# *MMG-Ego4D*: Multi-Modal Generalization in Egocentric Action Recognition

Xinyu Gong[2*†], Sreyas Mohan[1*], Naina Dhingra[1], Jean-Charles Bazin[1], Yilei Li[1],
Zhangyang Wang[2], Rakesh Ranjan[1]
[1]Meta Reality Labs, [2]The University of Texas at Austin

## Abstract

*In this paper, we study a novel problem in egocentric action recognition, which we term as "Multimodal Generalization" (**MMG**). MMG aims to study how systems can generalize when data from certain modalities is limited or even completely missing. We thoroughly investigate MMG in the context of standard supervised action recognition and the more challenging few-shot setting for learning new action categories. MMG consists of two novel scenarios, designed to support security, and efficiency considerations in real-world applications: (1) missing modality generalization where some modalities that were present during the train time are missing during the inference time, and (2) cross-modal zero-shot generalization, where the modalities present during the inference time and the training time are disjoint. To enable this investigation, we construct a new dataset MMG-Ego4D containing data points with video, audio, and inertial motion sensor (IMU) modalities. Our dataset is derived from Ego4D [27] dataset, but processed and thoroughly re-annotated by human experts to facilitate research in the MMG problem. We evaluate a diverse array of models on MMG-Ego4D and propose new methods with improved generalization ability. In particular, we introduce a new fusion module with modality dropout training, contrastive-based alignment training, and a novel cross-modal prototypical loss for better few-shot performance. We hope this study will serve as a benchmark and guide future research in multimodal generalization problems. The benchmark and code are available at https://github.com/facebookresearch/MMG_Ego4D*

## 1. Introduction

Action recognition systems are typically trained on data captured from a third-person or spectator perspective [37, 56]. However, in areas such as robotics and augmented reality, we capture data through the eyes of agents, *i.e.*, in a first-person or egocentric perspective. With head-mounted devices such Ray-Ban Stories becoming popular, action recognition from egocentric videos is critical to enable downstream applications, such as contextual recommendations or reminders. However, egocentric action recognition is fundamentally different and more challenging [6, 7, 43, 55]. While third-person video clips are often curated, egocentric video clips are uncurated and have low-level corruptions, such as large motion blur due to head motion. Moreover, egocentric perception requires a careful understanding of the camera wearer's physical surroundings, and must interpret the objects and interactions from the wearer's perspective.

Recognizing egocentric activity exclusively from one

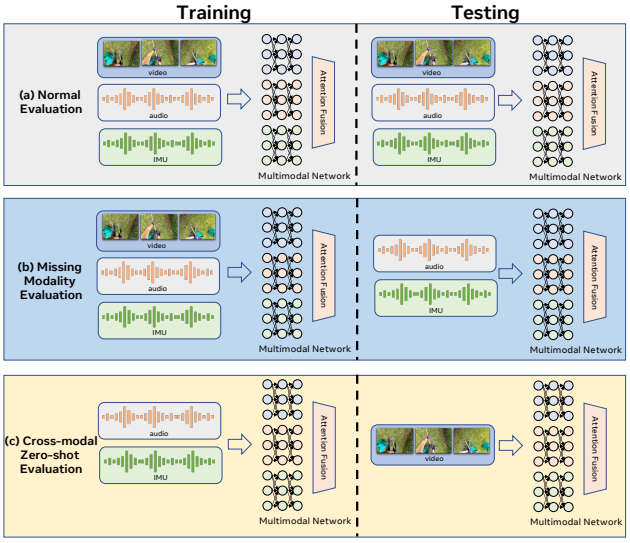


Figure 1. **Overview of *MMG-Ego4D* challenge**. In a typical evaluation setting (a) networks are trained for the supervised setting or the few-shot setting using training/support sets with data from all modalities and evaluated on data points with all modalities. However, there can often be a mismatch between training and testing modalities. Our proposed challenge contains two tasks to mimic these settings. In (b) *missing modality evaluation*, the model can only use a subset of training modalities for inference. In (c) *Cross-modal zero-shot evaluation*, the models are on modalities unseen during training.

*Equal contribution
†Work done during an internship at Meta Reality Labs.

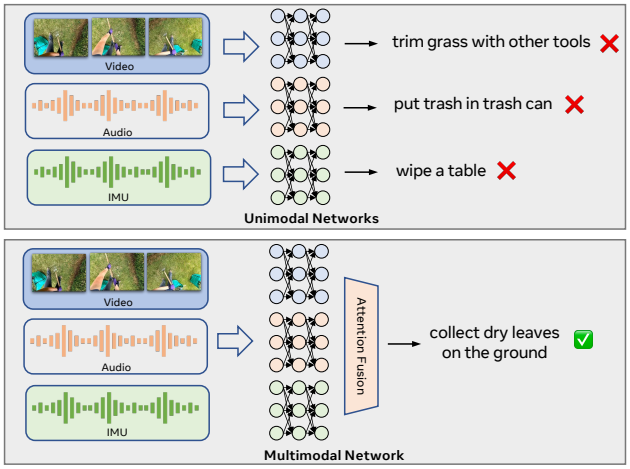


Figure 2. **Multimodal data is crucial for egocentric perception**. Input data consists of three modalities: video, audio, and IMU. (top) Video action recognition identifies the clip with a tool and much grass in the background as the class *trim grass with other tools*. Audio action recognition system classifies the periodic rubbing sound as *put trash in a trash can*. The IMU model classifies the head movement action into the class *wipe a table*. (bottom) Multimodal action recognition system correctly combines the video feed and audio feed and identifies the activity as *collect dry leaves on the ground*.

modality can often be ambiguous. This is because we want to perceive what the device's wearer is performing instead of what the camera feed is capturing. To this end, Multimodal information can be crucial for understanding and disambiguating the user's intent or action. We demonstrate it through an example in Fig. 2. In the example, the video feed shows a tool in the background of the grassland. An activity recognition model exclusively based on video recognizes it as the class *trim grass with other tools*. Similarly, a model exclusively trained in audio identifies the rubbing sounds in the clip as the class *put trash in a trash can*, and an IMU model mistakes the head motion as *wipe a table*. However, a multimodal system correctly identifies the class as *collect dry leaves on the ground* by combining video, audio, and IMU signals.

