MULTIMODAL OPEN-VOCABULARY VIDEO CLASSIFICATION VIA VISION AND LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Utilizing vision and language models (VLMs) pre-trained on internet-scale image-text pairs is becoming a promising paradigm for open-vocabulary vision tasks. This work conducts an extensive study for multimodal open-vocabulary video classification via pre-trained VLMs by leveraging motion and audio that naturally exist in the video. We design an asymmetrical cross-modal fusion mechanism to aggregate multimodal information differently for video and optical flow / audio. Experiments on Kinetics and VGGSound show that introducing more modalities significantly improves the accuracy on seen classes, while generalizing better to unseen classes over existing approaches. Despite its simplicity, our method achieves state-of-the-art results on UCF and HMDB zero-shot video action recognition benchmarks, significantly outperforming traditional zero-shot techniques, video-text pre-training methods and recent VLM-based approaches. Code and models will be released.

1 INTRODUCTION

Building open-vocabulary models capable of predicting beyond a fixed set of training classes is of crucial importance in computer vision. Recently, vision and language models (VLMs) pre-trained on internet-scale image-text pairs, *e.g.*, CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), demonstrate impressive transferability on a wide range of vision tasks. Utilizing strong pre-trained VLMs is becoming a promising paradigm for open-vocabulary vision tasks including object detection (Gu et al., 2022) and image segmentation (Ghiasi et al., 2021; Li et al., 2022a).

In this work, we focus on the novel challenging task of multimodal open-vocabulary video classification via pre-trained VLMs. We set up open-vocabulary video benchmarks by utilizing two existing large-scale multimodal video datasets: Kinetics-700 (Carreira et al., 2019) and VGGSound (Chen et al., 2020). Specifically, we constructed two sets of classes: base (seen) and novel (unseen). For base classes, we have both training and testing videos, aiming at helping the pre-trained VLMs adapt to the video domain. While for novel classes, we only have testing videos, mimicking the real-world challenge of open-vocabulary video classification. To the best of our knowledge, we are the first to study how to leverage pre-trained VLMs for multimodal open-vocabulary video classification.

We start with directly fine-tuning the vision encoder of CLIP (Radford et al., 2021) with the language encoder fixed, using the training videos from base classes. As shown in Fig. 1 (a-d), although there is a decent performance gain for base classes, the accuracy for novel classes decreases significantly. This observation corroborates with Zhou et al. (2022) on adapting pre-trained VLMs.

On the other hand, despite rich multimodal contents in internet videos, signals such as audio and motion are less explored in recent open-vocabulary models. This is in stark contrast to the human perception system that heavily relies on multimodal signals (Smith & Gasser, 2005). Can we leverage multimodal information to improve open-vocabulary models?

Instead of using specially designed modality-specific encoders (Wang et al., 2016; Hershey et al., 2017), we choose a more straightforward path by directly utilizing the pre-trained vision encoder from VLMs with minimal modifications to deal with optical flow and audio spectrogram.

We then conduct the same experiments by fine-tuning CLIP's vision encoder but instead using flow or audio as the input. As shown in Fig. 1 (e-h), surprisingly, we find that fine-tuning on base classes

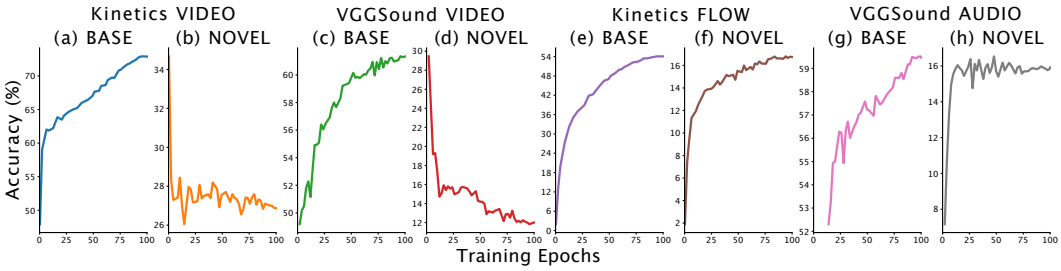


Figure 1: **Fine-tuning pre-trained CLIP with video, flow and audio modalities.** For all three modalities, fine-tuning on labeled base classes leads to significant accuracy improvement (a, c, e, g). However, when evaluating the same model on novel classes, the video modality shows decreasing performance (b, d), while the performance for both flow and audio modality is improving (f, h).

is able to also improve the performance on novel classes. This suggests that we may use flow or audio modality to improve the base to novel generalization of video modality.

In light of these observations, we propose **MOV**, a simple yet effective method for **M**ultimodal **O**pen-**V**ocabulary video classification. Fig. 2 shows an overview of our method. In MOV, we design a novel asymmetrical cross-modal fusion mechanism using cross-attention to leverage complementary multimodal information differently for video and optical flow / audio. The core idea is to exploit the strong transferability in the pre-trained vision encoder, while allowing greater flexibility in fine-tuning flow and audio encoders. MOV is trained using multimodal inputs from base classes and is able to predict both base and novel classes during inference.

We carry out extensive experiments and ablation studies on Kinetics-700 (Carreira et al., 2019) and VGGSound (Chen et al., 2020). MOV shows clear improvements over CLIP (Radford et al., 2021), recent CLIP adaptation techniques (Zhou et al., 2021; Gao et al., 2021), as well as video-text pre-training methods (Akbari et al., 2021) on both base and novel classes. MOV also achieves state-of-the-art results on UCF and HMDB zero-shot video action recognition benchmarks, significantly outperforming traditional zero-shot methods, state-of-the-art VLM adaption techniques, and a variety of video-text pre-training approaches. Furthermore, MOV is scalable with much stronger backbones, indicating its potential to be incorporated with large vision and language models.

2 Related Work

Vision and language models. Learning a joint embedding space from vision and language modalities has been extensively studied during the past decade. Early works usually first encode two modalities separately, using hand-crafted descriptors (Elhoseiny et al., 2013) or deep networks (Lei Ba et al., 2015) for images, and skip-gram text models for languages (Frome et al., 2013). The cross-modality alignment is then achieved by metric learning (Frome et al., 2013) or language concepts (Li et al., 2017). Recently, learning vision and language modalities jointly through contrastive learning (Hadsell et al., 2006; Oord et al., 2018) becomes a promising direction. Impressive performance has been achieved by utilizing stronger encoders for vision (Dosovitskiy et al., 2021), language (Vaswani et al., 2017) and web-scale pre-training data (Hinton et al., 2015; Radford et al., 2021). CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) are two representative approaches which show strong zero-shot [1] performance on various downstream tasks. Despite this strong baseline, adapting pre-trained VLMs to specific vision domains in a more effective way remains critical and is being actively studied. Examples abound, including image classification (Zhou et al., 2021; 2022; Gao et al., 2021), object detection (Gu et al., 2022; Zhong et al., 2022; Kamath et al., 2021; Li et al., 2022b), image segmentation (Ghiasi et al., 2021; Li et al., 2022a), audio classification (Guzhov et al., 2022) and video action recognition (Wang et al., 2021; Ju et al., 2021; Ni et al., 2022). Our method extends the existing research by adapting pre-trained VLMs to multimodal video and investigating the impact of additional input modalities like flow and audio.

