

Exploring Missing Modality in Multimodal Egocentric Datasets

Merey Ramazanova¹, Alejandro Pardo¹, Humam Alwassel², and Bernard Ghanem¹

¹ King Abdullah University of Science and Technology
 {merey.ramazanova, alejandro.pardo, bernard.ghanem}@kaust.edu.sa
² Intelmatix

Abstract. Multimodal video understanding is crucial for analyzing egocentric videos, where integrating multiple sensory signals significantly enhances action recognition and moment localization. However, practical applications often grapple with incomplete modalities due to factors like privacy concerns, efficiency demands, or hardware malfunctions. Addressing this, our study delves into the impact of missing modalities on egocentric action recognition, particularly within transformer-based models. We introduce a novel concept—Missing Modality Token (MMT)—to maintain performance even when modalities are absent, a strategy that proves effective in the Ego4D, Epic-Kitchens, and Epic-Sounds datasets. Our method mitigates the performance loss, reducing it from its original $\sim 30\%$ drop to only $\sim 10\%$ when half of the test set is modal-incomplete. Through extensive experimentation, we demonstrate the adaptability of MMT to different training scenarios and its superiority in handling missing modalities compared to current methods. Our research contributes a comprehensive analysis and an innovative approach, opening avenues for more resilient multimodal systems in real-world settings.

Keywords: Missing Modality · Multimodal Video Recognition · Egocentric Videos

1 Introduction

Multimodal video understanding has been the de facto approach for analyzing egocentric videos. Recent works have shown that the complimentary multisensory signals in egocentric videos are superior for understanding actions [24–26, 33, 37] and localizing moments [2, 41, 43, 47]. However, multimodal systems need to be practical for real-world applications that could suffer from the incompleteness of modality inputs due to privacy, efficiency, or simply device failures [30]. For example, when predicting in real-time using a wearable device, parts of the recordings might be scrapped to preserve the privacy of the bystanders/camera wearer [14, 16]. Furthermore, using all sensors could be expensive for a wearable device, opting for cheaper modalities such as audio or IMU [17]. Thus, studying the impact of missing modalities is crucial for realistic performance expectations.

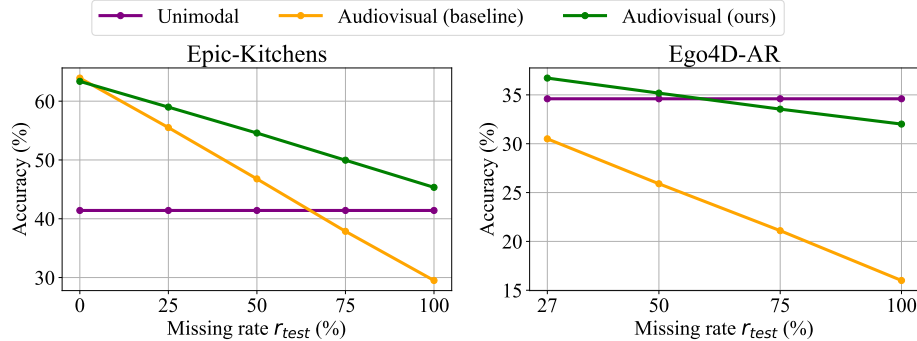


Fig. 1: Most commonly, we train the multimodal models on modal-complete data. These models (orange) fail when encountering modal-incomplete data at test time. Our proposed adaptation to the missing modality (green) significantly improves the performance across datasets. When all test inputs are modal-incomplete ($r_{test} = 100\%$), we surpass unimodal performance (purple) by 5 points in Epic-Kitchens, and double the baseline performance in Ego4D-AR.

Still, the current effort to study the impact of missing modalities in egocentric datasets remains rather limited. Most methods presume all modal inputs to be intact during training and inference. Recent works have studied the effect of missing modalities for different tasks varying from recommendation systems to emotion recognition [29, 35, 38, 40, 44, 46]. Notably, the majority of research concerning missing modalities has primarily addressed the issue during testing [3, 38, 40, 44, 46, 50], while just a handful studied it across both training and testing phases [29, 30, 36]. Similar to our setting, Lee *et al.* [30] propose a strategy to learn prompts for pre-trained backbones to deal with missing modalities. However, they analyze their method for image and text datasets only; we implement our version for action recognition and use it as a baseline in Sec. 4. More recently, Gong *et al.* [14] proposed a benchmark for multimodal generalization, focusing on few-shot learning recognition while considering missing modalities. Though the latter work proposes an interesting benchmark that includes a zero-shot and few-shot setup, no works have diagnosed how recent transformer-based approaches perform when modalities are missing for the action recognition setting.