While using multimodal information is essential to achieve state-of-the-performance, it also presents a unique challenge - *we may not be able to use all modalities in the real world* due to security or efficiency considerations. For example, a user might be located in a sensitive environment and decide to turn off the camera due to security concerns. Similarly, users may turn off microphones so that their voices are not heard. In these situations, multimodal systems must be able *generalize to missing modalities* (Fig. 1 (b)), *i.e.*, work with an incomplete set of modalities at inference, and make a robust prediction. These challenges are not just limited to inference time but could manifest in restrictions during training. For example, if a user has to train a system, often in a few-shot setting, computationally expensive modalities like video are best trained on the cloud. However, the user might prefer that their data stays on the device. However, the video will consume $60\times$ more storage, and $43\times$ more compute compared to cheaper modalities like IMU (see Tab. 1), significantly increasing the difficulty of training on devices with limited compute and storage. In this situation, we may want to enable training with computationally less demanding modalities like audio while maintaining the flexibility of performing inference on more informative modalities like video. Multimodal systems should *robustly generalize across modalities*.

| **Modality** | video | audio | IMU |
|---|---|---|---|
| **Memory per second of data (KB)** | 593.92 | 62.76 | 9.44 |
| **Typical model FLOPs (G)** | 70.50 | 42.08 | 1.65 |

Table 1. **Compute and memory cost for different modalities**. Memory used per second for each modality is computed by averaging the memory used by 1000 data points drawn randomly from Ego4D [27]. The provided compute number corresponds to the forward pass cost of MViT [15] for video, AST [26] for audio, and a ViT [11] based transformer model for IMU data.

In this work, we propose *MMG-Ego4D*: a challenge designed to measure the generalization ability of egocentric activity recognition models. Our challenge consists of two novel tasks: (1) *missing modality generalization* aimed at measuring the generalization ability of models when evaluated on an incomplete set of modalities (shown in Fig. 1 (b)), and (2) *cross-modal zero-shot generalization* aimed at measuring the generalization ability of models in generalizing to unseen modalities during test time (shown in Fig. 1 (c)). We evaluate several widely-used architectures using this benchmark and introduce a novel approach that enhances generalization capability in the *MMG-Ego4D* challenge, while also improving performance in standard full-modalities settings. Our primary contributions are:

- **MMG Problem**. We present *MMG*, a novel and practical problem with two tasks, *missing modality generalization* and *cross-modal zero-shot generalization*, for evaluating the generalization ability of multimodal action recognition models. These tasks are designed to support real-world security and efficiency considerations, and we define them in both supervised and more challenging few-shot settings.

- **MMG-Ego4D Dataset**. To facilitate the study of MMG problem in ego-centric action recognition task, we introduce a new dataset, *MMG-Ego4d*, which is derived from Ego4D [27] dataset by preprocessing the data points and thoroughly re-annotating by *human experts* to suit the task. To the best of our knowledge, this is the first work

to introduce these novel evaluation tasks and a benchmark challenge of its kind.

- **Strong Baselines**. We present a new method that achieves strong performance on the generalization ability benchmark and also improves the performance under the normal full-modalities setting. Our method employs a Transformer-based fusion module, which allows for flexible input of different modalities. We employ a cross-modal contrastive alignment loss to project features of different modalities into a unified space. Finally, a novel loss function is introduced, which is called *cross-modal prototypical loss*, achieving state-of-the-art results in multimodal few-shot settings. Extensive ablation studies are performed to identify each proposed component's contribution.

## 2. Related Work

**Multimodal egocentric action recognition**. Action recognition systems are typically trained on video [15–19, 25, 65, 70]. However, for *egocentric* activity recognition (*i.e.*, first-person perspective, or recognizing the activity the user wearing the capturing device is performing), complimentary multimodal information is essential for identifying the correct activity (see Fig. 2). Previous methods for multimodal fusion in egocentric activity recognition have ranged from simple concatenation [57, 73] to tensor decomposition [45], with some recent studies adopting transformer-based architectures [1,38,46,49] that have shown promising results. In this work, we utilize a Transformer-based fusion module and a modality dropout training strategy to further improve performance on *MMG* tasks.

**Generalizability of multimodal models**. As one of the tasks belonging to *MMG* problem, the missing modality problem has been studied by a few work recently [40, 46, 47, 53, 66, 67]. However, most work focuses on the bimodal situation. [67] solve the multimodal image (two domains) classification problem by learning factorized multimodal representations. [47] addresses the audio-visual classification problem leveraging a Bayesian meta-learning framework. [46] specifically investigate the robustness of the multimodal transformer model to missing-modality data on the text-visual classification task, and improve robustness via multi-task learning and a searched optimal fusion strategy. Cross-modal zero-shot action recognition is still an under-explored new problem. It is related to the cross-modal retrieval problem [63, 74, 75, 77], while the latter focuses on how to measure the feature similarity across different modalities.

**Multimodal few-shot learning**. Multimodal few-shot learning [13, 48, 52, 53, 68] is an emerging research area that aims to enable machine learning models to recognize and classify new objects based on limited examples from multiple modalities. Existing research in few-shot learning has predominantly focused on a single modality, like image [5, 8, 12, 23, 24, 41, 50, 59, 62, 64, 78] or language [2, 29, 71, 72, 76]. However, there has been an increasing interest in extending few-shot learning to multimodal scenarios. Pioneering work in this area includes using text-conditional GANs to augment data via hallucinating images, as demonstrated in [52,53]. Eloff *et al.* [13] utilize a siamese network for a one-shot cross-modal matching problem on speech and image modalities.

