JoVALE: Detecting Human Actions in Video Using Audiovisual and Language Contexts

Taein Son[1*], **Soo Won Seo**[2*], **Jisong Kim**[1], **Seok Hwan Lee**[1], **Jun Won Choi**[2†]

[1]Hanyang University
[2]Seoul National University
tison@spa.hanyang.ac.kr, swseo@spa.snu.ac.kr, jskim@spa.hanyang.ac.kr, shlee@spa.hanyang.ac.kr, junwchoi@snu.ac.kr

Abstract

Video Action Detection (VAD) entails localizing and categorizing action instances within videos, which inherently consist of diverse information sources such as audio, visual cues, and surrounding scene contexts. Leveraging this multi-modal information effectively for VAD poses a significant challenge, as the model must identify action-relevant cues with precision. In this study, we introduce a novel multi-modal VAD architecture, referred to as the Joint Actor-centric Visual, Audio, Language Encoder (JoVALE). JoVALE is the first VAD method to integrate audio and visual features with scene descriptive context sourced from large-capacity image captioning models. At the heart of JoVALE is the actor-centric aggregation of audio, visual, and scene descriptive information, enabling adaptive integration of crucial features for recognizing each actor's actions. We have developed a Transformer-based architecture, the Actor-centric Multi-modal Fusion Network, specifically designed to capture the dynamic interactions among actors and their multi-modal contexts. Our evaluation on three prominent VAD benchmarks—AVA, UCF101-24, and JHMDB51-21—demonstrates that incorporating multi-modal information significantly enhances performance, setting new state-of-the-art performances in the field.

Code — https://github.com/taeiin/AAAI2025-JoVALE

Introduction

Video action detection (VAD) is a challenging task that aims to localize and classify human actions within video sequences. VAD generates bounding boxes with action scores for a keyframe by analyzing the sequential frames around the keyframe. This task differs from *Action Recognition* task, which classifies the action for a given video clip, and from *Temporal Action Detection* task, which identifies the intervals of particular actions within a video clip.

Humans rely on various sources of information to detect actions, including visual appearance, motion sequences, actor postures, and interactions with their environment. Numerous studies have demonstrated that leveraging such multi-modal information can significantly enhance action recognition performance (Kazakos et al. 2019; Gao et al. 2020; Xiao et al. 2020; Nagrani et al. 2021). Audio, in particular, offers valuable information, providing both direct and indirect contextual cues for action recognition. For example, sounds directly linked to actions, like speech, gunshots, or music, can help identify corresponding actions. Additionally, environmental sounds can indirectly suggest relevant actions, such as the sound of waves indicating beach-related activities. Therefore, incorporating audio data alongside visual data can improve the performance and robustness of VAD. Several action recognition methods have successfully utilized both audio and visual data (Gao et al. 2020; Xiao et al. 2020; Nagrani et al. 2021).

While multi-modal information has shown promise for action recognition tasks, its application in VAD presents significant challenges. Action instances in videos are dispersed across both temporal and spatial dimensions, and the contextual cues necessary for their detection are similarly spread throughout the video. It is crucial to accurately link these actions and contextual features to ensure robust VAD performance. For example, the sound of a piano might help identify a 'playing piano' action but would be irrelevant for detecting a 'talking with others' action within the same scene. Therefore, the piano sound should be selectively used to detect the 'playing piano' action and not the 'talking with others' action. Moreover, effectively integrating multi-modal information from various sources is another key to enhancing VAD performance. Despite its potential, the use of audio-visual information for VAD is still relatively unexplored in current research.

Another valuable resource for VAD identified in this study is the prior general scene-descriptive knowledge gained through vision-language foundation models. Vision Language Pre-training (VLP) models have shown substantial success by leveraging extensive multi-modal data sourced from the web, public databases, and various corpora. These models excel in capturing complex relational structures between text and images, enabling them to adapt to a variety of downstream tasks in either a zero-shot or one-shot manner. With their ability to understand images, the features derived from VLP can significantly enhance VAD performances. In this research, we further investigate a VAD approach that capitalizes on the rich language context provided by VLP models.

[*]These authors contributed equally.
[†]Corresponding author. Email: junwchoi@snu.ac.kr

