VarCMP: Adapting Cross-Modal Pre-Training Models for Video Anomaly Retrieval

Peng Wu[1], Wanshun Su[1], Xiangteng He[2], Peng Wang[1]*, Yanning Zhang[1]

[1]School of Computer Science, Northwestern Polytechnical University, China
[2]Wangxuan Institute of Computer Technology, Peking University, China
{xdwupeng, suws0616, hexiangteng}@gmail.com, {peng.wang, ynzhang}@nwpu.edu.cn

Abstract

Video anomaly retrieval (VAR) aims to retrieve pertinent abnormal or normal videos from collections of untrimmed and long videos through cross-modal requires such as textual descriptions and synchronized audios. Cross-modal pre-training (CMP) models, by pre-training on large-scale cross-modal pairs, e.g., image and text, can learn the rich associations between different modalities, and this cross-modal association capability gives CMP an advantage in conventional retrieval tasks. Inspired by this, how to utilize the robust cross-modal association capabilities of CMP in VAR to search crucial visual component from these untrimmed and long videos becomes a critical research problem. Therefore, this paper proposes a VAR method based on CMP models, named **VarCMP**. First, a unified hierarchical alignment strategy is proposed to constrain the semantic and spatial consistency between video and text, as well as the semantic, temporal, and spatial consistency between video and audio. It fully leverages the efficient cross-modal association capabilities of CMP models by considering cross-modal similarities at multiple granularities, enabling VarCMP to achieve effective all-round information matching for both video-text and video-audio VAR tasks. Moreover, to further solve the problem of untrimmed and long video alignment, an anomaly-biased weighting is devised in the fine-grained alignment, which identifies key segments in untrimmed long videos using anomaly priors, giving them more attention, thereby discarding irrelevant segment information, and achieving more accurate matching with cross-modal queries. Extensive experiments demonstrates high efficacy of VarCMP in both video-text and video-audio VAR tasks, achieving significant improvements on both text-video (UCFCrime-AR) and audio-video (XDViolence-AR) datasets against the best competitors by 5.0% and 5.3% R@1.

Introduction

In recent years, with the exponential growth of video data and rapid advancements in artificial intelligence technology, video anomaly detection (VAD) has achieved significant progress, leading to the development of various works (Sultani, Chen, and Shah 2018; Wu, Liu, and Shen 2019; Wu et al. 2020; Park, Noh, and Ham 2020; Georgescu et al. 2021; Feng, Hong, and Zheng 2021; Wu et al. 2024c,d,b). However, events in videos typically capture evolving actions and interactions among entities over time, simply using single labels in VAD may be insufficient to comprehensively elucidate these sequential occurrences. In contrast, video anomaly retrieval (VAR), through detailed textual descriptions or synchronized audio, provides a more thorough account of the events transpiring within the video. Therefore, it also plays a crucial role in intelligent video analysis, such as, sifting through evidence of dangerous occurrence. Additionally, unlike traditional video retrieval, which targets short and trimmed clips, VAR is designed to handle long and untrimmed videos. It aims to retrieve both normal and abnormal videos, where anomalies may only appear in certain segments of untrimmed videos, presenting novel challenges.

ALAN (Wu et al. 2024a) is the first work for VAR, which utilizes anomaly-guided sampling to capture local anomalous segments and learn cross-modal fine-grained associations through video-prompt-based masked phrase modeling. However, this approach fails to achieve interaction at even finer granularity between visual and text features, as well as audio features. Furthermore, ALAN does not leverage the robust cross-modal semantic matching capabilities of the latest CMP models, leaving room for further improvement.

Recent advancements have witnessed the widespread adoption of CMP models, such as CLIP (Radford et al. 2021) and CLAP (Wu et al. 2023), in cross-modal retrieval tasks. CLIP4Clip (Luo et al. 2022) is a pioneering approach that integrated CLIP into video-text retrieval (VTR), which introduces a temporal fusion module designed to aggregate features across different video frames. This facilitates inter-modal alignment between video and text features, thereby enhancing the retrieval capability. However, this work directly aligns the fused temporal sequences of video and text features, which limits its ability to capture fine-grained interactions. Building upon the success of CLIP4Clip, subsequent research endeavors delve further into exploring alignment strategies that progress from coarse to fine granularity (Ma et al. 2022; Gorti et al. 2022; Wang et al. 2023; Yang et al. 2024). These methods aim to more accurately capture the intricate relationships between videos and their textual counterparts. On the other hand, CLAP has also made notable progress in audio-text retrieval by employing a contrastive learning mechanism to jointly train on large-scale audio-text pairs. Clearly, CMP models has achieved

*Corresponding Author

notable success in conventional video retrieval, thanks to its strong cross-modal association capabilities. However, it is worth conducting in-depth research on how to leverage these strengths to address unique challenges posed by VAR task.

In this paper, we propose VarCMP, adapting cross-modal pre-training models for VAR. Our goal is not simply to utilize CMP models; rather, we design a range of targeted optimizations to address the challenges posed by VAR. Specifically, we first leverage the image and text encoders of CLIP to extract visual and text features, and the audio encoder of CLAP to extract audio features. This approach not only transforms raw data into high-level representations but also dramatically enhances cross-modal alignment. Then, we propose a unified hierarchical alignment strategy. For video-text VAR, alignment is achieved at three levels: video-sentence, frame-sentence, and patch-word. For video-audio VAR, alignment occurs at the video-audio, frame-audio segment, and patch-audio segment levels. This video-frame-patch multi-granularity alignment strategy allows for the extraction of abundant and diverse information. However, due to the nature of VAR task, which consists of long and untrimmed videos, such fine-grained information can introduce substantial irrelevant content. To address this issue, we further incorporate anomaly-biased weighting mechanism within the hierarchical alignment framework. E.g., using anomaly priors to identify abnormal segments in the video, thereby highlighting key information and endowing higher weights in the fine-grained alignment. It is important to underscore that our proposed hierarchical alignment method differs from previous fine-grained alignment approaches in video-text VAR. It not only provides more detailed information for retrieval but also extracts key information from a plethora of irrelevant content, making the retrieval task more targeted. As a result, our approach achieves superior performance in both video-text and video-audio VAR tasks. Note that during training, the weights of the CLIP image and text encoders, as well as the CLAP audio encoder, are kept fixed. The gradients are back-propagated to optimize the learnable parameters of the devised modules.