[1] We use the term "zero-shot" when we need to align with settings described in some existing works. Otherwise, we would use "open-vocabulary" which we believe is a more precise term.

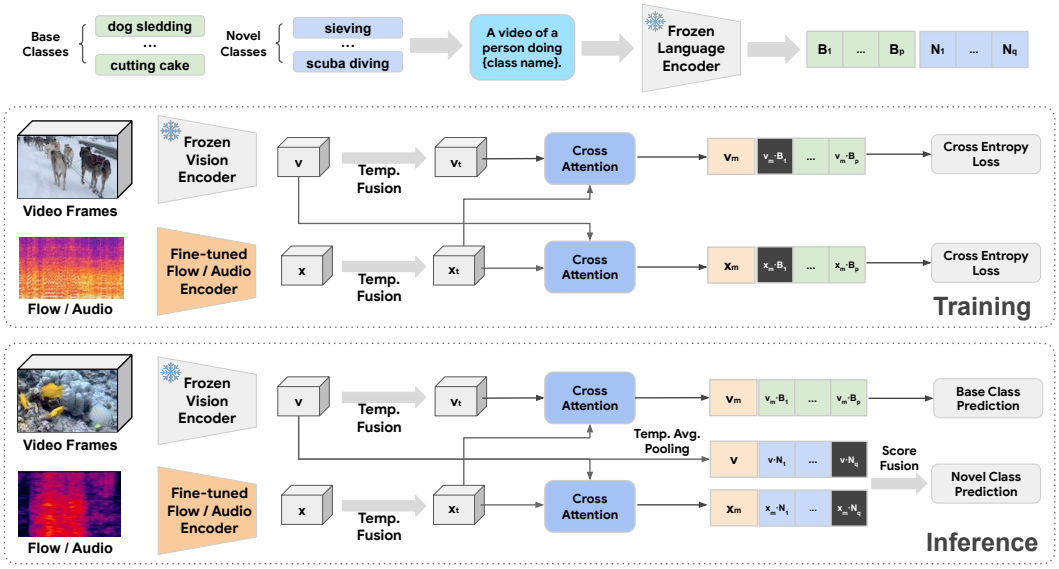


Figure 2: **Overview of the proposed multimodal open-vocabulary (MOV) method**. We use the same encoder architecture from the pre-trained vision and language model to encode the video frames, optical flow, and audio spectrogram. We then apply a transformer head for temporal fusion. We design an asymmetrical cross-attention mechanism for fusion across modalities. During training, we optimize different modalities simultaneously via maximizing their similarities with the corresponding text embeddings. During inference, we use different paths to leverage video and audio / flow modalities for base and novel class prediction.

Open-vocabulary video classification. Zero-shot or open-vocabulary video action recognition is a representative task in this domain. Similar to early works of vision and language learning, the video input and labeled texts are encoded with modality-specific pre-trained models such as S3D (Xie et al., 2018), R(2+1)D (Tran et al., 2018) for video, and Word2Vec (Mikolov et al., 2013) for text. Since the generated video and text embeddings are not aligned, various methods have been proposed to bridge the gap by mapping two modalities into a joint embedding space (Wang & Chen, 2017; Chen & Huang, 2021; Gao et al., 2019; Wu et al., 2016; Xu et al., 2016; Zhu et al., 2018), mapping vision modality to language space (Bishay et al., 2019; Brattoli et al., 2020; Hahn et al., 2019; Xu et al., 2017) or mapping language modality to vision space (Mandal et al., 2019; Zhang & Peng, 2018). These joint embedding mapping methods are further extended to audiovisual classification (Mercea et al., 2022; Mazumder et al., 2021; Parida et al., 2020). Our approach shows that we can significantly improve the performance of open-vocabulary video classification by leveraging strong pre-trained VLMs and other modalities like flow and audio. To our knowledge, this has not been done by prior works in this field.

Mutlimodal fusion for video. Videos are a natural source of multimodal data including motion and audio. Two-stream networks is used to model video and optical flow simultaneously for action recognition (Simonyan & Zisserman, 2014; Wang et al., 2016; Feichtenhofer et al., 2016; 2017). Late fusion is adopted (Simonyan & Zisserman, 2014; Wang et al., 2016) and then thoroughly studied (Feichtenhofer et al., 2016; 2017) on how to better perform spatio-temporal fusion from two modalities. As in the domain of audiovisual fusion, early methods (Chen & Rao, 1998) usually adopt straightforward score fusion or stacking input data for early fusion. Later research (Kazakos et al., 2019; Xiao et al., 2020; Fayek & Kumar, 2020; Nagrani et al., 2021; Chen & Ho, 2022; Chen et al., 2021; Zhao et al., 2022) focus on developing better mid or late fusion strategies to improve the final performance. Different from existing works focusing on a fixed set of classes, we use multimodal fusion to help open-vocabulary video models generalize better to novel classes.

3 METHOD

An overview of our proposed method is shown in Fig. 2. We next describe each component.

3.1 MODALITY-SPECIFIC ENCODING

Given a pre-trained vision and language model, *e.g.*, CLIP (Radford et al., 2021), we denote its vision encoder as $h(\cdot|\theta_h)$ and its language encoder as $g(\cdot|\theta_g)$. For a multimodal video input, we sample N RGB frames V and calculate the corresponding optical flow images F, resulting in $V = \{v_1, v_2, \ldots, v_N\}$ and $F = \{f_1, f_2, \ldots, f_N\}$. We also generate the spectrogram image A from the raw audio waveform. More implementation details can be found in Sec. 4. We use the same encoder architecture $h(\cdot|\cdot)$ to extract feature representations for video, flow and audio modalities, denoted as $h_v(\cdot|\theta_v)$, $h_f(\cdot|\theta_f)$, and $h_a(\cdot|\theta_a)$ respectively. Model parameters θ_v, θ_f and θ_a are all initialized with the weight θ_h from the pre-trained VLM's vision encoder. Apart from being simple and easy to implement, this design has two additional advantages: 1) the performance of adopting the pre-trained VLM's vision encoder to other modalities is competitive against in-domain methods (a detailed study in Appendix A); 2) the vision encoder is trained to align with the language encoder, potentially helping the generalization from base to novel classes. We encode each modality separately as:

$$\mathbf{v} = h_v(V|\theta_v),\ \mathbf{f} = h_f(F|\theta_f),\ \mathbf{a} = h_a(A|\theta_a), \tag{1}$$

where $\mathbf{v}$ and $\mathbf{f}$ are features from N frames, and $\mathbf{a}$ is the representation of a single spectrogram image.