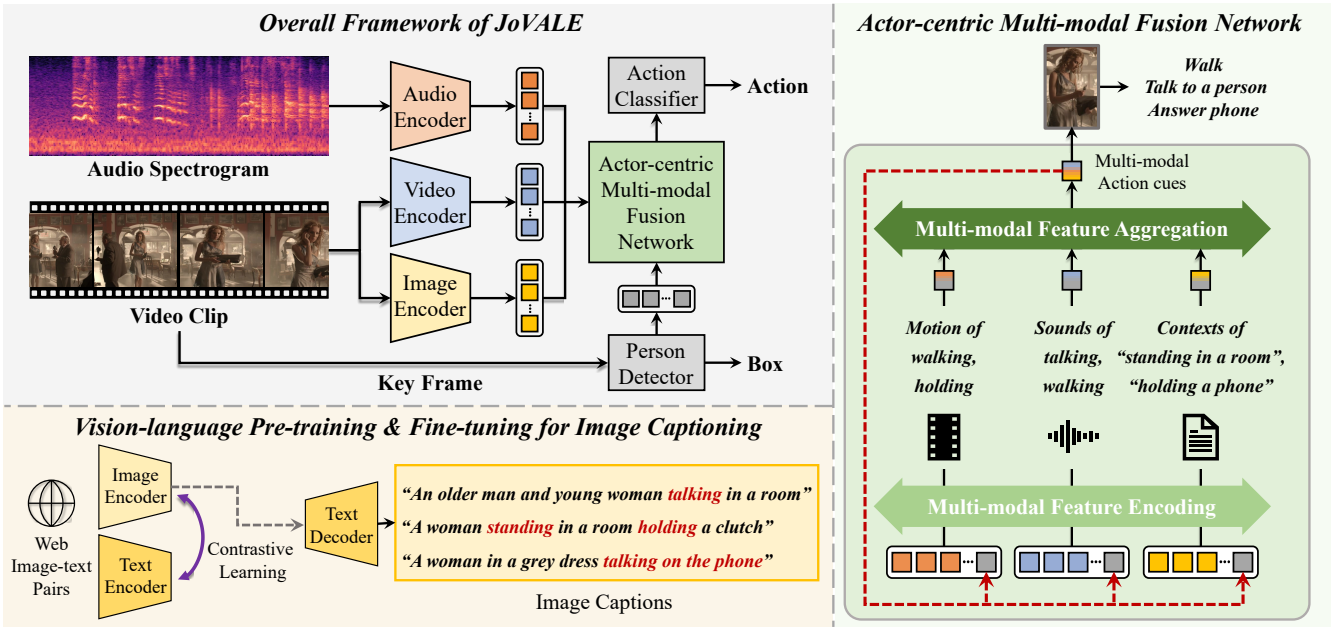


Figure 1: Overview of JoVALE: (top-left) The proposed JoVALE integrates audio, visual, and scene-descriptive features using an AMFN. (bottom-left) JoVALE leverages a VLP model fine-tuned on an image captioning task to generate scene-descriptive features. (right) AMFN encodes high-level interactions between multi-modal features through MFE and MFA.

This paper introduces a novel multi-modal VAD approach referred to as the Joint Actor-centric Visual, Audio, Language Encoder (JoVALE). JoVALE is the first method to leverage audio and visual modalities alongside language context to localize and classify actions in videos. At the core of JoVALE is the actor-centric modeling of multi-modal contextual information.

The key concepts of JoVALE are illustrated in Fig. 1. JoVALE begins by generating densely sampled actor proposal features using an off-the-shelf person detector. These actor proposal features are then processed by the Actor-centric Multi-modal Fusion Network (AMFN), which aggregates relevant contextual information from both audio and visual features. Furthermore, AMFN integrates scene-descriptive knowledge acquired from the VLP model, BLIP (Li et al. 2022a), to enrich the action representation.

To fully leverage multi-modal information for VAD, JoVALE effectively models the relationships among actors, temporal dynamics, and various modalities through AMFN. The AMFN captures their complex interactions through successive updates of Action Embeddings across multiple Transformer layers. It comprises two main components: Multi-modal Feature Encoding (MFE) and Multi-modal Feature Aggregation (MFA). The MFE module jointly encodes Action Embeddings and Multi-modal Context Embeddings for each modality, achieving computational efficiency through the use of Temporal Bottleneck Features. These Temporal Bottleneck Features provide a compact representation of the temporal changes across all actors. Following this, the MFA module aggregates the Action Embeddings from each modality in a weighted fashion. The resulting features are fed into the subsequent Transformer layer for final action detection.

We evaluated JoVALE on three popular VAD benchmarks: AVA (Gu et al. 2018), UCF101-24 (Soomro, Zamir, and Shah 2012), and JHMDB51-21 (Jhuang et al. 2013). By effectively combining audio, visual, and scene-descriptive context information, JoVALE significantly outperforms the baseline on these benchmarks. On the challenging AVA dataset, JoVALE records a mean Average Precision (mAP) of 40.1%, achieving a substantial improvement of 2.4% over the previous best method, EVAD (Chen et al. 2023).

Our contributions can be summarized as follows:

- We present a simple yet effective multi-modal VAD architecture that utilizes the audio-visual information present in videos. Our main approach is Actor-Centric Feature Aggregation, which adaptively attends to the multi-modal context essential for detecting each action instance. There are only a few studies that have explored the use of audio-visual context for VAD.
- We are the first to introduce a VAD approach that incorporates general scene-descriptive knowledge inferred from a Vision Language Foundation model.
- We propose an efficient architecture that effectively models complex relationships among actors, temporal dynamics, and modalities. Our modeling approach differs from existing VAD methods, which typically combine semantic actor features or predicted scores from each modality in a straightforward manner.

Related Work

Video Action Detection

Various VAD methods have been proposed, which can be broadly classified into two main approaches: end-to-end and two-stage methods. End-to-end methods predict both the action location and class simultaneously within a single network. These approaches often utilize a Transformer (Vaswani et al. 2017) to predict the set of actions present in a scene. Notable examples of these end-to-end VAD methods include VTr (Girdhar et al. 2019), TubeR (Zhao et al. 2022), STMixer (Wu et al. 2023), and EVAD (Chen et al. 2023)

In contrast, two-stage VAD methods first utilize a pre-trained person detector to localize the actors before classifying the actions. These two-stage VAD methods include AIA (Tang et al. 2020), ACAR (Pan et al. 2021), and JARViS (Lee et al. 2024). Recently, Vision Transformers (Tong et al. 2022; Wang et al. 2023a,b), pre-trained with Masked Autoencoders (MAE) (He et al. 2022), have shown excellent performance in the context of two-stage VAD.

Muti-modal Video Action Detection

Early multi-modal VAD methods (Gkioxari and Malik 2015; Saha et al. 2016; Zhao and Snoek 2019) leveraged both RGB and optical flow to capture appearance and motion information. Another research direction focused on utilizing human skeletal structures through pose estimation models. For example, JMRN (Shah et al. 2022) extracted individual joint features and captured inter-joint correlations. More recently, HIT (Faure, Chen, and Lai 2023) employed cross-attention mechanisms to capture interactions between key action-related components such as hands, objects, and poses.

Although various video classification methods have utilized both audio and visual information (Gao et al. 2020; Xiao et al. 2020; Nagrani et al. 2021; Gong et al. 2022; Georgescu et al. 2023; Huang et al. 2024), the application of multi-modal information for VAD has not been thoroughly explored. VAD poses unique challenges, as the relevant audio-visual context needed for accurate detection can vary depending on the specific action instance. This study aims to address this gap.