In this work, we study the problem of missing modalities in egocentric action recognition. First, we investigate how current transformer-based models are affected by incomplete modalities at test time. In Fig. 1, we observe how the current state-of-the-art audiovisual recognition model, Multimodal Bottleneck Transformer (MBT) [37], trained on modal-complete inputs, suffers from a critical degradation in performance when the missing modality rate increases. The advantage of the multimodal backbone (orange) is lost when the missing modality rate in the test set exceeds $\sim 27\%$ (Ego4D-AR) and $\sim 70\%$ (Epic-Kitchens). At this point, the unimodal model (purple) becomes a better alternative. To address this problem, we propose learning the missing modality "template" during

training to replace missing modalities at test time. We call this template the Missing Modality Token (MMT) and explain how to learn it in Sec. 3.4. Fig. 1 (Epic-Kitchens) also shows how our approach (green) dramatically improves the test accuracy and stays at least 5 points above the unimodal performance even when the test set is fully modal-incomplete. Furthermore, our methods enable better multimodal performance overall in modal-incomplete Ego4D-AR.

We verify the effectiveness of our method in 3 egocentric datasets, including Ego4D [16], which has a full coverage of RGB video, but only 70% of videos have audio. We extensively analyze our proposed training strategies, showing how to train with MMT under different missing modality scenarios. Our experiments show that our simple yet effective approach proposes a strong solution to this problem. **Our contributions** are threefold: **(1)** We present a thorough study of the challenge of missing modalities in egocentric action recognition. This involves exploring datasets with varying degrees of modal incompleteness and assessing the influence of the fusion layer. **(2)** We propose the Missing Modality Token (MMT) as a novel solution to address missing modalities during both the training and testing phases. Additionally, we propose a training strategy, termed *random-replace*, to enhance the efficacy of models utilizing MMT. **(3)** We extensively evaluate our method and demonstrate its notable improvement over existing baselines. Through our work, we provide valuable insights and lay the groundwork for developing multimodal backbones that exhibit robustness in the face of missing modalities.

2 Related work

Addressing missing modality. Addressing missing modalities presents a notable challenge, explored through various strategies by researchers from different areas. From medical applications [1] to sentiment analysis [3], missing modalities are a long-standing problem in multimodal understanding. Some methods [11, 12, 42] distill the knowledge from a multimodal teacher to an unimodal RGB model. Others are tailored for scenarios where test data is multimodal yet incomplete in terms of modalities. For example, Ma *et al.* [36] and Colombo *et al.* [3] investigate missing modalities within a Bayesian Meta-learning framework. Meanwhile, Tsai *et al.* [44], Zhao *et al.* [50], and Woo *et al.* [48] attempt to reconstruct missing inputs. Neverova *et al.* [38] focus on multimodal gesture recognition, employing depth, audio, and video streams, and updating network parameters based on different modality combinations. Most of these works rely on modality-specific architectures [20, 28] and/or use complex generative pipelines. Our approach uses a generic multimodal Transformer [45]. Furthermore, these methods assume that the training data is fully modal-complete, which is not the case in current large-scale datasets [17]. Instead, our method applies to modal-complete and modal-incomplete training sets.

Recent studies utilizing transformers, such as the work of Parthasarathy *et al.* [40], explore missing modalities at test time and propose training-time augmentations. Ma *et al.* [35] develop strategies for optimal fusion layers and class

tokens in the context of missing modalities, focusing on image-text datasets. Our research differs by demonstrating the effectiveness of our method across various fusion layers (Sec. 4.7), which avoids any expensive fusion policy learning. Lee *et al.* [30] proposes to learn to prompt large multimodal backbones for image and text classification when modalities are missing at train and test time. We adapt their method to our setting and show that ours is more practical and effective for dealing with missing modalities in egocentric videos. Lastly, Gong *et al.* [14] introduce a benchmark for handling missing modalities within the Ego4D dataset, tailored for few-shot classification³. Our work proposes to diagnose and study the problem in a simpler setting to understand the effect of missing modalities in egocentric video understanding. We want to note that most of related transformer-based methods [14, 35, 40] do not provide the code, which makes it challenging to compare to.