To better aggregate the temporal features of video and flow modalities, we attach temporal fusion networks $\phi_v(\cdot)$ and $\phi_f(\cdot)$, consisting of L transformer layers each, on top of $h_v(\cdot|\theta_v)$ and $h_f(\cdot|\theta_f)$. We denote the input of the l-th transformer layer as $\mathbf{z}^l$ and the input $\mathbf{z}^0$ can be either $\mathbf{v}$ or $\mathbf{f}$. Then the forward pass of the l-th layer in $\phi_v(\cdot)$ and $\phi_f(\cdot)$ can be formulated as:

$$\mathbf{y}^l = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}^l)) + \mathbf{z}^l, \tag{2}$$

$$\mathbf{z}^{l+1} = \mathrm{MLP}(\mathrm{LN}(\mathbf{y}^l)) + \mathbf{y}^l, \tag{3}$$

where LN stands for layer normalization, MSA represents multi-head self-attention, and MLP means multi-layer perceptron. For audio feature $\mathbf{a}$, we simply attach an MLP module upon the backbone. We obtain the temporally fused features as:

$$\mathbf{v_t} = \phi_v(\mathbf{v}),\ \mathbf{f_t} = \phi_f(\mathbf{f}),\ \mathbf{a_t} = \mathrm{MLP}(\mathbf{a}). \tag{4}$$

Finally, for the text modality, suppose we have p base classes with labels. We fill each of the class names into 28 video classification prompts provided by CLIP (Radford et al., 2021) like "a video of a person doing {class name}" and then encode the sentence using the pre-trained language encoder $g(\cdot|\theta_g)$ from VLM. The embedding of each class is averaged over all templates denoted as $\{\mathbf{B}_i\}_{i=1}^{p}$.

3.2 ASYMMETRICAL MULTIMODAL FUSION

We adopt an asymmetrical cross-attention mechanism to fuse mutlimodal features. Without loss of generality, as shown in Fig. 2, our method described here is for fusing one of {flow, audio} modality with video modality. The algorithm can be easily extended to fusing video with more modalities.

For the video modality, we extract the information from other modalities to enhance the performance of video feature. Thus we use $\mathbf{v_t}$ as the input for attention query, and $\mathbf{f_t}$ or $\mathbf{a_t}$ from the other modality as the input for attention key and value. The fused multimodal video feature $\mathbf{v_m}$ can be written as:

$$\mathbf{v_t} = \mathrm{MCA}(\mathrm{LN}(\mathbf{v_t}), \mathrm{LN}(\mathbf{x_t})) + \mathbf{v_t},\ \ \mathbf{x_t} \in \{\mathbf{f_t}, \mathbf{a_t}\}, \tag{5}$$

$$\mathbf{v_m} = \mathrm{AvgPool}\big(\mathrm{MLP}(\mathrm{LN}(\mathbf{v_t})) + \mathbf{v_t}\big), \tag{6}$$

where MCA denotes multi-head cross-attention, AvgPool denotes temporal average pooling.

For the audio and flow modalities, we adopt an asymmetrical design aiming at incorporating the information from video modality to enhance the *generalization ability* of the feature to novel classes. Since the video temporal fusion network $\phi_v(\cdot)$ for generating the video feature $\mathbf{v_t}$ are trained on base classes, $\mathbf{v_t}$ losses the generalization ability to novel classes (shown in Fig. 1). Therefore we choose to directly use the frozen vision encoder's output $\mathbf{v}$ instead of $\mathbf{v_t}$ for better generalization to novel classes. We obtain the fused multimodal flow and audio feature $\mathbf{f_m}$ and $\mathbf{a_m}$ as:

$$\mathbf{f_t} = \mathrm{MCA}(\mathrm{LN}(\mathbf{f_t}), \mathrm{LN}(\mathbf{v})) + \mathbf{f_t}, \qquad \mathbf{a_t} = \mathrm{MCA}(\mathrm{LN}(\mathbf{a_t}), \mathrm{LN}(\mathbf{v})) + \mathbf{a_t}, \tag{7}$$

$$\mathbf{f_m} = \mathrm{AvgPool}\big(\mathrm{MLP}(\mathrm{LN}(\mathbf{f_t})) + \mathbf{f_t}\big),\ \ \mathbf{a_m} = \mathrm{AvgPool}\big(\mathrm{MLP}(\mathrm{LN}(\mathbf{a_t})) + \mathbf{a_t}\big). \tag{8}$$

3.3 TRAINING AND INFERENCE ON BASE CLASSES

During training, each input multimodal video has a corresponding label y belonging to the base classes. We would optimize different modalities simultaneously via maximizing the video-text, flow-text and audio-text similarity. The training loss function can be formulated as:

$$\mathcal{L} = \alpha(-\log \frac{\exp(\text{sim}(\mathbf{v_m}, \mathbf{B}_y)/\tau)}{\sum_{i=1}^{p} \exp(\text{sim}(\mathbf{v_m}, \mathbf{B}_i)/\tau)}) + (1-\alpha)(-\log \frac{\exp(\text{sim}(\mathbf{x_m}, \mathbf{B}_y)/\tau)}{\sum_{i=1}^{p} \exp(\text{sim}(\mathbf{x_m}, \mathbf{B}_i)/\tau)}), \quad (9)$$

where $\mathbf{x_m} \in \{\mathbf{f_m}, \mathbf{a_m}\}$ is the final fused flow or audio feature, α is the weight for balancing two loss terms, $\text{sim}(\cdot, \cdot)$ is the cosine similarity, τ is a pre-defined temperature parameter. During training, we freeze the video encoder and the text encoder to retain their strong generalization to novel classes and save computation, while for other two modalities flow and audio, we fine-tune the encoder end-to-end. An ablation study on fine-tuning different number of layers can be found in Tab. 6.