Overall, the contributions of our work are threefold:

(1) We propose a unified hierarchical alignment mechanism based on the semantic and spatial consistency of video-text pairs and the semantic, temporal, and spatial consistency of video-audio pairs. This mechanism achieves unified retrieval for both video-text and video-audio tasks.

(2) We introduce the anomaly-biased weighting mechanism in hierarchical alignments, to extract key information from long and untrimmed videos, thereby endowing these dominant segments with greater weight in the fine-grained alignment. To our knowledge, this is the first attempt to efficiently implant anomaly priors to the fine-grained weighted alignment for VAR.

(3) We demonstrate the robust capabilities and effectiveness of VarCMP on two popular benchmarks. VarCMP achieves state-of-the-art performance, with an improvement of 5.0% R@1 in text-to-video retrieval on UCFCrime-AR and 5.3% R@1 in audio-to-video retrieval on XDViolence-AR, significantly surpassing current methods.

Related Work

Video Anomaly Analysis

With the advancement of deep learning, video anomaly analysis has made significant progress. Some VAD works use self-supervised manners (Georgescu et al. 2021; Wang et al. 2022a; Yan et al. 2023a,b) or reconstruction-based approaches (Yang et al. 2023; Ristea et al. 2024) to construct normal patterns. In this context, any test samples deviating from the established normal patterns are identified as anomalous. Some VAD works also use weakly supervised methods for binary classification through multiple instance learning (Wu et al. 2020; Wu, Liu, and Liu 2022; Zhou, Yu, and Yang 2023; Wu et al. 2024d). Zhou et al. (Zhou et al. 2024) proposes a human-centric video surveillance captioning dataset that provides detailed descriptions of abnormal behaviors occurring between individuals. Yuan et al (Yuan et al. 2024) focuses on action-centric approaches, dedicating their efforts to the precise detection of anomalous moments in videos. In addition, some works (Wu et al. 2024a; Zhang et al. 2024) focus on events and use subtitle information to assist in detecting anomalies, thereby improving the understanding and recognition accuracy of abnormal events. Our work provides an important supplement to the field of video anomaly analysis.

Cross-Modal Retrieval

We primarily introduce video-text retrieval and video-audio retrieval. Video-text retrieval (Yu, Kim, and Kim 218; Yang, Bisk, and Gao 2021; Wang, Zhu, and Yang 2021; Croitoru et al. 2021; Lin et al. 2023; Han et al. 2023, 2024) operates by analyzing a given sentence to identify the most corresponding video within a database, and vice versa. Early video-text retrieval works focus on enhancing the capabilities of text and visual encoders to improve performance. Dot-product interactions are widely used to compute the global similarity but lack consideration for fine-grained information (Li et al. 2019; Gabeur et al. 2020; Dong et al. 2021; Liu et al. 2021). Consequently, some research aims at aligning fine-grained details (Wray et al. 2019; Wu et al. 2021; Han et al. 2021; Yang, Bisk, and Gao 2021; Wang et al. 2021; Ge et al. 2022; Ma et al. 2022; Tian et al. 2024). With the tremendous success of CLIP, numerous CLIP-based video-text retrieval methods (Fang et al. 2021; Luo et al. 2022; Buch et al. 2022; Liu et al. 2022; Wang et al. 2023) have emerged. Video-audio cross-modal retrieval is a typical cross-modal perception task (Tian 2023; Hou et al. 2024). Its goal is to retrieve data in one modality based on information from another modality, making it a rapidly evolving research field in recent years (Zhang et al. 2013; Hong, Im, and Yang 2017; Yan, Gong, and Zhang 2018; Surís et al. 2018; Zeng, Yu, and Oyama 2020; Asano et al. 2020). For instance, Tian et al. (Tian et al. 2018) proposes an audio-to-video localization/retrieval task, which temporarily locates the corresponding visual sound source in the video given a sound segment, and vice versa. Hong et al. (Hong, Im, and Yang 2017) utilizes the sorting loss between patterns to improve the semantic similarity of video music pairs. Suris et al. (Surís et al. 2018) utilizes cosine similarity loss and

classification loss to project unimodal features into a common feature space. Unlike the aforementioned video-text and video-audio retrieval tasks, our focus is on videos where the majority contain anomalous events, and these videos are often long and untrimmed. This introduces new challenges for cross-modal retrieval.

Method

Overview

The feature encoders are first introduced to extract representations of different mod. Then, the unified hierarchical alignment are presented to match between different modalities at multiple granularities. Next, the anomaly-biased weighting mechanism is illustrated to mine key segments and assign them greater attention during the fine(r)-grained alignment. Finally, the model's training process is demonstrated. Our methodology is depicted in Figure 1.

Feature Encoders

Previous works rely on pre-training models such as I3D (Carreira and Zisserman 2017), BERT (Devlin et al. 2018), and VGGish (Gemmeke et al. 2017) to extract video, text, and audio features, which are subsequently aligned. Recently, large-scale pre-training cross-modal models like CLIP and CLAP have demonstrated exceptional generalization capabilities across various downstream tasks. Drawing inspiration from CLIP and related retrieval tasks, we utilize the encoders of CLIP and CLAP as backbone networks to extract video, text, and audio features, leveraging the robust correlations among visual content, textual descriptions, and synchronous audio.

Video Encoder. Following previous methods Liu et al. (2022); Luo et al. (2022); Wang et al. (2023), the pre-training CLIP visual encoder (Dosovitskiy et al. 2020) $\mathcal{F}_v$ is used to extract visual features for each video. A video with N_v frames can be represented as $[F_v^1, F_v^2, \ldots, F_v^{N_v}]$. For the n-th frame of a given video F_v^n, it is divided into non-overlapping patches, a [CLS] token is added, and the visual encoder $\mathcal{F}_v$ is used to obtain the patch representations $\mathbf{p}_v^n$, where $\mathbf{p}_v^n = \mathcal{F}_v\left(F_v^n\right) \in \mathbb{R}^{M \times C}$, with M denoting the number of patches within a video frame and C representing the dimensionality of the visual features. The [CLS] representations from each frame are then extracted and combined to form the frame representations $\mathbf{f}_v = [f_v^1; f_v^2; \ldots; f_v^{N_v}]$, with $\mathbf{f}_v \in \mathbb{R}^{N_v \times C}$. $\mathbf{f}_v$ through temporal encoder forms a video representation of $\mathbf{e}_v \in \mathbb{R}^C$, where a token shift module is leveraged to learn temporal information with minimal cost (Wang et al. 2023).

