rather than anomaly-related features on anomaly understanding (as exhibited in Fig. 1 (C)).

To solve the above challenges, we propose an interactive large visual-language model [21, 18, 29], **HAWK**, for precisely understanding video anomalies (as illustrated in Fig. 1 (D)). Considering that the motion in normal and abnormal videos is significantly different [44, 52], we explicitly integrate motion modality by a dual-branch framework in **HAWK** to enhance the understanding of anomalies (Section 4.1). Besides, to reinforce motion attention, we construct an auxiliary consistency loss based on the mutual information between the original video (appearance feature) and its motion in tight space (Section 4.2), to implicitly guide the video branch to focus on motion-related features. However, the interpretation of motion to the corresponding language remains unclear. Therefore, we extract the motion-related language (verbs and their entities) from the original description to directly supervise the visual and linguistic representations of motion, for accurately enhancing the interpretation of video anomaly in **HAWK** (Section 4.3).

Furthermore, we also collect seven video anomaly datasets from various scenarios and generate language descriptions for each video. Besides, to address the open-ended questions raised by users, we utilize language descriptions of the videos to generate potential question-answer pairs for training. Since these datasets cover a range of scenarios (Section 3), including crime (UCF-Crime [33]), campus environments (ShanghaiTech [22] and CUHK Avenue [23]), pedestrian walkways (UCSD Ped1 [6] and Ped2 [37]), traffic situations (DoTA [45]), and human behavior (UBnormal [2]), and finally, the model tends to generalize to open-world scenarios.

To train our framework, we initially pre-train it on WebVid [3] to equip it with the capability to understand general videos. Then, we fine-tuned it on our proposed video anomaly dataset to enhance its understanding of video anomalies across multiple scenarios. Compared to other baselines, our

model achieves SOTA performance in both Text-Level and GPT-Guided Metrics. Our contributions are summarized as follows:

- We propose a novel video-language framework, **HAWK**, aiming at understanding video anomalies, which incorporates motion modality to enhance its capability.
- We generate rich language descriptions for seven different video anomaly datasets. Meanwhile, considering the diversity of open-world problems, we also generate question-answer pairs to tackle potential user inquiries.
- Compared to other large video models, our framework demonstrates SOTA performance for video anomaly understanding and question-answering across multiple scenarios, which will help open-world anomaly understanding in the future.

2   Related Work

Video Anomaly Detection   Video Anomaly Detection (VAD) usually focuses on identifying unexpected events from the video and it has been widely applied in various fields, including autonomous driving [45], public surveillance [6, 23], and crime scene analysis [33] etc. Previous VAD methods [27, 33, 23, 44, 9, 13, 19, 34, 40, 48, 52] are designed in numerous pathways. Lu et al. [23] designed to learn video features only from normal videos, and hand-craft features or deep-learning-based features are leveraged. Sultani et al. [33] proposed multiple instance learning (MIL), which is the main paradigm for many weakly-supervised learning methods. Recently, Lv et al. [27] first proposed video-based large language models in the framework of VAD.

However, these methods lack sufficient semantic comprehension of scenes and offer inadequate user interaction. Several approaches [31, 42] have introduced multi-class classifiers to integrate semantic information with various types of anomaly information. Nevertheless, their output is still limited. In contrast, our framework not only integrates more comprehensive semantic information as a general video understanding system but also provides advanced interaction capabilities for users.

Large Model in Video Understanding   Recent studies have demonstrated the reliable capabilities of large models in video understanding. Beyond powerful vision-language models [16, 51, 21, 24], recent research has increasingly explored more modalities [27, 18, 28, 46, 26]. Bain et al.[3] introduced a large-scale dataset with general video content descriptions. Several LLM-based works[18, 28, 46, 26] aim to comprehend visual content. Additionally, Video-LLaMa [49] extends comprehension to both auditory and visual information, while Su et al.[32] utilize multi-modal encoders to understand across six modalities. Recently, Lv et al.[27] proposed video-based large language models for VAD tasks in a weakly supervised framework. In this paper, we introduce the *motion modality* in our proposed vision-language model, which enhances the model's ability to locate anomalies by prioritizing relevant video content.

3   Data Engineering

Previous datasets are inadequate for addressing our problem. Most existing VAD datasets, such as UBnormal [2] and DoTA [45], only contain **simple video category labels** and lack detailed language descriptions. This results in video understanding models lacking accurate and comprehensive supervision, creating a significant obstacle to identifying anomalies in videos. Recently, Lv et al.[27] attempted to create **pseudo language descriptions** for anomaly videos. However, these descriptions are naive combinations of labels and fixed text, relying on a rigid format that offers only limited information. Other datasets, like WebVid[3], include only **general descriptions** of video content, which may not direct the model's focus on anomalies.

Our Principle   To tackle the above problems, we annotate detailed language descriptions specifically for anomaly scenes in seven different existing <VIDEO> datasets. These seven datasets include a variety of anomalous scenarios such as crime (UCF-Cirme [33]), campus (ShanghaiTech [22] and CUHK Avenue [23]), pedestrian walkways (UCSD Ped1 [6] and Ped2 [37]), traffic (DoTA [45]), and human behavior (UBnormal [2]). With the support of these visual scenarios, we can perform comprehensive fine-tuning for various abnormal scenarios, being closer to open-world scenarios.
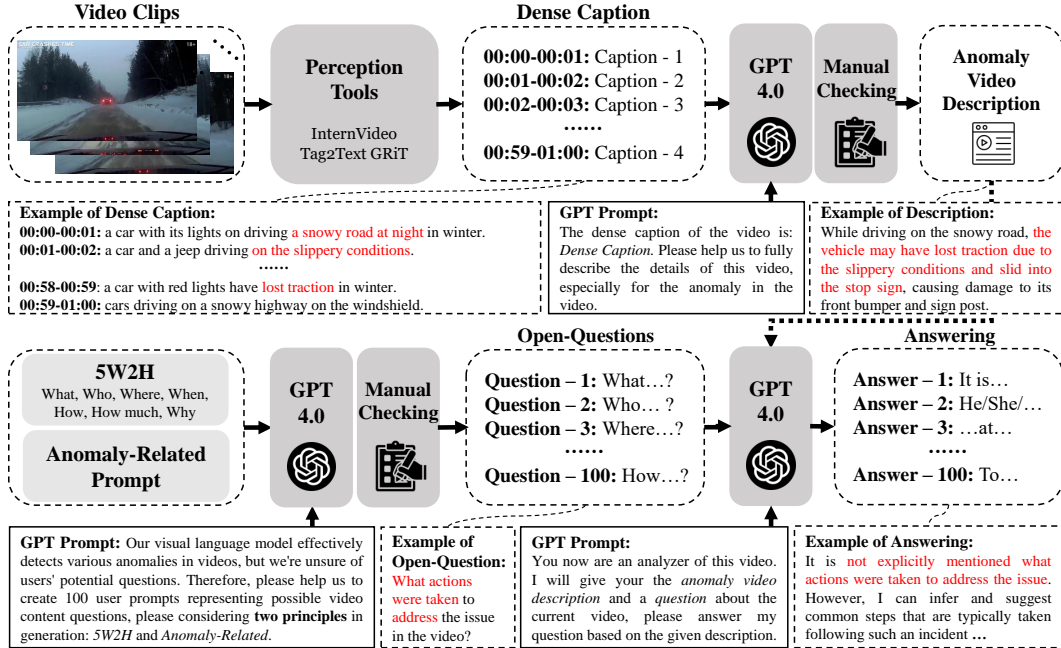
Figure 2: Generation pipeline of our dataset. In the first line, we first **segment videos into clips** and **generate dense captions** for each segment, including a comprehensive description of the video content. Then, we use GPT-4 to guide the **generation of corresponding anomalous video descriptions** based on these descriptions, which are then **manually checked to reduce mistakes**. In the second line, to generate user-centered QA pairs, we first use GPT-4 to **generate open-ended questions** based on the proposed two principles. Then, the questions and video descriptions are jointly input into GPT-4 to **provide possible answers**.

Moreover, to better account for real-world user situations, we believe that language descriptions should not only include **descriptions of the video anomalies** themselves, but also address **open questions asked by users**. Therefore, we construct open-ended question-answer pairs for each scenario to further enhance model's practical ability to answer users' varying questions. The procedure for answering users' questions is shown in Fig. 2. The data format of can be described by the Eq. (1),

$$\text{<VIDEO>}: \{\textbf{DIS}: \text{<DESCRIPTION>} \mid \textbf{QA}: \text{<QUESTION>} \rightarrow \text{<ANSWERING>}\}. \tag{1}$$

Anomaly Video Description Generation To construct natural language descriptions <DESCRIPTION> for anomalous video datasets, we refer to previous research such as LLaVa [21] and VideoChat [18], and employ GPT-4 [1] as an assistant. We first split the video into dense clips to ensure key information is captured. Following VideoChat [18], we use perception tools (InternVideo [38], Tag2Text [14], or GRiT [39]) to automatically generate captions for each key clip, obtaining a dense representation of the videos (except for the UCF-Crime dataset, which already has a dense representation built in [47]). Next, we use GPT-4 [1] to generate anomaly-related descriptions based on the captions for each video. Unlike other general video understanding datasets [21, 18], we provide prompts for GPT-4 to generate specific descriptions closely related to video anomalies. Finally, due to varying quality of dense captions, some videos may have incorrect annotations. Thus, we manually recheck the final generated video anomaly descriptions to ensure label accuracy.

Human-Centric Question-Answering Pairs Generation So far, we have obtained nearly accurate descriptions of anomaly videos. However, our framework may still face challenges with more open-ended questions from users. Therefore, anomaly-related question-answering is a significant practical requirement. Given the diversity of open-world scenes, users may ask questions involving various pronouns. Thus, we mainly consider these two principles: ① **Anomaly-related**, our questions should be strongly related to the anomaly in the video. ② **5W2H**, we introduce seven different question pronouns (What, Who, Where, When, How, How much, and Why) to simulate various question formats that users may employ. This enables us to address a wide range of open questions related to

video anomalies. We input these two principles into GPT-4 [1] to generate open questions for anomaly videos. We then manually review and select the 100 most suitable questions, which are randomly assigned to each video. Finally, GPT-4 [1] will generate <ANSWERS> to these <QUESTIONS>.

Our data is more practical compared to previous ones: it not only understands multiple anomalies in videos but also supports question-answering in open scenarios (More details in Appendix D).

4 Methodology

To construct a practical framework for understanding video anomalies, our goal is to accurately interpret these anomalies into natural language. However, most previous studies [18, 49, 29, 20, 27] focus on enhancing general video understanding capabilities while neglecting video anomalies. This oversight results in equal attention being given to all parts of the video, such as the background and human appearances, often at the expense of key anomaly features, as shown in Fig. 1 (C). Consequently, these approaches are not effective in accurately focusing on anomaly-related features.