For inference on base classes, we compute the probability belonging to the j-th class by:

$$P(j) = \frac{\exp(\text{sim}(\mathbf{v_m}, \mathbf{B}_j)/\tau)}{\sum_{i=1}^{p} \exp(\text{sim}(\mathbf{v_m}, \mathbf{B}_i)/\tau)}, \; j \in \{1, 2, \ldots, p\}. \quad (10)$$

3.4 GENERALIZATION TO NOVEL CLASSES

Similar to base classes, we obtain the text embeddings for novel classes as $\{\mathbf{N}_i\}_{i=1}^{q}$, where q is the number of novel classes. In addition to fused features $\mathbf{f_m}$ or $\mathbf{a_m}$, we also incorporate the video feature $\mathbf{v}$ extracted from the frozen video encoder, followed by a temporal average pooling. Similar to Eq. 10, we compute the probability predictions as (here we only show flow modality for simplicity):

$$P_f(j) = \frac{\exp(\text{sim}(\mathbf{f_m}, \mathbf{N}_j)/\tau_f)}{\sum_{i=1}^{q} \exp(\text{sim}(\mathbf{f_m}, \mathbf{N}_i)/\tau_f)}, \; P_v(j) = \frac{\exp(\text{sim}(\mathbf{v}, \mathbf{N}_j)/(\tau_v)}{\sum_{i=1}^{q} \exp(\text{sim}(\mathbf{v}, \mathbf{N}_i)/(\tau_v))}, \; j \in \{1, 2, \ldots, q\}. \quad (11)$$

We denote the probability distribution followed by $\{p_f(j)|_{j=1}^{q}\}$ and $\{p_v(j)|_{j=1}^{q}\}$ as D_f and D_v. In our experiments we find the curve of D_v tends to be much flatter (or have higher information entropy) than D_f when the temperatures τ_v and τ_f are both set to the CLIP's default value of 0.01, resulting in poor performance. We find simply lowering τ_v to 0.003 while keeping τ_f and τ_a as 0.01 solves this issue. A detailed ablation study about the temperature can be found in Appendix B.

The final probability predictions for novel classes are calculated by a weighted sum:

$$P(j) = \beta P_f(j) + (1-\beta) P_v(j). \quad (12)$$

4 EXPERIMENTS

4.1 DATASETS

We describe the details of dataset splits for benchmarking multimodal open-vocabulary video classification and preparing flow and audio modalities.

Kinetics-700 (Carreira et al., 2019) splits. Kinetics-700 contains around 650k video clips annotated with 700 human action classes. Apart from the visual appearance, motion plays an important role for distinguishing different action classes. For dataset split, we randomly select 400 classes as base classes and the testing videos of the rest 300 classes are used for novel class evaluation.

Kinetics-700 optical flow. We follow a standard procedure (Xie et al., 2018; Han et al., 2020a;b) to use the TV-L1 algorithm (Zach et al., 2007) to extract optical flow in an unsupervised manner. To accommodate for pre-trained vision encoders, we first truncate the vertical and horizontal motion values to $[-20, 20]$, then append a third all-zero channel. Finally we do a shift and scale transformation to map $[-20, 20]$ to $[0, 255]$.

VGGSound (Chen et al., 2020) splits. VGGSound contains around 200k video clips belonging to a total number of 309 classes. Different from other audiovisual datasets like AudioSet (Gemmeke et al., 2017), VGGSound ensures the source of the sound is visually present inside the same video. Thus we consider this dataset as an excellent test bed for our proposed method. We randomly select 154 base classes for training and leave the rest 155 classes for novel classes evaluation.

VGGSound audio spectrogram. We use the pre-processing practice of audio spectrogram transformer (AST) (Gong et al., 2021) to convert wavforms to spectrogram images. Each raw audio signal is re-sampled to 16kHz and converted to mono channel. We then calculate the log Mel spectrogram with 128 frequency bins. The processing Hamming window is 25ms with a hop length set to 10ms. For t second audio input, the generated 2D spectrogram would have the shape of $128 \times 100t$. We normalize the spectrogram by subtracting its mean value and dividing its standard deviation.

4.2 Implementation

Data augmentation and tokenization. For each video, we first randomly sample 16 frames with a stride of 4 from the whole video sequence. We then apply the standard image augmentation used on ImageNet (He et al., 2016; 2019) with the same augmentation parameters across all frames to keep temporal consistency (Qian et al., 2021). For optical flow, we follow the practice of Xie et al. (2018) and Han et al. (2020a;b) by directly treating it as images and apply the same augmentation with the video. The augmented output tensors have the shape of $(16, 224, 224, 3)$ from both modalities which can be directly fed into CLIP's ViT encoder. For audio, we apply specialized augmentations designed for spectrogram following Gong et al. (2021) and Nagrani et al. (2021). As the videos in VGGSound are all 10-second long, the generated spectrogram has a shape of $(128, 100 \times 10)$. We first conduct random cropping of $(128, 800)$, sampling all frequency bands with a time duration of 8 seconds. SpecAugment (Park et al., 2019) is applied subsequently with a time masking range of 192 frames and frequency masking of 48 bins. Finally, to accommodate this single channel output with the pre-trained tokenization layer, we make two necessary changes as in Gong et al. (2021): 1) expanding the spectrogram to three duplicated channels, 2) bilinearly interpolating the original positional encoding for spectrogram images with a different spatial resolution.

Network architecture. We adopt CLIP's ViT encoder for video, flow, and audio and the transformer encoder for text. We use 2 transformer layers for temporal fusion ($L = 2$) and 1 transformer layer for cross-attention, each layer has an embedding dimension of 512 and 8 attention heads. For cross-attention, query and key-value inputs use separate layer normalization.

Training hyper-parameters. We adopt the same hyper-parameters for experiments on Kinetics-700 and VGGSound except for training epochs. We use a batch size of 1024 on 128 Cloud TPUv3 cores, AdamW (Loshchilov & Hutter, 2017) optimizer with a weight decay of 0.05 and an initial learning rate of 1e-4 followed by half-cosine decay (He et al., 2019). We set the weight in Eq. 9 as $\alpha = 0.5$. We train 100 epochs on Kinetics-700 and 20 epochs on VGGSound since we observe an overfit using audio modality when trained longer.

Inference hyper-parameters. For video and flow, we use 4×3 views following Arnab et al. (2021) and Liu et al. (2022), where a video is uniformly sampled into 4 clips temporally, and 3 spatial crops for each clip. For audio, we use 12 temporal views without spatial cropping. The final score is averaged over 12 views. For novel classes, we set the weight β in Eq. 12 to 0.25.

4.3 Multimodal Open-Vocabulary Video Classification

We evaluate MOV on Kinetics-700 to utilize modalities of video, optical flow, and text, and on VGGSound to explore the combination of video, audio and text.