**Datasets and benchmarks**. Availability of datasets and clearly defined benchmark tasks have been a driving factor in improving performance in use cases like classification [9], detection [14], segmentation [44] and action recognition [3, 14, 28, 37]. While performance on these tasks has often surpassed human performance [31], researchers have shown that state-of-the-art methods are often fragile [35] and do not generalize well to slightly different data points like corruptions [10, 21] or adversarial examples [4, 39]. Having clearly defined benchmarks, datasets, and tasks to measure the generalization ability has greatly contributed to driving robustness research [32–34]. Our proposed benchmark and dataset, *Ego4D-MMG*, is the first benchmark designed specifically to measure multimodal generalization ability in egocentric action recognition. We hope this benchmark will spur progress in MMG tasks and encourage the development of safety-aware generalizable models.

## 3. Proposed Benchmark: *MMG-Ego4D*

### 3.1. Overview

**Preliminaries**. We use the term "supervised setting" to refer to the regular action recognition task where a large number of labeled training data (training set) are available. During test time, the goal is to classify each testing data point (testing set) into one of the training labels. In contrast, in the few-shot setting, only a few labeled training data are available. The goal at test time is the same as in the supervised setting. In practice, we refer to the training and testing sets in the few-shot setting as the support and query sets, respectively. The collection of support and query sets together is called an "episode" [20, 61, 69]. The terms training modalities and testing modalities refer to the available modalities during supervised training and testing, respectively, in the supervised setting. In the few-shot setting, they refer to the support and query modalities.

**Data**. The *MMG-Ego4D* dataset comprises data points with three modalities - video, audio, and inertial motion sensors (IMU) sourced from the Ego4D dataset [27] (we illustrate why we do not choose other datasets in supplementary). The IMU data contains signals obtained from the accelerometer and gyroscope. We use approximately 202 hours of data from the Ego4D dataset to create our benchmark: 167 hours of unlabelled temporal-aligned Video-

Audio-IMU data and 35 hours of labeled temporal-aligned data. We perform several steps to make the data suitable for our benchmark.

First, we identify timestamps in the data where the activity occurs and standardize each data point to five seconds, drawing data from the Ego4D Moments track [27]. IMU and video data are subsampled to 200 Hz and 4 FPS.

Second, several data points in Ego4D contain multiple labels, primarily resulting from (1) multiple activities being performed and (2) activities being related to each other in a hierarchy (e.g., mixing ingredients vs. cooking). We used the WordNet hierarchy as a heuristic to consolidate the label space, using human annotators to scrutinize the labels. If annotators could not conclusively identify a single correct label, we discarded that data point from our benchmark.

Finally, we created the *MMG-Ego4D* few-shot benchmark with two main criteria. Firstly, it must consist of two semantically-disjoint class sets, namely the base classes and novel classes. Secondly, data points from the same original clip cannot be present in both the base and novel classes. We accomplished this in two steps. Initially, the annotators manually split the 79 labels into 65 base classes and 14 novel classes, ensuring that no semantically similar labels were included in the few-shot evaluation benchmark. Then, we confirmed that data points from the same underlying clip were not present in both the base and novel classes. We used the base classes as the training set for the supervised task and drew additional data from the Ego4D Moments track to form the corresponding testing set for the supervised task.

### 3.2. Proposed *MMG* Tasks

The goal of *MMG-Ego4D* is to evaluate the generalization ability of machine learning algorithms in situations where there is a mismatch between training and testing modalities. Humans can deal with missing modalities quite well. For example, we can identify an action from just a video. Similarly, even if a concept was introduced to us only using video, we can often identify this concept using another modality like audio (e.g., a crying baby). In this section, we describe two novel tasks designed to evaluate the generalization ability of multimodal activity recognition systems. These tasks reflect real-world security considerations while using wearable devices. Further, performance on these tasks could also measure how close our current multimodal machine perception is to human perception. The overview of *MMG* tasks is presented in Fig. 1.

**Missing modality evaluation**. This task measures how well models can perform inference using only a *subset* of modalities that were used for training. During inference, maybe due to power or computational constraints, we may only use a subset of the modalities to perform the evaluation. This presents us with variable evaluation settings, and we select some of them to report their results on the benchmark. In the context of the few-shot setting, this task reduces to using a subset of the support modalities as the query modalities.

**Zero-Shot Cross-Modal generalization**. This task measures how well models can generalize to unseen modalities. The training and testing modalities are disjoint. In our context, the models may only use IMU and audio for training (training video models is extremely expensive), but they could use video data at test time (the budget for video inference is acceptable and may yield better results). Similarly, in the context of few-shot learning, the support and query modalities are disjoint in this task.

## 4. Improving Generalization Performance

This section introduces a strong baseline that achieves high performance on the proposed *MMG-Ego4D* benchmark. Our method comprises three novel components designed to improve the generalization ability of multimodal systems. We begin by presenting an overview of our proposed method, emphasizing the significance of each component, and providing a detailed description of their implementation.

### 4.1. Method Overview

We illustrate the overview of our proposed method pipeline in this section. Under the few-shot setting, all evaluation tasks adopt the same training pipeline, composed of three stages. (1) *unimodal supervised pre-training*: feature extractor for each modality is trained separately. (2) *multimodal supervised pre-training*: a fusion module is attached at the end of unimodal networks to form a multimodal system, which is then trained with a cross-entropy loss and a cross-modal contrastive alignment loss. The latter loss term aims to enhance the multimodal generalizability of the model by constructing a unified feature space for all modalities. (3) *multimodal meta-training*: the multimodal network is meta-trained with prototypical-based loss to further improve the model's cross-modal generalizability. It's worth noting that data of all modalities are used in the above training pipeline of the few-shot setting. The modality restriction of *MMG-Ego4D* tasks is only applied in the support and query set during the few-shot evaluation.