Multimodal egocentric video understanding. Egocentric perception faces distinct challenges compared to traditional video understanding benchmarks such as ActivityNet [7] and Kinetics [23]. The nature of how egocentric datasets are captured means that they usually feature strongly aligned and synchronized audiovisual signals. Key benchmarks in this field, including Epic-Kitchens [4], Epic-Sounds [22], and the more recent and extensive Ego4D [16], have demonstrated the importance of audiovisual learning for understanding egocentric videos due to the complementary nature of the audio and visual modalities [24, 25, 43]. These datasets have facilitated the creation of several audiovisual backbones tailored for video understanding. Xiao *et al.* [49] introduced a CNN-based dual-stream architecture, utilizing SlowFast networks for the visual component [8] and a separate stream for audio [26]. With the advent and adaptability of transformer architectures, several studies have treated different modalities as input tokens for a multimodal transformer encoder [10, 27, 31, 32]. However, self-attention mechanisms can become prohibitively expensive as the number of tokens increases. To address this, Nagrani *et al.* [37] offered an efficient Multimodal Bottleneck Transformer (MBT) that avoids costly self-attention. We build atop MBT and introduce Missing Modality Token to make it robust for missing modalities at train and test times.

3 Dealing with missing modalities

This section details the aspects we consider while addressing the missing modality problem. Namely, the scenarios and evaluation (3.1), the multimodal design and fusion (3.2), the possible naive solutions to the problem (3.3), and our proposed method (3.4).

3.1 Problem statement, setup, and evaluation

Given the training and testing multimodal data samples, let us denote the missing modality rates in each set with r_{train} and r_{test} , respectively. These rates

³ Code and data are not available

are computed by dividing the number of modal-incomplete samples by the total number of samples. Note that we only consider the setting of one modality being missing in the dataset. Our work discusses the strategies of training a multimodal model under two scenarios: **modal-incomplete training set** *i.e.*, some samples in the training set are modal-incomplete ($r_{train} \neq 0\%$), or **modal-complete training set** *i.e.*, all samples in training set all modal-complete ($r_{train} = 0\%$).

To observe the trained model’s behavior under different missing modality severity levels, we create several variants of the test set by manually removing the modality information from the samples until $r_{test} = 100\%$.

Following previous works [35,36] when experimenting with fully modal-complete datasets ($r_{train} = 0\%$ and $r_{test} = 0\%$), we assume the modality with the best unimodal performance (dominant) to be incomplete at test-time (*e.g.*, audio for Epic-Sounds). Unlike previous works [30,35,36], we also validate our adaptation strategies on a dataset with naturally incomplete modalities in train and test splits. We use two modalities commonly available in the egocentric video datasets: visual (RGB frames) and audio. We evaluate classification accuracy on egocentric action recognition datasets.

3.2 Efficient and effective multimodal fusion

We deal with missing modalities while considering the methods proven to be the most effective for multimodal fusion. Previous transformer-based methods addressing missing modalities looked mostly at basic methods, such as early or mid-fusion with cross-modal self-attention, where all tokens are concatenated at the fusion layer. This does not scale well in videos due to the attention mechanism’s quadratic complexity (to the input size) [30,37]. Instead, we use the current state-of-the-art audiovisual fusion, MBT [37], which proved to be more efficient and effective. The bottleneck transformer in the MBT design allows the model to distill and propagate the most essential information across modalities where each modality performs self-attention only with a small number of learnable "bottleneck" tokens. Such design is especially useful for information-dense (redundant) modalities like video.

Fusion Layer. A key aspect of a multimodal fusion approach is the design of the fusion layer L_f . L_f is the layer at which the cross-modal interactions happen. We observe the performance of the original bottleneck model with modal-complete audiovisual inputs and train the model with different fusion layers. We show the test accuracy for these models in Figure 6 (marked as *baseline*) for the Epic-Kitchens and Epic-Sounds datasets. We find that the performance does not change significantly with different fusion layers when $r_{test} = 0$. However, the fusion layer does make a difference when inputs are incomplete (*e.g.*, 35% test accuracy with $L_f = 0$ *vs.* 42% with $L_f = 11$ in Epic-Sounds at $r_{test} = 50\%$), as shown in Section 4.7. Overall, fusing earlier is preferred in Epic-Kitchens, but fusing later gives better results in Epic-Sounds. This outcome is consistent with the observations from the previous work [35]: the best fusion strategy is dataset-specific. While this observation might be intuitive, it is impractical when the model is very sensitive to the fusion layer, as searching for the best layer might