Comparison baselines. We evaluate four baselines: **1)** CLIP (Radford et al., 2021), which directly encodes the video and class names into embeddings with pre-trained encoders. The final prediction is given by comparing similarity scores between video and text embeddings; **2)** CoOp (Zhou et al., 2021), which learns continuous text prompt embeddings instead of manually selected templates for better adaptation to downstream tasks; **3)** CLIP-Adapter (Gao et al., 2021), which attaches adapter heads to both video and text encoder; **4)** VATT (Akbari et al., 2021), which is a state-of-the-art multimodal video pre-trainning method and can do zero-shot inference for video classification. We use the same datasets, backbone and hyper-parameters as ours introduced in Sec. 4.2 to train (CLIP and VATT do not require training) and evaluate all methods.

Results. Tab. 1 shows results on Kinetics-700. Both CoOp and CLIP-Adapter achieve better performance than CLIP on base class prediction. While for novel classes, we observe a large accuracy drop compared with CLIP. The degraded performance in harmonic mean of these two methods indicates their loss of the generalization ability on novel classes outweigh their improvement on base classes.

Table 1: **Open-vocabulary video classification on Kinetics-700 (Carreira et al., 2019)**. Modalities are V: Vision, F: Optical Flow and T: Text. MOV obtains the best performance on both base and novel classes, surpassing CLIP (Radford et al., 2021) by 24.1% and 1.4%, respectively.

method	modalities	base acc.	novel acc.	harmonic mean
VATT (Akbari et al., 2021)	V, T	19.8	21.6	20.7
CLIP (Radford et al., 2021)	V, T	51.2	56.7	53.8
CoOp (Zhou et al., 2021)	V, T	58.9	45.7	51.5
CLIP-Adapter (Gao et al., 2021)	V, T	66.5	36.2	46.9
MOV (Ours)	V, F, T	(+8.8) **75.3**	(+1.4) **58.1**	(+11.8) **65.6**

Table 2: **Open-vocabulary video classification on VGGSound (Chen et al., 2020)**. Modalities are V: Vision, A: Audio and T: Text. MOV achieves the best performance on both base and novel classes, surpassing CLIP (Radford et al., 2021) by 19.9% and 2.7%, respectively.

method	modalities	base acc.	novel acc.	harmonic mean
VATT (Akbari et al., 2021)	V, A, T	21.6	23.7	22.6
CLIP (Radford et al., 2021)	V, T	48.5	48.8	48.6
CoOp (Zhou et al., 2021)	V, T	56.9	42.0	48.3
CLIP-Adapter (Gao et al., 2021)	V, T	60.0	27.5	37.7
MOV (Ours)	V, A, T	(+8.4) **68.4**	(+2.7) **51.5**	(+10.2) **58.8**

Our method outperforms CLIP-Adapter by 8.8% on base classes, demonstrating the effectiveness of leveraging multimodal information. On novel classes, we observe an improvement of 1.4% over CLIP, indicating that bringing in flow modality improves the generalization of the video model.

We observe similar trends on VGGSound in Tab. 2. CoOp and CLIP-Adapter improve base classes but fail to generalize to novel classes, resulting in a lower harmonic mean of accuracy compared with CLIP. MOV, when fused with rich audio information, obtains a performance gain of 2.7% over CLIP on novel classes. We also conduct an additional study of generalized open-vocabulary prediction in Appendix C, where the information of whether a class is from base or novel is not known.

Backbone scaling. It is also important to analyze the scalability of MOV with stronger backbones. We experiment with the largest ViT-L/14 model released by CLIP as the vision encoder and a text encoder with embedding dimension increased to 768 and attention heads increased to 12. ViT-L/14 contains 3$\times$ more parameters than ViT-B/16 and we observe around 8% improvement on direct CLIP zero-shot evaluation on Kinetics-700 and 5% improvement on VGGSound, as indicated in the first 2 rows of Tab. 3. MOV is able to preserve the performance gain brought by using stronger CLIP models (last 2 rows of Tab. 3). Despite the significantly stronger CLIP baseline, MOV still improves 20.5% and 1.6% on Kinetics-700, and 19.3% and 2.0% on VGGSound, when comparing row 2 and row 4 of Tab. 3. The scaling performance shows that MOV has a great potential to be incorporated into recent giant vision and language models (Yuan et al., 2021; Yu et al., 2022).

Table 3: **Scalability of MOV**. MOV scales well with a stronger ViT-L/14 backbone.

method	backbone	Kinetics-700		VGGSound	
		base acc.	novel acc.	base acc.	novel acc.
CLIP (Radford et al., 2021)	ViT-B/16	51.2	56.7	48.5	48.8
CLIP (Radford et al., 2021)	ViT-L/14	59.6	65.3	52.6	54.1
MOV (Ours)	ViT-B/16	75.3	58.1	68.4	51.5
MOV (Ours)	ViT-L/14	(+4.8) **80.1**	(+8.8) **66.9**	(+3.5) **71.9**	(+4.6) **56.1**

4.4 Cross-Dataset Transfer

Pre-training an open-vocabulary or zero-shot video classification model on large datasets like Kinetics (Carreira et al., 2019), ImageNet (Deng et al., 2009) or Sports-1M (Karpathy et al., 2014) and evaluating on UCF101 (Soomro et al., 2012) and HMDB51 (Kuehne et al., 2011) is the most

Table 4: **Cross-dataset zero-shot transfer on UCF101 and HMDB51**. We evaluate our proposed MOV without any additional training on two classic video action classification benchmarks. MOV shows the best performance, outperforming classic zero-shot video classification methods, a variety of video and language pre-training approaches, as well as recent CLIP adaptation methods ActionCLIP and X-CLIP, demonstrating a strong cross-dataset generalization ability.

