A Multimodal Benchmark and Improved Architecture for Zero Shot Learning

Keval Doshi, Amanmeet Garg, Burak Uzkent, Xiaolong Wang, Mohamed Omar
Amazon Prime Video
{kcdos, amanmega, burauzke, xiaowanf, omarmk}@amazon.com

Abstract

In this work, we demonstrate that due to the inadequacies in the existing evaluation protocols and datasets, there is a need to revisit and comprehensively examine the multimodal Zero-Shot Learning (MZSL) problem formulation. Specifically, we address two major challenges faced by current MZSL approaches; (1) Established baselines are frequently incomparable and occasionally even flawed since existing evaluation datasets often have some overlap with the training dataset, thus violating the zero-shot paradigm; (2) Most existing methods are biased towards seen classes, which significantly reduces the performance when evaluated on both seen and unseen classes. To address these challenges, we first introduce a new multimodal dataset for zero-shot evaluation called MZSL-50 with 4462 videos from 50 widely diversified classes and no overlap with the training data. Further, we propose a novel multimodal zero-shot transformer (MZST) architecture that leverages attention bottlenecks for multimodal fusion. Our model directly predicts the semantic representation and is superior at reducing the bias towards seen classes. We conduct extensive ablation studies, and achieve state-of-the-art results on three benchmark datasets and our novel MZSL-50 dataset. Specifically, we improve the conventional MZSL performance by a margin of 2.1%*,* 9.81% *and* 8.68% *on VGG-Sound, UCF-101 and ActivityNet, respectively. Finally, we expect the introduction of the MZSL-50 dataset will promote the future in-depth research on multimodal zero-shot learning in the community.* [1]

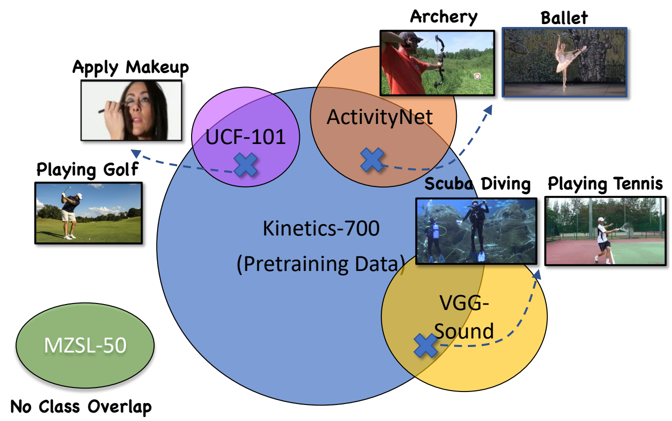


Figure 1. Comparison between existing and proposed benchmark towards multimodal zero shot learning (MZST). Unlike existing methods, our benchmarks has no overlap.

1. Introduction

In existing literature, multimodal zero-shot learning (MZSL) can be broadly classified into two settings, the *conventional* zero-shot setup which assumes only previously unseen classes are available at test time, and the *generalized* zero-shot setup where the test samples belong to both seen and unseen classes. To address practical challenges such as domain adaptation and out-of-distribution examples, recent works employ off-the-shelf pre-trained action recognition models to extract video and audio features [3, 11, 18–20, 23, 31, 34, 35, 51]. Moreover, due to the lack of a dataset specifically designed for MZSL, most existing works evaluate the performance on small scale datasets such as UCF-101 [43], HMDB-51 [28], VGG-Sound [6] and ActivityNet [12]. However, this setup has several limitations. As shown in [3, 9], classes that are considered as *unseen*, are also present in the training set, which clearly violates the conventional zero-shot paradigm. To circumvent this problem, recent research works propose to remove the overlapping classes from large scale datasets such as Kinetics-400/600/700, and train action recognition models on these modified datasets [3, 34, 35]. Considering the lack of general consensus on a fair zero-shot setup, there have been several formulations proposed leading to multiple evaluation setups, as shown in Table 1. To tackle these problems and provide a consistent framework for zero-shot evaluation, we introduce a novel dataset called MZSL-50, which consists of 4462 videos from 50 classes. MZSL-50 does not overlap with any of existing benchmark datasets a shown in Fig. 1, thus eliminating the need for creating dataset splits.

[1]The proposed dataset will be released in public domain for future research use upon publication of the manuscript.

Formulation	Method	Multimodal Split	No Overlap with Kinetics 400/600/700	Type	UCF (Seen/Unseen)	ActivityNet (Seen/Unseen)
Remove training classes	E2E [3]	✗	✓	$\mathcal{R}$	50/51	100/100
	ViSET [9]	✗	✓	$\mathcal{R}$	50/51	100/100
	ConSE [41]	✗	✓	$\mathcal{R}$	50/51	100/100
Remove evaluation classes	TrueZe [22]	✗	✗	$\mathcal{D}$	-/31	-
	Maltoni [10]	✗	✗	$\mathcal{D}$	-/78	-/135
	ViSET [9]	✗	✓	$\mathcal{D}$	-/8	-/19
	AVCA [34]	✓	✗	$\mathcal{D}$	42/9	150/50
	TCAF [35]	✓	✗	$\mathcal{D}$	45/6	152/48
Proposed	**Ours**	✓	✓	$\mathcal{O}$	101/50†	200/50†

Table 1. Multiple ZSL splits proposed in existing works. $\mathcal{R}$ indicates the dataset classes were randomly divided to generate train/test splits, whereas $\mathcal{D}$ represents a deterministic split based on class overlap. In our proposed formulation ($\mathcal{O}$), all classes from the training dataset are considered as seen, and MZSL-50 classes ($\dagger$) are considered as unseen.