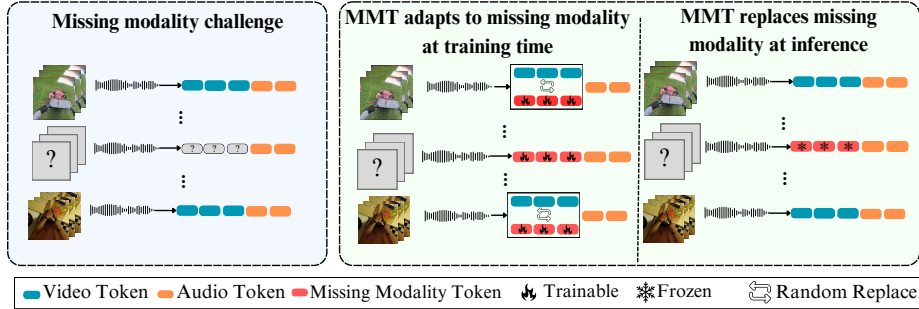


Fig. 2: Learning and Predicting with Missing Modalities. **Left:** Given modal-incomplete data, it is still unclear how to effectively train and predict with a multimodal model (we present some naive baseline methods in Sec. 3.3). **Right:** To address this issue, we introduce a Missing Modality Token (MMT). During training, MMT learns the representation of missing inputs from modal-incomplete samples and modal-complete samples. For the latter, we use *random-replace* to let the network observe the missing inputs and thus learn better representations (Sec. 3.4). At test time, we replace the tokens of missing inputs with MMT to effectively represent them.

be computationally expensive. Thus, we analyze the effect of the fusion layer when modalities are missing and show the effect of our approach in Sec. 4.7.

3.3 Intuitive baselines

While designing effective solutions for missing modalities is challenging (Fig. 2), one could suggest simple and intuitive ways of adapting to missing inputs at test time. We propose the following *training-free baselines* to deal with missing modality at inference time.

Passing missing inputs as tensors with zeros. We employ a straightforward approach of substituting missing modality inputs with tensors filled with zeroes. This method, widely acknowledged in the literature [1, 30, 40], is favored for its simplicity and ease of implementation in practice. Thus, unless stated otherwise, we adopt this method as our primary baseline and refer to it as *baseline*.

Only pass complete inputs. As transformers can process sequences of varying lengths, we can selectively omit tokens corresponding to missing signals and exclusively supply the transformer with non-missing modality tokens. While intuitive, this approach becomes less practical when inference involves batch sizes greater than 1 and not all inputs within the batch exhibit modal incompleteness.

3.4 Our approach to dealing with missing modalities

We suggest a simple and generic way to deal with missing modalities. Instead of passing tensors filled with zeroes or discarding the tokens of the missing inputs at test time, we propose to learn a "template" for the missing inputs. We introduce

a learnable Missing Modality Token (MMT), which is trained to represent the tokens of the missing modality by learning from the tokens of the non-missing modality.

Figure 2 (right) illustrates how MMT represents the tokens of the missing modality at train and test time. When the model encounters a modal-incomplete sample, MMT, a shared parameter, is repeated to match the missing modality token number and passed to the transformer, along with the tokens of non-missing modalities. Note that each token has a different positional embedding added, similar to the mask token in [18]. With this intuitive way, MMT observes all modal-incomplete samples and leverages the non-missing modality tokens to learn.

If trained with modal-incomplete samples only, the model never encounters the inputs it tries to mimic. Furthermore, MMT is limited in the number of training samples (*e.g.* in a dataset with $r_{train} = 10\%$, MMT will only encounter 10% of the data). We try to facilitate it and suggest *random-replace* strategy. With *random-replace*, input tokens of modal-complete samples can also be used to train MMT. Namely, for each modal-complete training sample, with probability p , the tokens of one modality will be replaced with MMT. When $p = 0$, the MMT learns with naturally missing inputs only. If p is set to a non-zero value, the strategy provides more training samples for learning MMT and lets the network observe the same sample as modal-incomplete and modal-complete; thus encouraging the model to understand the relationships and dependencies of the modalities. However, setting p too high might cause high information loss and hinder the performance, especially if we drop the tokens of a more information-dense modality, as we will show in Sec. 4.4

Thus, there are 2 ways of providing training samples for MMT: (1) Using samples with naturally missing modalities and (2) Randomly replacing the tokens of complete samples with MMT. Below, we discuss how both ways are used with modal-complete and modal-incomplete training sets.

Modal-incomplete training set. Recall that r_{train} of the training samples are modal-incomplete and $100\% - r_{train}$ are modal-complete. Thus, MMT can:

1. *Learn from modal-incomplete only.* During training, we use MMT to represent the missing modality inputs for modal-incomplete samples. The tokens of modal-complete samples are never replaced with MMT *i.e.*, $p = 0$.

2. *Learn from modal-complete and modal-incomplete.* Use modal-incomplete samples as in (1), and *random-replace* with non-zero p .

Modal-complete training set. As **all** training samples have complete multi-modal inputs, MMT is trained using *random-replace* with $p > 0$.

Inference. Regardless of the strategy, we replace the missing input tokens with the learned MMT at test time.