Text Encoder. Given a text query T (with an [EOS] token appended at the end), the pre-training CLIP (Radford et al. 2021) text encoder $\mathcal{F}_t$ is used to produce word features $\mathbf{w}_t$, where $\mathbf{w}_t = \mathcal{F}_t\left(T\right) \in \mathbb{R}^{N_t \times C}$, with N_t denoting the length of the word sequence. The representation of the [EOS] token is then used as the sentence feature $\mathbf{s}_t \in \mathbb{R}^C$.

Audio Encoder. Given an audio query A, the pre-training CLAP (Wu et al. 2023) audio encoder $\mathcal{F}_a$ is used to produce audio segment features $\mathbf{e}_a$, where $\mathbf{e}_a = \mathcal{F}_a\left(A\right) \in \mathbb{R}^{N_a \times C}$, with N_a denoting the length of the audio segment sequence. Subsequently, a Transformer encoder (Vaswani et al. 2017) is employed to obtain the [CLS] token representation, which is then used as the audio feature $\mathbf{m}_a \in \mathbb{R}^C$. Note that visual, text, and audio features are all embedded into the same dimensional space C.

Unified Hierarchical Alignment

Semantic consistency refers to the alignment of the core intentions conveyed across different modalities. Spatial consistency pertains to the congruence of content within the spatial dimension across modalities. In video-text VAR, the positions of objects, scenes, or actions depicted in the video should correspond to the spatial information described in the text. Temporal consistency denotes the alignment of the chronological sequence across modalities, a concept articulated by (Wei et al. 2022). For instance, in video-audio VAR, the audio content should be synchronized with the video timeline, ensuring the sequence of events remains consistent. Based on these characteristics, we propose a unified hierarchical alignment strategy for both video-text and video-audio VAR tasks, which offers the advantage of significantly improving retrieval accuracy by systematically aligning semantic, spatial, and temporal elements across multiple levels. Such a strategy leverages the efficient cross-modal matching capabilities of CMP models from multiple granularities, addressing the weaknesses of considering only coarse-grained or fine-grained alignments. Note that, this approach is unified in form and represents a generalized expression, enabling not only multi-modal alignment between video and text but also cross-modal alignment between video and audio.

Coarse-Grained Alignment. Coarse-grained alignment reflects global semantic consistency. In video-text VAR, it is obtained by computing the similarity score between the whole video and the sentence. Similarly, in video-audio VAR, the similarity score between the whole video and the whole audio is computed. The specific formulas are as follows:

$$\mathbf{S}_{vg} = \frac{1}{2}\left(f_{gw}(\mathbf{l})(\tilde{\mathbf{l}})^\top \tilde{\mathbf{e}}_v + f_{ew}(\mathbf{e}_v)(\tilde{\mathbf{l}})^\top \tilde{\mathbf{e}}_v\right) \tag{1}$$

where $\mathbf{l}$ represents the sentence $\mathbf{s}_t$ in video-text VAR and audio $\mathbf{e}_a$ in video-audio VAR. $f_{gw}(\mathbf{l})$ denotes the weight for sentence/audio-level feature. $f_{ew}(\mathbf{e}_v)$ denotes the weight for video-level feature. f_{gw} and f_{ew} are both composed of classic MLP and Softmax functions. The normalization operation for $\tilde{\mathbf{e}}_v$ is defined as $\tilde{\mathbf{e}}_v = \mathbf{e}_v^i / |\mathbf{e}_v^i|_2$, applied along the channels, and $\tilde{\mathbf{l}}$ is normalized in the same manner.

Fine-Grained Alignment. Fine-grained alignment here refers to local semantic consistency in video-text and unique temporal consistency in video-audio. For video-text VAR, fine-grained alignment is performed at the video frame and sentence levels. Specifically, the similarity between frame-level visual features and textual query features is calculated to obtain the frame-sentence similarity score. Inspired

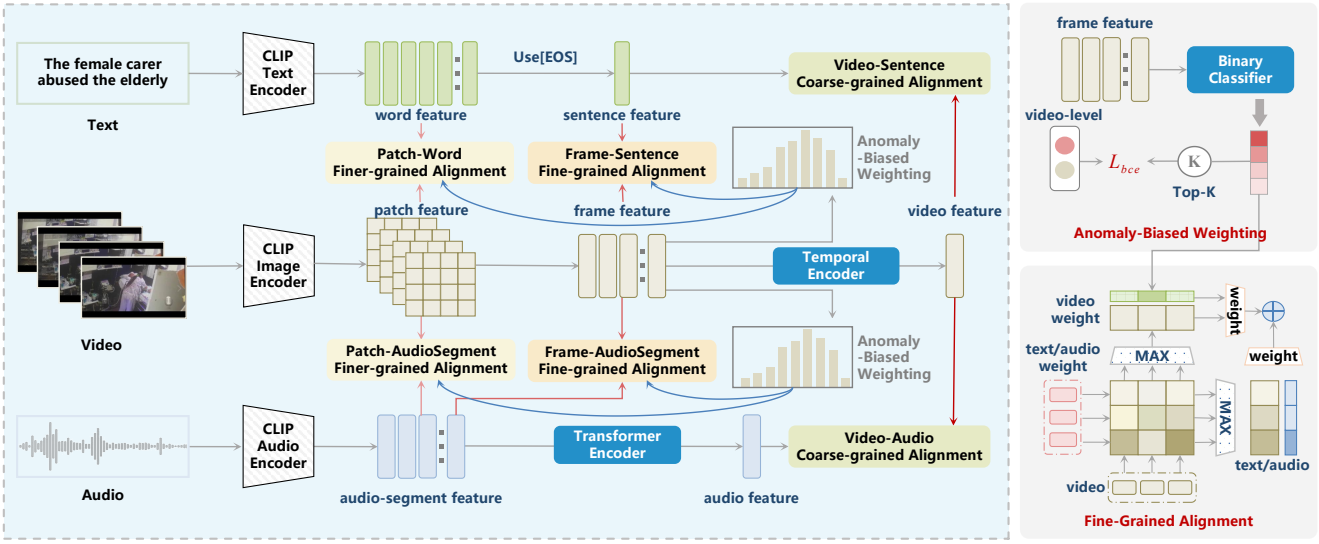

Figure 1: The framework of our proposed VarCMP. For ease of viewing, we duplicated the anomaly-biased weighting twice.