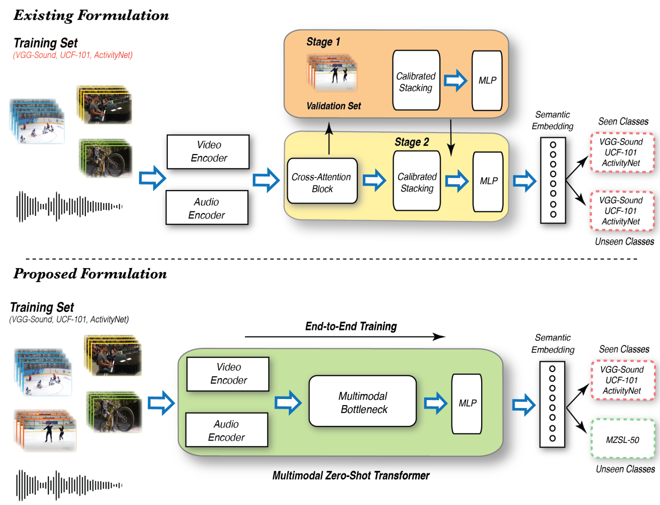


Figure 2. Comparison between existing and proposed approach towards multimodal zero shot learning (MZST). Unlike existing methods, our model performs end-to-end training for MZST.

In addition to the lack of a distinctive evaluation dataset, current SOTA methods [33–35] suffer from using two-stage training procedures to perform well under the generalized zero-shot setup. Two key limitations arise in this setting: (1) they need a complicated training setup to train parts of the model in stages, (2) they introduce bias towards the previously seen classes as direct prediction (Fig. 2). This further introduces bias, where, in an evaluation set with unseen classes the prediction is forced towards the seen classes. To the best of our knowledge, all existing approaches overcome this problem by using some form of *calibrated stacking* [39], which requires multiple stages and a validation set to tune hyperparameters. To mitigate these problems, (1) we propose an architecture, called multimodal zero-shot transformer (MZST), which consists of a multiscale video transformer and an audio spectrogram transformer [21]. (2) To circumvent the bias issue, we propose to project the output of the proposed model to a semantic representation space, and design a loss function to reduce the bias towards seen classes. This allows our model to learn improved representations and enhance the performance in the conventional zero-shot setup, without any performance drop in the generalized zero-shot setting.

To summarize, our intent in this work is to reformulate multimodal zero-shot learning for action recognition by proposing a novel dataset and unified formulation that enables fair comparison across multiple settings. The key contributions in this work are as follows:

- *A new dataset and unified formulation:* We propose MZSL-50, the first dataset specifically designed for evaluating MZSL performance along with a novel training and evaluation protocol to enable fair comparison across different approaches. [2]
- *Novel Architecture:* We propose a novel end-to-end learning model for multimodal zero-shot learning. As shown in Fig. 2, compared to other state-of-the-art models, our model reduces the model intricacy and outperforms current works by $2-9\%$.
- *In-depth analysis:* We perform an in-depth analysis of the proposed model and a pretrained baseline. In a series of guided experiments, we explore the characteristics of a fair zero-shot setup and answer several pertinent questions.

[2] We will release our novel dataset MZSL-50 publicly upon publication of our paper.

2. Related Works

Zero Shot Learning: Action recognition has been extensively studied over the past several years [5, 14–16, 42, 47]. In contrast, ZSL for action recognition has only recently started gaining attention. Broadly, ZSL can be classified into the conventional setting [3, 7, 9, 11, 22, 31], where approaches are evaluated only on the unseen classes, and the generalized setting [27, 33–35, 40] .

Multimodal Representation Learning: Action recognition research works [30, 42, 46] have mostly focused on learning representations from video frames. Only recently the interest in the multimodal domain learning for action recognition has picked up [1, 4]. Recently, transformer based architectures have shown strong performance in learning from multiple modalities. In [37], an attention bottleneck architecture is proposed to fuse feature tokens from all modalities. Similarly, there has been recent interest in the zero-shot action recognition community to leverage multiple modalities. Specifically, Mercea et al. [34, 35] jointly learn audio and video features using cross-attention blocks, whereas Mazumder et al. [33] propose using a composite triplet loss for aligning audio and video embeddings. In [31], a cross-modal approach is proposed to jointly learn the visual and textual features using a transformer architecture, similar to BERT [8].

Visual Feature Extraction: To extract the visual features, most recent approaches [2, 3, 11, 18–20, 23, 48–51] propose using a 3D-CNN, which takes 16 frames sampled from a video as input. In [3], Brattoli et al. propose training a C3D [44] and a R(2+1)D model [45] in an end-to-end fashion for ZSL. On the other hand, Gowda et al. [11] propose a reinforcement learning based clustering approach, which uses a two-stream I3D [5] model for learning visual features. Similarly, it is well understood that hierarchical representations capture scale variation in the data and learn concepts that vary with scale. Recent works learn multi scale feature representation from video [30] and show strong performance in video classification.

Zero Shot Evaluation: Several approaches have extended the work of Roitberg et al. [41] to formulate a novel evaluation protocol that satisfies the ZSL paradigm. Particularly, Brattoli et al. [3] proposes removing certain classes from the training set which overlap with the test set by using semantic embedding matching. However, Doshi et al. [9] show that such an approach fails to remove all the overlapping classes, and propose a new evaluation split called *Fair ZSL*, composed of non-overlapping classes from benchmark datasets. Alternatively, Gowda et al. [22] also propose a *TruZe* split for the UCF-101 [43] and HMDB-51 [28] datasets, by manually removing all classes which overlap with the Kinetics-400 dataset. Unfortunately, these works fail to reach a common consensus regarding the evaluation splits, leading to multiple evaluation setups as shown in Table 1.

3. The MZSL-50 dataset

In this section, we present MZSL-50, a multimodal dataset composed of carefully chosen action classes that has *no overlap* with existing benchmark datasets. It is, to the best of our knowledge, the first dataset that has been specifically tailored for evaluation in multimodal zero-shot learning. We begin by discussing the motivation for proposing a new dataset, followed by a thorough analysis of the annotation protocol and dataset statistics. Finally, we present the evaluation procedure for both conventional and generalized zero-shot setups.