JoVALE Method

Overview

Fig. 1 illustrates the overall structure of JoVALE. The model takes audio samples and image frames as input. Audio and visual backbone features are extracted from these inputs, while scene-descriptive features are obtained using the BLIP image encoder, pre-trained on an image captioning task. Together, these features form the Multi-modal Embeddings $f_a^{(l)}$, $f_v^{(l)}$, and $f_s^{(l)}$ used for the VAD architecture.

JoVALE detects actions through the following steps. Using the keyframe image, an off-the-shelf person detector generates K actor proposals along with their corresponding Region of Interest (RoI) features, referred to as Actor Proposal Features. The AMFN employs a Transformer to aggregate action-related information from the separately encoded Multi-modal Embeddings $f_a^{(l)}$, $f_v^{(l)}$, and $f_s^{(l)}$. To achieve this, the AMFN jointly encodes the action queries $a_a^{(l)}$, $a_v^{(l)}$, and $a_s^{(l)}$ associated with audio, visual, and scene-descriptive modalities, respectively, where l indicates the layer index. In the first layer, these action queries are all initialized using the Actor Proposal Features.

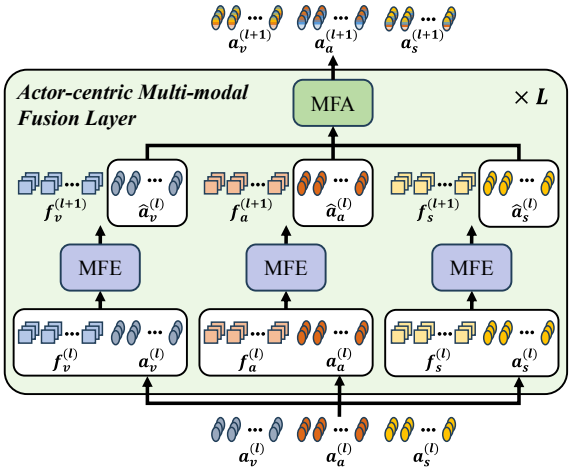


Figure 2: Structure of AMFN: Three independent MFEs encode the context features within each modality. Then, MFA combines Action Embeddings derived from each modality.

The structure of AMFN is depicted in Fig.2. The AMFN comprises two modules: MFE and MFA. For each modality, the MFE module jointly encodes the Action Embeddings, $a_{mod}^{(l)}$ and the Multi-modal Embeddings, $f_{mod}^{(l)}$, producing updated representations $\hat{a}_{mod}^{(l)}$ and $f_{mod}^{(l+1)}$, where $mod \in \mathcal{M} = \{a, v, s\}$. Subsequently, the MFA module employs an adaptive gated fusion mechanism (Kim et al. 2018) to perform a weighted combination of the three Action Embeddings, $\hat{a}_a^{(l)}$, $\hat{a}_v^{(l)}$, and $\hat{a}_s^{(l)}$, resulting in the combined Action Embeddings $a_a^{(l+1)}$, $a_v^{(l+1)}$, and $a_s^{(l+1)}$. These embeddings are then propagated to the next layer. This process is repeated over L iterations, progressively refining the Action Embeddings. Finally, the refined Action Embeddings from the L-th layer are input into a classifier to predict the action instances.

Generation of Multi-modal Features

Visual Embeddings. We encode an input video clip using a video backbone network such as SlowFast (Feichtenhofer et al. 2019) or ViT (Dosovitskiy et al. 2020). This process generates the spatio-temporal visual features $F_v \in \mathbb{R}^{T_v \times H \times W \times C}$, where T_v, H, W, and C represent temporal, height, width, and channel dimensions, respectively. These features are then reshaped into the visual embeddings $f_v \in \mathbb{R}^{T_v \times N_v \times D}$, where $N_v = HW$ and D denotes the embedding dimension.

Audio Embeddings. Following existing audio preprocessing techniques (Gong, Chung, and Glass 2021), we transform the audio waveform samples into a log-mel-

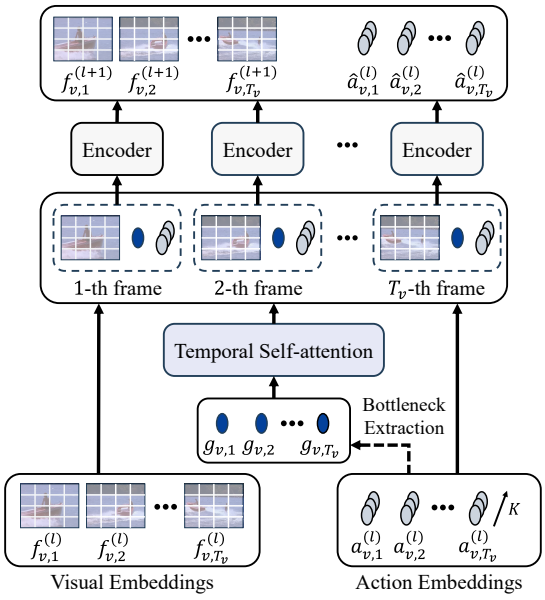


Figure 3: Structure of Multi-modal Feature Encoding. This illustration depicts the process for the visual modality. Identical structures are applied individually to other modalities.

spectrogram in time and frequency bins. This spectrogram is fed into a convolution layer with kernel size $P \times P$ and stride S, then reshaped into a temporal sequence of N_a feature vectors. This process results in audio embeddings $f_a \in \mathbb{R}^{T_a \times N_a \times D}$.

Scene-Descriptive Embeddings. Scene-descriptive features are generated using the BLIP captioner (Li et al. 2022a), a vision-language foundation model that is fine-tuned on an image captioning task. The BLIP captioner takes each image frame as input, encodes it with an image encoder, and produces a text description of the image through a text decoder. Since the output of the image encoder contains high-level semantic scene information that can be readily translated into the text, we can use it as scene-descriptive features. We first uniformly sample T_s image frames from a video clip. Then, we apply the image encoder of the BLIP captioner to each of T_s image frames. The resulting feature maps are then linearly projected into the scene-descriptive embeddings $f_s \in \mathbb{R}^{T_s \times N_s \times D}$.