Overview of Solution   The core of our solution is guiding visual instruction to focus on anomalies. Previous studies in video anomaly detection [44, 52] have demonstrated that *motion-related feature* help identify multiple anomalies. Therefore, in Section 4.1, we first explicitly integrate a motion modality into our proposed framework to target anomaly-related features. Subsequently, in Section 4.2, we maintain mutual information consistency between the appearance and motion modalities within a tight feature space, implicitly guiding the appearance branch to reinforce motion attention. Finally, in Section 4.3, to improve the interpretation of motion-to-language, we extract motion-related language descriptions to directly match the motion and its corresponding motion-related language.

4.1   Explicit Motion Modality Integration

To enhance the capability of interpreting anomalies, we build a framework, **HAWK**, to explicitly integrate motion modality. **HAWK** has a dual-branch architecture, with f_v as the original video understanding network and f_m for motion understanding. Inspired by Video-LLaMA [49], f_v and f_m share the same architecture but separate parameters in Fig. 3. Eq. (2) denotes our framework as,

$$\mathbf{Y} = \text{LLAMA}\left([P_v(f_v(\mathbf{X_v})), P_m(f_m(\mathbf{X_m}))] \oplus f_t(\mathbf{T})\right), \tag{2}$$

where $\mathbf{X_v} \in \mathrm{R}^{T \times C \times H \times W}$ represents the <VIDEO> input for extracting appearance feature, and T denotes the temporal dimension. $\mathbf{X_m} = M(\mathbf{X_v})$, with $M(\cdot)$ being the motion extractor.

$f_v(\cdot)$ and $f_m(\cdot)$ are the frozen pre-trained video encoders from BLIP-2 [17], which consist of one EVA-CLIP [10] and one pre-trained Video Q-Former to output embeddings. Then, the output embeddings from $f_v(\cdot)$ and $f_m(\cdot)$ are passed through learnable projection networks for video and motion, $P_v(\cdot)$ and $P_m(\cdot)$, respectively. These networks aim to project visual (video and motion) embedding into the language feature space for interpreting. $f_t(\cdot)$ is the frozen text token to embedding projection, that makes textual information can be inputted into LLaMA-2 [35]. $\oplus$ is for combining our input prompt, we define our prompt as: "*Here is the input video embedding:* <VIDEO_EMBEDDING> *and motion embedding* <MOTION_EMBEDDING> *in different frames, please help me to* <DESCRIBE_VIDEO> | <QUESTION>.". <DESCRIBE_VIDEO> and <QUESTION> are the question classes for video description generation and video question answering respectively (Details see Appendix D). By combining the visual token embedding with the textual embedding, $f_t(\mathbf{T})$, LLaMA-2 [35], is employed to generate the final language response, $\mathbf{Y}$. This framework explicitly integrates the motion modality during visual instruction tuning, significantly targeting anomaly-related features.

4.2   Implicitly Motion Attention Reinforcement

Although we integrate the motion modality to facilitate fine-tuning of **HAWK**, motion and video branches operate independently. Therefore, we cannot expect the original video branch to extract appearance features that focus on the region where the anomaly occurred (i.e., motion). To help **HAWK** focus more on these regions, we observed the containment relationship in mutual information between motion and the original video. We use this relationship to construct an auxiliary consistency loss function, implicitly reinforcing the motion attention (Fig. 4 ②).
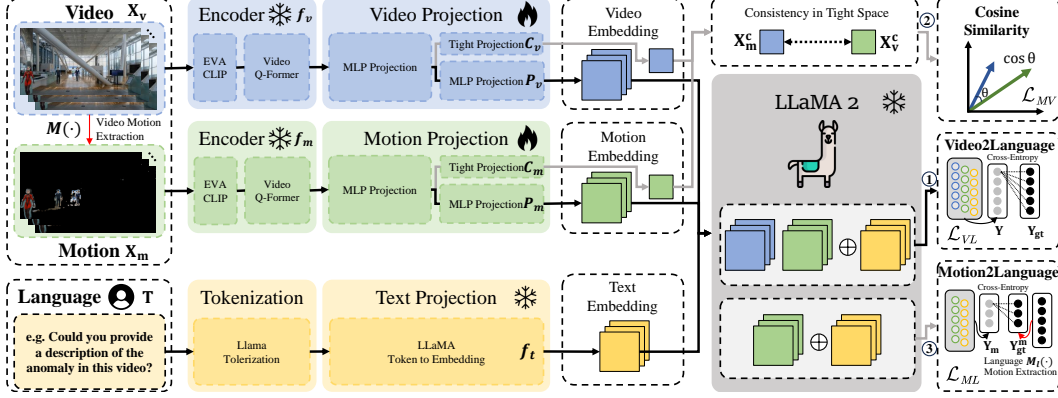
Figure 3: Overview of **HAWK**. During training (**Black** and Gray path), we aim to optimize for video-language matching loss, along with Video-Motion Consistency and Motion-Language Matching. During inference (only Gray path), we generate language descriptions using video, motion, and text.

Extract Motion   Specifically, to obtain motion, we employ a motion describer $M(\cdot)$, which generates motion between two successive frames as shown in Eq. (3),

$$\mathbf{X}_{\mathbf{Motion}}^{(\mathbf{t})} = M^{(\mathbf{t})}(\mathbf{X}_{\mathbf{v}}^{(\mathbf{t})}, \mathbf{X}_{\mathbf{v}}^{(\mathbf{t-1})}), \tag{3}$$

where $M^{(\mathbf{t})}(\cdot)$ is the motion describer at the time step $\mathbf{t}$, we currently use Gunnar Farneback's algorithm [11], and $\mathbf{X}_{\mathbf{v}}^{(\mathbf{t})}, \mathbf{X}_{\mathbf{v}}^{(\mathbf{t-1})} \in \mathrm{R}^{1 \times C \times H \times W}$ denote the video frames at time steps $\mathbf{t}$ and $\mathbf{t-1}$.

$\mathbf{X}_{\mathbf{Motion}}^{(\mathbf{t})} \in \mathbb{R}^{2 \times H \times W}$ includes two channels motion vector in $\mathbf{X}$ (horizontal) and $\mathbf{Y}$ (vertical) directions. We use the optical flow magnitude from these channels as a **Mask**, normalized to $[0, 1]$ and multiplied with the original video appearance, to hide other non-motion regions, as Eq. (4),

$$\mathbf{X}_{\mathbf{m}}^{(\mathbf{t})} = \underbrace{\mathrm{NORM}(\sqrt{(\mathbf{X}_{\mathbf{Motion}}^{(\mathbf{t})}(\mathbf{X}))^2 + (\mathbf{X}_{\mathbf{Motion}}^{(\mathbf{t})}(\mathbf{Y}))^2})}_{\mathbf{Mask}} \times \mathbf{X}_{\mathbf{v}}^{(\mathbf{t})}, \tag{4}$$

where $\times$ is the operator of pixel-wise multiplication. $\mathbf{X}_{\mathbf{v}}^{(\mathbf{t})}, \mathbf{X}_{\mathbf{m}}^{(\mathbf{t})} \in \mathrm{R}^{1 \times C \times H \times W}$ donate the original video and our input motion information at time step $\mathbf{t}$, respectively. We usually extract T frames as motion input $\mathbf{X}_{\mathbf{m}} \in \mathrm{R}^{T \times C \times H \times W}$, same as $\mathbf{X}_{\mathbf{v}}$.

Build $\mathcal{L}_{MV}$ Loss   Then, we consider that $\mathbf{X}_{\mathbf{m}}$ only contains key information for anomaly and it is contained in $\mathbf{X}_{\mathbf{v}}$, and feature space from $\mathbf{X}_{\mathbf{v}}$ is more sparse. Therefore, we compact features from $\mathbf{X}_{\mathbf{m}}$ and $\mathbf{X}_{\mathbf{v}}$ into a tight space. At this space, we aim to maintain the mutual information between $\mathbf{X}_{\mathbf{m}}$ and $\mathbf{X}_{\mathbf{v}}$ consistency, and in this way, the appearance feature can be focused on the motion region. Therefore, we construct an auxiliary loss to promote $\mathbf{X}_{\mathbf{v}}$'s motion attention, as in Eq. (5),

$$\mathcal{L}_{MV} = 1 - \mathrm{COS\_SIM}(\mathbf{X}_{\mathbf{m}}^{\mathbf{c}}, \mathbf{X}_{\mathbf{v}}^{\mathbf{c}}) = 1 - \frac{\mathbf{X}_{\mathbf{m}}^{\mathbf{c}} \cdot \mathbf{X}_{\mathbf{v}}^{\mathbf{c}}}{\|\mathbf{X}_{\mathbf{m}}^{\mathbf{c}}\| \|\mathbf{X}_{\mathbf{v}}^{\mathbf{c}}\|}, \quad (5)$$

where $\mathbf{X}_{\mathbf{v}}^{\mathbf{c}} = C_v(f_v(\mathbf{X}_{\mathbf{v}}))$ and $\mathbf{X}_{\mathbf{m}}^{\mathbf{c}} = C_m(f_m(\mathbf{X}_{\mathbf{m}}))$ denote the tightly compressed representations of $\mathbf{X}_{\mathbf{v}}$ and $\mathbf{X}_{\mathbf{m}}$, respectively, by the compression functions C_v and C_m. C_v and C_m share some initial shallow layer parameters with P_v and P_m (as Fig. 3). Then, following a subsequent tight projection to compresses both $\mathbf{X}_{\mathbf{v}}$ and $\mathbf{X}_{\mathbf{m}}$ into a more compacted space.

Figure 4: Visualization of **HAWK**'s loss. ① is the original video-to-language loss. ② is the cosine similarity loss for motion modality adaptation. ③ is the motion-to-language loss.

Finally, with this auxiliary loss, we can reinforce the motion attention in the appearance feature, and **HAWK**'s feature space will focus on more abnormal related features, which will promote the understanding of anomalies in the whole framework.
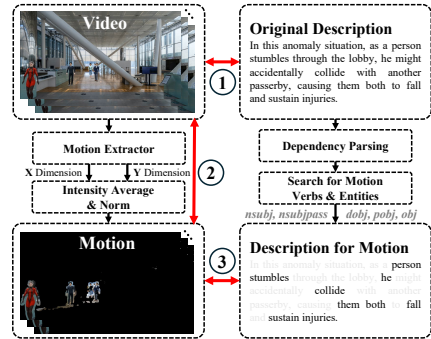
4.3 Interpreting Motion-to-Language

Although **HAWK** has already accommodated the motion modality in visual input, the corresponding motion from language is still unclear. This limitation hinders **HAWK**'s interpretation in motion modality. Hence, to augment this relationship, we aim to reinforce the correspondence between motion and their linguistic representation.

Extract Motion-related Language Previous studies [5, 36, 43, 15] have proved that the representation of motion in the language is predominantly from **verbs** and their corresponding **entities**. Therefore, to extract linguistic representation, the first step is to do dependency parsing for the original sentences, as Eq. (6),

$$\mathbf{G_{gt}} = D(\mathbf{Y_{gt}}), \tag{6}$$

where $D(\cdot)$ is the dependency parsing and $\mathbf{Y_{gt}}$ is the ground truth. $\mathbf{G_{gt}}$ represents the graph of the dependency structure, which symbolizes the syntactic relationships among the words in a sentence.