Motivation: The purpose of MZSL-50 is to establish a unified evaluation benchmark for multimodal zero-shot learning, such that it does not violate the zero-shot paradigm. As shown in Table 1, existing approaches lack a general consensus regarding classes that *overlap with benchmark training datasets*, which has lead to several proposed splits, increasing ambiguity in comparing related works. It is also important to take into account, in case of an unseen class, completely unrelated to any other seen classes (e.g. cooking omelette vs. playing tennis). Intuitively, even humans struggle to understand novel activities when they include unfamiliar relationships and objects. Thus, we also consider the *semantic relatedness* of the classes in the proposed dataset with respect to the 4 benchmark training datasets used in existing works. We hope that the straightforward way in which we have formulated the training and evaluation protocols will make it simple for future researchers to assess and compare their MZSL performance.

Design: To ensure a realistic MZSL setting and *avoid an overlap with seen classes*, we specifically take benchmark datasets used in the research community as our reference. Specifically, we consider the actions included in Kinetics-400/600/700, VGG-Sound, UCF-101 and ActivityNet as the seen classes, and collect videos for classes that are semantically related, but not identical to these seen classes. Formally, we define the semantic relatedness (SR) score as:

$$SR(class) = \min D_{\cos}(\phi(class), \phi(X^S)), \tag{1}$$

where $D_{\cos}$ is the cosine distance, $\phi(class)$ is the semantic embedding of an unseen class and $\phi(X^S)$ is the set of semantic embeddings of all the classes in the benchmark datasets. We use Word2Vec [36] to parameterize function ϕ for extracting the embeddings. The semantic similarity of classes in the training and evaluation sets can directly impact the model performance on individual classes. In order to quantify this impact, we assign a SR score to all the classes in the proposed evaluation set. A low SR value implies high semantic similarity, thus easy for the model to recognise, similarly, a high SR value would imply a difficult case for the model to predict on. Empirically, we see that SR should be higher than 0.1 to avoid including identical classes, and less than 0.8 to avoid classes that are significantly different from the seen classes. Hence, we divide our classes into 3 sets based on the SR score; easy ($0.1 < SR \leq 0.33$), medium ($0.33 < SR \leq 0.66$) and hard ($0.66 < SR \leq 0.8$) classes.

Statistics: The final dataset consists of 4462 videos for a total duration of 405 hours covering 50 classes. Following our dataset design, we further divide the videos into *easy*, *medium*, and *hard* sub-groups based on the semantic relatedness score. The table 2 outlines the data in each subclass.

Difficulty	# Classes	# Videos
Easy	9	926
Medium	31	2811
Hard	10	725
Total	50	4462

Table 2. Statistics for difficulty levels based on semantic relatedness score (SR) on MZSL-50.

Annotation protocol: Given a video clip, the goal of an-

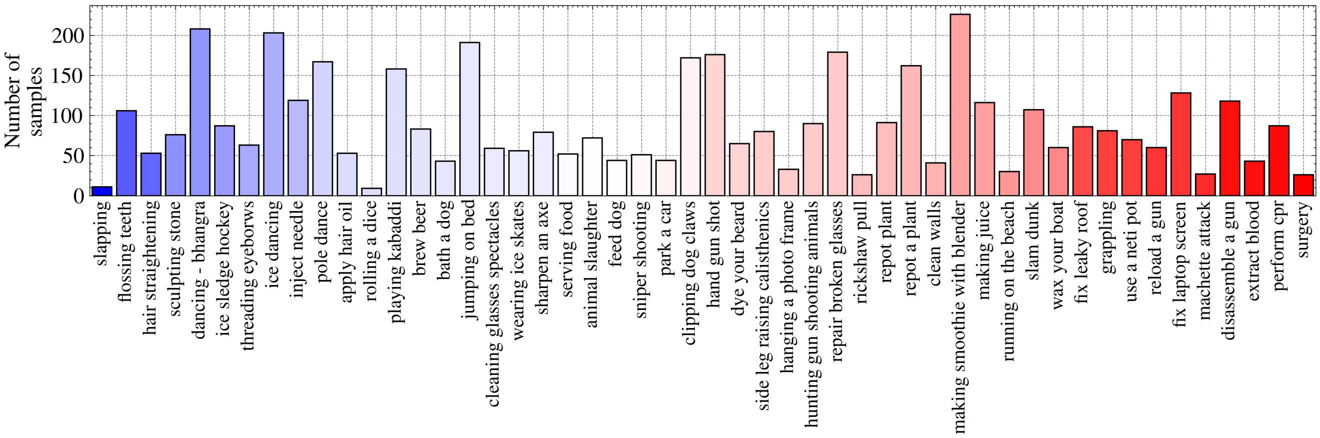


Figure 3. The sample distribution of all classes in MZSL-50 dataset. The color of the bars represent the *Semantic Relatedness* score ranging from 0.1 (blue, slapping) to 0.8 (red, surgery) across the 50 classes. Semantic Relatedness represents the maximal similarity value of an unseen class to a seen class in the training dataset.

(a) Example from the *making juice* class. (b) Example from the *flossing teeth* class. (c) Class distribution w.r.t hardness score.

Figure 4. Example filmstrips and statistics visualization for the dataset.

notators is to validate the presence of a sequence of frames related to the class of interest. Hence, to create this new evaluation dataset, we source public videos from top 300 results for each class from YouTube. In order to improve the annotation efficiency, we create a filmstrip using frames of a video sampled at 1 *fps* arranged into a 10x10 grid. The annotators are instructed to visually check the filmstrip and assert the presence of frames representing the action category of the class. For example, in *making juice* class (Fig. 4a) the presence of row 6-8 with juice frames would qualify this video as a positive sample. The videos were annotated by multiple reviewers to avoid negative examples, and we remove videos which had mixed annotations.