by (Wang et al. 2022b, 2023), the token-wise interaction is performed to adaptively exploit fine-grained correlations:

$$\mathbf{S}_{fs} = \frac{1}{2}(f_{sw}(\mathbf{s}_t) \max_{j=1}^{N_v}(\tilde{\mathbf{s}}_t)^\top \tilde{\mathbf{f}}_v^j + \sum_{j=1}^{N_v} f_{fw}^j(\mathbf{f}_v)(\tilde{\mathbf{s}}_t)^\top \tilde{\mathbf{f}}_v^j) \quad (2)$$

where, $f_{fw}(\mathbf{f}_v)$ and $f_{sw}(\mathbf{s}_t)$ denote the weights for frame-level features and sentence-level features.

As for video-audio VAR, the slightly different is, we compute the similarity score between frame-level visual features and audio segment features to get the frame-audio segment similarity score. The specific formula is presented as follows:

$$\mathbf{S}_{fe} = \frac{1}{2}\left(\sum_{i=1}^{N_a} f_{ew}^i(\mathbf{e}_a) \max_{j=1}^{N_v}(\tilde{\mathbf{e}}_a^i)^\top \tilde{\mathbf{f}}_v^j + \sum_{j=1}^{N_v} f_{fw}^j(\mathbf{f}_v) \max_{i=1}^{N_a}(\tilde{\mathbf{e}}_a^i)^\top \tilde{\mathbf{f}}_v^j\right) \quad (3)$$

where, $f_{ew}(\mathbf{e}_a)$ and $f_{fw}(\mathbf{f}_v)$ denote the weights for audio segment level features and frame-level features. It is particularly worth noting that, considering the characteristics of long and untrimmed videos in VAR tasks, the weights here are not entirely derived from the network f_{fw}. For videos identified as anomalous, we apply anomaly priors to obtain anomaly-biased weights, thereby assigning greater attention to potential key segments during the fine-grained alignment stage. We elaborate on this mechanism in the next section.

Finer-Grained Alignment. Finally, finer-grained alignment represents the consistency across different modalities at the local spatial level. That is, we calculate the similarity score between patch-level visual features and word-level features to achieve finer-grained matching in video-text VAR. Similarly, in video-audio VAR, we obtain the similarity score between patch-level feature and audio segment level feature. The specific formula is shown as follows:

$$\mathbf{S}_{po} = \frac{1}{2}\left(\sum_{i=1}^{N_y}\left[f_{ew}(\mathbf{o}_y^i) \max_{j=1}^{N_v \times C}(\tilde{\mathbf{o}}_y^i)^\top \tilde{\mathbf{p}}_v^j\right] + \sum_{j=1}^{N_v \times M}\left[f_{pw}^j(\mathbf{p}_v^j) \max_{i=1}^{N_y}(\tilde{\mathbf{o}}_y^i)^\top \tilde{\mathbf{p}}_v^j\right]\right) \quad (4)$$

where $\mathbf{o}_y$ represents the word in video-text VAR and audio segment in video-audio VAR. $f_{ew}(\mathbf{o}_y)$ denotes the weight for word-level feature in video-text VAR and the weight for audio segment level feature in video-audio VAR. $f_{pw}(\mathbf{p}_v)$ denotes the weight for patch-level feature. N_y represents N_t in video-text VAR and N_a in video-audio VAR.

Anomaly-Biased Weighting Mechanism

For fine-grained alignment within the hierarchical framework, we introduce a module. This module utilizes anomaly prior information to locate key segments in untrimmed long videos and proposes an anomaly-oriented fine-grained alignment method. The following sections elaborate on these two components respectively.

We first propose incorporating key frame detection to extract key information from videos. Specifically, we take inspirations from the weakly supervised VAD method, i.e., VadCLIP (Wu et al. 2024d), where the video features $\mathbf{f}_v \in \mathbb{R}^{N_v \times C}$ encoded by the visual encoder are fed into a binary classifier that includes a feed-forward network (FFN), a fully connected (FC) layer, and a Sigmoid activation function. This process computes the frame-level abnormal confidence $\mathbf{A} \in \mathbb{R}^{N_v \times 1}$ for the visual features:

$$\mathbf{A} = \text{Sigmoid}\left(FC\left(FFN\left(f_v\right) + f_v\right)\right) \quad (5)$$

During the training phase, we determine whether the given video is anomalous based on the labels. If it is anomalous, we use the anomaly confidence to obtain the corresponding weights, i.e., the frame-level and patch-level

Method	Text → Video				Video → Text			
	R@1↑	R@5↑	R@10↑	MdR↓	R@1↑	R@5↑	R@10↑	MdR↓
Non-CLIP Methods								
Random Baseline	0.3	2.1	3.4	144.0	0.3	1.0	3.1	145.5
CE(2019)	6.6	19.7	32.4	23.5	5.5	19.7	32.4	21.0
MMT(2020)	8.3	26.2	39.3	16.0	7.2	23.1	39.0	16.0
HL-Net(2020)	5.5	20.2	38.3	19.5	5.5	22.8	35.5	20.0
XML(2020)	6.9	24.1	42.4	14.0	6.6	25.9	43.4	13.0
T2VLAD(2021)	7.6	23.4	39.7	15.5	6.2	27.9	43.1	14.0
X-CLIP*(2022)	8.2	27.2	41.7	16.0	6.9	25.8	40.3	15.0
ALAN(2024a)	9.0	27.9	44.8	14.0	7.3	24.8	46.9	12.0
CLIP-based Methods								
CLIP4Clip(2022)	17.8	40.1	57.5	8.0	15.3	33.0	44.6	13.0
TS2-Net(2022)	23.4	54.1	70.7	4.0	17.7	53.7	68.0	5.0
UCoFiA(2023)	23.6	53.8	69.2	5.0	22.8	54.1	67.7	4.0
ALAN†(2024a)	10.3	36.9	57.9	8.0	14.1	42.4	57.9	8.0
VarCMP(Ours)	**28.6**	**60.3**	**74.8**	**4.0**	**24.1**	**56.8**	**72.1**	**4.0**