Actor-centric Multi-modal Fusion Network

AMFN updates Action Embeddings $a_a^{(l)}$, $a_v^{(l)}$, and $a_s^{(l)}$ by applying MFE and MFA in an iterative fashion.

Multi-modal Feature Encoding. The structure of MFE is shown in Fig. 3. The MFE performs the following operation

$$\hat{a}_{mod}^{(l)}, f_{mod}^{(l+1)} = \text{MFE}\left(a_{mod}^{(l)}, f_{mod}^{(l)}\right). \quad (1)$$

Applying self-attention to the combination of $a_{mod}^{(l)}$ and $f_{mod}^{(l)}$ can result in high computational complexity, especially when the number of embeddings is large. To address

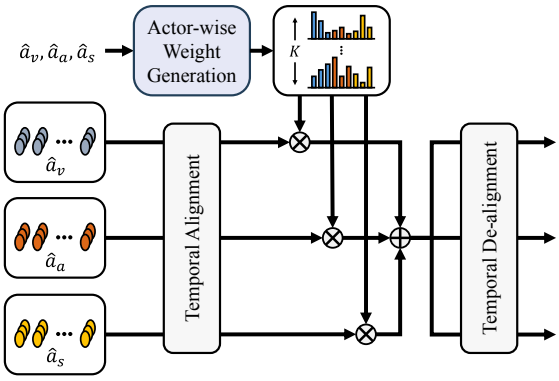


Figure 4: Structure of Multi-modal Feature Aggregation.

this, we generate Temporal Bottleneck Features, which compress the input embeddings across actors at each time step, effectively reducing the computational overhead. By taking $a_{mod}^{(l)} \in \mathbb{R}^{K \times T_{mod} \times D}$ as an input, MFE computes the Temporal Bottleneck Features, $b_{mod}^{(l)} \in \mathbb{R}^{T_{mod} \times D}$ from

$$b_{mod}^{(l)} = \text{SA}(\text{Pool}(a_{mod}^{(l)})), \quad (2)$$

where Pool refers to the average pooling over the actor dimension, and SA denotes the multi-head self-attention. Note that this SA operation encodes the Action Embeddings in the time domain. Finally, the Temporal Bottleneck Features are merged into the Multi-modal Embeddings $f_{mod,t}^{(l)}$ and the Action Embeddings $a_{mod,t}^{(l)}$. Then, MFE jointly encodes the merged embeddings for each time step

$$\hat{a}_{mod,t}^{(l)}, f_{mod,t}^{(l+1)} = \text{Encoder}([a_{mod,t}^{(l)}, f_{mod,t}^{(l)}, b_{mod,t}^{(l)}]), \quad (3)$$

where $t \in [1, T_{mod}]$, Encoder consists of a SA, two normalization layers, and an FFN. Finally, the updated Action Embeddings $\hat{a}_a^{(l)}$, $\hat{a}_v^{(l)}$, and $\hat{a}_s^{(l)}$ are delivered to MFA module.

Multi-modal Feature Aggregation. The structure of MFA is depicted in Fig. 4. The MFA operates as follows

$$a_v^{(l+1)}, a_a^{(l+1)}, a_s^{(l+1)} = \text{MFA}(\hat{a}_v^{(l)}, \hat{a}_a^{(l)}, \hat{a}_s^{(l)}). \quad (4)$$

MFA starts with Temporal Alignment, which aligns time sampling between the Action Embeddings $\hat{a}_v^{(l)}$, $\hat{a}_a^{(l)}$, and $\hat{a}_s^{(l)}$. Following (Cooper 2019), the Action Embeddings of size T_{mod} in time dimension are resized to those of the fixed size T_c. Then, MFA adaptively integrates query features $\hat{a}_v^{(l)}, \hat{a}_a^{(l)}, \hat{a}_s^{(l)}$ using an adaptive gated fusion mechanism (Kim et al. 2018). We compute the combining weights $w_v^{(l)}$, $w_a^{(l)}$, and $w_s^{(l)}$ for each actor. We first concatenate Action Embeddings $\hat{a}_v^{(l)}, \hat{a}_a^{(l)}, \hat{a}_s^{(l)}$ and applies average pooling over the time dimension

$$a_p^{(l)} = \text{Pool}([\hat{a}_v^{(l)} \parallel \hat{a}_a^{(l)} \parallel \hat{a}_s^{(l)}]), \quad (5)$$

where $\parallel$ denotes channel-wise concatenation. Then, we obtain the combining weights by passing the pooled features

Model	Input	Backbone	Pre-train	Val mAP
Models with 3D-CNN backbones				
WOO (Chen et al. 2021)	32×2	SF-R101	K600	28.3
SlowFast (Feichtenhofer et al. 2019)	32×2	SF-R101	K600	29.0
AIA (Tang et al. 2020)	32×2	SF-R101	K700	32.3
ACAR (Pan et al. 2021)	32×2	SF-R101	K700	33.3
TubeR (Zhao et al. 2022)	32×2	CSN-152	K400	33.6
HIT (Faure, Chen, and Lai 2023)	32×2	SF-R101	K700	32.6
STMixer (Wu et al. 2023)	32×2	SF-R101	K700	30.9
JoVALE	32×2	SF-R101	K700	**35.5**
Models with ViT backbones				
VideoMAE (Tong et al. 2022)	16×4	ViT-B	K400	31.8
MViTv2 (Li et al. 2022b)	32×3	MViTv2-B	K700	32.3
MeMViT (Wu et al. 2022)	32×3	MViTv2-B	K700	34.4
MVD (Wang et al. 2023b)	16×4	ViT-B	K400	34.2
STMixer (Wu et al. 2023)	16×4	ViT-B†	K710	36.1
EVAD (Chen et al. 2023)	16×4	ViT-B†	K710	37.7
JoVALE	16×4	ViT-B†	K710	**40.1**

Table 1: Performance comparison evaluated on the AVA 2.2 dataset. ViT-B marked with † is initialized with pre-trained weights from VideoMAE v2 (Wang et al. 2023a).