MZSL-50 Evaluation: For conventional and generalized zero-shot setups, we present a unified training and evaluation protocol. Instead of splitting the dataset and training on a subset of classes [3, 9, 41], we train the model on all classes from the benchmark datasets, such as (VGG-Sound, UCF-101 and ActivityNet). The model is then evaluated solely on MZSL-50 using the standard zero-shot setup. The final metric on the MZSL-50 dataset is reported as the weighted average of the performance in the easy, medium and hard subsets (weighted on number of classes in each subset). We evaluate the generalized zero-shot setup on both MZSL-50 and the respective benchmark dataset.

4. Multimodal Zero Shot Transformer

4.1. Preliminaries

We can formally define MZSL as a classification problem, where given a tuple of video, audio and text class labels (V^s,A^s,T) as training data from S seen classes $\{(v_1^s, a_1^s, t_1), \ldots, (v_N^s, a_N^s, t_N)\}$, we aim to accurately classify video and audio $X^u = \{(v_1^u, a_1^u), \ldots, (v_M^u, a_M^u)\}$ from previously unseen classes U, where N and M are the number of training and testing videos respectively. Ideally, to satisfy the conventional zero-shot learning paradigm, there should be no overlap between the seen and unseen classes, i.e., $(S \cap U = \varnothing)$. On the other hand, in the generalized zero-shot setup, a model is evaluated on both seen and unseen classes, which requires a broad generalization capacity. Conventionally, semantic embeddings are used as

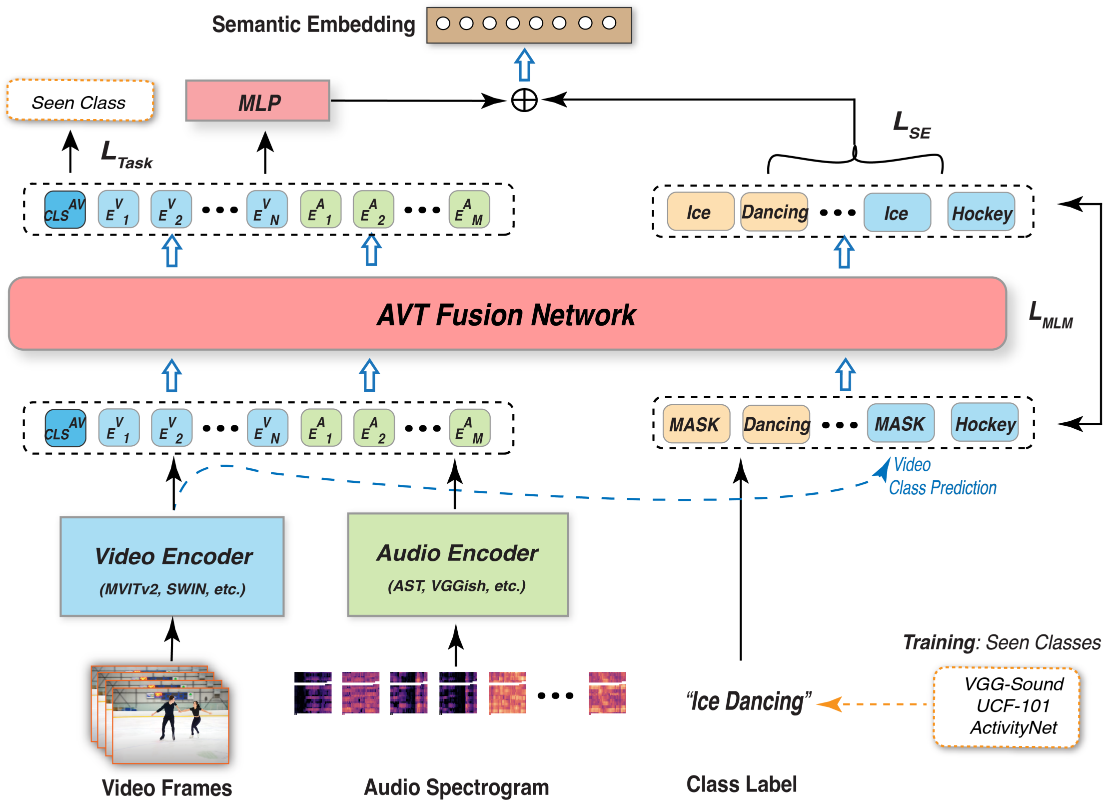


Figure 5. The framework overview of Multiscale Zero-Shot Transformer, MZST, where multimodal inputs are frame sequence V_i and audio spectrogram A_i from the i-th video. We leverage the recently proposed MViT-v2 and AST architectures along with an audio-video-text (AVT) fusion network to directly predict the semantic embedding.

a mapping between the input videos and the class labels, which are mainly composed of words. The idea behind this popular approach is to learn a semantic embedding model $f(x)$ for the input videos and then select the class that is semantically closest.

4.2. Proposed Model Architecture

MViT & AST Architecture: Hierarchical representations often outperform their single scale counterparts in recognition tasks [30]. We take a pretrained MVIT [29] model as our backbone to provide multiscale embeddings for the video frames. Inspired by the recent strong performance by the MVIT model [30], we extend the model architecture to learn hierarchical representations for the underlying data for video inputs. Recently, Audio Spectrogram Transformer (AST) [21] has utilized audio spectrograms as images to extract patch based embedding tokens from patches of the spectrograms. We leverage the AST architecture to obtain token embeddings for each audio sample. Specifically, we extract audio tokens $[E_1^A, \cdots, E_M^A]$ using M non-overlapping patches from the input spectrogram and N video tokens $[E_1^V, \cdots, E_N^V]$ from the input video.

AVT Fusion Network: Previous cross-modality transformers either simply concatenated multimodal representations [1], or exchanged the key and value matrices between the two modalities in the attention block [24]. On the other hand, inspired by the cross modality fusion between audio and image transformers [37], we construct an audio-video-text (AVT) fusion network by leveraging bottleneck transformers, which handles varied lengths of modality tokens efficiently as illustrated in Fig. 5.