Table 1: Comparisons with the state-of-the-art methods on UCFCrime-AR.

weights are assigned according to the abnormal confidence of each frame to highlight the key information; otherwise, we continue to use the weights learned by the network. During the testing phase, we use the anomaly confidence to assess whether the given test video is anomalous, thereby assigning different weights accordingly. This can be specifically represented as follows:

$$f_{fw}(\mathbf{f}_v) = \begin{cases} f_{fw}(\mathbf{f}_v), & y/\hat{y} = 0 \\ \mathrm{S}_\tau(\mathbf{A}), & y/\hat{y} = 1 \end{cases} \tag{6}$$

$$\mathrm{S}_\tau(\mathbf{A}_i) = \frac{\exp(\mathbf{A}_i/\tau)}{\sum_j \exp(\mathbf{A}_j/\tau)} \tag{7}$$

where $y/\hat{y} = 0$ indicates that the video is a normal video, and $y/\hat{y} = 1$ indicates that the video is an abnormal video, $\hat{y}$ indicates the video-level prediction based on $\mathbf{A}$. The patch-level weights of the abnormal video are assigned based on the frame-level weights in a broadcast manner.

Training and Inference

The similarity score $\mathbf{R}(v_i, t_j)$ and $\mathbf{R}(v_i, a_j)$ measure the semantic similarity between two instances. The similarity score $\mathbf{R}(v_i, t_j)$ in video-text VAR is:

$$\mathbf{R}(v_i, t_j) = \frac{1}{3}(\mathbf{S}_{vg} + \mathbf{S}_{fs} + \mathbf{S}_{po}) \tag{8}$$

The similarity score $\mathbf{R}(v_i, a_j)$ in video-audio VAR is:

$$\mathbf{R}(v_i, a_j) = \frac{1}{3}(\mathbf{S}_{vg} + \mathbf{S}_{fe} + \mathbf{S}_{po}) \tag{9}$$

To ensure consistency, we will use $\mathbf{R}(v_i, q_j)$ to represent both the video-text similarity score $\mathbf{R}(v_i, t_j)$ and the video-audio similarity score $\mathbf{R}(v_i, a_j)$. In multi-granularity alignment, during training, given a batch of B video-text pairs, the model will generate a $B \times B$ similarity matrix. We adopt the symmetric InfoNCE loss over the similarity matrix to optimize the retrieval model, which can be formulated as:

$$\mathcal{L}_{v2t/v2a} = -\frac{1}{B}\sum_i^B \log \frac{\exp\left(\mathbf{R}(v_i, q_i)\right)}{\sum_{j=1}^B \exp\left(\mathbf{R}(v_i, q_j)\right)} \tag{10}$$

$$\mathcal{L}_{t2v/a2v} = -\frac{1}{B}\sum_i^B \log \frac{\exp\left(\mathbf{R}(v_i, q_i)\right)}{\sum_{j=1}^B \exp\left(\mathbf{R}(v_j, q_i)\right)} \tag{11}$$

For the weakly supervised VAD in anomaly-biased weighting, the classification loss $\mathcal{L}_{bce}$ can be computed using binary cross-entropy.

Overall, the final objective of the VarCMP method is given by the following formula:

$$\mathcal{L} = \mathcal{L}_{bce} + \mathcal{L}_{v2t/v2a} + \mathcal{L}_{t2v/a2v} \tag{12}$$

During inference, to perform video-to-text and video-to-audio retrieval, the similarity between all videos and queries is computed, and the text or audio with high similarity is retrieved. For text-to-video and audio-to-video retrieval, the same process is conducted in a similar manner but in the opposite direction.

Experiments

Datasets and Evaluation Metrics

We conduct experiments on two popular VAR datasets, i.e., UCFCrime-AR and XDViolence-AR (Wu et al. 2024a) for video-text and video-audio VAR tasks. Following (Wu et al. 2024a), we use the rank-based metric for evaluation, i.e., Recall at K (R@K, K=1, 5, 10), Median Rank (MdR) to measure the overall performance.

Method	Audio → Video				Video → Audio			
	R@1↑	R@5↑	R@10↑	MdR↓	R@1↑	R@5↑	R@10↑	MdR↓
Non-CLIP Methods								
Random Baseline	0.4	0.6	2.5	399.5	0.1	0.6	0.8	399.5
CE(2019)	11.4	33.3	47.0	12.5	13.0	34.3	46.4	13.0
MMT(2020)	20.5	53.5	68.0	5.0	23.0	54.6	69.5	5.0
HL-Net(2020)	12.4	36.6	48.3	11.0	13.4	38.3	52.1	10.0
XML(2020)	22.9	55.6	70.3	5.0	22.6	57.4	71.4	4.0
T2VLAD(2021)	22.4	56.1	71.0	4.0	23.2	57.1	73.5	4.0
X-CLIP*(2022)	26.4	61.1	73.9	3.0	26.4	61.3	73.8	4.0
ALAN(2024a)	29.8	68.0	82.0	3.0	32.3	70.0	82.3	3.0
CLIP-based Methods								
CLIP4Clip(2022)	5.9	21.5	34.6	21.0	9.2	30.5	45.0	13.0
TS2-Net(2022)	15.5	47.8	63.6	6.0	12.0	41.6	61.4	7.0
UCoFiA(2023)	26.2	59.1	71.9	4.0	25.9	60.1	75.1	4.0
ALAN†(2024a)	32.9	70.5	82.9	3.0	34.3	70.4	82.1	3.0
VarCMP(Ours)	**38.2**	**79.8**	**92.5**	**2.0**	**38.5**	**76.4**	**91.1**	**2.0**

Table 2: Comparisons with the state-of-the-art methods on XDViolence-AR.