through a bottleneck MLP followed by a sigmoid function

$$[w_v^{(l)} \parallel w_a^{(l)} \parallel w_s^{(l)}] = \sigma(\text{MLP}(a_p^{(l)})), \tag{6}$$

where MLP consists of two fully connected layers with an activation function and $\sigma(\cdot)$ denotes sigmoid function. Then, the updated Action Embeddings $a_a^{(l+1)}, a_v^{(l+1)}$, and $a_s^{(l+1)}$ are obtained by a weighted summation

$$a_{fuse}^{(l)} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} w_m^{(l)} \otimes \hat{a}_m^{(l)} \tag{7}$$

$$a_{mod}^{(l+1)} = a_{fuse}^{(l)} + w_{mod}^{(l)} \otimes \hat{a}_{mod}^{(l)}, \tag{8}$$

where $\otimes$ denotes the element-wise multiplication.

The updated Action Embeddings $a_a^{(l+1)}, a_v^{(l+1)}, a_s^{(l+1)}$ are converted back to those of their own temporal sampling. After L layers, the classification head is applied to $a_{fuse}^{(L)}$ to predict action scores $c \in \mathbb{R}^{K \times N_{\text{cls}}}$, where N_{cls} denotes the number of target classes and K is the number of actor proposals. These scores, along with the corresponding actor bounding boxes $b \in \mathbb{R}^{K \times 4}$, form a set of action instances (b, c).

Experiments

Datasets and Metrics

We evaluate JoVALE on three standard VAD datasets: AVA (Gu et al. 2018), UCF101-24 (Soomro, Zamir, and Shah 2012), and JHMDB51-21 (Jhuang et al. 2013). AVA consists of 299 15-minute movie clips, with 235 for training and 64 for validation. We evaluate our approach on 60 action classes in AVA v2.2.

UCF101-24, a subset of UCF101, contains 24 sport action classes with 3,207 instances, and our method is evaluated on the first split.

JHMDB51-21, a subset of JHMDB51, includes 928 trimmed video clips spanning 21 action classes.

We report the average performance across the three standard splits of the dataset. The evaluation metric is frame-level mAP at an Intersection over Union (IoU) threshold of 0.5 for all datasets.

Implementation Details

In this section, we describe the implementation details of our proposed JoVALE.

Hyperparameters. The AMFN consists of $L = 6$ Transformer layers. When conducting temporal alignment and dealignment within the MFA module, the temporal dimension T_c is aligned to match that of the visual images, T_v. Following the approach in (Cooper 2019), we apply temporal average pooling when $T_{mod} \geq T_c$ and temporal repetition when $T_{mod} < T_c$. The temporal length of the visual features T_v varies based on the chosen backbone architecture. Specifically, T_v is set to 4 when utilizing the SlowFast (Feichtenhofer et al. 2019) architecture and 8 with the ViT (Dosovitskiy et al. 2020). For audio data, the spectrograms are processed through a convolutional layer with a kernel size of $P = 16$ and a stride of $S = 10$, yielding audio embeddings with a temporal length of $T_a = 20$. Regarding the scene-descriptive features, the number of input frames inputted into BLIP is set at $T_s = 4$. The hyperparameter D, representing the Transformer embedding size, is set to 256.

Generation of Multi-modal Features. Visual features were extracted using one of the following backbones: 1) SlowFast-R50 pre-trained on Kinetics-400 (Kay et al. 2017), 2) SlowFast-R101 pre-trained on Kinetics-700 (Carreira et al. 2019), or 3) ViT-B with pre-trained weights from VideoMAE v2.

Model	Input	Backbone	Pre-train	UCF	JHMDB
AVA	20×1	I3D	K400	76.3	73.3
AIA	32×1	C2D	K400	78.8	-
ACRN	20×1	S3D-G	K400	-	77.9
CARN	32×2	I3D-R50	K400	-	79.2
YOWO	16×1	3D-RX-101	K400	75.7	80.4
WOO	32×2	SF-R101	K600	-	80.5
TubeR	32×2	CSN-152	K400	83.2	-
ACAR*	32×1	SF-R50	K400	84.3	-
HIT*	32×2	SF-R50	K700	84.8	83.8
STMixer	32×2	SF-R101	K700	83.7	86.7
JoVALE	32×2	SF-R101	K700	**84.9**	**91.0**

Table 2: Performance comparison on UCF101-24 and JHMDB51-21. The models marked with $*$ employ YOWO (Köpüklü, Wei, and Rigoll 2019) as a person detector.

Model	Input	Backbone	GFLOPs	mAP
SlowFast	32×2	SF-R101-NL	199 + NA	29.0
ACAR	32×2	SF-R101	160 + NA	31.7
VideoMAE	16×4	ViT-B	180 + NA	31.8
MeMViT	32×3	MViTv2-B	212 + NA	33.5
WOO	32×2	SF-R101-NL	252	28.3
TubeR	32×2	CSN-152	240	32.0
STMixer	16×4	ViT-B†	355	36.1
EVAD	16×4	ViT-B†	243	37.7
JoVALE	16×4	ViT-B†	495	**40.1**
JoVALE@288	16×4	ViT-B†	387	39.8
JoVALE@192	16×4	ViT-B†	314	39.3

Table 3: Computational costs comparison evaluated on AVA 2.2 dataset. **JoVALE@**N indicates $N \times N$ input size for BLIP. 'NA' is the person detector's cost, which was not reported in the corresponding paper.

Audio preprocessing followed the approach in (Gong, Chung, and Glass 2021), where log-mel-spectrograms were extracted from raw audio waveforms. The waveforms, sampled at 16kHz, were converted into 128 Mel-frequency bands using a 25ms Hamming window with a 10ms stride. For an input audio clip of t seconds, this process produced spectrograms of dimension $100t \times 128$.