Based on this graph, we can extract predicates (verbs) $\mathbf{V}$, and also entities closely related to these predicates, such as subjects $\mathbf{S}$, objects $\mathbf{O}$, indirect subjects $\mathbf{S_i}$, and indirect objects $\mathbf{O_i}$. These elements are then combined to form short phrases representing motion, as in Eq. (7),

$$\mathbf{Y_{gt}^m} = \{\mathbf{V}, \mathbf{S}, \mathbf{O}, \mathbf{S_i}, \mathbf{O_i}\} = M_l(\mathbf{G_{gt}}), \tag{7}$$

where $M_l(\cdot)$ is the language motion extraction operator, and $\mathbf{Y_{gt}^m}$ is the motion-related language.

Build $\mathcal{L}_{ML}$ Loss After obtaining motion-related language, we can establish strong supervision between motion in both vision and linguistic representation (as Fig. 4 ③), significantly enhancing the ability to interpret motion to language in **HAWK**. Consequently, we design a motion-language matching as an auxiliary loss, as Eq. (8),

$$\mathcal{L}_{ML}^m(\mathbf{Y_m}, \mathbf{Y_{gt}^m}) = -\sum_{i=1}^{N} \mathbf{Y_{gt}^{m\,i}} \log(\mathbf{Y_m}^i)$$
$$\text{s.t. } \mathbf{Y_m} = \text{LLAMA}\left(P_m(f_m(\mathbf{X_m})), f_t(\mathbf{T})\right), \tag{8}$$

where $\mathcal{L}_{ML}(\cdot)$ is the cross-entropy loss, which contains N words.

Optimization Goal Finally, our total loss $\mathcal{L}$ shows as, $\mathcal{L} = \mathbf{t_0} \times \mathcal{L}_{VL} + \mathbf{t_1} \times \mathcal{L}_{MV} + \mathbf{t_2} \times \mathcal{L}_{ML}$, where $\mathcal{L}_{VL}$ is original video to language loss (as Fig. 4 ①), and $\mathbf{t_0}$, $\mathbf{t_1}$ and $\mathbf{t_2}$ is the hyper-parameter.

5 Experiments

In this section, we will provide a comprehensive introduction to the experiments, including the processes of training and testing, the establishment of baselines, the methods of evaluations, and the detailed examination of ablation experiments pertaining to **HAWK**.

Training & Testing To enhance our framework's anomaly understanding capabilities, we've structured our training and testing process into three stages, as Fig. 5. **Stage 1** involves pre-training on the WebVid dataset [3] to acquire a general understanding of video content. In **Stage 2**, we finetune the model's focus towards video anomaly understanding by employing a specially curated dataset described in Section 1, consisting of over $8,000$ videos. We use 90% of these videos for training and allocate the remaining 10% for testing purposes. We jointly train on two tasks: video <DESCRIPTION> generation and video <QUESTION>→<ANSWERING>. In **Stage 3**, we evaluate these two tasks independently in the testing set to ensure our model's effectiveness.
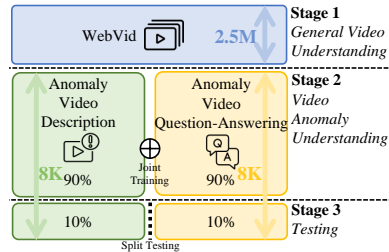

Figure 5: Training & Testing.

Baselines To evaluate the anomaly understanding performance of our proposed framework, we conduct comparisons with SOTA video understanding baselines. We select five baselines: Video-ChatGPT [29], VideoChat [18], Video-LLaMA [49], LLaMA-Adapter [50], and Video-LLaVA [20]. Our comparison aims to determine whether these baselines can fully understand and interpret video

Table 1: Quantitative performance of (A) anomaly video description generation and (B) video question-answering. Red indicates the best performance, while blue denotes the second best.

(A) Anomaly Video Description Generation

Method	Backbones	Text-Level (↑) [30]				GPT-Guided (↑) [21]		
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	Reasonability	Detail	Consistency
Video-ChatGPT [29]	LLaMA 7B	0.107	0.046	0.017	0.008	0.084	0.108	0.055
VideoChat [18]	Vicuna 7B	0.053	0.023	0.008	0.003	0.107	0.205	0.054
Video-LLaMA [49]	LLaMA-2 7B	0.062	0.025	0.009	0.004	0.120	0.217	0.066
LLaMA-Adapter [50]	LLaMA-2 7B	0.132	0.052	0.018	0.008	0.060	0.091	0.038
Video-LLaVA [20]	Vicuna 7B	0.071	0.030	0.012	0.005	0.077	0.115	0.038
Ours	LLaMA-2 7B	0.270	0.139	0.074	0.043	0.283	0.320	0.218

(B) Anomaly Video Question-Answering

Method	Backbones	Text-Level (↑) [30]				GPT-Guided (↑) [21]		
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	Reasonability	Detail	Consistency
Video-ChatGPT [29]	LLaMA 7B	0.177	0.096	0.058	0.038	0.508	0.430	0.421
VideoChat [18]	Vicuna 7B	0.261	0.133	0.074	0.043	0.699	0.631	0.598
Video-LLaMA [49]	LLaMA-2 7B	0.156	0.081	0.045	0.027	0.586	0.485	0.497
LLaMA-Adapter [50]	LLaMA-2 7B	0.199	0.109	0.067	0.043	0.646	0.559	0.549
Video-LLaVA [20]	Vicuna 7B	0.094	0.054	0.034	0.023	0.393	0.274	0.316
Ours	LLaMA-2 7B	0.319	0.179	0.112	0.073	0.840	0.794	0.753

anomalies. To ensure the fairness of our experiments, we employed the baselines with the same size (7B parameters) as the backbone.

Evaluation Metrics   To accurately evaluate our model's performance in understanding video anomalies, we firstly adopt four **Text-Level** metrics, from BLEU (Bilingual Evaluation Understudy) [30]-1 to BLEU-4 to measure word overlap between the model-generated text and the ground truth. This approach enables us to objectively assess the similarity and also take into account various levels of granularity at the text-level, thus providing a clear indicator of how well the model understands and describes anomalies.

Besides, we expand our evaluation framework by incorporating insights from recent research in LLaVa [21] or Video-ChatGPT [29], utilizing **GPT-Guided** [1] methods to assess the quality of the generated text. GPT [1] serves as a critical evaluator, generating scores for three key aspects of the language produced, with each aspect scored on a scale from 0 to 1. These three aspects are as,

- **Reasonability:** evaluates the logical reasoning and coherence of the generated language.
- **Detail:** assesses the level of detail and specificity of the generated language.
- **Consistency:** evaluates the coherence and consistency of the generated language.

By leveraging GPT [1] as an evaluative tool, we aim to provide a nuanced understanding of the text's quality, focusing on aspects that traditional metrics may overlook.

Quantitative Evaluation   Table 1 (A) and (B) demonstrate the effectiveness of our model to describe abnormal phenomena. Our proposed model significantly outperforms the previous baselines, achieving SOTA performance in every metric for both Text-level and GPT-guided metrics, thus it can generate text that more closely aligns with actual scenarios.

Qualitative Evaluation   Table 2 (A) and (B) demonstrate that our proposed framework achieves optimal qualitative performance in video description generation and question-answering, respectively. Compared with other baselines, **HAWK** can accurately understand and focus on video anomalies. For example, in Table 2 (A) - Video-LLaMa [49], it pays more attention to the clothing information from the people (*wearing blue and red jacket*), while ignoring the motion-related anomaly (*slipping*). In Table 2 (B) - Video-ChatGPT, it may produce hallucinations (*two people... who were hit by the car*), which differ from the original video anomaly (*car suddenly braking*). In contrast, **HAWK** generates descriptions that are close to the real semantics (*driver losing control*).

Table 2: Qualitative performance on (A) anomaly video description generation, and (B) question-answering. Red texts indicate key semantic inconsistencies, whereas Green texts signify that the generated results are closely aligned with the Ground Truth. [YELLOW] indicates the text problem.

(A) Anomaly Video Description Generation

| | 00:02 | 00:06 | 00:14 |

Prompt
Please help me to fully describe the details of this video, especially for the anomaly in the video.

Video-ChatGPT [29]:	[NOT FOCUS ON ANOMALY] The most interesting part of the video is a person walking on the snow-covered sidewalk, which is not a common sight on a snowy street.
VideoChat [18]:	The video captures a pedestrian crossing in a snowy town at night. The crossing features a white and red crossing sign, and there are some street lights on the road. [NOT FOCUS ON ANOMALY] A few people are walking on the sidewalk, while others are crossing the street.
Video-LLaMA [49]:	First, the video shows a city street with cars parked on both sides. There are also people walking on the sidewalk. Next, the view changes to [FOCUS ON PEOPLE'S CLOTHING] a man and a woman standing on a crosswalk, wearing blue and red jackets respectively.
LLaMA-Adapter [50]:	There is an anomaly in the video: a person appears to be walking on the street, but they are not visible in the blurred image. This person seems to be walking in the midst of the snowstorm, [HALLUCINATION] which is unusual as most people would be staying indoors during such weather conditions.
Video-LLaVA [20]:	The video shows a man walking across a street while a woman is lying on the ground. [UNCLEAR] The woman lying on the ground is not moving, and it is unclear what happened to her.
Ours:	In this video, the man and woman crossing the street at a crosswalk might be impacted by the anomaly. The woman could slip on the snowy sidewalk and be injured, which would be the anomaly in the video as they should be mindful of their surroundings.
Ground Truth:	In this anomaly situation, one of the people walking across the snowy crosswalk suddenly falls down probably due to a medical emergency, possibly suffering an injury. This could also lead to a chain reaction, with other pedestrians trying to avoid the fallen person and potentially slipping themselves.

(B) Anomaly Video Question-Answering

| | 00:01 | 00:02 | 00:04 |

Question
Why is the vehicle in the video moving erratically?

Video-ChatGPT [29]:	The other people affected by the anomaly in the video are the [HALLUCINATION] two people in the car who were hit by the car. They are seen lying on the ground, and the car is seen driving away.
VideoChat [18]:	There are [WRONG DESCRIPTION] no other people or vehicles shown in the video. The only thing that is shown in the video is the black car and the street and trees on either side of it.
Video-LLaMA [49]:	Since the video is not providing any additional context, it is not possible to determine if there are other people affected by the anomaly. It could be [NO IDENTIFYING] a minor problem that only affects the car, or it could be something more significant that affects other drivers or pedestrians on the road.
LLaMA-Adapter [50]:	[WRONG] The other people affected by the anomaly are the driver of the car and the motorcyclist.
Video-LLaVA [20]:	[HALLUCINATION] The driver of the car is seen driving with the hood open.
Ours:	The vehicle in the video is moving erratically likely due to the driver losing control. The speed and steering of the vehicle may have been affected by the uneven road conditions, causing the car to swerve unpredictably and almost hitting the white car. The abrupt maneuvers might be indicative of the driver's attempts to regain control or avoid collisions with other vehicles or obstacles on the road.
Ground Truth:	Based on the description you've provided, it seems the erratic movement of the vehicle in the video may be a consequence of the white car suddenly braking on an empty highway. This abrupt action could have led to the driver of the vehicle behind reacting in a panicked or abrupt manner to avoid a collision, resulting in erratic movements such as swerving or harsh braking.