Implementation Details

For network structure, image and text encoders are adopted from pre-training CLIP (VIT-B/32), audio encoder is adopted from pre-training CLAP (630k-audioset-fusion-best). The dimension C of visual, text, and audio features is set to 512. We choose $M = 4$ most salient patches for each frame in our patch selection module on the datasets. The transformer layer is set to 1, the number of heads to 8, and the dimension of FFN to 1024. We sample $N_v = 32$ frames per video, set the max length of text and audio query as 32. For model training, VarCMP is trained on a single NVIDIA RTX 4090 GPU using PyTorch. We use AdamW as the optimizer with batch size of 8 with the learning rate of 1e-4 and total epoch of 15.

Comparison with State-of-the-Art Methods

Comparative results of our VarCMP versus existing methods are presented in Tables 1 and 2. Comparison methods include both Non-CLIP based methods and CLIP based methods. Among them, the symbol "*" represents the I3D features used by the CLIP method, and "†" represents the method is re-implemented with CLIP features. We found that VarCMP consistently outperforms existing methods across all metrics in text-to-video, video-to-text, audio-to-video, and video-to-audio retrieval metrics. Specifically, on UCFCrime-AR, VarCMP achieved an 18.3% improvement in text-to-video retrieval R@1 compared to ALAN (Wu et al. 2024a), the first VAR work. When compared to UCoFiA (Wang et al. 2023), a recent coarse-to-fine alignment method for video-text VAR, VarCMP shows a 5.0% improvement in text-to-video retrieval R@1. This performance enhancement partly reflects the robust capabilities of CLIP and underscores the efficacy of our proposed multi-granularity alignment and anomaly-biased weighting techniques. Similarly, on XDViolence-AR, we observe that VarCMP also excels in handling cross-modal retrieval between video and audio. Compared to the previous state-of-the-art baselines, it achieves a 5.3% improvement in audio-to-video retrieval R@1 and a 9.6% improvement in R@10. This indicates that our proposed VarCMP more effectively maintains temporal consistency between video and audio.

Coarse	Fine	Finer	Text → Video		Video → Text	
			R@1↑	R@10↑	R@1↑	R@10↑
√			24.7	72.3	20.7	64.6
√	√		26.0	72.9	21.8	68.0
√	√	√	**28.6**	**74.8**	**24.1**	**72.1**

Table 3: The effect of different level alignments on UCFCrime-AR.

Coarse	Fine	Finer	Audio → Video		Video → Audio	
			R@1↑	R@10↑	R@1↑	R@10↑
√			31.4	86.1	28.6	87.0
√	√		34.0	90.2	33.1	89.2
√	√	√	**38.2**	**92.5**	**38.5**	**91.1**

Table 4: The effect of different level alignments on XDViolence-AR.

Ablation Studies

Effectiveness of Unified Hierarchical Alignment. First, we verify the effectiveness of our approach at different levels of granularity. As shown in Tables 3 and 4, the results that fully consider both coarse-grained and fine-grained alignments outperform those that only consider coarse-grained alignment, with significant improvements in R@1 and R@10. This reflects that fine-grained alignment better captures subtle differences within videos, allowing for more

Dataset →	UCFCrime-AR				XDViolence-AR			
Setting ↓	Text → Video		Video → Text		Audio → Video		Video → Audio	
	R@1↑	R@10↑	R@1↑	R@10↑	R@1↑	R@10↑	R@1↑	R@10↑
w/o Anomaly-biased weighting	27.9	73.1	**25.9**	69.4	35.9	91.2	34.4	89.6
w Anomaly-biased weighting	**28.6**	**74.8**	24.1	**72.1**	**38.2**	**92.5**	**38.5**	**91.1**

Table 5: Comparisons with and without the anomaly-biased weighting mechanism on UCFCrime-AR and XDViolence-AR.

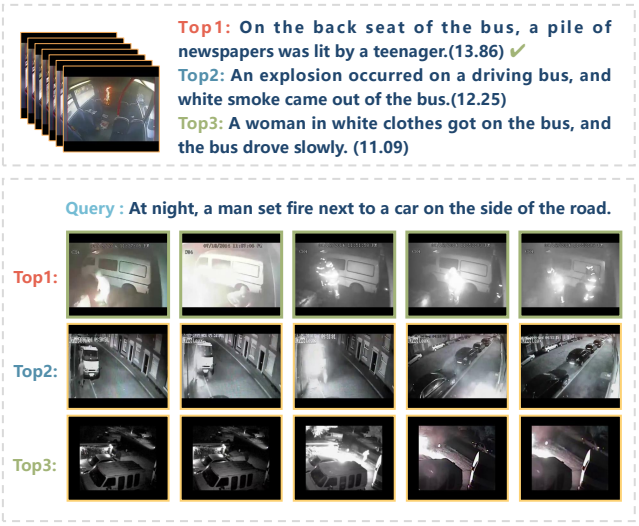

Figure 2: Top-3 v2t and t2v retrieval results on UCFCrime-AR.

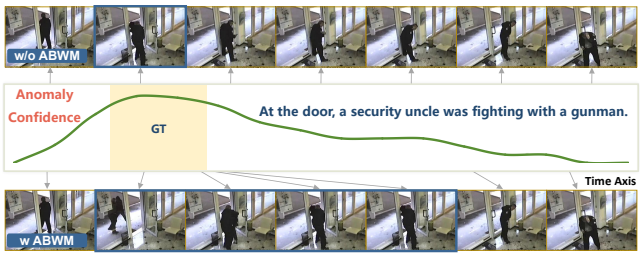

Figure 3: Visualization results with and without anomaly-bias weighting mechanism.

precise differentiation between them. This outcome effectively demonstrates the validity of our proposed multi-level alignment approach.

Effectiveness of Anomaly-Biased Weighting. As previously stated, we propose integrating anomaly-biased weighting into the video anomaly retrieval method to better emphasize abnormal events. To evaluate the efficacy of this module, we conduct experiments on two benchmarks and present results in Table 5. Omitting the anomaly-biased weighting mechanism leads to a significant decline in retrieval performance since the equal alignment strategy leads to the loss of key segment information.

Qualitative Analyses

Visualization of Retrieval Results. Figure 2 illustrates the visualized results of retrieving the three most relevant texts for a given video and retrieving the three most relevant videos for a given text on the UCFCrime-AR dataset. In both visualizations, the retrieved texts and videos are highly matching. For example, in the visualization results of a given text retrieval video, the video clips contain information about "car", our fine-grained alignment effectively considers details such as "car-road-fire", thereby retrieving information that best matches the textual description.