Scene-descriptive features were extracted using the ViT-B BLIP captioner, which was pre-trained on images from COCO, Visual Genome (Krishna et al. 2017), and web datasets (Changpinyo et al. 2021; Ordonez, Kulkarni, and Berg 2011; Schuhmann et al. 2021), and subsequently fine-tuned on the COCO Caption dataset.

Initializing Action Embeddings. We employed a Faster-RCNN (Ren et al. 2015) with a ResNeXt-101-FPN (Lin et al. 2017; Xie et al. 2017) as a person detector. The detector was pre-trained on ImageNet (Russakovsky et al. 2015) and COCO human keypoint images (Lin et al. 2014) and fine-tuned on each target VAD dataset. The top $K = 15$ actor features were extracted from the penultimate layer based on human confidence scores.

Training. The pre-trained person detector and image captioner were kept frozen during both training and inference. The entire model was trained using sigmoid focal loss for action classification. The AdamW optimizer was employed with a weight decay of 1e-4. Initial learning rates were set to 1e-5 for the video backbone and 1e-4 for the other networks, with a tenfold reduction applied at the 7th epoch. Training was conducted for 8 epochs with a batch size of 16, utilizing four NVIDIA GeForce RTX 3090 GPUs.

For data augmentation, we applied random horizontal flipping to RGB frames. For audio, we utilized SpecAugment (Park et al. 2019) with time and frequency masking, following the approach in AST (Gong, Chung, and Glass 2021).

Main Results

Performance Comparison. We compare JoVALE against existing VAD methods across three widely used datasets. As shown in Table 1, on the AVA dataset, JoVALE outperforms all other methods on the AVA dataset when employing both 3D-CNN and ViT backbones. With a 3D-CNN backbone, JoVALE surpasses the previously leading method, TubeR (Zhao et al. 2022), by 1.9% mAP. When utilizing a ViT backbone, JoVALE establishes a new state-of-the-art on AVA, outperforming EVAD (Chen et al. 2023) by 2.4% mAP.

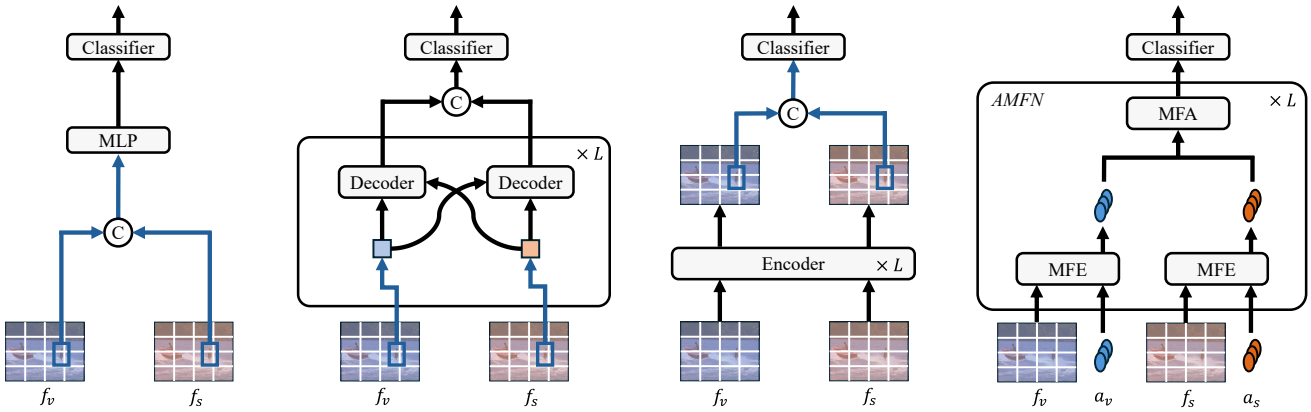
The results evaluated on the UCF101-24 and JHMDB51-21 datasets are shown in Table 2. Here, we used SF-101 as the visual backbone. Given that over 80% of video clips in these datasets lack audio, JoVALE relies solely on visual and scene-descriptive features for input. Even without audio, JoVALE achieves state-of-the-art performance, with mAP scores of 84.9% on UCF101-24 and 91.0% on JHMDB51-21.

Computational Analysis. Table 3 compares the computational costs of JoVALE with those of other methods. JoVALE exhibits higher computational complexity compared to other methods, largely due to its incorporation of audio and video data, along with the use of the BLIP model. However, the increased complexity is justified by JoVALE's superior performance and remains within a reasonable range. It has been observed that reducing the input resolution of BLIP can lower computational costs, albeit at the expense of a slight decrease in performance.

Ablation Study

Our ablation studies were conducted on AVA v2.2 using the SlowFast-R50 configuration. Unless stated otherwise, all other settings were consistent with the main experiments. Detailed model configurations used in these ablations are provided in the *Supplementary Material*.

Multi-modalities. In Table 4, we assess the performance of JoVALE using various modality combinations. Using only the video modality, we achieve the highest mAP of 28.0%, highlighting its essential role in VAD. In contrast, relying solely on audio results in significantly lower perfor-



(a) RoI feature fusion (b) RoI relation modeling (c) Global relation modeling (d) JoVALE

Figure 5: Different multi-modal fusion strategies: The symbol ⓒ denotes the channel-wise concatenation.

Method	Modality			mAP
	Video	Audio	Scene-desc.	
Uni-modal	✓			28.0
		✓		11.5
			✓	26.9
Multi-modal	✓	✓		28.6
	✓		✓	32.7
		✓	✓	27.8
	✓	✓	✓	**34.0**

Table 4: Performance evaluated when various combinations of modalities are used.

Methods	mAP
RoI feature fusion	29.1
RoI relation modeling	30.4
Global relation modeling	32.1
JoVALE	**32.7**

Table 5: Performance of different multi-modal fusion strategies.

mance, illustrating its limitations as an independent modality.

By integrating video with scene descriptive embeddings, a notable improvement to an mAP of 32.7% is observed. This enhancement underscores the effectiveness of combining visual and scene-descriptive contexts to boost VAD performance. Notably, when all three modalities—audio, video, and scene descriptive—are utilized together, JoVALE achieves the highest mAP of 34.0%. This underscores the benefits of leveraging the complementary nature of diverse modalities in action detection.