Let $\{E_1^F, \cdots, E_L^F\}$ be the L initial multimodal fusion tokens. During training, the fusion tokens are input to the transformer block alternatively with the joint video-audio tokens and text tokens. The model is forced to learn the fusion tokens with attention based token update along all three modalities. For example, for the joint video-audio token update we perform following operations formulated as

$$\begin{aligned} \tilde{E}^{AVF} &= \text{MSA}(\text{LN}(E^{AVF})) + E^{AVF}, \\ \hat{E}^{AVF} &= \text{FFN}(\text{LN}(\tilde{E}^{AVF})) + \tilde{E}^{AVF}, \end{aligned} \tag{2}$$

where $E^{AVF} = [E_{CLS}^V, E_1^V, \cdots, E_N^V, E_1^A, \cdots, E_M^A, E_1^F, \cdots, E_L^F]$ whereas FFN represents the Feed Forward Neural Network and LN represents the layer norm.

We repeat the above operations for the text tokens such that $E^{TF} = [E^T_{CLS}, E^T_1, \cdots, E^T_M, \hat{E}^F_1, \cdots, \hat{E}^F_L]$. Given modality-specific tokens, multimodal tokens can be updated by averaging the multimodal tokens along the AVT bottleneck blocks. The AVT fusion network consists of K stacked blocks.

4.3. Loss Function

Masked Language Modeling Loss: Inspired by the Masked Language Modeling (MLM) task of BERT transformer [8], we apply the MLM loss to the discrete token sequence for input text. We randomly mask the input text tokens so as to force the bottleneck architecture to predict these masked tokens (E^{TF}_{α}) based on their surrounding word tokens ($E^{TF}_{s\alpha}$) and joint audio-visual feature tokens E^{AVF} by minimizing the negative log-likelihood:

$$\mathcal{L}_{MLM} = -\frac{1}{n}\sum_{i=1}^{n} \log p_i^{AV}(E^{TF}_{\alpha} | E^{TF}_{s\alpha}, E^{AVF}). \tag{3}$$

The purpose of using MLM loss is to align and learn the dependencies between visual and audio content and semantic concepts.

Semantic Embedding Loss: The supervised semantic embedding loss is formulated as:

$$\mathcal{L}_{SE} = \| f(x_i^s) - (\phi(s_i^v) + \phi(s_i^t)) \|^2, \tag{4}$$

where $f(x_i^s)$ is the output of the FFN and $\phi(s_i^v)$, $\phi(s_i^t)$ are the semantic embedding for the training video x_i^s from the class label of the pretrained video model and the ground truth label, respectively, extracted using the Word2Vec model [38].

Task Loss: We perform multimodal classification by passing the joint video-audio representation $[E^{AV}_{CLS}]$ through a fully connected layer. We use cross entropy loss, $\mathcal{L}_{TASK}$, for this objective.

Combined Loss Function: Finally, we combine all the objectives linearly as

$$\mathcal{L} = \mathcal{L}_{TASK} + \mathcal{L}_{SE} + \sigma \cdot \mathcal{L}_{MLM} \tag{5}$$

where $\sigma \in [0, 1]$. We provide the sensitivity analysis for σ in the **supplementary** material.

5. Experiments

We evaluate the proposed approach and compare with other state-of-the-arts both on three popular benchmarks and newly collected MZSL-50. In this section, we first give the description of dataset, then provide the implementation details. After that, we list the experimental results and ablation studies.

5.1. Experimental Setup

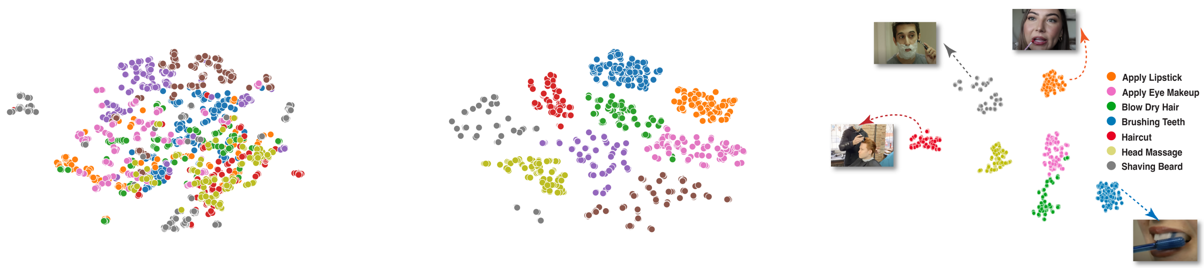
Benchmark Datasets: We first evaluate our proposed approach on our new MZSL-50 dataset and three benchmark datasets, namely VGGSound [6], UCF-101 [43] and ActivityNet [12] datasets. **VGGSound** is a large scale action recognition dataset, which consists of about 200K 10-second clips and 309 categories ranging from human actions and sound-emitting objects to human-object interactions. Like other YouTube datasets, *e.g.*, K400 [26], some clips are no longer available. After removing invalid clips, we collect 159,223 valid training multimodal videos and 12,790 valid test multimodal videos. **UCF-101** consists of over 13k videos in 101 classes. We only consider classes which include the audio modality, leading to 6,816 videos from 51 classes. **ActivityNet** consists of videos from 200 classes related to daily activities and is considerably more comprehensive, consisting of 27,801 videos from 200 classes.

For a fair comparison with existing state-of-the-art approaches, we use the same training and evaluation splits as proposed in [34].