In the supervised setting, the regular and missing modality evaluation settings adopt the same training pipeline, containing (1) *unimodal supervised pre-training* and (2) *multimodal supervised pre-training*, the same as the first two stages of the few-shot setting training pipeline. In contrast, the zero-shot cross-modal setting has a different two-stage training pipeline. (1) *multimodal unsupervised pre-training*: the multimodal network is trained with a cross-modal contrastive alignment loss using unlabeled data, to establish a modality-agnostic unified feature space. (2) *multimodal supervised pre-training*: the multimodal network is

| Setting | Task | Multimodal unsupervised pre-train | Unimodal supervised pre-train | Multimodal supervised train | Multimodal meta-train |
|---|---|---|---|---|---|
| Supervised | Regular | - | $\mathcal{L}_{\text{CE}}$ | $\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{align}}$ | - |
| | Missing Modal | - | $\mathcal{L}_{\text{CE}}$ | $\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{align}}$ | - |
| | Zero-Shot | $\mathcal{L}_{\text{align}}$ | - | $\mathcal{L}_{\text{CE}}$ | - |
| Few-shot | Regular | - | $\mathcal{L}_{\text{CE}}$ | $\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{align}}$ | $\mathcal{L}_{\text{proto}}$ |
| | Missing Modal | - | $\mathcal{L}_{\text{CE}}$ | $\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{align}}$ | $\mathcal{L}_{\text{proto}}$ |
| | Zero-Shot | - | $\mathcal{L}_{\text{CE}}$ | $\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{align}}$ | $\mathcal{L}_{\text{proto}}$ |

Table 2. **Training pipelines of supervised & few-shot settings**. $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss. $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{proto}}$ are cross-modal contrastive alignment loss and cross-modal prototypical loss, which will be explained in Sec. 4.3 and 4.4.

trained using a cross-entropy loss, without the contrastive alignment loss term used in previous settings, as the evaluation modality is absent in the labeled data, due to the restriction of this setting. It is meaningless to construct an alignment between the training modalities. Therefore we choose to build such an alignment using the modality-complete unlabeled data, which also does not violate the rule of this setting. It should be noted that the modality restriction in the *MMG-Ego4D* tasks applies to the labeled training data used in the *multimodal supervised pre-training* stage and the evaluation stage. This is different from the few-shot setting. We have summarized the training pipeline in Tab. 2.

### 4.2. Multimodal Network with a Transformer-based Fusion Module

Our proposed multimodal network consists of two main components: unimodal backbones and a Transformer-based fusion module. The unimodal backbones consist of three separate feature extractors, which extract features from different input modalities. The fusion module aims to fuse and aggregate the features of different modalities from unimodal backbones and output the fused feature. There are two widely-used options for fusing modalities: using an MLP to process the concatenated representations of different modalities [51, 54, 57], or utilizing a Transformer-based fusion module to take a series of tokens from different modalities [46, 49, 60]. We adopt the Transformer-based fusion design as it can easily scale to an arbitrary number of input tokens using attention modules. This is especially important as the multimodal model is expected to handle data with a varying number of modalities in the context of our proposed task. The final output of the fusion module is obtained by averaging the output tokens instead of using the CLS token [11, 46]. Formally, the output of the fusion module $\boldsymbol{z}_{\text{fuse}}$ can be written as follows:

$$\boldsymbol{z}_{\text{fuse}} = f\left(\left[\boldsymbol{x}^{m}_{\text{output}} + \boldsymbol{e}^{m}; |m \in \{\text{audio, video, IMU}\}\right]\right), \tag{1}$$

where $\boldsymbol{x}^{m}_{\text{output}}$ represents the output representation of the feature extractor for modality $m$. $f$ is the fusion module that takes a sequence of input tokens from different modalities. Tokens from each modality are augmented with a modality-specific learnable embedding $\boldsymbol{e}^{m}$, which is used to disambiguate input tokens' modality information.

During the training of the fusion module, we applied a technique named *modality drop*. A subset of modalities is randomly dropped out with a probability $p$ during training, to ensure the robustness of the fusion module to a varying number of input modalities.

### 4.3. Cross-Modal Alignment Multimodal Training

In the zero-shot cross-modal setting, the multimodal model is required to learn and infer from disjoint modalities. One approach to achieving this is to construct a unified feature space that captures representations from different modalities. The feature space should ensure that features from the same data point but different modalities are in close proximity to each other. This allows knowledge learned from one modality to be applied to inference in other modalities. To achieve this, we propose to align features from the same data point but different modalities in multimodal training with contrastive loss. Specifically, the unimodal feature output by the fusion module is represented as follows:

$$\boldsymbol{z}_{m} = f\left(\boldsymbol{x}^{m}_{\text{output}} + \boldsymbol{e}^{m}\right), \quad m \in \{\text{audio, video, IMU}\}, \tag{2}$$

which is expected to lie in the unified feature space. We impose Noise Contrastive Estimation (NCE) [58] loss to align video-audio and video-IMU pairs, drawn from different time stamps of video-audio-IMU data. Positive pairs consist of different modalities pairs from the same temporal location, while negative pairs are from different temporal locations. Our NCE alignment loss $\mathcal{L}_{\text{NCE}}$ is written as follows:

$$\mathcal{L}_{\text{align}}\left(\boldsymbol{z}_{\text{video}}, \boldsymbol{z}_{m}\right) = \sum_{m \in \{\text{audio, IMU}\}} - \log\left(\frac{\exp\left(\boldsymbol{z}^{\top}_{\text{video}} \boldsymbol{z}_{m}/\tau\right)}{\exp\left(\boldsymbol{z}^{\top}_{\text{video}} \boldsymbol{z}_{m}/\tau\right) + \sum_{z' \in \mathcal{N}} \exp\left(\boldsymbol{z}^{\top}_{\text{video}} \boldsymbol{z}'_{m}/\tau\right)}\right), \tag{3}$$

where $\mathcal{N}$ are negative pairs in a batch. We use cosine similarity as the feature distance measurement metric in our NCS loss. $\tau$ is a temperature parameter controlling the softness. Unlike previous methods that build a hierarchical common space [1], our approach defines a unified feature space for all modalities.