Multi-modal Fusion Strategies. In Table 5, we compare the performance of various multi-modal fusion strategies commonly used in VAD, with Fig. 5 illustrating the different strategies considered. Specifically, we evaluate (a) RoI feature fusion (Gkioxari and Malik 2015), (b) RoI relation modeling (Faure, Chen, and Lai 2023), and (c) Global relation modeling. For a fair comparison, audio features were excluded from the fusion process.

RoI feature fusion (Gkioxari and Malik 2015) directly combines RoI actor features from each modality for action classification. RoI relation modeling (Faure, Chen, and Lai 2023) uses cross-attention to capture relationships among RoI features from different modalities. In global relation modeling, a transformer encoder is utilized to capture holistic dependencies across multi-modal embeddings. Notably, the feature fusion strategy employed in JoVALE achieves the best performance among all evaluated strategies.

MFE Structure. We explored various MFE structures for spatio-temporal feature extraction, with results presented in Table 6. We first established a baseline using joint space-time attention that encodes spatio-temporal features within a single encoder. While achieving 33.8 mAP, this approach incurs substantial computational overhead. Factorized encoder (Arnab et al. 2021) first extracts features from individual frames, then captures temporal relationships between them. While this sequential encoding reduces complexity, it yields lower performance at 30.6 mAP. Divided space-time attention (Bertasius, Wang, and Torresani 2021), which applies temporal and spatial attention separately within each Transformer layer, achieves 33.1 mAP with 31.2 GFLOPs. Cross-frame attention (Ni et al. 2022), which uses randomly initialized tokens for inter-frame information exchange, resulting in a 1.4 mAP decrease compared to the baseline. In contrast, our MFE leverages bottleneck features derived from actor features, enabling the exchange of crucial actor-centric information. By focusing on actor-relevant information, our method effectively balances the trade-off between model complexity and performance, achieving superior performance with 34.0 mAP and maintaining computational efficiency at 25.4 GFLOPs.

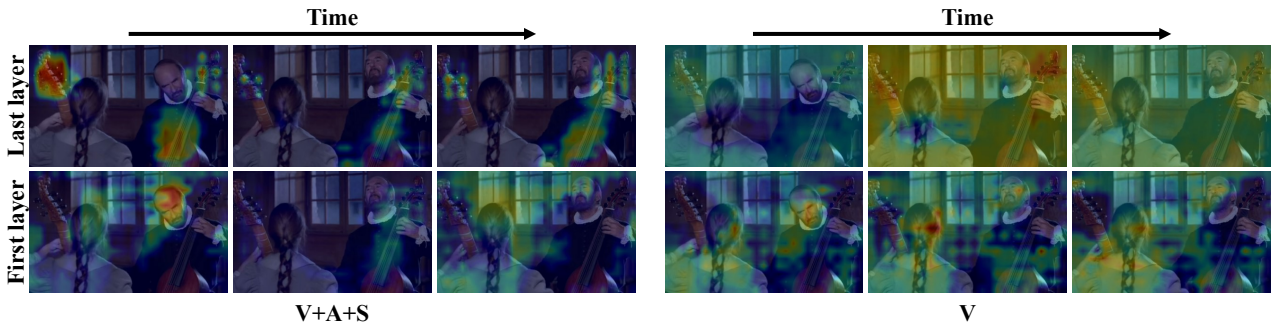


Figure 6: Visualization of activation maps: The left side displays heatmaps when using audio, visual, and scene-descriptive features for JoVALE, while the right side shows heatmaps based solely on visual features.

Multi-modal Feature Encoding	GFLOPs	mAP
Joint space-time attention (Baseline)	41.7	33.8
Factorized encoder	23.0	30.6
Divided space-time attention	31.2	33.1
Cross-frame attention	24.4	32.4
MFE	25.4	**34.0**

Table 6: Performance of different spatio-temporal feature encoding approaches.

Methods	mAP
Late score fusion	29.4
Uniform weighted fusion	31.9
Adaptive Gated Fusion	**34.0**

Table 7: Effect of adaptive gated fusion in MFA.

Effect of Adaptive Gated Fusion in MFA. Table 7 compares the performance of the model with Adaptive Gated Fusion enabled against other prevalent multi-modal fusion techniques. Initially, we established a baseline using Late score fusion, which achieves an mAP of 29.4. We then experimented with the case where action embeddings are combined using equal weights, yielding an mAP of 31.9. We confirm that Adaptive Gated Fusion provides a 2.1% mAP improvement over the baseline.

Qualitative Results

Fig. 6 compares the activation maps when using visual, audio, and scene-descriptive features together versus using only visual features. The top row shows activation maps from the first layer, while the bottom row presents maps from the final layer. With multi-modal input, JoVALE demonstrates improved localization of regions of interest compared to using visual input alone. Notably, JoVALE successfully focuses on a cello when both visual and audio data are provided but fails to do so with visual input only. These results highlight the interactions between visual and audio features, which leads to better extraction of visual cues.

Conclusions

In this paper, we introduced JoVALE, a multi-modal VAD network that effectively extracts audio, visual, and scene-descriptive contexts from the input. JoVALE selectively integrates critical information from each modality to detect various actions within a scene. Built on a Transformer architecture, JoVALE attends to features from each modality using actor features, identified by a person detector, as queries. The AMFN module facilitates computationally efficient modeling of high-level relationships among actors and the temporal dynamics across different modalities. It jointly encodes visual, audio, and scene-descriptive embeddings through the MFE and aggregates them with adaptive weights for each actor through the MFA. Evaluations on challenging VAD benchmarks demonstrate that JoVALE achieves state-of-the-art performance, significantly outperforming existing VAD methods by notable margins.