4 Experiments

We present a detailed analysis of our MMT under both modal-complete and modal-incomplete training sets. For both scenarios, we explore the usage of

random-replace. Namely, in Sec. 4.4 we ablate the effect of p . In Sec. 4.5, we study how the severity of missing modality in the training data affects the performance. Sec. 4.6 we extend the setup to multiple missing modalities. Additionally, we study the effect of fusion layers L_f in Sec. 4.7. Then, we compare our method to the baselines we proposed and [30] in Sec. 4.8.

4.1 Datasets

We use videos from **Ego4D** [16] for pre-training the MBT backbone. Due to privacy and regulations, only 2.5K of 3.7K video hours have original audio in this dataset. We trim 450K 10-second modal-complete audiovisual clips spanning all 2.5K modal-complete hours. For the downstream tasks, we use the following egocentric action recognition benchmarks:

Epic-Kitchens-100 [5] has 90K trimmed clips of variable length, spanning 100 video hours. Each clip is labeled with a noun + verb pair, which describes the camera-wearer action. In total, there are 300 noun and 97 verb classes in the dataset. We train the model with 2 heads to jointly predict verb and noun classes. All videos in the dataset have complete visual and audio streams ($r_{train} = r_{test} = 0\%$).

Epic-Sounds [22] spans the same 100 video hours as Epic-Kitchens but is annotated with sound labels. This dataset does not follow the noun and verb annotations from Epic-Kitchens; instead, it has 44 unique class labels. The dataset is composed of 79K annotated clips.

Ego4D-AR: We use the annotated clips from the Short-Term Action Anticipation task in Ego4D benchmark [16] to create an action recognition dataset that we dub as Ego4D-AR⁴. Specifically, we use the provided *time-to-contact* timestamps to trim the clips and the anticipated actions as labels. Ego4D-AR contains 142K clips annotated as noun and verb pairs. Overall, there are 128 noun and 81 verb classes. We find that the verb classes are highly imbalanced in this dataset. Therefore, we balance the class weights in the cross-entropy loss during training. We provide more details on the dataset in Supplementary. Similarly to Epic-Kitchens, we use 2 heads to predict the nouns and verbs. As we didn't filter the Ego4D videos for this dataset, it has naturally missing modality. Only 71% of training clips and 73% of test clips have audio ($r_{train} = 29\%$, $r_{test} = 27\%$).

For clarity and due to space constraints, we report the verb accuracy for Epic-Kitchens, the class accuracy for Epic-Sounds, and the noun accuracy for Ego4D-AR in this section. We report the rest of the metrics for Supplementary.

4.2 Implementations details

Pre-training. We use the audiovisual MAEs [9, 13, 18] protocols and train our own implementation of Audiovisual Bottleneck MAE. We use the trimmed Ego4D clips and train for 200 epochs. We mask 70% of audio and 90% of video

⁴ Ego4D does not have an action recognition benchmark

tokens. We use the same pre-trained model for all experiments. More details of the decoder used in the pre-training can be found in the supplementary material.

Architecture. Following [21], we use ViT-Base [6] with 12 transformer layers, 12 attention heads, and embedding dimension 768 as the encoder for each modality. For the fusion design, we follow MBT [37] and fix the number of bottlenecks to $B = 4$ and the fusion layer to $L_f = 8$ (except for Sec. 4.7).

Inputs. Following [15, 21], we convert an audio waveform of t seconds to log Mel-filterbank with 128 Mel-frequency bins, with a Hanning window of 25ms, shifting every 10ms. The output is a spectrogram of $128 \times 100t$. We use 8-second audio and the patch size of 16×16 , resulting in $(128 \times 100 \times 8)/256 = 400$ audio tokens. For video, we sample 16 RGB frames at 8 fps of 224×224 . Similarly to [21], we tokenize the frames with 3D convolutions, using the spacetime patch size of $16 \times 16 \times 2$. Each video input produces $(16 \times 224 \times 224)/(256 \times 2) = 1568$ tokens.

Finetuning. We train for 50 epochs in Epic-Kitchens experiments, 20 in Epic-Sounds, and 15 in Ego4D-AR. We use SpecAugment [39] for audio augmentation and Augmix [19] for video augmentation. We use AdamW [34] optimizer with half-cycle cosine learning rate decay.

Table 1: The performance of the audio, video, and bottleneck audiovisual models on each dataset. We train and evaluate all models with $r_{train} = 0$ and $r_{test} = 0\%$. In Ego4D-AR, this is done by filtering out the modal-incomplete samples. In all datasets, multimodal performance beats unimodal performance.