method	vision† + text‡	pre-train§	UCF*/UCF	HMDB*/HMDB
GA (Mishra et al., 2018)	C3D + W2V	S1M	17.3±1.1 / -	19.3±2.1 / -
TARN (Bishay et al., 2019)	C3D + W2V	S1M	19.0±2.3 / -	19.5±4.2 / -
CWEGAN (Mandal et al., 2019)	I3D + W2V	IN, K400	26.9±2.8 / -	30.2±2.7 / -
TS-GCN (Gao et al., 2019)	GLNet + W2V	IN-shuffle	34.2±3.1 / -	23.2±3.0 / -
PS-GNN (Gao et al., 2020)	GLNet + W2V	IN-shuffle	36.1±4.8 / -	25.9±4.1 / -
E2E (Brattoli et al., 2020)	R(2+1)D + W2V	K700	48.0 / 37.6	32.7 / 26.9
DASZL (Kim et al., 2021)	TSM + Attributes	IN, K400	48.9±5.8 / -	- / -
ER (Chen & Huang, 2021)	TSM + BERT	IN, K400	51.8±2.9 / -	35.3±4.6 / -
ResT (Lin et al., 2022)	RN101 + W2V	K700	58.7±3.3 / 40.6	41.1±3.7 / 34.4
MIL-NCE (Miech et al., 2020)	S3D + W2V	HT100M	- / 29.3	- / 10.4
VideoCLIP (Xu et al., 2021)	S3D + TSF	HT100M	- / 22.5	- / 11.3
VATT (Akbari et al., 2021)	ViT + TSF	HT100M	- / 18.4	- / 13.2
CLIP (Radford et al., 2021)	ViT-B/16 + TSF	WIT	79.9±3.8 / 73.0	54.0±4.1 / 46.1
ActionCLIP (Wang et al., 2021)	ViT-B/16 + TSF	WIT$^+$	- / 69.5	- / 50.5
X-CLIP (Ni et al., 2022)	ViT-B/16 + TSF	WIT$^+$	- / 72.0	- / 44.6
MOV (Ours)	ViT-B/16 + TSF	WIT$^+$	**82.6±4.1 / 76.2**	**60.8±2.8 / 52.1**
MOV (Ours)	ViT-L/14 + TSF	WIT$^+$	**87.1±3.2 / 80.9**	**64.7±3.2 / 57.8**

† vision encoder: C3D (Tran et al., 2015), I3D (Carreira & Zisserman, 2017), GLNet (Szegedy et al., 2015), R(2+1)D (Tran et al., 2018), TSM (Lin et al., 2019), RN101 (He et al., 2016), S3D (Xie et al., 2018), ViT (Dosovitskiy et al., 2021).
‡ text encoder: W2V (Mikolov et al., 2013), BERT (Devlin et al., 2019), TSF (Vaswani et al., 2017).
§ pre-train data: S1M (Karpathy et al., 2014), IN (Deng et al., 2009), K400 (Kay et al., 2017), IN-shuffle (Mettes et al., 2016), K700 (Carreira et al., 2019), HT100M (Radford et al., 2021), WIT (Radford et al., 2021), WIT$^+$ has additional training on Kinetics.

common paradigm in the literature. Two settings are used for performance evaluation (Brattoli et al., 2020). The first is randomly choosing half of the classes in the test set and evaluate on the selected subset. To alleviate fluctuations caused by randomness, the evaluation is conducted independently for 10 times and we report the mean accuracy with standard deviation of all trials. We donate this setting as UCF* and HMDB* in Tab. 4. The second evaluation setting is directly evaluating on the whole dataset, which is suitable for methods pre-trained purely on other datasets (Brattoli et al., 2020; Wang et al., 2021; Lin et al., 2022). We train MOV only using 400 base classes subsampled from Kinectis-700, with video, flow and text. For evaluating on UCF and HMDB, we also use the same three modalities. The flow processing follows the same procedure described in Sec. 4.1.

We present a comprehensive comparison in Tab. 4. As in Lin et al. (2022), we list the vision and text encoder and pre-train data used. We compare with three types of state-of-the-art methods: **1)** zero-shot video classification approaches (top part), **2)** video and language pre-training methods (Miech et al., 2020; Xu et al., 2021; Akbari et al., 2021) (middle part), **3)** CLIP adaptation methods (Wang et al., 2021; Ni et al., 2022) (bottom part). Compared to these methods, we find utilizing pre-trained vision and language models like CLIP yield much stronger performance. MOV achieves performance gains over CLIP with around 3% on UCF101 and around 6% on HMDB51. Compared with recently proposed adaptation methods like ActionCLIP and X-CLIP, MOV performs 4.2% to 6.7% better on UCF101 and 1.6% to 7.5% better on HMDB51.

4.5 Ablation Study

Multimodal fusion for base classes. As demonstrated in Fig. 1 and Fig. 2, the asymmetrical cross-attention mechanism is proposed to improve the generalization to novel classes. Here we justify cross-attention also has the advantage for base classes. Tab. 5 shows, for Kinetics-700, simply using the optical flow as input obtains 54.2% on base classes. When using score fusion, compared with video modality, we observe identical performance on base classes. Equipped with the proposed multimodal cross-attention fusion mechanism, we obtain 2.6% improvement on base classes. For VGGSound, the performance of audio only is quite close to video only, and the score fusion facilitates base classes with a significant 6.5% improvement. Our cross-attention mechanism is able to further improve upon this strong baseline by 0.7%.

Table 5: **Ablation on multimodal fusion**. Mutlimodal fusion improves upon using single modality, and the proposed cross-attention works better than score fusion.

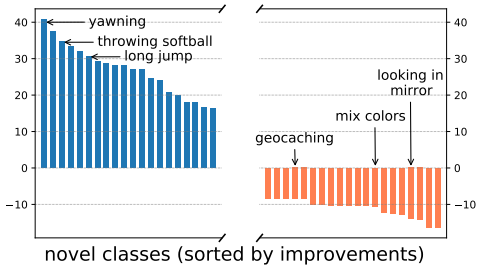
fusion method	Kinetics-700 base acc.	VGGSound base acc.
Flow or Audio only	54.2	59.5
Video only	72.7	61.2
Mutlimodal score fusion	72.7	67.7
Mutlimodal cross-attention	(+2.6) **75.3**	(+0.7) **68.4**

Fine-tuning. We fine-tune different layers of the encoder for flow and audio modality and show results in Tab. 6. As mentioned in Sec. 3, we use the same ViT-B/16 encoder and the same initialization weight for video, flow and audio. We iterate choices of fine-tuning the last 1, 3, 6, 9, and all 12 layers and find consistent performance gains with the increasing number of trainable layers on both modalities. Thus, we adopt the setting of fine-tuning all layers for flow and audio modality.

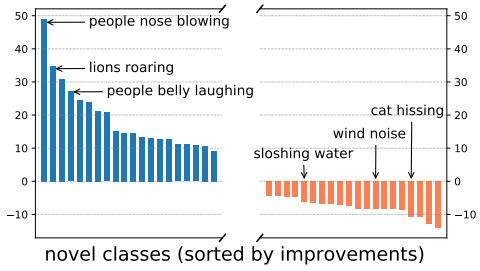
Table 6: **Ablation on fine-tuning the vision encoder with flow and audio**. We report results on base classes of both datasets. The best setting is fine-tuning all layers of the vision encoder.

trainable layers	Kinetics-700		VGGSound	
	modality	base acc.	modality	base acc.
Last 1 layer	Flow	30.5	Audio	40.1
Last 3 layers	Flow	38.3	Audio	47.8
Last 6 layers	Flow	46.0	Audio	50.8
Last 9 layers	Flow	51.6	Audio	57.1
All 12 layers	Flow	**54.2**	Audio	**59.5**

Per-class accuracy analysis. We analyze and interpret class-wise performance differences between MOV and CLIP baseline, which only uses video and text. As illustrated in Fig. 3a, we observe strong gains on classes that require motion understanding, *e.g.* yawning and long jump. While we also find decreased performance on classes with subtle or ambiguous motions, *e.g.* look in mirror and geocaching. In Fig. 3b, we observe audio modality can significantly help disambiguate classes sharing similar visual contents, *e.g.* people nose blowing and people laughing. For classes being difficult in the audio domain, *e.g.* sloshing water and wind noise, we observe decreased performances.