Ablation Study   We conducted ablation experiments on three key structures proposed in this paper and analyzed their impact on the overall performance in Table 3 (A) and (B).

- **Effectiveness of Motion Information:** We ablate all the motion components, including f_m, P_m and the motion input $\mathbf{X_m}$ for proving the effectiveness of introducing motion modality.

9

Table 3: Ablation study of (A) anomaly video description generation and (B) video question-answering. Red indicates the best performance, while blue denotes the second best.

(A) Anomaly Video Description Generation

	Text-Level (↑) [30]				GPT-Guided (↑) [21]		
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Reasonability	Detail	Consistency
w/o Motion Information	0.249	0.121	0.062	0.034	0.253	0.306	0.189
w/o Video-Motion Consistency	0.249	0.123	0.064	0.036	0.261	0.295	0.194
w/o Motion-Language Matching Loss	0.270	0.135	0.073	0.041	0.276	0.320	0.212
Ours	0.270	0.139	0.074	0.043	0.283	0.320	0.218

(B) Anomaly Video Question-Answering

	Text-Level (↑) [30]				GPT-Guided (↑) [21]		
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Reasonability	Detail	Consistency
w/o Motion Information	0.309	0.171	0.105	0.065	0.837	0.790	0.743
w/o Video-Motion Consistency	0.313	0.172	0.105	0.066	0.833	0.784	0.742
w/o Motion-Language Matching Loss	0.316	0.176	0.109	0.069	0.836	0.788	0.752
Ours	0.319	0.179	0.112	0.073	0.840	0.794	0.753

Table 4: Qualitative evaluation of ablation study. Red/Green texts indicate language semantic inconsistency/consistency with the Ground Truth respectively. [YELLOW] indicates the text problem.

w/o Motion Information:	While driving down the [UNRELATED INFORMATION] snowy street, the vehicle suddenly skids on a patch of ice , causing it to veer into another car that was [HALLUCINATION] parked beside a tree on the side of the road.
w/o Video-Motion Consistency:	In this video, the car fails to stop in time due to the [WRONG CONDITION] slippery road caused by the snowy conditions. As a result, it skids and sideswipes the white car parked on the side of the road, causing minor damage to both vehicles.
w/o Motion-Language Matching:	As the white truck drives down the street, it suddenly skids on the wet road surface, losing control, and sideswiping the parked white car. The impact results in visible damage to both vehicles, [UNCLEAR CONDITION] with smoke emitting from the truck's side and the car's mirrors shattering.
Ours:	While driving down a narrow street with cars parked on both sides, the current vehicle's front right side scrapes against a parked car, causing minor damage to both vehicles.
Ground Truth:	While driving down the street, the silver car suddenly swerves to avoid a parked car, but clips its rear bumper, causing minor damage to both vehicles.

When explicit motion information is lacking, the model's ability to describe the motions-related anomaly diminishes, leading to inaccurate descriptions or even hallucinations (Table 4 w/o Motion Information), then impedes the overall performance (Table 3).

- **Effectiveness of Video-Motion Consistency:** The absence of video-motion consistency constraints reduces the generative model's ability to adapt to the motion modality, causing difficulties in accurately understanding motion scenes (Table 4 w/o Video-Motion Consistency), then impedes the overall performance (Table 3).

- **Effectiveness of Motion-Language Matching:** Without motion-language matching loss, the correlation between motion and language becomes unclear. This ambiguity leads to the generation of language that includes unspecified motion information (Table 4 w/o Motion-Language Matching), subsequently degrading the overall performance (Table 3).

6 Conclusion

In conclusion, we have developed a novel video-language framework for understanding video anomalies across various scenarios. By incorporating motion features and constructing rich linguistic descriptions, our model demonstrates SOTA performance in the open world. It has the potential to benefit practical applications in diverse domains and paves the way for improving the model's interactivity with users, enabling more efficient and effective communication in addressing user-specific inquiries related to video anomalies.

Acknowledgments and Disclosure of Funding

References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 4, 5, 8

[2] Acsintoae, A., Florescu, A., Georgescu, M.I., Mare, T., Sumedrea, P., Ionescu, R.T., Khan, F.S., Shah, M.: Ubnormal: New benchmark for supervised open-set video anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20143–20153 (2022) 2, 3, 16, 26

[3] Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: IEEE International Conference on Computer Vision (2021) 2, 3, 7, 15

[4] Brox, T., Malik, J.: Large displacement optical flow: descriptor matching in variational motion estimation. IEEE transactions on pattern analysis and machine intelligence **33**(3), 500–513 (2010) 19

[5] Cadiot, P., Lebas, F., Visetti, Y.M.: The semantics of the motion verbs. Space in Languages: Linguistic Systems and Cognitive Categories **66**, 175 (2006) 7

[6] Chan, A.B., Vasconcelos, N.: Modeling, clustering, and segmenting video with mixtures of dynamic textures. IEEE transactions on pattern analysis and machine intelligence **30**(5), 909–926 (2008) 1, 2, 3, 16, 27

[7] Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2758–2766 (2015) 19

[8] Du, H., Zhang, S., Xie, B., Nan, G., Zhang, J., Xu, J., Liu, H., Leng, S., Liu, J., Fan, H., Huang, D., Feng, J., Chen, L., Zhang, C., Li, X., Zhang, H., Chen, J., Cui, Q., Tao, X.: Uncovering what, why and how: A comprehensive benchmark for causation understanding of video anomaly. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 21

[9] Dubey, S., Boragule, A., Jeon, M.: 3d resnet with ranking loss function for abnormal activity detection in videos. In: 2019 International Conference on Control, Automation and Information Sciences (ICCAIS). pp. 1–6. IEEE (2019) 2, 3

[10] Fang, Y., Wang, W., Xie, B., Sun, Q.S., Wu, L.Y., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 19358–19369 (2022) 5

[11] Farneback, G.: Fast and accurate motion estimation using orientation tensors and parametric motion models. In: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. vol. 1, pp. 135–139. IEEE (2000) 6

[12] Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13. pp. 363–370. Springer (2003) 19

[13] He, C., Shao, J., Sun, J.: An anomaly-introduced learning method for abnormal event detection. Multimedia Tools and Applications **77**, 29573–29588 (2018) 2, 3

[14] Huang, X., Zhang, Y., Ma, J., Tian, W., Feng, R., Zhang, Y., Li, Y., Guo, Y., Zhang, L.: Tag2text: Guiding vision-language model via image tagging. arXiv preprint arXiv:2303.05657 (2023) 4

[15] Langacker, R.W.: Nouns and verbs. Language pp. 53–94 (1987) 7

[16] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023) 3

[17] Li, J., Li, D., Savarese, S., Hoi, S.C.H.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning (2023) 5

[18] Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023) 2, 3, 4, 5, 7, 8, 9, 22, 23, 24, 25, 26, 27

[19] Li, S., Liu, F., Jiao, L.: Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1395–1403 (2022) 2, 3

[20] Lin, B., Zhu, B., Ye, Y., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122 (2023) 5, 7, 8, 9, 22, 23, 24, 25, 26, 27

[21] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems 36 (2024) 2, 3, 4, 8, 10

[22] Liu, W., W. Luo, D.L., Gao, S.: Future frame prediction for anomaly detection – a new baseline. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 2, 3, 15, 25

[23] Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 fps in matlab. In: Proceedings of the IEEE international conference on computer vision. pp. 2720–2727 (2013) 1, 2, 3, 15, 23

[24] Lu, H., Niu, X., Wang, J., Wang, Y., Hu, Q., Tang, J., Zhang, Y., Yuan, K., Huang, B., Yu, Z., et al.: Gpt as psychologist? preliminary evaluations for gpt-4v on visual affective computing. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) workshop (2024) 3

[25] Lu, H., Tang, J., Xu, X., Cao, X., Zhang, Y., Wang, G., Du, D., Chen, H., Chen, Y.: Scaling multi-camera 3d object detection through weak-to-strong eliciting. arXiv (2024) 1

[26] Luo, R., Zhao, Z., Yang, M., Dong, J., Qiu, M., Lu, P., Wang, T., Wei, Z.: Valley: Video assistant with large language model enhanced ability. arXiv preprint arXiv:2306.07207 (2023) 3

[27] Lv, H., Sun, Q.: Video anomaly detection and explanation via large language models. arXiv preprint arXiv:2401.05702 (2024) 2, 3, 5

[28] Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023) 3

[29] Muhammad Maaz, Hanoona Rasheed, S.K., Khan, F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. ArXiv 2306.05424 (2023) 2, 5, 7, 8, 9, 22, 23, 24, 25, 26, 27

[30] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: 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) 8, 10

[31] Pu, Y., Wu, X., Wang, S.: Learning prompt-enhanced context features for weakly-supervised video anomaly detection. arXiv preprint arXiv:2306.14451 (2023) 2, 3

[32] Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: Pandagpt: One model to instruction-follow them all. arXiv preprint arXiv:2305.16355 (2023) 3

[33] Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6479–6488 (2018) 1, 2, 3, 15, 24

[34] Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4975–4986 (2021) 2, 3

[35] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 5

[36] Vo, N.P.A., Manotas, I., Sheinin, V., Popescu, O.: Identifying motion entities in natural language and a case study for named entity recognition. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 5250–5258 (2020) 7

[37] Wang, S., Miao, Z.: Anomaly detection in crowd scene. In: IEEE 10th International Conference on Signal Processing Proceedings. pp. 1220–1223. IEEE (2010) 2, 3, 16

[38] Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al.: Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022) 4

[39] Wu, J., Wang, J., Yang, Z., Gan, Z., Liu, Z., Yuan, J., Wang, L.: Grit: A generative region-to-text transformer for object understanding. arXiv preprint arXiv:2212.00280 (2022) 4

[40] Wu, P., Liu, J.: Learning causal temporal relation and feature discrimination for anomaly detection. IEEE Transactions on Image Processing **30**, 3513–3527 (2021) 2, 3

[41] Wu, P., Liu, j., Shi, Y., Sun, Y., Shao, F., Wu, Z., Yang, Z.: Not only look, but also listen: Learning multimodal violence detection under weak supervision. In: European Conference on Computer Vision (ECCV) (2020) 20

[42] Wu, P., Zhou, X., Pang, G., Sun, Y., Liu, J., Wang, P., Zhang, Y.: Open-vocabulary video anomaly detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 2, 3