Dataset	Audio	Video	Audiovisual
Epic-Kitchens	40.0%	63.2%	64.0%
Epic-Sounds	46.5%	41.4%	55.2%
Ego4D-AR	26.3%	34.6%	36.4%

4.3 Unimodal and baseline multimodal models

In Table 1, we show the unimodal and multimodal performance for each downstream dataset. As the original MBT, the multimodal models are trained with fully modal-complete samples. For our *baseline* results across all datasets, we employ these multimodal models. In scenarios where r_{test} is small, indicating that the testing data is nearly modal-complete, it is desired that the adapted model exhibits performance more closely aligned with the multimodal baseline model. Consequently, *this adaptation should uphold the multimodal reasoning capabilities of the model.*

As expected due to the annotation strategy, video serves as the dominant modality in Epic-Kitchens and Ego4d-AR, and audio takes precedence in Epic-Sounds. Consequently, as detailed in Sec. 3.1, we train our models to be robust to missing video in Epic-Kitchens and Ego4D-AR and to missing audio in Epic-Sounds (except for Sec. 4.6 where we assume that any modality could be absent).

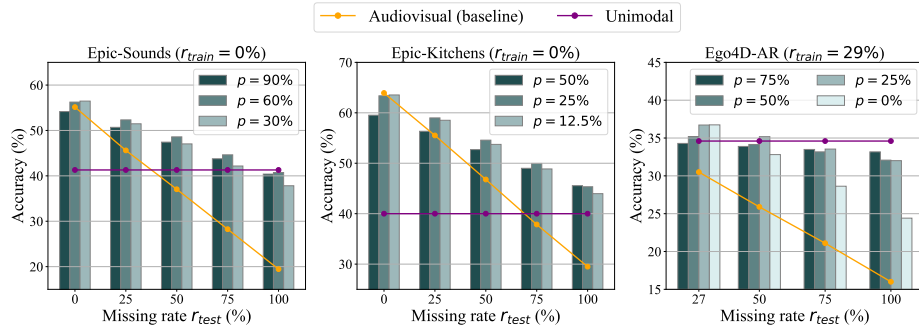


Fig. 3: Modality drop probability p vs. accuracy for modal-complete Epic-Sounds, Epic-Kitchens and Ego4D-AR. In all datasets, our method dramatically improves the performance of the baseline (orange).

Additionally, we report the unimodal performance of the non-missing modality (referred to as *unimodal*, i.e., video in Epic-Sounds) as *we expect the adapted models to converge towards this performance at higher values of r_{test}* .

Since Epic-Sounds has fewer training samples and exhibits a more balanced unimodal performance across each modality, we leverage it more extensively in Sec. 4.5 and Sec. 4.6.

4.4 Results with MMT.

We apply MMT as mentioned in Sec. 3.4. Fig. 3 shows changes in performance with different p in *random-replace* and compares it with the unimodal (purple) and baseline multimodal (orange) performance in each dataset. We use $p \in \{12.5\%, 25\%, 50\%\}$ for Epic-Kitchens. Ego4D-AR has naturally missing modality with $r_{train} = 29\%$, and we ablate with $p \in \{0\%, 25\%, 50\%, 75\%\}$. We observed that higher values of p yield better performance in Epic-Sounds; hence, we use $p \in \{30\%, 60\%, 90\%\}$ in this dataset.

By looking at the model’s performance across all p values, we notice the importance of picking p large enough for the model to adapt well but not too large to avoid high information loss. On the one hand, if there are insufficient training samples for MMT (smaller p), the model does not perform optimally at higher r_{test} . For example, in Epic-Sounds, the model trained with $p = 30\%$ reaches 42.7% at $r_{test} = 25\%$ while increasing p to 60% gives 44.6%. Similarly, if we only use naturally modal-incomplete samples to train MMT in Ego4D-AR ($p = 0\%$), the model reaches 28.6% at $r_{test} = 25\%$, while if we increase to $p = 25\%$, the performance increases to 33.5%. On the other hand, if p is too large, the model performs suboptimally for more modal-complete test data (smaller r_{test}). For example, while the model trained with $p = 75\%$ achieves higher accuracy at $r_{test} = 100\%$ in Ego4D-AR, it seems that it learns to ignore the audio completely, as the performance at $r_{test} = 0\%$ drops to the unimodal video performance in this dataset.