### 4.4. Cross-Modal Prototypical Loss

*What properties can help representations better generalize in the few-shot task?* We design a novel extension of prototypical loss [61] that takes into account the alignment between features of different modalities.

The prototypical loss aims to minimize the distance between the centroid of support embeddings and the query embeddings in the feature space, where the labels for query data points are assigned according to their distance to every support centroid. In our proposed approach, support and query examples can belong to different modalities, allowing for cross-modal alignment (see Fig. 3). We use $\boldsymbol{z}_m^k$ to denote a unified space support feature of class $k$ and $\hat{\boldsymbol{z}}_n$ to represent a unified space query feature, where they might belong to different modalities $m, n \in \{\text{audio, video, IMU}\}$. The centroid unified space support feature $c_m^k$ is calculated by averaging:

$$c_m^k = \frac{1}{|\mathcal{Z}_m^k|} \sum_{\boldsymbol{z}_m^k \in \mathcal{Z}_m^k} \boldsymbol{z}_m^k, \tag{4}$$

where $\mathcal{Z}_m^k$ is the set of support features of class $k$ with modality $m$.

The predicted probability of a query example $\hat{\boldsymbol{z}}_n$ belonging to class $k$ is computed using the negative exponential of the $\ell_2$ distance $d$ between the query feature and the centroid of the unified space support feature for class $k$:

$$P_k = \frac{\exp\left(-d\left(\hat{\boldsymbol{z}}_n, c_m^k\right)\right)}{\sum_{k'} \exp\left(-d\left(\hat{\boldsymbol{z}}_n, c_m^{k'}\right)\right)}, m, n \in \{\text{audio, video, IMU}\} \tag{5}$$

Our proposed cross-modal prototypical loss $\mathcal{L}_{\text{proto}}$ is then formulated as the negative log-likelihood loss between the predicted probability and the ground truth class for the query example $\hat{y}$:

$$\mathcal{L}_{\text{proto}} = \text{NLL}\left(\log\left[P_0, P_1, ...P_{N-1}\right], \hat{y}\right). \tag{6}$$

In summary, our cross-modal prototypical loss extends the prototypical loss by enabling the cross-modal alignment between support and query features in the unified feature space. This loss can improve the generalization ability of representations in the zero-shot cross-modal task under the few-shot setting.

# 5. Experimental Setup

## 5.1. Architecture Details

**Unimodal backbones**. We use MViT-B ($16 \times 4$) [15] as the feature extractor for video modality, which is pre-trained on Kinetics-400 [37]. Audio Spectrogram Transformer (AST) [26] is used as the audio feature extractor, and it is pre-trained on AudioSet [22]. For IMU feature extractor, we designed a ViT [11] based transformer network.
**Fusion module**. Our fusion module is a transformer network with two layers. Each layer contains a self-attention block with 12 heads. The embedding dimension is 768.

## 5.2. Training & Evaluation Details

We illustrate some basic details of the model training and evaluation. Hyper-parameters like learning rate and batch size are detailed in our supplementary material.

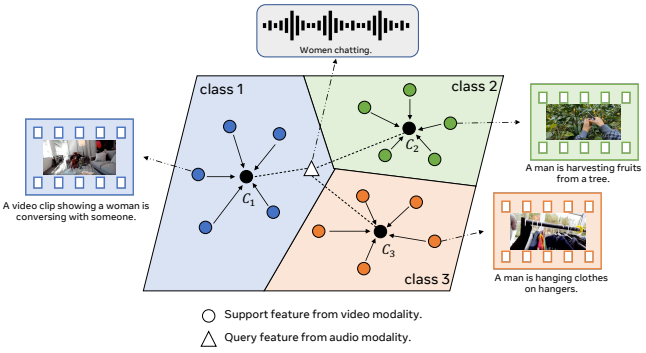


Figure 3. **Cross-modal prototypical loss**. Few-shot prototypes centroid $C_k$ computed by averaging support examples' feature. In contrast to the vanilla prototypical loss, our approach allows support and query examples to belong to different modalities. The figure shows an example where the support examples are video data, and the query example is audio data.

| Model | FLOPs (G) | Param (M) | Modality | 5 Way 5 Shot Accuracy | Top-1 Accuracy |
|---|---|---|---|---|---|
| MViT-B [15] | 70.50 | 36.50 | video | 58.89 | 52.40 |
| AST [26] | 42.08 | 87.03 | audio | 31.06 | 39.48 |
| IMU Transformer | 1.65 | 15.55 | IMU | 40.07 | 29.78 |

Table 3. **Unimodal few-shot & supervised evaluation results**. Networks are trained on each modality independently. Video achieves the best performance, while also consuming more computational resources.

**Supervised setting**. Our model uses *MMG-Ego4D* base classes for multimodal supervised training. Under the zero-shot cross-modal setting, our model also utilizes *MMG-Ego4D* unlabeled data to do the multimodal unsupervised pre-training. We use Top-1 Accuracy to measure model performance.