[43] Wunderlich, D.: Cause and the structure of verbs. Linguistic inquiry pp. 27–68 (1997) 7

[44] Xu, D., Ricci, E., Yan, Y., Song, J., Sebe, N.: Learning deep representations of appearance and motion for anomalous event detection. arXiv preprint arXiv:1510.01553 (2015) 2, 3, 5

[45] Yao, Y., Wang, X., Xu, M., Pu, Z., Wang, Y., Atkins, E., Crandall, D.J.: Dota: unsupervised detection of traffic anomaly in driving videos. IEEE transactions on pattern analysis and machine intelligence **45**(1), 444–459 (2022) 1, 2, 3, 16, 22

[46] Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023) 3

[47] Yuan, T., Zhang, X., Liu, K., Liu, B., Chen, C., Jin, J., Jiao, Z.: Towards surveillance video-and-language understanding: New dataset, baselines, and challenges (2023) 4

[48] Zaheer, M.Z., Mahmood, A., Astrid, M., Lee, S.I.: Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. pp. 358–376. Springer (2020) 2, 3

[49] Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023) 3, 5, 7, 8, 9, 22, 23, 24, 25, 26, 27

[50] Zhang, R., Han, J., Liu, C., Gao, P., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Qiao, Y.: Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199 (2023) 7, 8, 9, 22, 23, 24, 25, 26, 27

[51] Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In: The Twelfth International Conference on Learning Representations (2024) 3

[52] Zhu, Y., Newsam, S.: Motion-aware feature for improved video anomaly detection. arXiv preprint arXiv:1907.10211 (2019) 2, 3, 5

A Summary of Appendix

This appendix provides supplementary information that was not included in the main paper. Firstly, we address the security statement of our study, ensuring the confidentiality and integrity of the data used. Additionally, we provide detailed explanations of the training and testing resources utilized, including information on the hardware and software configurations. We also present statistics and distribution of the training data, along with the costs associated with human resources involved in the study. Furthermore, we describe the evaluation metrics employed to assess the performance of our method. Moreover, we present additional qualitative results comparisons, showcasing the effectiveness of our approach. Additionally, we provide an open-world demo, demonstrating the real-world applicability of our method. Finally, we discuss the existing limitations of our paper and propose potential avenues for future research.

B Security Statement

To prevent any potential misuse and ensure responsible use, we have strictly limited the application scope of our proposed method, **HAWK**. Unless authorized, **HAWK** is only permitted for use in research domains.

Additionally, access to the proposed dataset is restricted to qualified institutions and organizations, who must provide a clear purpose for its use. We explicitly prohibit the application of the dataset in situations that may cause potential danger or have a significant social impact.

These measures are in place to ensure the ethical and responsible use of our research.

C Details in Training and Testing

Computational Resource   During the pre-training phase, we utilized four Nvidia RTX A6000 GPUs[*] to train on the WebVid dataset [3] for approximately 120 hours. In the fine-tuning phase, we employed two Nvidia RTX A6000 GPUs to fine-tune on our proposed dataset for about 80 hours.

Efficiency   During testing, the average model response time for each round of conversation with **HAWK** is approximately 2ms. Additionally, considering the available graphics memory, the model can handle video clips of up to 32 frames. Therefore, it is necessary to extract different frames from longer videos.

Hyper-parameters   In the loss function, t_0 is set to 1 for our main task, video-to-language, and t_1 and t_2 are set to 0.1, as two auxiliary tasks for balancing different loss values.

D Details in Dataset

Dataset Introduction and Statistics   Our study utilizes seven video anomaly datasets, each encompassing different scenes. The detailed statistics and introduction of these datasets are as follows:

- UCF-Cirme [33]: The UCF-Crime dataset comprises an extensive collection of 128 hours of video. It consists of 1,900 long and untrimmed real-world surveillance videos, featuring 13 distinct classes of realistic anomalies. These anomalies are carefully chosen due to their notable implications for public safety.
- ShanghaiTech [22]: The ShanghaiTech Campus dataset comprises 13 scenes characterized by complex light conditions and varied camera angles. It encompasses 130 instances of abnormal events and encompasses over 270,000 training frames. Notably, this dataset includes annotations for both frame-level and pixel-level ground truth of abnormal events, providing comprehensive insight into anomaly detection and localization tasks.
- CUHK Avenue [23]: The CUHK Avenue Dataset comprises 16 training and 21 testing video clips designed for abnormal event detection. Captured within the CUHK campus avenue,

---

[*] https://www.nvidia.com/en-us/design-visualization/rtx-a6000/

these videos encompass a total of 30,652 frames, divided into 15,328 frames for training and 15,324 frames for testing. The training videos capture normal situations, while the testing videos include both normal and abnormal events.

- UCSD Dataset [6, 37]: The UCSD Anomaly Detection Dataset was captured using a stationary camera positioned at an elevation, providing an overhead view of pedestrian walkways. The crowd density within these walkways exhibits variability, spanning from sparsely populated areas to densely crowded environments. It is split into 2 subsets, each corresponding to a different scene. Ped1 [6] includes a total of 34 training video samples and 36 testing video samples, while Ped2 [37] consists of 16 training video samples and 12 testing video samples.

- DoTA [45]: The Detection of Traffic Anomaly (DOTA) Dataset introduces the When-Where-What pipeline with temporal, spatial, and categorical annotations. It contains 4677 videos, all with a resolution of 1280 x 720 pixels. Notably, the original videos were extracted at a frame rate of 10 fps in this dataset.

- UBnormal [2]: The UBnormal dataset is a supervised open-set benchmark designed explicitly for video anomaly detection, comprising diverse virtual scenes. It introduces abnormal events annotated at the pixel level during training, which enables the utilization of fully-supervised learning techniques for abnormal event detection.

In our study, we extend upon these existing datasets by implementing our data engineering pipeline. This pipeline generates comprehensive descriptions of video anomalies and formulates open questions derived from these anomalies.

Data Distribution   To demonstrate the applicability of our data in an open-world scenario, we conducted a statistical analysis of the data distribution. Figure 6 illustrates the data distribution of all the datasets we utilized, indicating that our method can effectively support various open-world datasets. Besides, we acknowledge the need to expand our dataset further to enhance the model's applicability in this task.
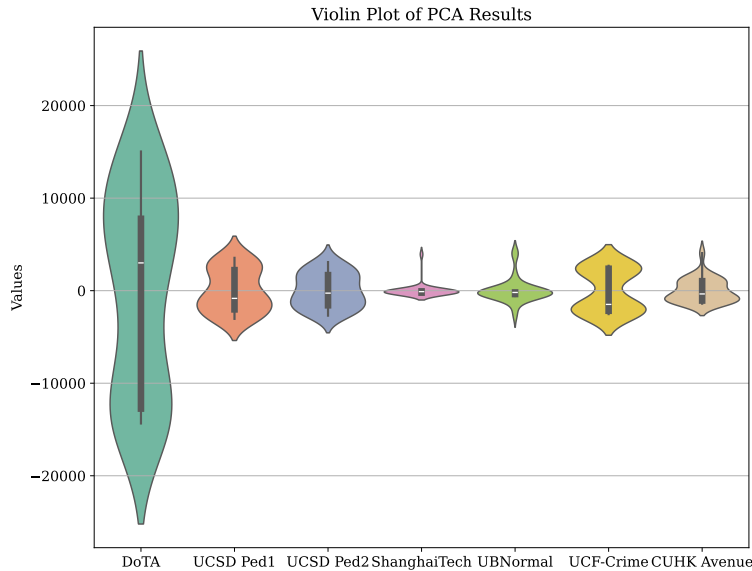

Figure 6: Violin plot of data distribution. We use PCA dimensional reduction to measure the feature distribution of different datasets, where there are significant differences in the feature distribution.

Manual Checking   Before conducting the experiments, we manually checked the textual descriptions generated for the videos. Specifically, we consider the following aspects:

16

1. Error Correction: We removed text descriptions that contained obvious errors about the video content and supplemented the correct object, behavior, and scene information. (For instance, GPT tends to misidentify dogs in videos, describe running pedestrians as skateboards and motorcycles, and mistake scenes containing water as rainy days.)

2. Detail Enhancement: We provided more detailed textual descriptions of anomalies in the video (such as pedestrians lingering or jumping in the middle of the road).

3. Human Resource Cost: We formed a team of five annotators to conduct Manual Checking on all the videos. Since most of the videos already had automatically generated annotations, each annotator invested approximately 30 hours of work during the labeling process, processing about 1700 videos.

Table 5 below provides an example of before and after manual checking.

Table 5: An Example of Manual Checking.

Before Manual Checking (GPT-4)	After Manual Checking
The video depicts different scenes of people walking and sitting in front of a building and a train station. There are also shots of a woman with a red bag sitting on the ground, a man walking by with a shopping bag, and a young boy in a red jacket standing outside a building. There are also several shots of people standing at a bus stop, and a sign with a glowing red hand.	The video shows scenes in front of the subway entrance, with students and pedestrians coming and going or entering and exiting the subway. In addition, two school security guards patrol through, and at the subway entrance, two students are chatting and discussing.

In this example, the caption generated by GPT-4 included hallucinations (such as "a man walking by with a shopping bag" or "a young boy in a red jacket"), which were corrected after Manual Checking. Additionally, GPT-4's description was inaccurate (for instance, "bus stop" should have been more accurately described as "subway entrance").