Comparing with State-of-the-Art: We compare our proposed approach to recent state-of-the-art multimodal ZSL approaches TCaF [35], AVCA [34], AVGZSLNet [33] and CJME [39]. AVCA leverages cross-attention mechanism and combines information from video and audio modalities by temporally averaging them. TCaF builds upon AVCA and applies an improved cross-attention mechanism which introduces temporal attention in addition to spatial attention. Furthermore, we also compare to image based approaches Attention Fusion [13], Perceiver [25] and DeViSE [17], which are adapted for multimodal inputs by [34, 35].

Implementation Details: We employ 16 frames for multiscale video Transformer [29] along with 3 spatial crops and 4 ensemble views during inference. We are able to train the model using a batch size of 64 on 8 NVIDIA A100 GPUs, each with 40 GB of memory. AdamW [32] is used in the backpropagation and the learning rate is set as 0.0001. The number of epochs is set as 100. We set λ_1 and λ_2 as $0.25, 0.25$ for the first 20 epochs, $0.1, 0.1$ from the 21- to 40-th epochs, and $0.05, 0.05$ after the 40-th epochs. These hyperparameters are generally set to tune the loss values into the same scale. We sample audio clips at 16kHz and convert them to mono channel. We extract log mel spectrograms with a frequency dimension of 128. The AST model is initialized with ImageNet weights. The bootlneck tokens are initialized using a Gaussian distribution of mean 0 and standard deviation of 0.02. More details are available in the **supplementary** material.

Evaluation metrics: We follow the evaluation protocol discussed in [34], and propose to evaluate all models using the mean class accuracy. For generalized MZSL, we eval-



(a) Audio features TSNE. (b) Video features TSNE. (c) MZST features TSNE.

Figure 6. TSNE visualization of different modalities. We see that the joint features (c) extracted using our MZST model is semantically more separable.

uate the proposed model on both seen and unseen classes, and report the harmonic mean. The harmonic mean is given by:

$$HM = \frac{2 \times S \times U}{S + U} \tag{6}$$

where S,U are the mean class accuracies on the seen and unseen classes, respectively.

5.2. Results on MZSL-50

In Table 3, we show the performance of our approach based on the semantic relatedness of the classes, specifically the easy, medium and hard cases. We can clearly observe that the semantic relatedness metric is an excellent indicator for estimating the performance of a model on a certain class, since all approaches notice a drop in performance as the classes become more difficult. This shows that the semantic relatedness can be used as a reliable metric for designing datasets which have classes between a certain range, unlike existing datasets such as UCF-101 which has 80% of the classes with a SR score of less than 0.2.

Model	Modality	VGG-Sound/MZSL-50				ActivityNet/MZSL-50			
		E	M	H	Avg	E	M	H	Avg
CJME [39]	A,V	25.37	13.48	3.20	14.01	28.95	12.23	3.02	14.73
AVCA [34]	A,V	32.13	15.29	4.78	17.4	31.78	14.12	5.03	16.97
MZST (Ours)	A,V	**46.74**	**24.98**	**8.82**	**26.84**	**52.29**	**25.24**	**5.72**	**27.75**

Table 3. MZSL performance on the MZSL-50 dataset when using audio and visual features as inputs on the VGG-Sound and ActivityNet datasets. We report the performance on easy (E), medium (M) and hard (H) classes, as well as the average (Avg) class accuracy. We can clearly see that MZST outperforms the recent SOTA methods by a wide margin.

Moreover, we also compare our proposed framework to recent approaches under the conventional zero-shot and the generalized zero-shot setup. As shown in Fig. 2, under the conventional setup, we finetune our model on all the classes of the benchmark dataset and evaluate on the MZSL-50 dataset. We only finetune on classes from VGG-Sound and ActivityNet since very few classes in UCF-101 contain both video and audio. We observe that the proposed approach achieves a significantly higher performance (Table 4) across all settings. We were unable to compare our model performance to TCaF [35], since it required features extracted at separate temporal intervals, whereas all the other recent approaches use averaged features.

Model	Modality	VGG-Sound/MZSL-50				ActivityNet/MZSL-50			
		S	U	HM	ZS	S	U	HM	ZS
AVCA	A,V	13.23	7.15	9.28	9.48	6.14	5.28	5.68	6.99
AST	A	18.27	9.15	12.19	11.37	8.05	7.25	7.66	9.03
MVIT-v2	V	18.11	12.54	14.82	15.64	20.18	18.26	19	19.15
MZST (Ours)	A,V	**28.31**	**15.85**	**20.32**	**25.36**	**24.57**	**19.23**	**21.58**	**29.29**

Table 4. Audio-visual MZSL results under the generalized zero-shot setting when using audio and visual features as inputs on the VGG-Sound, UCF and ActivityNet datasets.

5.3. Results on Benchmark Dataset

We compare our proposed framework to recent approaches under the conventional zero-shot setup in Table 5 and under the generalized zero-shot setup in Table 6. For a fair comparison, we use the same splits proposed in [34]. The multiscale video model is trained on 595 classes from Kinetics, after removing the overlapping classes, as proposed in [3]. We observe that the proposed approach achieves a significantly higher performance, by a margin of 18.03%, 20.23% and 20.58% as compared to [34]. To have a fair comparison with recent state-of-the-art approaches, we also evaluate the proposed approach using the features extracted using [34], and yet have a significant performance gap. Specifically, we outperform recent approaches by a margin of 2.1%, 9.81% and 8.68%, which shows the efficacy of our approach.