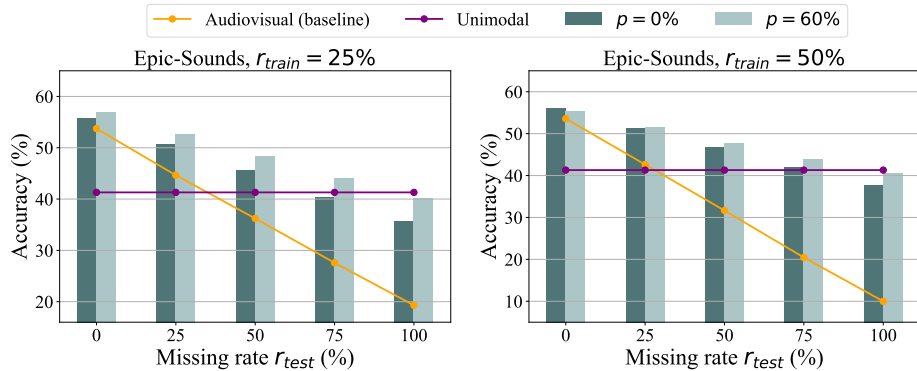


Fig. 4: Results with the modal-incomplete training data. As Epic-Sounds does not naturally have missing modality in the training data, we manually remove the audio from **(left)** $r_{train} = 25\%$ and **(right)** $r_{train} = 50\%$ of samples in the train set.

We find that the models trained with $p = 60\%$, $p = 25\%$, and $p = 25\%$ in Epic-Sounds, Epic-Kitchen, and Ego4D-AR, respectively, yield the best overall performance in each dataset. These models *perform significantly better than the baseline models (orange)* in the missing modality scenarios. For instance, at $r_{test} = 50\%$, the adapted models improve the baseline performance by 11.5% points in Epic-Sounds, 7.8% points in Epic-Kitchens, and 9.3% points in Ego4D-AR. Furthermore, for Epic-Sound and Epic-Kitchens, the models trained with MMT *reach or surpass the unimodal performance* at extremely severe $r_{test} = 100\%$. Nevertheless, the adapted models maintain the multimodal baseline performance at $r_{test} = 0\%$, showing that *the adaptation strategy does not harm* the model’s capabilities.

Note how in Ego4D-AR, the baseline fails to reach the unimodal accuracy even at the lowest $r_{test} = 27\%$, making our models trained with MMT a significantly better choice. This happens because $r_{train} = 29\%$ of samples were filtered out to train the multimodal baseline, causing information loss. This demonstrates how MMT enables better leverage of data in modal-incomplete datasets.

In Epic-Kitchens, the model performs better with $p = 25\%$ much smaller than in Epic-Sounds. We believe this is because video (the missing modality) produces almost $4\times$ more tokens than audio; thus, replacing video tokens causes more information loss.

4.5 Datasets with modal-incomplete training data.

In Sec. 4.4, Ego4D-AR is the only naturally modal-incomplete dataset, and we want to know how our method generalizes in other datasets with different r_{train} . What happens if a dataset has even higher r_{train} than Ego4D-AR? Filtering out noisy or corrupted data causes significant information loss if r_{train} is high. By using MMT, we can mitigate that and learn from all instances.

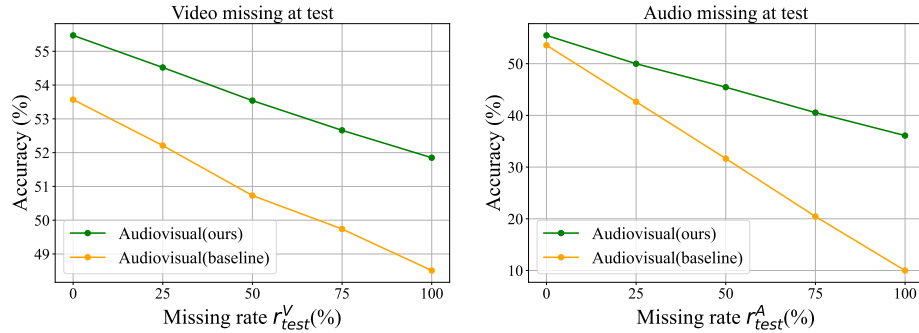


Fig. 5: Results on Epic-Sounds with $r_{train}^A = 25\%$, $r_{train}^V = 25\%$. We train our model with two MMTs: one for missing video and one for audio. We run the inference twice: (left) with missing video and (right) missing audio.

We show the results with our modal-incomplete version of Epic-Sounds with $r_{train} = 25\%$ and $r_{train} = 50\%$ in Fig. 4. We discuss the details of creating this version of the dataset in Supplementary. We can see how, indeed, the baseline model performs suboptimally, similarly as in Ego4D-AR in Sec. 4.4. Using *random-replace* with $p = 60\%$ significantly improves the baseline accuracy at $r_{test} = 100\%$ by 20 points when $r_{train} = 25\%$ and 30 points when $r_{train} = 50\%$. Furthermore, using modal-complete samples to train MMT ($p = 60\%$) causes a significant performance boost compared to the model trained with modal-incomplete samples only ($p = 0\%$).

4.6 Both modalities missing

In our setup, we train one MMT per dataset because we assume that only one modality could be missing. However, in more realistic scenarios, either modality could be missing for each sample in the dataset. Training a conventional multi-modal baseline involves filtering out all modal-incomplete samples, causing the training data to shrink to a small intersection subset where all modalities are available, which could negatively affect the model’s performance.