(a) **Kinetics with additional flow modality**



(b) **VGGSound with additional audio modality**

Figure 3: **Per-class improvement.** We show top 20 classes with the most improvement (%) and top 20 classes with the most degradation (%) when compare the proposed MOV against CLIP.

5 Conclusion

We propose a multimodal open-vocabulary video classification method named MOV via adopting pre-trained vision and language models. Motivated by observing drastic performance differences when using video, audio, and optical flow to generalize from base to novel classes, we design a novel asymmetrical cross-modal fusion mechanism to aggregate multimodal information. Extensive experiments on Kinetics, VGGSound, UCF, and HMDB benchmarks demonstrate the effectiveness of our method and the potential of scaling to giant vision and language models.

6 REPRODUCIBILITY STATEMENT

We plan to release our code, dataset splits, and models to facilitate reproducibility. We provided details of our model, data, implementation and experiments in Sec. 3, Sec. 4 and Appendix B. The CLIP model (Radford et al., 2021) and all datasets used in this work (Carreira et al., 2019; Chen et al., 2020; Soomro et al., 2012; Kuehne et al., 2011) are publicly available.

7 ETHICS STATEMENT

The proposed method shows better classification performance on multimodal videos with novel classes on Kinetics, VGGSound, UCF, and HMDB datasets, indicating its potential for real world applications. Our method is built upon vision and language models pre-trained on large-scale data from the internet, which may contain deficiencies and biases. Our models are used only for the purpose of evaluating research ideas. More rigorous studies for bias, fairness, *etc*., are required before using our models for any other purposes.

REFERENCES

Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *NeurIPS*, 2021. 2, 6, 7, 8

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021. 6

Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. In *BMVC*, 2019. 3, 8

Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *CVPR*, 2020. 3, 8

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 8

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 1, 2, 5, 7, 8, 10

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 1, 2, 5, 7, 10, 15

Jiawei Chen and Chiu Man Ho. Mm-vit: Multi-modal video transformer for compressed video action recognition. In *CVPR*, 2022. 3

Shizhe Chen and Dong Huang. Elaborative rehearsal for zero-shot action recognition. In *ICCV*, 2021. 3, 8

Tsuhan Chen and Ram R Rao. Audio-visual integration in multimodal communication. *Proceedings of the IEEE*, 1998. 3

Yi-Wen Chen, Yi-Hsuan Tsai, and Ming-Hsuan Yang. End-to-end multi-modal video temporal grounding. *NeurIPS*, 2021. 3

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 7, 8

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 8

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 8

Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *ICCV*, 2013. 2

Haytham M Fayek and Anurag Kumar. Large scale audiovisual learning of sounds with weakly labeled data. *arXiv preprint arXiv:2006.01595*, 2020. 3

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 3

Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *CVPR*, 2017. 3

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2013. 2

Junyu Gao, Tianzhu Zhang, and Changsheng Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *AAAI*, 2019. 3, 8

Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Learning to model relationships for zero-shot video classification. In *TPAMI*, 2020. 8

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 2, 6, 7

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017. 5

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021. 1, 2, 15

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. In *Interspeech*, 2021. 6

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 1, 2, 15

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP*, 2022. 2

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 2

Meera Hahn, Andrew Silva, and James M Rehg. Action2vec: A crossmodal embedding approach to action learning. *arXiv preprint arXiv:1901.00484*, 2019. 3

Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *ECCV*, 2020a. 5, 6

Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *NeurIPS*, 2020b. 5, 6

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 8

Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *CVPR*, 2019. 6

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. CNN architectures for large-scale audio classification. In *ICASSP*, 2017. 1, 15

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1, 2

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. *arXiv preprint arXiv:2112.04478*, 2021. 2

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021. 2

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 7, 8

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 8

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *ICCV*, 2019. 3

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *ICASSP*, pp. 855–859. IEEE, 2021. 15

Tae Soo Kim, Jonathan D Jones, Michael Peven, Zihao Xiao, Jin Bai, Yi Zhang, Weichao Qiu, Alan Yuille, and Gregory D Hager. Daszl: Dynamic action signatures for zero-shot learning. In *AAAI*, 2021. 8

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 7, 10

Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *ICCV*, 2015. 2

Ang Li, Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Learning visual n-grams from web data. In *ICCV*, 2017. 2

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022a. 1, 2

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022b. 2

Chung-Ching Lin, Kevin Lin, Linjie Li, Lijuan Wang, and Zicheng Liu. Cross-modal representation learning for zero-shot action recognition. In *CVPR*, 2022. 8

Ji Lin, Chuang Gan, and Song Han. Temporal shift module for efficient video understanding. In *ICCV*, 2019. 8

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022. 6

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed, Fahad Shahbaz Khan, and Ling Shao. Out-of-distribution detection for generalized zero-shot action recognition. In *CVPR*, 2019. 3, 8

Pratik Mazumder, Pravendra Singh, Kranti Kumar Parida, and Vinay P Namboodiri. Avgzslnet: Audio-visual generalized zero-shot learning by reconstructing label features from multi-modal embeddings. In *WACV*, 2021. 3

Otniel-Bogdan Mercea, Lukas Riesch, A Koepke, and Zeynep Akata. Audio-visual generalised zero-shot learning with cross-modal attention and language. In *CVPR*, 2022. 3

Pascal Mettes, Dennis C Koelma, and Cees GM Snoek. The imagenet shuffle: Reorganized pre-training for video event detection. In *ICMR*, 2016. 8

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020. 8

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 3, 8

Ashish Mishra, Vinay Kumar Verma, M Shiva Krishna Reddy, S Arulkumar, Piyush Rai, and Anurag Mittal. A generative approach to zero-shot and few-shot action recognition. In *WACV*, 2018. 8

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *NeurIPS*, 2021. 3, 6

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, 2022. 2, 8

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

Kranti Parida, Neeraj Matiyali, Tanaya Guha, and Gaurav Sharma. Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos. In *WACV*, 2020. 3