While we have utilized fine-tuned image captioning models to extract scene-context information, exploring the potential of generic pre-trained VLMs to further enhance VAD is an exciting direction. These models could offer a high-level understanding of a scene, which may significantly boost VAD performance. This enhancement could be achieved by designing effective prompting strategies and integrating tokens generated by foundation models into VAD architectures. We plan to pursue this line of research in future work.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No.2020R1A2C2012146), and the Technology Innovation Program (RS-2024-00468747, Development of AI and Lightweight Technology for Embedding Multisensory Intelligence Modules) funded By the Ministry of Trade Industry & Energy (MOTIE, Korea).

References

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6836–6846.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *ICML*, volume 2, 4.

Carreira, J.; Noland, E.; Hillier, C.; and Zisserman, A. 2019. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3558–3568.

Chen, L.; Tong, Z.; Song, Y.; Wu, G.; and Wang, L. 2023. Efficient video action detection with token dropout and context refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10388–10399.

Chen, S.; Sun, P.; Xie, E.; Ge, C.; Wu, J.; Ma, L.; Shen, J.; and Luo, P. 2021. Watch only once: An end-to-end video action detection framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8178–8187.

Cooper, A. 2019. Hear me out: hearing each other for the first time: the implications of cochlear implant activation. *Missouri medicine*, 116(6): 469.

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*.

Faure, G. J.; Chen, M.-H.; and Lai, S.-H. 2023. Holistic interaction transformer network for action detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3340–3350.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6202–6211.

Gao, R.; Oh, T.-H.; Grauman, K.; and Torresani, L. 2020. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10457–10467.

Georgescu, M.-I.; Fonseca, E.; Ionescu, R. T.; Lucic, M.; Schmid, C.; and Arnab, A. 2023. Audiovisual masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16144–16154.

Girdhar, R.; Carreira, J.; Doersch, C.; and Zisserman, A. 2019. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 244–253.

Gkioxari, G.; and Malik, J. 2015. Finding action tubes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 759–768.

Gong, Y.; Chung, Y.-A.; and Glass, J. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*.

Gong, Y.; Rouditchenko, A.; Liu, A. H.; Harwath, D.; Karlinsky, L.; Kuehne, H.; and Glass, J. 2022. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839*.

Gu, C.; Sun, C.; Ross, D. A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6047–6056.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Huang, P.-Y.; Sharma, V.; Xu, H.; Ryali, C.; Li, Y.; Li, S.-W.; Ghosh, G.; Malik, J.; Feichtenhofer, C.; et al. 2024. Mavil: Masked audio-video learners. *Advances in Neural Information Processing Systems*, 36.

Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; and Black, M. J. 2013. Towards understanding action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 3192–3199.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kazakos, E.; Nagrani, A.; Zisserman, A.; and Damen, D. 2019. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5492–5501.

Kim, J.; Koh, J.; Kim, Y.; Choi, J.; Hwang, Y.; and Choi, J. W. 2018. Robust deep multi-modal learning based on gated information fusion network. In *Asian Conference on Computer Vision*, 90–106. Springer.

Köpüklü, O.; Wei, X.; and Rigoll, G. 2019. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. *arXiv preprint arXiv:1911.06644*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73.

Lee, S. H.; Son, T.; Seo, S. W.; Kim, J.; and Choi, J. W. 2024. JARViS: Detecting actions in video using unified actor-scene context relation modeling. *Neurocomputing*, 610: 128616.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Li, Y.; Wu, C.-Y.; Fan, H.; Mangalam, K.; Xiong, B.; Malik, J.; and Feichtenhofer, C. 2022b. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4804–4814.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; and Sun, C. 2021. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34: 14200–14213.

Ni, B.; Peng, H.; Chen, M.; Zhang, S.; Meng, G.; Fu, J.; Xiang, S.; and Ling, H. 2022. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, 1–18. Springer.

Ordonez, V.; Kulkarni, G.; and Berg, T. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

Pan, J.; Chen, S.; Shou, M. Z.; Liu, Y.; Shao, J.; and Li, H. 2021. Actor-context-actor relation network for spatio-temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 464–474.

Park, D. S.; Chan, W.; Zhang, Y.; Chiu, C.-C.; Zoph, B.; Cubuk, E. D.; and Le, Q. V. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Saha, S.; Singh, G.; Sapienza, M.; Torr, P. H.; and Cuzzolin, F. 2016. Deep learning for detecting multiple space-time action tubes in videos. *arXiv preprint arXiv:1608.01529*.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Shah, A.; Mishra, S.; Bansal, A.; Chen, J.-C.; Chellappa, R.; and Shrivastava, A. 2022. Pose and joint-aware action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3850–3860.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Tang, J.; Xia, J.; Mu, X.; Pang, B.; and Lu, C. 2020. Asynchronous interaction aggregation for action detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, 71–87. Springer.

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35: 10078–10093.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, L.; Huang, B.; Zhao, Z.; Tong, Z.; He, Y.; Wang, Y.; Wang, Y.; and Qiao, Y. 2023a. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14549–14560.

Wang, R.; Chen, D.; Wu, Z.; Chen, Y.; Dai, X.; Liu, M.; Yuan, L.; and Jiang, Y.-G. 2023b. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6312–6322.

Wu, C.-Y.; Li, Y.; Mangalam, K.; Fan, H.; Xiong, B.; Malik, J.; and Feichtenhofer, C. 2022. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13587–13597.

Wu, T.; Cao, M.; Gao, Z.; Wu, G.; and Wang, L. 2023. Stmixer: A one-stage sparse action detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14720–14729.

Xiao, F.; Lee, Y. J.; Grauman, K.; Malik, J.; and Feichtenhofer, C. 2020. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1492–1500.

Zhao, J.; and Snoek, C. G. 2019. Dance with flow: Two-in-one stream action detection. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 9935–9944.

Zhao, J.; Zhang, Y.; Li, X.; Chen, H.; Shuai, B.; Xu, M.; Liu, C.; Kundu, K.; Xiong, Y.; Modolo, D.; et al. 2022. Tuber: Tubelet transformer for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13598–13607.