**Few-Shot setting**. We use the finetune-based method to perform few-shot evaluation, where a small neural network is trained on the support set and is used to classify data points in the query set [30,36,42]. We adopt the standard N-way K-shot setting [20, 69] as the evaluation setting. Top-1 Accuracy is used to measure model performance. The final number is obtained by averaging the results on 10 000 episodes.

# 6. Results on *MMG-Ego4D* Benchmark

## 6.1. *MMG-Ego4D* Few-Shot Setting Results

**Multimodal system outperforms unimodal system significantly.** Tab. 3 presents the few-shot classification results for individual modalities. Notably, the video modality achieves the highest accuracy, which is anticipated since most classes can be easily recognized using visual information. However, as illustrated in Fig 2, fusing information from different modalities is critical to achieving better per-

| Eval. Setting | Support Modalities | | | Query Modalities | | | 5 Way 5 Shot Accuracy |
|---|---|---|---|---|---|---|---|
| | Video | Audio | IMU | Video | Audio | IMU | |
| Regular | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 63.00 |
| Missing Modality | ✓ | ✓ | ✓ | ✓ | ✓ | | 61.76 |
| | ✓ | ✓ | ✓ | | ✓ | ✓ | 50.77 |
| | ✓ | ✓ | ✓ | ✓ | | ✓ | 62.79 |
| | ✓ | ✓ | ✓ | ✓ | | | 62.68 |
| | ✓ | ✓ | ✓ | | ✓ | | 43.65 |
| | ✓ | ✓ | ✓ | | | ✓ | 47.48 |
| Zero-Shot Evaluation | | ✓ | | ✓ | | | 46.90 |
| | | | ✓ | ✓ | | | 42.07 |
| | | ✓ | ✓ | ✓ | | | 50.80 |
| | ✓ | | | | ✓ | | 44.01 |
| | ✓ | | | | | ✓ | 46.56 |
| | ✓ | | | | ✓ | ✓ | 49.37 |

Table 4. **Multimodal few-shot evaluation results**. These results are obtained with a single network that works across all three evaluation settings. We show the *regular evaluation* results in the first block, where the model is trained and evaluated with all the modalities. The second block presents *missing-modality results*, where the model is trained on all modalities but evaluated only on a subset. The last block is the result of *cross-modal zero-shot* evaluation, where the training and evaluation modalities are disjoint. Note that all results are obtained using the same model weight. Our supplementary material provides results with more training and test modalities configurations.

formance in egocentric action detection. Our proposed multimodal system outperforms the best-performing unimodal system by 4.11 in terms of accuracy (Tab. 4 block 1).

**Missing modality generalization**. We present the results of the missing modality evaluation in the second block of Tab. 4, where the query modality is a subset of the support modality. Our model exhibits good generalizability even when some modalities are missing during evaluation, achieving solid accuracy. Notably, including the video modality in the query set yields a slight change in performance compared to the multimodal case. When video modality is not included, there is a 19.41% drop in accuracy, indicating that video modality is the most informative. Surprisingly, when queries have only one cheap modality (audio or IMU), our method outperforms unimodal results (Tab. 3) by a large margin of 19.24% on IMU and 40.53% on audio modality, demonstrating the effectiveness of our approach.

**Zero-shot cross-modal generalization**. This task presents a more significant challenge than missing modality generalization as the support and query modalities are disjoint. We select a few combinations and present the results in the last block of Tab. 4. To enable efficient training, we choose a setting where the support modality is computationally cheap, such as IMU and Audio, while the query modality is relatively more informative, such as video, to achieve high performance. Our model significantly outperforms the audio and IMU unimodal settings using this evaluation setting. We also present the results of using video as the support modality and IMU and/or audio as the query modality, where our model still obtains decent accuracy. While we did not include all support-query modality combinations in the paper due to space limitations, readers can refer to our supplementary materials for additional results.

| Eval. Setting | Train Modalities | | | Test Modalities | | | Top-1 Accuracy |
|---|---|---|---|---|---|---|---|
| | Video | Audio | IMU | Video | Audio | IMU | |
| Regular | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 55.66 |
| Missing Modality | ✓ | ✓ | ✓ | ✓ | ✓ | | 55.47 |
| | ✓ | ✓ | ✓ | | ✓ | ✓ | 37.07 |
| | ✓ | ✓ | ✓ | ✓ | | ✓ | 54.57 |
| Zero-Shot Evaluation | | ✓ | | ✓ | | | 30.98 |
| | | | ✓ | ✓ | | | 20.00 |
| | | ✓ | ✓ | ✓ | | | 25.03 |
| | ✓ | | | | ✓ | | 43.43 |
| | ✓ | | | | | ✓ | 35.67 |
| | ✓ | | | | ✓ | ✓ | 41.02 |

Table 5. **Supervised setting evaluation results**. Results are organized following the same structure as in Tab. 4. The model has the same weight in regular and missing modality evaluation.

## 6.2. *MMG-Ego4D* Supervised Setting Results

The results of the supervised settings are presented in Tab. 5. Our multimodal model outperforms each unimodal model in Tab. 3 significantly in the regular setting. Regarding the missing modality evaluation, our method exhibits strong generalization ability in the presence of missing modalities. If the video modality is preserved in the evaluation modality, the performance only experiences a minor drop. However, when video data is missing during evaluation, the performance drops by around 33%, suggesting that the video modality is more informative than the other two modalities. The last block of Tab. 5 shows the zero-shot cross-modal results. We explore two cases: using expensive modalities for training and cheap modalities for inference, and using cheap modalities for training and expensive modalities for inference. We observe that the model performs better in the latter case, indicating that learning from informative modalities benefits the model more.