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019. 6

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *CVPR*, 2021. 6

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 4, 6, 7, 8, 10, 16

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *NeurIPS*, 2014. 3

Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 2005. 1

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 7, 10

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 8

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 8

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 3, 8

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 8

Helin Wang, Dading Chong, and Yuexian Zou. Acoustic scene classification with multiple decision schemes. *Tech. Rep., DCASE2020 Challenge*, 2020. 15

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 1, 3, 15

Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 2, 8

Qian Wang and Ke Chen. Zero-shot visual recognition via bidirectional latent embedding. *IJCV*, 2017. 3

Zuxuan Wu, Yanwei Fu, Yu-Gang Jiang, and Leonid Sigal. Harnessing object and scene semantics for large-scale video understanding. In *CVPR*, 2016. 3

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 2018. 16

Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 3

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018. 3, 5, 6, 8

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, 2021. 8

Xun Xu, Timothy M Hospedales, and Shaogang Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *ECCV*, 2016. 3

Xun Xu, Timothy Hospedales, and Shaogang Gong. Transductive zero-shot action recognition by word-vector embedding. *IJCV*, 2017. 3

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 7

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 7

Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *PR*, 2007. 5

Chenrui Zhang and Yuxin Peng. Visual data synthesis via GAN for zero-shot video classification. In *IJCAI*, 2018. 3

Wangbo Zhao, Kai Wang, Xiangxiang Chu, Fuzhao Xue, Xinchao Wang, and Yang You. Modeling motion with multi-modal features for text-based video segmentation. In *CVPR*, 2022. 3

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 2

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021. 2, 6, 7, 15

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 1, 2, 15

Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, and Ling Shao. Towards universal representation for unseen action recognition. In *CVPR*, 2018. 3

APPENDIX

A COMPARISON WITH MODALITY-SPECIFIC PRE-TRAINED NETWORKS

As we haved metioned in the introduction, instead of using modality-specfic pre-trained encoder networks or methods (Wang et al., 2016; Hershey et al., 2017), we choose a more straightforward path by directly utilizing the pre-trained vision encoder from VLMs with minimal modifications to deal with optical flow and audio spectrogram. Here we list the experimental results using the audio of VGGSound in Tab. 7 to show the effectiveness of our design choice. All methods only use the audio training data and evaluate on audios. Our MOV based on CLIP's vision encoder shows competitive performance compared to other audio specific encoders.

Table 7: **Comparison with audio-specific pre-trained networks.** MOV shows competitive performance compared to other audio specific encoders.

method	audio specific	VGGSound acc.
Acoustic scene classifier (Wang et al., 2020)	✓	39.7
Best baseline of (Chen et al., 2020)	✓	51.0
Audio-Slowfast (Kazakos et al., 2021)	✓	52.5
MOV (Ours)	✗	(+2.1) **54.6**

B TEMPERATURE TUNING

As described in Sec. 3.4, in addition to fused flow and audio features of $\{\mathbf{f_m}, \mathbf{a_m}\}$, we also incorporate the video feature $\mathbf{v}$ extracted from the frozen video backbone to enhance the performance of generalization to novel classes. We denote the probability distribution followed by $\{p_f(j)|_{j=1}^q\}$, $\{p_a(j)|_{j=1}^q\}$ and $\{p_v(j)|_{j=1}^q\}$ as D_f, D_a and D_v. In our experiments we find the curve of D_v tends to be much flatter (or have higher information entropy) than D_f and D_a when the temperatures τ_v, τ_f and τ_a are all set to the CLIP's default value of 0.01. Neglecting this difference and combining the scores as in Eq. 12 would lead to poor performance. We address this problem by lowering τ_v so that the distribution of D_v would be more similar to D_f and D_a (or having similar information entropy). As shown in Tab. 8, adjusting τ_v to 0.003 while keeping τ_f and τ_a as 0.01 greatly improves the performance by 20.1% on Kinetics-700 and 15.8% on VGGSound.

Table 8: **Ablation on temperature tuning**. Compared with using CLIP's default temperature (the first row), using a smaller temperature of 0.003 could greatly improve the performance by 20.1% on Kinetics-700 and 15.8% on VGGSound.

(a) τ_v **tuning on Kinectics-700.**

$\mathbf{v}$ acc.	$\mathbf{f_m}$ acc.	τ_v	final acc.
56.7	30.4	0.01	38.0
56.7	30.4	0.003	**58.1**
56.7	30.4	0.001	57.1
56.7	30.4	0.0003	56.0
56.7	30.4	0.0001	56.4

(b) τ_v **tuning on VGGSound.**

$\mathbf{v}$ acc.	$\mathbf{a_m}$ acc.	τ_v	final acc.
48.8	24.8	0.01	35.7
48.8	24.8	0.003	**51.5**
48.8	24.8	0.001	49.5
48.8	24.8	0.0003	49.1
48.8	24.8	0.0001	49.0

C DISCUSSION ON GENERALIZED OPEN-VOCABULARY PREDICTION

Our model adopt different inference paths for base and novel classes. The evaluation setting of dividing classes into base and novel is a very common practice in existing open-vocabulary literature (Zhou et al., 2021; 2022; Gu et al., 2022; Ghiasi et al., 2021). We follow this established open-vocabulary setting to conduct experiments and evaluate our method.

If label category information isn't given, evaluating purely on unseen classes is the classic setting of zero-shot evaluation (Xian et al., 2018). We benchmark our method in this zero-shot setting in Sec. 4.4 Cross-Dataset Transfer. Our method achieves state-of-the-art performance on commonly used UCF and HMDB zero-shot video classification benchmarks.

Here we consider another setting of generalized open-vocabulary prediction where we train our model on base classes but the model doesn't know whether a class is from base or not during inference. A simple solution is to treat all classes as novel (*i.e.*, use only the "Novel Class Prediction" path illustrated in Fig. 2). We conduct such experiment on Kinetics-700 by training MOV on 400 base classes and evaluating on all 700 classes by treating all of them as novel classes. In this scenario, we observe detrimental performances for both our method MOV and the CLIP baseline. Since the number of classes is 2× more (300 to 700), we consider it a reasonable result. MOV improves upon CLIP in both original (+1.4%) and generalized (+0.7%) open-vocabulary settings for predicting novel classes.

Table 9: **Original and generalized open-vocabulary settings on Kinetics-700**. MOV outperforms CLIP in both settings for predicting novel classes.

method	original (300-class)	generalized (700-class)
CLIP (Radford et al., 2021)	56.7	46.0
MOV (Ours)	(+1.4) **58.1**	(+0.7) **46.7**