<DESCRIBE_VIDEO> and Generated Open-World <QUESTION> We set 20 problems for <DESCRIBE_VIDEO>, and during each iteration in training, we randomly select one of them.

```
1.      Can you describe the anomaly in the video?
2.      How would you detail the anomaly found in the video?
3.      What anomaly can you identify in the video?
4.      Could you explain the anomaly observed in the video?
5.      Can you point out the anomaly in the video?
6.      What's the anomaly depicted in the video?
7.      Could you specify the anomaly present in the video?
8.      How do you perceive the anomaly in the video?
9.      Can you highlight the anomaly within the video?
10.      What anomaly is noticeable in the video?
11.      Could you characterize the anomaly seen in the video?
12.      Can you detail the specific anomaly encountered in the video?
13.      How would you describe the particular anomaly in the video?
14.      What details can you provide about the anomaly in the video?
15.      Could you elucidate on the anomaly detected in the video?
16.      Can you illustrate the nature of the anomaly in the video?
17.      What features of the anomaly in the video can you describe?
18.      Could you outline the anomaly observed in the video?
19.      How does the anomaly in the video manifest?
20.      Can you clarify the aspects of the anomaly in the video?
```

We have also generated 100 <QUESTIONS> for open-world anomalies. To mimic user behavior, some of these questions are closely related to the video scene, while others are less closely related. However, all of these questions are potential inquiries in an open-world scenario.

```
1.      Who is causing the disturbance in the video?
2.      What is the unusual activity happening in the video?
3.      When did the anomaly occur in the video?
4.      Where is the strange event taking place in the video?
5.      Why is the object in the video behaving abnormally?
6.      How is the anomaly in the video affecting the surroundings?
7.      How much damage was caused by the incident in the video?
8.      Who is the main person involved in the unusual event?
9.      What is the cause of the sudden change in the video?
```

10. When does the suspicious activity start in the video?
11. Where can I find more information about the incident in the video?
12. Why are the people in the video reacting in that way?
13. How can I identify the source of the problem in the video?
14. How much time does the abnormal event last in the video?
15. Who are the other people affected by the anomaly in the video?
16. What actions were taken to address the issue in the video?
17. When was the video recorded, and is it a recent event?
18. Where else can I find similar incidents in other videos?
19. Why is the vehicle in the video moving erratically?
20. How can I prevent such anomalies from occurring in the future?
21. How much impact does the abnormal event have on the overall situation?
22. Who should I contact if I notice a similar anomaly in another video?
23. What steps can I take to investigate the issue further?
24. When is the best time to report an unusual event in a video?
25. Where can I find resources to help me understand the anomaly better?
26. Why did the equipment in the video malfunction?
27. How can I differentiate between normal and abnormal behavior in a video?
28. How much does it cost to implement a system that detects anomalies in videos?
29. Who can provide expert advice on handling video anomalies?
30. What is the most common type of anomaly found in videos?
31. When should I be concerned about an anomaly in a video?
32. Where can I find a list of known video anomalies and their descriptions?
33. Why is it important to detect and analyze anomalies in videos?
34. How can I improve my ability to spot anomalies in videos?
35. How much training is required to become proficient in detecting video anomalies?
36. Who can I collaborate with to better understand video anomalies?
37. What are the potential consequences of ignoring an anomaly in a video?
38. When did the trend of analyzing anomalies in videos begin?
39. Where can I find examples of successfully resolved video anomaly cases?
40. Why do some anomalies in videos go unnoticed?
41. How can I report a video anomaly to the appropriate authorities?
42. How much time is needed to thoroughly analyze a video anomaly?
43. Who is responsible for monitoring and addressing video anomalies?
44. What are the best tools to use for detecting anomalies in videos?
45. When is it necessary to escalate a video anomaly for further investigation?
46. Where can I find guidelines on how to handle video anomalies?
47. Why do some video anomalies lead to serious consequences?
48. How can I ensure the accuracy of my video anomaly detection system?
49. How much effort is needed to maintain a video anomaly detection system?
50. Who should be informed when a video anomaly is detected?
51. What are the signs that indicate a potential anomaly in a video?
52. When should I perform a follow-up analysis on a detected video anomaly?
53. Where can I find support for dealing with video anomalies?
54. Why is it crucial to act quickly when a video anomaly is detected?
55. How can I improve the efficiency of my video anomaly detection process?
56. How much data is needed to accurately detect anomalies in videos?
57. Who can help me fine-tune my video anomaly detection system?
58. What are the key factors to consider when analyzing video anomalies?
59. When should I update my video anomaly detection system?
60. Where can I find the latest research on video anomaly detection techniques?
61. Why is it necessary to have a video anomaly detection system in place?
62. How can I minimize false alarms in my video anomaly detection system?
63. How much does it cost to maintain a video anomaly detection system?
64. Who can I consult if I encounter difficulties with my video anomaly detection system?
65. What are the best practices for dealing with video anomalies?
66. When is it appropriate to involve law enforcement in a video anomaly case?
67. Where can I find a community of professionals who specialize in video anomaly detection?
68. Why do some video anomalies require immediate attention?
69. How can I enhance the performance of my video anomaly detection system?
70. How much should I invest in a video anomaly detection system?
71. Who can provide training on how to detect and analyze video anomalies?
72. What are the most effective methods for detecting anomalies in videos?
73. When should I seek external help for a video anomaly case?
74. Where can I find a comprehensive database of video anomalies?
75. Why is it important to continuously monitor videos for anomalies?
76. How can I validate the results of my video anomaly detection system?
77. How much influence do external factors have on video anomalies?
78. Who can I reach out to for assistance with a complex video anomaly case?
79. What are the main challenges in detecting and analyzing video anomalies?
80. When is it necessary to involve other stakeholders in a video anomaly case?
81. Where can I find case studies on successful video anomaly detection projects?
82. Why is it essential to have a systematic approach to video anomaly detection?
83. How can I optimize my video anomaly detection system for different scenarios?
84. How much storage is needed to archive video anomalies for future analysis?
85. Who should be held accountable for undetected video anomalies?
86. What are the most common reasons for video anomalies to occur?
87. When should I reevaluate my video anomaly detection system?
88. Where can I find information on the latest video anomaly detection technologies?
89. Why is it beneficial to collaborate with others in the field of video anomaly detection?
90. How can I ensure the confidentiality of video anomaly cases?

```
91.      How much should I rely on automated systems for video anomaly detection?
92.      Who can I contact for technical support with my video anomaly detection system?
93.      What are the ethical considerations when dealing with video anomalies?
94.      When should I notify the public about a video anomaly case?
95.      Where can I find reliable sources of information on video anomalies?
96.      Why is it important to have a backup plan for dealing with video anomalies?
97.      How can I customize my video anomaly detection system for specific use cases?
98.      How much time should I allocate for analyzing video anomalies?
99.      Who can I turn to for guidance on handling sensitive video anomaly cases?
100.      What are the most critical factors to consider when choosing a video anomaly detection system?
```

Demonstrating that Our model outperforms GPT-4-based Data Generation.   To demonstrate that our model has better detection capabilities than GPT-4, we compared our model's results with the unchecked labels generated by GPT-4 on the same testset, as shown in the following Table 6.

Table 6: Comparison with GPT-based Data Engineering Pipeline.

Methods	Backbones	LLM Size	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Reasonability	Detail	Consistency
Data Engineering Pipeline	GPT-4 [1]	-	0.188	0.098	0.056	0.034	0.189	0.313	0.158
Ours	LLaMA-2 7B	7B	**0.270**	**0.139**	**0.074**	**0.043**	**0.283**	**0.320**	**0.218**

Clearly, the results indicate that our model can better assist in understanding video anomalies compared to GPT-4, achieving state-of-the-art (SOTA) performance in both text-level and GPT-guided evaluations.

E   Efficiency of Gunnar Farneback's Algorithm for Motion Modality Extraction

Firstly, Gunnar Farneback's algorithm demonstrates remarkable efficiency in generating video optical flows—even on CPU platforms. For each frame, the efficiency of this algorithm surpasses that of other widely deployed methods, as illustrated in the table below:

Table 7: Performance Comparison of Different Optical Flow Methods.

Methods	Gunnar Farneback [12]	LDOF [4]	FlownetS [7]	FlownetC [7]
Seconds Per Frame	**0.02 (CPU)**	65 (CPU) 2.5 (GPU)	0.08 (GPU)	0.15 (GPU)

Secondly, for processing one video, the motion of multiple rounds of dialogue necessitates just a single iteration of video motion extraction. Consequently, the response time for processing one video (about 0.72 seconds) is significantly shorter than the time required to generate a single round of dialogue (about 1.5 seconds). This cost is deemed acceptable for a practical anomaly detection system for users.

Certainly, we concur that future research into more efficient methods for optical flow extraction, including end-to-end optical flow extraction strategies, will likely further augment the efficiency of our system.

F   Details in GPT-Guided Metrics

In the GPT-Guided metrics, we employ GPT-4 as an auxiliary tool to evaluate the generated response of **HAWK**. Our evaluation focuses on three primary dimensions: Reasonability, Detail, and Consistency.

We first set the system prompt as follows: Initially, we establish the system prompt as shown below:

```
{"role": "system", "content":
"You are an intelligent chatbot designed for evaluating the generative
↪    outputs for video-based pairs. you will be given two answers, one
↪    reference ground truth and one our generated, but this does not mean
↪    that the reference GT is the only answer. Your task is to give the
↪    score of the predicted answers."}
```

Our system prompt is designed to compare the degree of matching between image pairs. However, this does not imply fine-grained matching at the text level. Instead, it emphasizes the semantic information-related aspects.

To assess a particular dimension of the metric, we employ the following prompt:

```
{"role": "user", "content":
"### Video Description Generation
Please evaluate the following video-based video description pair:
Reference: <DESCRIPTION_GT>
Ours: <DESCRIPTION_Ours>

### Video Question-Answering
Please evaluate the following video-based video question-answer pair:
Question: <QUESTION>
Reference: <ANSWER_GT>
Ours: <ANSWER_Ours>

Provide your evaluation only as a <Reasonability|Detail|Consistency> score
↪   where the <Reasonability|Detail|Consistency> score is a FLOAT value
↪   between 0 and 1, with 1 indicating the highest level of
↪   <Reasonability|Detail|Consistency>. Please generate the response in the
↪   form of a Python dictionary string with key 'score', where its value is
↪   the <Reasonability|Detail|Consistency> score in FLOAT, not STRING. DO
↪   NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the
↪   Python dictionary string. For example, your response should look like
↪   this: {'score': 0.675}."}
```

We have developed distinct prompts for two tasks: Video Description Generation and Video Question-Answering. The primary difference is the addition of the <QUESTION> field in Video Question-Answering, which indicates what kind of question the model should answer. <DESCRIPTION_GT> and <DESCRIPTION_Ours> represent the Ground Truth and our generated video description, respectively. Similarly, <ANSWER_GT> and <ANSWER_Ours> signify the Ground Truth and our generated video answers, respectively. < Reasonability | Detail | Consistency > represents the three dimensions we aim to evaluate. Lastly, besides the essential reminders, we have constrained GPT's output format to {'score': 0.675}.

G   More Results

Table (A), (B), (C), (D), (E), and (F) below present additional qualitative results from different datasets. In the tables, red texts indicate key semantic inconsistencies with the Ground Truth, while green texts signify that the generated results closely align with the Ground Truth.

H   Open-World Video Anomaly Understanding Demo

We present a demo showcasing the use of **HAWK** in an open-world scenario, using XD-Violence [41] (which is not included in our dataset). The practical capability of the system in an unknown scenario in the open world is depicted in Fig.7 and Fig.8. Furthermore, **HAWK** can provide accurate answers to users' questions and engage in long dialogues in the open world.

I   Limitations

Hallucination   Although most of the hallucinations can be decreased through motion, some error motion may still also cause hallucinations. Future work may need to consider the connection between the hallucination and the abnormal region more precisely.

Background Information   Leveraging background information can provide a robust prior for understanding video anomalies. We will try to integrate information related to scenes, backgrounds, and objects in a large-model-based video anomaly understanding model.

Video-level v.s. Streaming Data   The goal of this paper is video-level video anomaly understanding. However, for a video anomaly detection system, anomaly detection in streaming is essential, so to increase the practical application ability, we need to design a more practical system for streaming data.

Data Limitations   While our dataset includes multiple anomaly scenarios and our framework is designed for an open-world setting, the limitations of our data make it difficult to fully support open-world scenarios. This is a significant drawback of our study. To address this limitation, we recommend building larger and more diverse open datasets.