## 6.3. Insights from Ablation Study

In this section, we carefully ablate the effect of each component in our designed multimodal system under various evaluation settings, including the regular, missing modality, and cross-modal zero-shot evaluations.

**Fusion module**. In this study, we propose the use of a Transformer-based fusion module as an alternative approach to integrating information from different modalities in a multimodal network. To evaluate its performance, we conduct a comparative analysis against an MLP-based fusion module that concatenates representations from diverse

| Eval. Setting | Train/Support Modal. | | | Test/Query Modal. | | | Fusion Module | Contrastive Alignment | Top-1 Accuracy | Cross-Modal Proto. Loss | 5 Way 5 Shot Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Video | Audio | IMU | Video | Audio | IMU | | | | | |
| Regular | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Attention | ✓ | **55.66** | ✓ | **63.00** |
| | | | | | | | Attention | × | 52.18 | ✓ | 61.16 |
| | | | | | | | MLP | ✓ | 52.79 | ✓ | 58.67 |
| | | | | | | | Attention | ✓ | - | × | 62.37 |
| Missing Modality | ✓ | ✓ | ✓ | | ✓ | ✓ | Attention | ✓ | **37.07** | ✓ | **50.77** |
| | | | | | | | Attention | × | 21.32 | ✓ | 40.87 |
| | | | | | | | MLP | ✓ | 32.89 | ✓ | 49.00 |
| | | | | | | | Attention | ✓ | - | × | 50.03 |
| Zero-shot Cross-Modal | | ✓ | ✓ | ✓ | | | Attention | ✓ | **25.03**$^*$ | ✓ | **51.40** |
| | | | | | | | Attention | × | 2.37 | ✓ | 33.93 |
| | | | | | | | MLP | ✓ | 24.54$^*$ | ✓ | 51.08 |
| | | | | | | | Attention | ✓ | - | × | 50.80 |

Table 6. **Ablation study of each design component under supervised & few-shot settings**. Our proposed components improve the performance under all evaluation settings. Note that cross-modal prototypical loss is only applied under the few-shot setting. $^*$Different from other settings, the cross-modal contrastive alignment loss is applied at the unsupervised multimodal pre-training stage in the supervised zero-shot cross-modal setting.

modalities and processes them using an MLP. To ensure a fair comparison, we maintain the dimensionality of input and output representations of both modules to be consistent, with a similar number of parameters. In situations where some modalities are not present in the input of the MLP-based fusion module, we replace their representations with zero vectors. The results of the ablation study presented in Tab. 6 demonstrate that the Transformer-based fusion module outperforms the MLP-based fusion module in both few-shot and supervised learning scenarios across all tasks. We also investigate three decision choices empirically. Specifically, we examine the efficacy of using the CLS token or averaging all output tokens for the final prediction. We find that averaging all output tokens produces better performance. Additionally, we evaluate the inclusion of modality-specific embeddings before fusion and find that it is effective in aiding the model's ability to differentiate between modalities. Finally, we experiment with various dropout rates ($p$) for modality dropout and find that consistent performance is obtained across a range of values ($0.3$ to $0.8$), with the best results achieved at $p = 0.6$.

**Cross-modal contrastive alignment loss**. Our motivation for incorporating cross-modal alignment loss into our pipeline is rooted in the desire to enhance cross-modal zero-shot generalization performance. In Tab. 6, the inclusion of this component resulted in a remarkable improvement of $22.66$ and $17.47$ in cross-modal zero-shot generalization performance in supervised and few-shot learning settings, respectively. Additionally, we observed that the incorporation of cross-modal alignment loss also yielded performance gains in regular and missing modality tasks. These results underscore the importance of cross-modal alignment in succeeding in the *MMG-Ego4D* benchmark.

**Cross-Modal prototypical loss**. In our study, we proposed the incorporation of cross-modal prototypical loss as a means of enhancing few-shot performance in MMG tasks. Our experimental results, as demonstrated in Tab. 6, reveal that this novel component contributes to performance improvements of $0.74$ and $0.6$ points in missing modality and zero-shot scenarios, respectively, while also yielding an enhancement of $0.63$ points in the regular modality complete evaluation setting. These findings attest to the efficacy of cross-modal prototypical loss as a valuable addition to the MMG task performance optimization strategy.

## 7. Conclusions

In this paper, we introduced the first comprehensive benchmark for multimodal generalization (MMG) and proposed three components to improve the generalization performance of models. Our benchmark, *MMG-Ego4D*, includes two new tasks and a new dataset. The evaluation of different baseline architectures showed that the generalization ability of current systems is limited. Therefore, benchmarking and improving generalization ability deserve attention, especially as models are deployed into more sensitive use cases. Through extensive experiments and ablation study, we demonstrated that our proposed attention-based fusion mechanism with modality dropout training and alignment of unimodal representation during fusion could improve the performance of supervised and few-shot tasks in *MMG-Ego4D*. Our proposed cross-modal prototypical loss also improves the performance of few-shot tasks in *MMG-Ego4D*. We created a new dataset and introduced novel experiments for the rigorous study of multimodal generalization problems. These methods can increase generalizability and are essential for real-world settings where secure environemnts are important.