5.4. Ablation Studies

Impact of pretraining datasets: We experiment with various backbones to study the impact of pretraining on the zero-shot performance. As shown in Table 5, we see that pretraining the video model on K595 significantly increases

Models	Modality	Pretraining	VGG-Sound^{ZS}	UCF^{ZS}	ActivityNet^{ZS}
Att. Fusion	A, V	VGG	2.38	12.54	2.63
Perceiver	A, V	VGG	2.93	18.77	5.37
DeViSE	A, V	VGG	2.59	16.09	8.53
CJME	A, V	VGG	5.16	8.29	5.84
AVGZSLNet	A, V	VGG	5.28	13.65	5.40
TCaF	A, V	VGG	6.06	24.81	7.91
AVCA	A	VGG	5.01	16.05	8.78
AVCA	V	VGG	4.78	17.22	6.89
AVCA	A, V	VGG	6.00	20.01	9.13
AST	A	ImageNet	17.23	18.01	12.03
SeLaVi	V	VGG	14.07	24.39	17.31
MVIT-v2	V	K595	21.99	35.23	21.03
MZST (Ours)	A, V	VGG†	16.23	31.35	19.3
MZST (Ours)	A, V	K595+ImageNet	**24.09 (2.1% ↑)**	**45.04 (9.81% ↑)**	**29.71 (8.68% ↑)**

Table 5. Audio-visual MZSL results under the zero shot setting on benchmark datasets. VGG† pretraining implies self-supervised SeLAVi features used by state-of-the-art approaches [34, 35]

Model	Modality	VGG-Sound^{GZS}			UCF-101^{GZS}			ActivityNet^{GZS}		
		S	U	HM	S	U	HM	S	U	HM
Att. Fusion	A,V	6.12	2.26	3.308	35.47	11.26	17.10	6.49	2.04	3.11
Perceiver	A,V	7.92	2.72	4.05	34.10	18.18	23.72	7.22	5.16	6.02
DeViSE	A,V	**36.22**	1.07	2.08	55.59	14.94	23.56	3.45	8.53	4.91
CJME	A,V	8.69	4.78	6.17	26.04	8.21	12.48	5.55	4.75	5.12
AVGZSLNet	A,V	18.05	3.48	5.83	52.52	10.90	18.05	8.93	5.04	6.44
AVCA	A,V	14.90	4.00	6.31	51.53	18.43	27.15	24.86	8.02	12.13
TCaF	A,V	9.64	5.91	7.33	58.60	21.74	31.72	18.70	7.50	10.71
MZST (VGG)	A,V	15.73	6.01	8.69	53.86	20.67	30.08	28.32	14.17	18.88
MZST (Ours)	A,V	32.13	**22.08**	**26.17**	**90.357**	**39.387**	**54.85**	**32.43**	**20.79**	**25.33**

Table 6. Audio-visual MZSL results under the generalized zero-shot setting when using audio and visual features as inputs on the VGG-Sound, UCF and ActivityNet datasets.

Models	Multimodal Fusion	Pretraining	VGG-Sound^{ZS}	UCF^{ZS}	ActivityNet^{ZS}
AVCA	Cross-Attention	VGG	6.00	20.01	9.13
MZST	Feature Concatenation	VGG	15.1	32.6	19.2
MZST	Feature Concatenation	K595+ImageNet	21.42	**49.72**	28.13
MZST	Bottleneck Transformer	K595+ImageNet	21.73	43.21	27.98
MZST	AVT Fusion + $\mathcal{L}_{Task}$	K595+ImageNet	21.35	45.38	27.22
MZST	AVT Fusion + $\mathcal{L}_{Task}$ + $\mathcal{L}_{SE}$	K595+ImageNet	23.68	46.85	28.19
MZST	AVT Fusion + $\mathcal{L}_{Task}$ + $\mathcal{L}_{SE}$ + $\mathcal{L}_{MLM}$	K595+ImageNet	**24.09**	45.04	**29.71**

Table 7. Results on comparing the impact of different fusion techniques on the performance in the MZSL task.

the performance as compared to using self-supervised pretraining on VGG-Sound. Moreover, it is interesting to observe that audio model initialized with imagenet weights performs better than AVCA pretrained on VGG-Sound, which further demonstrates the efficacy of our proposed model (Table 7).

Evaluating fusion mechanisms: Next, we investigate the effect of using our AVT fusion network and loss functions in Table 7. To obtain results without audio-video fusion, each branch is optimised individually. For evaluation, we simply concatenate the video and audio embeddings. We observe that except for UCF-101, audio-video fusion consistently performs better than feature concatenation. The reason for feature concatenation performing better on UCF-101 can be attributed to the dataset being spatially heavy and video being the dominant modality.

Impact of variation in semantic embedding: As illustrated in Fig. 7, we compare and visualize the projected semantic embedding of the proposed approach to the Word2Vec semantic embedding of the recent SOTA approach. We see that normalizing the semantic feature representation deteriorates it's inherent relatedness. For example, before normalizing the embeddings, *Shaving Beard* is semantically similar to *Haircut* and *Brushing Teeth*, but normalizing the embeddings cause it to be similar to *Javelin Throw*. We can also see that there is a noticeable improvement in the performance (Table 8) when non-normalized semantic embeddings are used, which ascertains our conjecture.

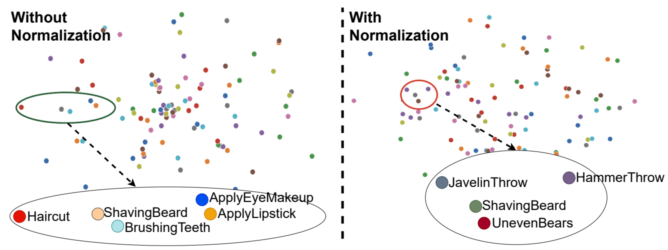


Figure 7. Embedding TSNE comparison. We show that normalizing the semantic embedding skews the natural semantic distribution, which directly affects the performance on unseen classes. By avoiding normalization, similar classes are projected closely to each other in the latent space.

Models	Pretraining	Semantic Embedding	Normalized	VGG-Sound^{ZS}	UCF^{ZS}
AVCA	VGG	W2V	Yes	6.00	20.01
MZST	VGG	W2V	Yes	9.83	24.38
MZST	VGG	W2V	No	16.23	31.35
MZST	K595+ImageNet	W2V	Yes	17.11	24.86
MZST	K595+ImageNet	W2V	No	**23.76**	**42.35**
MZST	VGG	S2V	Yes	9.89	23.62
MZST	VGG	S2V	No	16.40	23.39
MZST	K595+ImageNet	S2V	Yes	17.35	24.98
MZST	K595+ImageNet	S2V	No	**24.09**	**45.04**

Table 8. Results on comparing the impact of different semantic embeddings on the performance in the MZSL task.