Keyframes   First, our current methodology for dataset construction involves sampling the video at consistent one-second intervals. This technique is strategically chosen to ensure that all possible anomalies within the videos are comprehensively captured (Some anomalies only happened in 1-2 seconds), thereby significantly reducing the likelihood of missing critical accidents. While we realize that this may lead to a degree of caption redundancy, we still prioritize the facilitation of thorough annotation to ensure that all anomalies are detected.

In addition, we have leveraged the capabilities of GPT-4 for generating captions, especially for anomalous events. Due to GPT-4's advanced text generation and summarization abilities, it serves as an effective tool in minimizing redundancy, ensuring that the extracted captions are both high-quality and succinct.

After the initial processing with GPT-4, we also undertake a manual checking process. This step is crucial for further reducing any residual redundancy and correcting possible errors within the captions, thereby ensuring the quality and accuracy of our dataset.

Certainly, we agree that the utility of keyframes is an effective strategy, especially for much longer videos, and believe its potential to significantly enhance data annotation efficiency. This will be the future work.

J   Future Work

In this section, we discuss potential avenues for future research to build upon this paper.

Expanding Applicability   One possible direction for future work is to expand the applicability of the model to a wider range of scenarios (like most recent research by Du et al. [8]). While our current dataset includes multiple anomaly scenarios, there are still many diverse scenarios the model may not handle. We could explore ways to incorporate additional data sources or create new datasets that cover a broader range of scenarios. Additionally, we could investigate ways to deploy the model in cloud, edge, and hybrid computing environments to support more diverse deployment scenarios.

Enriching Task Capabilities   While our current model is able to detect anomalies in videos, it does not provide location information where the anomaly occurs. This information could be valuable for certain applications, such as security monitoring or surveillance. Therefore, a valuable direction for future work would be to investigate ways to improve the model's ability to localize anomalies. By improving the model's ability to locate the scene of an anomaly, we could further enhance its usability and expand its range of applications.

(A) Anomaly Video Description Generation in DoTA [45].

Prompt
Please help me to fully describe the details of this video, especially for the anomaly in the video.

Video-ChatGPT [29]:	The video shows a car driving down a busy street, with a person sitting in the passenger seat. The car is surrounded by other cars, and the driver is seen looking at the camera. The car is seen turning a corner, and the person sitting in the passenger seat is looking at the camera. The video also shows a person walking in the street, and a truck driving by.
VideoChat [18]:	The video is a car driving on the road with the camera mounted inside the car. It is a clear day with some clouds in the sky. There are cars and other vehicles on the road, with one vehicle parked on the side of the road. The road is lined with trees on both sides. The car is driving along the road, passing other vehicles, and crossing roads. The anomaly in the video occurs when the car suddenly stops in front of a bus, blocking the road.
Video-LLaMA [49]:	The car is driving down the street, and there are buildings and cars parked along the side of the road. Finally, the car is seen driving down the street at night. The camera is shown outside the car, and there are white lights on the road. There are trees and buildings in the distance, and the cars on the road are all different colors. The video ends with the car turning a corner and driving down the street at night.
LLaMA-Adapter [50]:	The video shows a car driving down a road, with a person inside the vehicle. Suddenly, the car appears to be hit by a large object, causing it to flip over and crash. The person inside the car is thrown out, and the car is left in a wrecked state on the road. The scene is captured from a dash cam perspective, providing a clear view of the incident.
Video-LLaVA [20]:	The video starts with a car driving down a street, and the camera captures the view from inside the car. The car is seen driving past a blue truck and a red car. The camera then pans to the side of the road, where a man is seen walking down the sidewalk. The camera then pans back to the street, where a blue truck is seen driving down the road.
Ours:	While driving down the street, the current vehicle is involved in an accident when a sudden stop causes it to rear-end the white car in front of it, resulting in minor damage to both vehicles and a startled reaction from the occupants.
Ground Truth:	As the car continues down the busy tree-lined street, the traffic light ahead suddenly turns red. The white car in front of the vehicle abruptly comes to a stop, causing the driver recording the video to brake hard but ultimately colliding into the rear of the white car with a loud impact.

(B) Anomaly Video Description Generation in CUHK Avenue [23].

00:00   00:02   00:04

Prompt
Please help me to fully describe the details of this video, especially for the anomaly in the video.

Video-ChatGPT [29]:	The video shows a man sitting on a bench in a busy city street, surrounded by many people. Suddenly, a dog appears and starts walking towards the man. The dog is wearing a red collar and is accompanied by a person. The dog is walking towards the man, who is sitting on the bench, and the man is watching the dog as it approaches. The dog is walking towards the man, and it appears to be a unique and unexpected event in the busy city street.
VideoChat [18]:	The video shows a group of students walking outside the entrance to an office building located in a park. There are two white benches in front of the entrance, one on the left side and one on the right side. The entrance has a large blue sign above it that reads Science and Engineering. People are standing and walking around the entrance while the students are walking towards the building.
Video-LLaMA [49]:	First, the video shows a busy street with many people walking around. There is a sign for a bus stop at one corner. Next, the video focuses on a line of people standing in front of a blue building, with a man in a black shirt and black jacket walking towards them. Finally, the video shows a crowd of people at the bus stop, with people walking to and from the bus. There is a green park bench and a building visible behind the people. The street is also filled with trees and a black car is visible in the background.
LLaMA-Adapter [50]:	The video shows a group of people walking down a hallway, with some of them carrying backpacks. There are at least 12 people in the scene, with some walking in pairs or small groups. The hallway is long and has a few benches along the way. One of the people in the group appears to be wearing a tie, which stands out as an anomaly in the otherwise casual attire of the other individuals.
Video-LLaVA [20]:	In the video, a man is seen standing in a hallway while talking on his cell phone. He is wearing a red shirt and is standing in front of a group of people who are walking by. The man's red shirt stands out against the crowd, making him easily noticeable. The man's behavior is quite unusual, as he is not paying attention to his surroundings while talking on the phone. It is unclear what the man is doing or why he is not paying attention to the people around him.
Ours:	The video is set at a subway entrance near campus, with pedestrians entering and exiting the area from different directions. A crowd of students, teachers, and tourists is standing at the subway entrance, watching a performance by a street musician. They are throwing coins and applauding. Other students and teachers are walking and talking near the subway entrance.
Ground Truth:	The video is set at a subway entrance where some pedestrians are hurriedly entering and exiting the subway or passing through nearby pathways. Another group of people is captivated by a performance and stands at the subway entrance to watch.

(C) Anomaly Video Description Generation in UCF-Crime [33].

01:06   01:15   01:36

Prompt
Please help me to fully describe the details of this video, especially for the anomaly in the video.

Video-ChatGPT [29]:	The video shows a group of people riding motorcycles and a car in a city street. The video also shows a man walking in the street, and a man riding a motorcycle. The video is taken at night, and the street is dimly lit. The anomaly in the video is a person riding a motorcycle who is not wearing a helmet. This is a safety concern as it is illegal and can lead to severe injuries in case of an accident.
VideoChat [18]:	This video appears to be a surveillance video of a residential street at night in India. The video shows a street with some motorcycles parked along the side of the road and a few people walking around. However, the two anomalies of the man riding the motorcycle and the man walking down the street make it stand out.
Video-LLaMA [49]:	First, the video shows a street in an Indian city at night. There are a few buildings on the street, and people are walking around. Next, there is a close-up of a street at night with cars parked on it. The street is narrow, and there are many buildings on both sides. Then, the video shows a street at night with a motorcycle parked on it. There is also a bicycle parked on the sidewalk and a white car parked next to a bike.
LLaMA-Adapter [50]:	The video shows a man walking down a street at night, with a car parked nearby. Suddenly, the scene shifts to a different location, where a woman is walking down the street. The man and the woman appear to be walking in opposite directions, and the man is carrying a backpack. The video ends with the man walking down the street, still carrying the backpack.
Video-LLaVA [20]:	The video shows a person riding a motorcycle down a narrow alley at night. The person is wearing a white shirt and black pants. The person rides the motorcycle quickly, and the camera follows them until they reach the end of the alley. However, there is an anomaly in the video. The person on the motorcycle is not visible in the video, and the camera only shows the motorcycle and the surrounding environment.
Ours:	The video shows what appears to be a hit-and-run incident or an attempted theft. A man in blue pants and shorts is shown walking along the side of the road, where he is subsequently struck by a motorcycle traveling from behind. The man, possibly injured, then falls to the ground, which could be indicative of either the collision itself or a previous event that has left him unconscious but not necessarily injured.
Ground Truth:	The video portrays routine urban activity that people and vehicles moving along a street. The abnormal part unfolds when a motorcycle stops next to a pedestrian, leading to a confrontation. The altercation becomes physical when the person who dismounted from the motorcycle starts arguing with the passerby, eventually escalating as the passerby is pushed to the ground. Following the altercation, the individuals involved in the aggressive act hastily escape the scene on the motorcycle, which adds a hit-and-run element to the incident, indicating a consciousness of guilt or an attempt to evade law enforcement.

(D) Anomaly Video Question-Answering in ShanghaiTech [22].

Question

Who are the people affected by the anomaly in the video?

|	
Video-ChatGPT [29]:	The other people affected by the anomaly in the video are the people walking on the sidewalk. They are walking in a line, and suddenly, a person starts running in the opposite direction. The person running is wearing a red shirt.
VideoChat [18]:	It is possible that there may be other people in the video who cross the yellow line, although they are not visible in the frame at that point in the video. The pedestrian who crosses the traffic line appears to be walking away from the camera and appears to be aware of the traffic line, which suggests that they are not necessarily unaware of the rules of the road.
Video-LLaMA [49]:	It only talks about a group of people walking down the street. It does not mention any specific people being affected by the anomaly.
LLaMA-Adapter [50]:	The other people affected by the anomaly in the video are the people walking on the sidewalk, who are unaware of the man's disappearance.
Video-LLaVA [20]:	The video shows that the other people are also affected by the anomaly. They are seen running and jumping over the obstacle in the video.
Ours:	The anomaly in the video that I described involves a man walking on a pedestrian walkway while other people are nearby, which could potentially cause a collision or accident. Additionally, other pedestrians or passersby in the vicinity may also be affected by the situation, potentially being startled or distracted by the incident. Additionally, other pedestrians or passersby in the vicinity may also be affected by the situation, potentially being startled or distracted by the incident.
Ground Truth:	The other people affected by the anomaly in the video would include: 1. The pedestrian who was collided with by the running man, as they might sustain injuries from the impact. 2. Bystanders or other individuals in the park who may witness the incident and could potentially be involved in the ensuing chaos or need to provide assistance. 3. Any friends or family members of the individuals directly involved in the collision who might be present or affected indirectly by the consequences of the incident.

(E) Anomaly Video Question-Answering in UBNormal [2].

00:01   00:02   00:04

Question

How much impact does the abnormal event have on the overall situation?