## 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. 3, 5

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, 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. 3

[4] Nicholas Carlini, Guy Katz, Clark Barrett, and David L Dill. Ground-truth adversarial examples. 2018. 3

[5] Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu, and Martial Hebert. Image deformation meta-networks for one-shot learning. In *CVPR*, 2019. 3

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 1

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130(1):33–55, 2022. 1

[8] Rajshekhar Das, Yu-Xiong Wang, and Jose MF Moura. On the importance of distractors for few-shot classification. In *ICCV*, 2021. 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[10] Samuel Dodge and Lina Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)*, pages 1–7. IEEE, 2017. 3

[11] 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. *arXiv preprint arXiv:2010.11929*, 2020. 2, 5, 6

[12] Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Diversity with cooperation: Ensemble methods for few-shot classification. In *ICCV*, 2019. 3

[13] Ryan Eloff, Herman A Engelbrecht, and Herman Kamper. Multimodal one-shot learning of speech and images. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8623–8627. IEEE, 2019. 3

[14] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 3

[15] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 2, 3, 6

[16] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020. 3

[17] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 3

[18] 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. 3

[19] 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. 3

[20] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 3, 6

[21] Robert Geirhos, David HJ Janssen, Heiko H Schütt, Jonas Rauber, Matthias Bethge, and Felix A Wichmann. Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*, 2017. 3

[22] 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 *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 6

[23] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Perez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *ICCV*, 2019. 3

[24] Spyros Gidaris and Nikos Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *CVPR*, 2019. 3

[25] Xinyu Gong, Heng Wang, Mike Zheng Shou, Matt Feiszli, Zhangyang Wang, and Zhicheng Yan. Searching for two-stream models in multivariate space for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8033–8042, 2021. 3

[26] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. 2, 6

[27] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 2, 3, 4

[28] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 3

[29] Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. Ppt: Pre-trained prompt tuning for few-shot learning. *arXiv preprint arXiv:2109.04332*, 2021. 3

[30] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *European conference on computer vision*, pages 124–141. Springer, 2020. 6

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 3

[32] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 3

[33] Guenter Hirsch. Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task. *ETSI STQ Aurora DSR Working Group, Dec. 2002*, 2002. 3

[34] Hans-Günter Hirsch and David Pearce. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *ASR2000-Automatic speech recognition: challenges for the new Millenium ISCA tutorial and research workshop (ITRW)*, 2000. 3

[35] Hossein Hosseini, Baicen Xiao, and Radha Poovendran. Google's cloud vision api is not robust to noise. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 101–105. IEEE, 2017. 3

[36] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9068–9077, 2022. 6

[37] 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. 1, 3, 6

[38] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 3

[39] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 3

[40] Hu-Cheng Lee, Chih-Yu Lin, Pin-Chun Hsu, and Winston H Hsu. Audio feature generation for missing modality problem in video action recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3956–3960. IEEE, 2019. 3

[41] Aoxue Li, Tiange Luo, Tao Xiang, Weiran Huang, and Liwei Wang. Few-shot learning with global class representations. In *ICCV*, 2019. 3

[42] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Improving task adaptation for cross-domain few-shot learning. *arXiv preprint arXiv:2107.00358*, 2021. 6

[43] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 619–635, 2018. 1

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[45] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*, 2018. 3

[46] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186, 2022. 3, 5

[47] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310, 2021. 3

[48] Yao Ma, Shilin Zhao, Weixiao Wang, Yaoman Li, and Irwin King. Multimodality in meta-learning: A comprehensive survey. *Knowledge-Based Systems*, page 108976, 2022. 3

[49] 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. 3, 5

[50] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv*, 2018. 3

[51] Juan DS Ortega, Mohammed Senoussaoui, Eric Granger, Marco Pedersoli, Patrick Cardinal, and Alessandro L Koerich. Multimodal fusion with deep neural networks for audio-video emotion recognition. *arXiv preprint arXiv:1907.03196*, 2019. 5

[52] Frederik Pahde, Patrick Jähnichen, Tassilo Klein, and Moin Nabi. Cross-modal hallucination for few-shot fine-grained recognition. *arXiv preprint arXiv:1806.05147*, 2018. 3

[53] Frederik Pahde, Oleksiy Ostapenko, Patrick Jä Hnichen, Tassilo Klein, and Moin Nabi. Self-paced adversarial training for multimodal few-shot learning. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 218–226. IEEE, 2019. 3

[54] Yagya Raj Pandeya and Joonwhoan Lee. Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, 80(2):2887–2905, 2021. 5

[55] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2847–2854. IEEE, 2012. 1

[56] Ronald Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010. 1

[57] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 439–448. IEEE, 2016. 3, 5

[58] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5

[59] Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Few-shot learning with embedded class models and shot-free meta training. In *ICCV*, 2019. 3

[60] Tim Siebert, Kai Norman Clasen, Mahdyar Ravanbakhsh, and Begüm Demir. Multi-modal fusion transformer for visual question answering in remote sensing. In *Image and Signal Processing for Remote Sensing XXVIII*, volume 12267, pages 162–170. SPIE, 2022. 5

[61] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 3, 5

[62] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *CVPR*, 2019. 3

[63] Christopher Thomas and Adriana Kovashka. Emphasizing complementary samples for non-literal cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4632–4641, 2022. 3

[64] Pavel Tokmakov, Yu-Xiong Wang, and Martial Hebert. Learning compositional representations for few-shot recognition. In *ICCV*, 2019. 3

[65] 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

[66] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1405–1414, 2017. 3

[67] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018. 3

[68] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 3

[69] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 3, 6

[70] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1430–1439, 2018. 3

[71] Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*, 2021. 3

[72] Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. Language models are few-shot multilingual learners. *arXiv preprint arXiv:2109.07684*, 2021. 3

[73] 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

[74] Hong Xuan and Xi Stephen Chen. Dissecting deep metric learning losses for image-text retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2164–2173, 2023. 3

[75] Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. X 2-vlm: All-in-one pre-trained model for vision-language tasks. *arXiv preprint arXiv:2211.12402*, 2022. 3

[76] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021. 3

[77] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10394–10403, 2019. 3

[78] Xiatian Zhu, Antoine Toisoul, Juan Perez-Rua, Li Zhang, Brais Martinez, and Tao Xiang. Few-shot action recognition with prototype-centered attentive learning. In *BMVC*, 2021. 3