6. Conclusion

We first propose a novel dataset called MZSL-50 for multimodal zero-shot action recognition. The MZSL-50 allows for a unified formulation and proposes a comprehensive evaluation protocol that strictly adheres to the zero-shot premise. Additionally, we propose an end-to-end multimodal transformer called MZST that outperforms existing approaches by a wide margin on both existing datasets and on MZSL-50. Specifically, we outperform the state-of-the-art approaches by 2.1%, 9.81% and 8.68% using the conventional zero-shot setup on the VGG-Sound, UCF-101 and ActivityNet datasets, respectively.

References

[1] Hassan Akbari, Liangzhe 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. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021. 2, 5

[2] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. *arXiv preprint arXiv:1907.09021*, 2019. 3

[3] Biagio Brattoli and others. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *CVPR*, 2020. 1, 2, 3, 4, 7

[4] XB Bruce, Yan Liu, Xiang Zhang, Sheng-hua Zhong, and Keith CC Chan. Mmnet: A model-based multimodal network for human action recognition in rgb-d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2, 3

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 1, 6

[7] Shizhe Chen and Dong Huang. Elaborative rehearsal for zero-shot action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13638–13647, 2021. 2

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 6

[9] Keval Doshi and Yasin Yilmaz. Zero-shot action recognition with transformer-based video semantic embedding. *arXiv preprint arXiv:2203.05156*, 2022. 1, 2, 3, 4

[10] Valter Estevam, Helio Pedrini, and David Menotti. Zero-shot action recognition in videos: A survey. *Neurocomputing*, 439:159–175, 2021. 2

[11] Gowda et al. Claster: Clustering with reinforcement learning for zero-shot action recognition. *arXiv preprint arXiv:2101.07042*, 2021. 1, 2, 3

[12] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 1, 6

[13] Haytham M Fayek and Anurag Kumar. Large scale audio-visual learning of sounds with weakly labeled data. *arXiv preprint arXiv:2006.01595*, 2020. 6

[14] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4768–4777, 2017. 2

[15] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016. 2

[16] Basura Fernando, Efstratios Gavves, Jose M Oramas, Amir Ghodrati, and Tinne Tuytelaars. Modeling video evolution for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5378–5387, 2015. 2

[17] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013. 6

[18] Chuang Gan, Ming Lin, Yi Yang, Gerard De Melo, and Alexander G Hauptmann. Concepts not alone: Exploring pairwise relationships for zero-shot video activity recognition. In *Thirtieth AAAI conference on artificial intelligence*, 2016. 1, 3

[19] Chuang Gan, Ming Lin, Yi Yang, Yueting Zhuang, and Alexander G Hauptmann. Exploring semantic inter-class relationships (sir) for zero-shot action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015. 1, 3

[20] Chuang Gan, Tianbao Yang, and Boqing Gong. Learning attributes equals multi-source domain generalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 87–97, 2016. 1, 3

[21] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. 2, 5

[22] Shreyank Gowda et al. A new split for evaluating true zero-shot action recognition. *preprint arXiv:2107.13029*, 2021. 2, 3

[23] Meera Hahn, Andrew Silva, and James M Rehg. Action2vec: A crossmodal embedding approach to action learning. *arXiv preprint arXiv:1901.00484*, 2019. 1, 3

[24] Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585, 2021. 5

[25] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 6

[26] 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. 6

[27] Tae Soo Kim, Jonathan 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 *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1817–1826, 2021. 2

[28] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 1, 3

[29] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Improved multiscale vision transformers for classification and detection. *arXiv preprint arXiv:2112.01526*, 2021. 5, 6

[30] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 2, 3, 5

[31] Chung-Ching Lin, Kevin Lin, Lijuan Wang, Zicheng Liu, and Linjie Li. Cross-modal representation learning for zero-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19978–19988, 2022. 1, 2

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

[33] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3090–3099, 2021. 2, 6

[34] Otniel-Bogdan Mercea et al. Audio-visual generalised zero-shot learning with cross-modal attention and language. In *CVPR*, 2022. 1, 2, 6, 7, 8

[35] Otniel-Bogdan Mercea, Thomas Hummel, A Koepke, and Zeynep Akata. Temporal and cross-modal attention for audio-visual zero-shot learning. In *European Conference on Computer Vision*, pages 488–505. Springer, 2022. 1, 2, 6, 7, 8

[36] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 3

[37] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021. 2, 5

[38] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*, 2018. 6

[39] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3251–3260, 2020. 2, 6, 7

[40] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *European Conference on Computer Vision*, pages 292–308. Springer, 2020. 2

[41] Alina Roitberg, Manuel Martinez, Monica Haurilet, and Rainer Stiefelhagen. Towards a fair evaluation of zero-shot action recognition using external data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2, 3, 4

[42] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014. 2

[43] 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. 1, 3, 6

[44] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 3

[45] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 3

[46] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018. 2

[47] Yunbo Wang, Mingsheng Long, Jianmin Wang, and Philip S Yu. Spatiotemporal pyramid network for video action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1529–1538, 2017. 2

[48] Xun Xu, Timothy Hospedales, and Shaogang Gong. Semantic embedding space for zero-shot action recognition. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 63–67. IEEE, 2015. 3

[49] Xun Xu, Timothy Hospedales, and Shaogang Gong. Transductive zero-shot action recognition by word-vector embedding. *International Journal of Computer Vision*, 123(3):309–333, 2017. 3

[50] Xun Xu, Timothy M Hospedales, and Shaogang Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *European Conference on Computer Vision*, pages 343–359. Springer, 2016. 3

[51] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 374–390, 2018. 1, 3