Video-ChatGPT [29]:	The abnormal event of a person falling on the ground in the middle of the train station has a significant impact on the overall situation. It disrupts the flow of people walking around the station, as they are forced to stop and look at the person lying on the ground. This creates a sense of confusion and concern among the people, as they may not know what to do or how to help the person.
VideoChat [18]:	Based on the information provided, it's unclear if there is any abnormal event occurring in the video. I cannot assess the overall situation without further context or information.
Video-LLaMA [49]:	The abnormal event of a person lying on the ground could disrupt the normal functioning of the metro station. It may cause delays or disruptions in the train service, potentially affecting passengers' plans and travel experiences. Additionally, first responders may have to intervene to ensure the person's safety.
LLaMA-Adapter [50]:	The abnormal event of a train passing through a tunnel with no passengers or train cars significantly impacts the overall situation. In the image, the tunnel is empty, which is unusual as trains are typically filled with passengers or cargo. This lack of activity could be due to various reasons, such as maintenance work, a temporary suspension of services, or even an emergency situation.
Video-LLaVA [20]:	The abnormal event of the man falling down has a significant impact on the overall situation. It disrupts the normal flow of the game and adds an element of surprise and unpredictability. The players' reactions to the event, such as the man getting up and continuing to play, show that the game continues despite the unexpected interruption.
Ours:	Based on the description provided, the abnormal event in the video, which is the man in blue lying on the ground, could potentially have a significant impact on the overall situation. This is because the man's unexpected appearance on the ground could cause a distraction or confusion among the other passengers, which could in turn affect the smooth operation of the train and the evacuation process if necessary. Additionally, if the man is injured, it could also impact the response time and resources needed to address the situation.
Ground Truth:	Based on the description provided, the abnormal event of a person collapsing or someone appearing intoxicated and losing their balance has a significant impact on the overall situation. Such an event can lead to serious consequences, including severe injury or fatality, if it occurs near train tracks and coincides with the arrival of a train. The risk of a train not being able to stop in time and the individual's inability to move out of harm's way due to their condition magnifies the potential danger and impact of the event. Therefore, the impact is considered critical to the safety of the individuals involved and any potential bystanders.

(F) Anomaly Video Question-Answering in UCSD Ped1 [6].

00:00	00:01	00:02
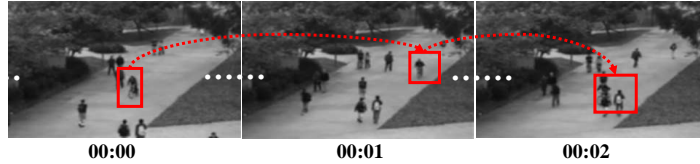
Question

Who is causing the disturbance in the video?

Video-ChatGPT [29]:	The disturbance in the video is caused by a group of people walking down a sidewalk.
VideoChat [18]:	There is no clear indication of any disturbance in the video. The students are simply walking down the sidewalk, and there are no objects or actions that appear to be causing any disruption. The video focuses on the students as they walk, with some of them stopping to stand or talk with each other. There are no sounds or other indications of any disturbance in the video.
Video-LLaMA [49]:	There is no disturbance in the video. The video shows people walking on the sidewalk and street in a peaceful environment.
LLaMA-Adapter [50]:	The disturbance in the video is caused by a group of people walking and talking, creating a crowd.
Video-LLaVA [20]:	The video does not provide any information about the cause of the disturbance. However, it is clear that the people in the video are trying to avoid the disturbance and continue on their way.
Ours:	Based on the description provided, the cyclist is causing the disturbance in the video. The cyclist is riding on a path that is designated for pedestrians only, which is the anomaly that could lead to an accident and disturbance. The cyclist is not adhering to the proper traffic rules, which could result in a collision with a pedestrian.
Ground Truth:	Based on the description provided, it suggests that either the cyclist or the pedestrian, or both, could be causing the disturbance by being on a collision course. Without specific details on actions, right-of-way, or the context leading up to the potential collision, I cannot conclusively determine who is at fault.

User

Can you help me understand this video?

HAWK

Of course, I'd be happy to help you understand a video! Can you please provide more information about the video, such as a description, the context, or a specific part you're trying to understand? Additionally, please let me know what kind of help you need – is it an analysis of the content, a breakdown of the video's meaning, or something else?

What anomaly can you identify in the video?

Based on the description you provided, there are several anomalies that can be identified in the video:

1. **Car fire**: The most obvious anomaly is the car fire that occurs near the end of the video. A car catching fire without any prior indication of a malfunction or mechanical issue is unusual and potentially dangerous.
2. **Lights flickering:** The lights on the car and the building nearby flicker, which could be indicative of an electrical issue or a malfunction in the camera equipment.
3. ....

How can we solve this problem?

To solve the problem of the car fire and the sudden stop in the middle of the road, it's important to gather more information and context. Here are some steps that can be taken:

1. **Check for any mechanical issues: ...**
2. **Look for any external factors: ...**
3. **Consult with experts: ...**
4. **Conduct a thorough investigation: ...**

Figure 7: Demo on Open-World Video Anomaly Understanding. (Part 1)

HAWK

User

Where did the fire start?

Based on the information provided in the video, it is difficult to determine with certainty where the fire started. However, there are a few possibilities:

1. **Engine compartment: ...**
2. **Battery: ...**
3. **Electrical system: ...**
4. **External sources: ...**

When did the fire break out? What's the impact?

Based on the information provided in the video, the fire broke out suddenly and without warning in the middle of the night. The exact time of the fire is not specified, but it is clear that it occurred while the car was parked in a secure area.
The impact of the fire was significant, as it caused significant damage to the car and resulted in the sudden stop in the middle of the road. The people in the area ran towards the car fire, which suggests that the fire was sudden and unexpected. The fire also caused the car to be engulfed in flames, which could have potentially spread to other nearby vehicles or structures.

I'm a fireman. How do I solve this problem?

As a fireman, you have the training and expertise to handle emergency situations like the one depicted in the video. Here are some steps you can take to solve the problem:
1. **Assess the situation:** When you arrive at the scene, take a moment to assess the situation and determine the best course of action. Check if there are any hazards present, such as flames or smoke, and if anyone is in danger.
2. **Call for backup:** If the situation is too intense or dangerous for you to handle alone, call for backup from other firefighters. This will ensure that you have enough personnel to handle the situation effectively.
3. **Use firefighting equipment: ...**
4. **Evacuate people: ...**
5. **Contain the fire: ...**
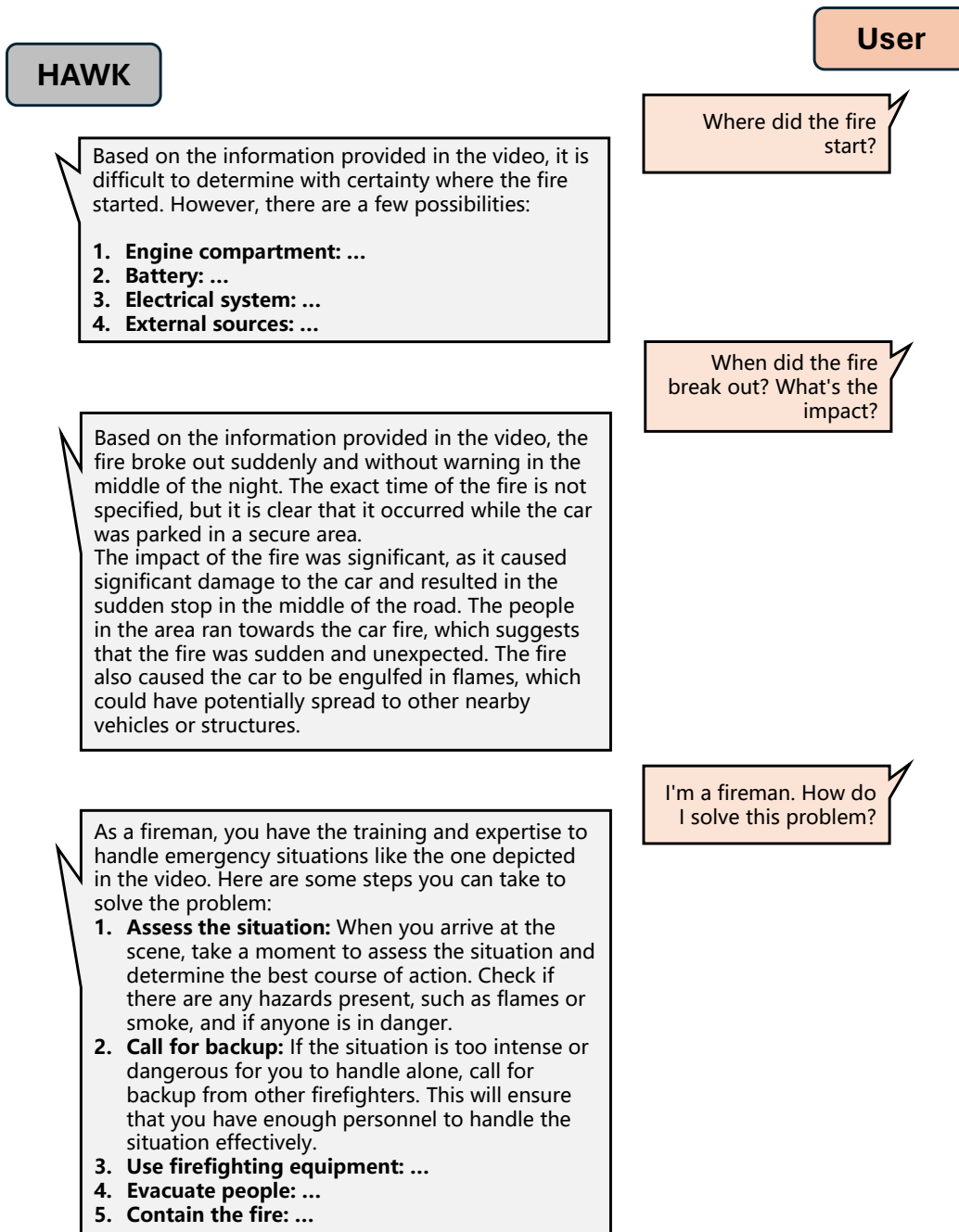
Figure 8: Demo on Open-World Video Anomaly Understanding. (Part 2)

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and title accurately reflect the content of the article.

   Guidelines:
   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the paper's limitations in the Appendix.

   Guidelines:
   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our paper does not require mathematical theory proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We will release all our code, data, and demo and most of the key information is included in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will make the data and code publicly available, but to prevent misuse, our data is subject to requests from eligible research institutions.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We present most of the key parameters in the paper, and the remaining details will be open-sourced.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Our paper does not include the results of significance statistics, we used a given random seed to maintain a random initialization, and the results are fixed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Details in Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: Details in Appendix.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Details in Appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Details in Appendix.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We strictly follow the principle of open data.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Details in Appendix.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: These are not included in our research.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: These are not included in our research.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.