Visualization of Anomaly-Biased Weighting Mechanism. We present in the Figure 3 a visual comparison of the video frames focused on by the model with and without the anomaly-biased weighting mechanism. Although anomalous events account for less than one-third of the video, by introducing the anomaly-biased weighting mechanism, the model assigns weights to frames based on the anomaly confidence, selectively focusing more on frames related to anomalous events while reducing attention to normal frames. This makes the retrieval process more precise, as clearly demonstrated in the second row.

Conclusion

In this paper, we propose VarCMP for video anomaly retrieval task. Considering the advantages of cross-modal pre-training models in learning different modal information, we exploit fixed pre-training models such as CLIP and CLAP to extract cross-modal features. Based on the semantic, temporal, and spatial consistency, we propose a unified hierarchical alignment strategy, which effectively unifies video-text and video-audio matching by considering cross-modal similarity at different granularities. Then, we proposed the anomaly biased weighting mechanism to effectively identify key frames in videos and endow these frames with high weights. We empirically validate the effectiveness of VarCMP through state-of-the-art performance and sufficient ablations on two VAR benchmarks. In future work, we strive toward a more unified VAR framework that simultaneously incorporates the ability for semantic alignment across multiple modalities.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62306240, U23B2013, 62272013), China Postdoctoral Science Foundation (No.2023TQ0272), and Beijing Natural Science Foundation(No. 4232005).

References

Asano, Y.; Patrick, M.; Rupprecht, C.; and Vedaldi, A. 2020. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *NeurIPS*, volume 33, 4660–4671.

Buch, S.; Eyzaguirre, C.; Gaidon, A.; Wu, J.; Fei-Fei, L.; and Niebles, J. C. 2022. Revisiting the "video" in video-language understanding. In *CVPR*, 2917–2927.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 6299–6308.

Croitoru, I.; Bogolin, S.-V.; Leordeanu, M.; Jin, H.; Zisserman, A.; Albanie, S.; and Liu, Y. 2021. Teachtext: Cross-modal generalized distillation for text-video retrieval. In *ICCV*, 11583–11593.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, J.; Li, X.; Xu, C.; Yang, X.; Yang, G.; Wang, X.; and Wang, M. 2021. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8): 4065–4080.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fang, H.; Xiong, P.; Xu, L.; and Chen, Y. 2021. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*.

Feng, J.-C.; Hong, F.-T.; and Zheng, W.-S. 2021. Mist: Multiple instance self-training framework for video anomaly detection. In *CVPR*, 14009–14018.

Gabeur, V.; Sun, C.; Alahari, K.; and Schmid, C. 2020. Multi-modal transformer for video retrieval. In *ECCV*, 214–229.

Ge, Y.; Ge, Y.; Liu, X.; Li, D.; Shan, Y.; Qie, X.; and Luo, P. 2022. Bridging video-text retrieval with multiple choice questions. In *CVPR*, 16167–16176.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 776–780.

Georgescu, M.-I.; Barbalau, A.; Ionescu, R. T.; Khan, F. S.; Popescu, M.; and Shah, M. 2021. Anomaly detection in video via self-supervised and multi-task learning. In *CVPR*, 12742–12752.

Gorti, S. K.; Vouitsis, N.; Ma, J.; Golestan, K.; Volkovs, M.; Garg, A.; and Yu, G. 2022. X-pool: Cross-modal language-video attention for text-video retrieval. In *CVPR*, 5006–5015.

Han, N.; Chen, J.; Xiao, G.; Zhang, H.; Zeng, Y.; and Chen, H. 2021. Fine-grained cross-modal alignment network for text-video retrieval. In *ACM MM*, 3826–3834.

Han, N.; Yang, X.; Lim, E.-P.; Chen, H.; and Sun, Q. 2024. Efficient cross-modal video retrieval with meta-optimized frames. *IEEE Transactions on Multimedia*, 1–14.

Han, N.; Zeng, Y.; Shi, C.; Xiao, G.; Chen, H.; and Chen, J. 2023. Bic-net: Learning efficient spatio-temporal relation for text-video retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3): 1–21.

Hong, S.; Im, W.; and Yang, H. S. 2017. Content-based video-music retrieval using soft intra-modal structure constraint. *arXiv preprint arXiv:1704.06761*.

Hou, W.; Li, G.; Tian, Y.; and Hu, D. 2024. Toward Long Form Audio-Visual Video Understanding. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(9): 1–26.

Lei, J.; Yu, L.; Berg, T. L.; and Bansal, M. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 447–463.

Li, X.; Xu, C.; Yang, G.; Chen, Z.; and Dong, J. 2019. W2vv++ fully deep learning for ad-hoc video search. In *ACM MM*, 1786–1794.

Lin, Y.-B.; Sung, Y.-L.; Lei, J.; Bansal, M.; and Bertasius, G. 2023. Vision transformers are parameter-efficient audio-visual learners. In *CVPR*, 2299–2309.

Liu, S.; Fan, H.; Qian, S.; Chen, Y.; Ding, W.; and Wang, Z. 2021. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In *ICCV*, 11915–11925.

Liu, Y.; Albanie, S.; Nagrani, A.; and Zisserman, A. 2019. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*.

Liu, Y.; Xiong, P.; Xu, L.; Cao, S.; and Jin, Q. 2022. Ts2-net: Token shift and selection transformer for text-video retrieval. In *ECCV*, 319–335.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304.

Ma, Y.; Xu, G.; Sun, X.; Yan, M.; Zhang, J.; and Ji, R. 2022. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACM MM*, 638–647.

Park, H.; Noh, J.; and Ham, B. 2020. Learning memory-guided normality for anomaly detection. In *CVPR*, 14372–14381.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.

Ristea, N.-C.; Croitoru, F.-A.; Ionescu, R. T.; Popescu, M.; Khan, F. S.; Shah, M.; et al. 2024. Self-distilled masked auto-encoders are efficient video anomaly detectors. In *CVPR*, 15984–15995.

Sultani, W.; Chen, C.; and Shah, M. 2018. Real-world anomaly detection in surveillance videos. In *CVPR*, 6479–6488.

Surís, D.; Duarte, A.; Salvador, A.; Torres, J.; and Giró-i Nieto, X. 2018. Cross-modal embeddings for video and audio retrieval. In *ECCV workshops*, 0–0.

Tian, K.; Cheng, Y.; Liu, Y.; Hou, X.; Chen, Q.; and Li, H. 2024. Towards Efficient and Effective Text-to-Video Retrieval with Coarse-to-Fine Visual Representation Learning. In *AAAI*, volume 38, 5207–5214.

Tian, Y. 2023. Towards unified, explainable, and robust multisensory perception. In *AAAI*, volume 37, 15456–15456.

Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018. Audio-visual event localization in unconstrained videos. In *ECCV*, 247–263.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NuerIPS*, volume 30.

Wang, G.; Wang, Y.; Qin, J.; Zhang, D.; Bao, X.; and Huang, D. 2022a. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *ECCV*, 494–511.

Wang, Q.; Zhang, Y.; Zheng, Y.; Pan, P.; and Hua, X.-S. 2022b. Disentangled Representation Learning for Text-Video Retrieval. *arXiv preprint arXiv:2203.07111*.

Wang, W.; Zhang, M.; Chen, R.; Cai, G.; Zhou, P.; Peng, P.; Guo, X.; Wu, J.; and Sun, X. 2021. Dig into Multi-modal Cues for Video Retrieval with Hierarchical Alignment. In *IJCAI*, 1113–1121.

Wang, X.; Zhu, L.; and Yang, Y. 2021. T2vlad: global-local sequence alignment for text-video retrieval. In *CVPR*, 5079–5088.

Wang, Z.; Sung, Y.-L.; Cheng, F.; Bertasius, G.; and Bansal, M. 2023. Unified coarse-to-fine alignment for video-text retrieval. In *ICCV*, 2816–2827.

Wei, Y.; Hu, D.; Tian, Y.; and Li, X. 2022. Learning in Audio-visual Context: A Review, Analysis, and New Perspective. *arXiv preprint arXiv:2208.09579*.

Wray, M.; Larlus, D.; Csurka, G.; and Damen, D. 2019. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *ICCV*, 450–459.

Wu, P.; He, X.; Tang, M.; Lv, Y.; and Liu, J. 2021. Hanet: Hierarchical alignment networks for video-text retrieval. In *ACM MM*, 3518–3527.

Wu, P.; Liu, J.; He, X.; Peng, Y.; Wang, P.; and Zhang, Y. 2024a. Toward Video Anomaly Retrieval From Video Anomaly Detection: New Benchmarks and Model. *IEEE Transactions on Image Processing*, 33: 2213–2225.

Wu, P.; Liu, J.; and Shen, F. 2019. A deep one-class neural network for anomalous event detection in complex scenes. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7): 2609–2622.

Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; and Yang, Z. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *ECCV*, 322–339.

Wu, P.; Liu, X.; and Liu, J. 2022. Weakly supervised audio-visual violence detection. *IEEE Transactions on Multimedia*, 25: 1674–1685.

Wu, P.; Pan, C.; Yan, Y.; Pang, G.; Wang, P.; and Zhang, Y. 2024b. Deep Learning for Video Anomaly Detection: A Review. *arXiv preprint arXiv:2409.05383*.

Wu, P.; Zhou, X.; Pang, G.; Sun, Y.; Liu, J.; Wang, P.; and Zhang, Y. 2024c. Open-vocabulary video anomaly detection. In *CVPR*, 18297–18307.

Wu, P.; Zhou, X.; Pang, G.; Zhou, L.; Yan, Q.; Wang, P.; and Zhang, Y. 2024d. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *AAAI*, 6074–6082.

Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; and Dubnov, S. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 1–5.

Yan, C.; Zhang, S.; Liu, Y.; Pang, G.; and Wang, W. 2023a. Feature prediction diffusion model for video anomaly detection. In *ICCV*, 5527–5537.

Yan, Q.; Gong, D.; and Zhang, Y. 2018. Two-stream convolutional networks for blind image quality assessment. *IEEE Transactions on Image Processing*, 28(5): 2200–2211.

Yan, Q.; Hu, T.; Sun, Y.; Tang, H.; Zhu, Y.; Dong, W.; Van Gool, L.; and Zhang, Y. 2023b. Towards high-quality hdr deghosting with conditional diffusion models. *IEEE Transactions on Circuits and Systems for Video Technology*.

Yang, J.; Bisk, Y.; and Gao, J. 2021. Taco: Token-aware cascade contrastive learning for video-text alignment. In *ICCV*, 11562–11572.

Yang, X.; Zhu, L.; Wang, X.; and Yang, Y. 2024. DGL: Dynamic Global-Local Prompt Tuning for Text-Video Retrieval. In *AAAI*, volume 38, 6540–6548.

Yang, Z.; Liu, J.; Wu, Z.; Wu, P.; and Liu, X. 2023. Video event restoration based on keyframes for video anomaly detection. In *CVPR*, 14592–14601.

Yu, Y.; Kim, J.; and Kim, G. 218. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, 471–487.

Yuan, T.; Zhang, X.; Liu, K.; Liu, B.; Chen, C.; Jin, J.; and Jiao, Z. 2024. Towards Surveillance Video-and-Language Understanding: New Dataset Baselines and Challenges. In *CVPR*, 22052–22061.

Zeng, D.; Yu, Y.; and Oyama, K. 2020. Deep triplet neural networks with cluster-cca for audio-visual cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(3): 1–23.

Zhang, H.; Xu, X.; Wang, X.; Zuo, J.; Han, C.; Huang, X.; Gao, C.; Wang, Y.; and Sang, N. 2024. Holmes-VAD: Towards Unbiased and Explainable Video Anomaly Detection via Multi-modal LLM. *arXiv preprint arXiv:2406.12235*.

Zhang, Y.; Zhang, H.; Nasrabadi, N. M.; and Huang, T. S. 2013. Multi-metric learning for multi-sensor fusion based classification. *Information Fusion*, 14(4): 431–440.

Zhou, H.; Yu, J.; and Yang, W. 2023. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *AAAI*, volume 37, 3769–3777.

Zhou, L.; Gao, Y.; Zhang, M.; Wu, P.; Wang, P.; and Zhang, Y. 2024. Human-Centric Behavior Description in Videos: New Benchmark and Model. *IEEE Transactions on Multimedia*, 26: 10867–10878.