Luckily, it is quite straightforward to extend our approach to two MMTs, one for each modality. To simulate this scenario, we use Epic-Sounds and remove audio from $r_{train}^A = 25\%$ ⁵ of the training samples and video from another $r_{train}^V = 25\%$ of the training samples. For simplicity, we set $p = 0$ for both MMTs when training our adapted model. To train a multimodal baseline, we filter out all modal-incomplete samples *i.e.*, half of the training samples. Fig. 5, we show how our model’s performance compares to the baseline. We evaluate it for missing video (left) and audio (right). As can be seen, training with MMT dramatically improves the model performance.

⁵ We use superscripts $\{A,V\}$ to refer to the missing modality

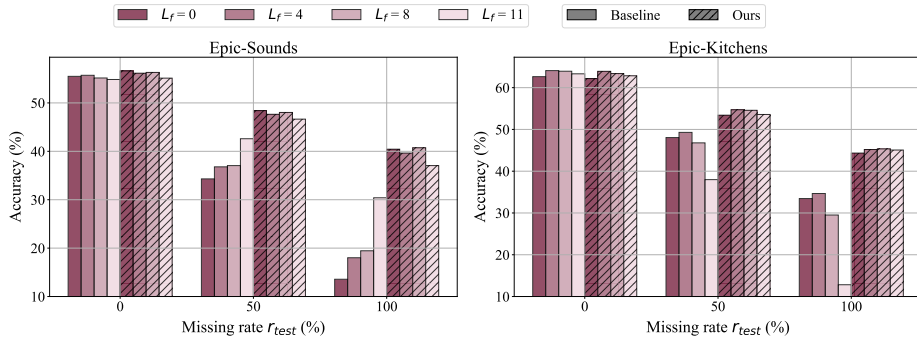


Fig. 6: Fusion Layer L_f vs. accuracy in the models trained with no adaptation strategy (Baseline) and trained with MMT (Ours). For training with MMT, we use *random-drop* strategy with $p = 60\%$ for Epic-Sounds and $p = 25\%$ for Epic-Kitchens. This strategy makes MBT more robust to missing modalities across all L_f and significantly reduces the negative effect of missing modalities.

4.7 The effect of the fusion layer.

As we mentioned in Sec. 3.2, the fusion layer does affect the performance of the bottleneck model, especially when test inputs are modal-incomplete. In Fig. 6, we examine whether training with MMT enhances robustness to a missing modality at various fusion layers and whether models trained with MMT (ours) show the same sensitivity to the fusion layer as those trained without it (baseline). As we observe, **the introduction of MMT makes the model more robust to the missing modality across all fusion layers** in both Epic-Sounds and Epic-Kitchens. With extremely severe $r_{test} = 100\%$, the adapted models perform with $\sim 45\%$ accuracy in Epic-Kitchens, while the unimodal audio model achieves 40% in this dataset. Furthermore, these adapted models do not exhibit similar sensitivity to the fusion layer as the baseline. For example, in Epic-Sounds, the baseline models trained with $L_f = 11$ exhibit superior performance, which is not the case for the adapted models (at $r_{test} = 100\%$, 37% accuracy with $L_f = 11$ but $\sim 40\%$ for other L_f). Overall, all adapted models perform consistently well, each providing decent performance in a modal-incomplete inference. Our approach effectively addresses a long-standing issue [35] of selecting the appropriate fusion layer when faced with missing modalities.

4.8 MMT vs. other missing modality representations

In Tables 2, we report the results using our *random-replace* and the baselines from Sec. 3.3 for Epic-Sounds, Epic-Kitchens, and Ego4D-AR. We also report the accuracy of the unimodal model in each dataset. Interestingly, passing zeroes as missing modality works better for Epic-Sounds and skipping missing modality tokens works better for Epic-Kitchens and Ego4D-AR.

To compare to the baseline tailored to the missing modality problem, we implement the multimodal prompts method [30]. It was originally based on ViLT [27] for image-text classification, so we implemented our version for action recognition based on MBT. Note that this method relies on a strong pre-trained backbone to efficiently finetune it by optimizing a few network parameters. As we deal with audiovisual learning in egocentric videos, we rely on large-scale egocentric pre-training using MAEs. As seen in Tab. 2, our method outperforms the one proposed by [30].

Table 2: Comparison of our method with baselines in Epic-Sounds, Epic-Kitchens, and Ego4D-AR. We demonstrate the accuracy across different missing modality ratios r_{test} . We show in **bold** the best result and underline the runner-up. We mentioned the way of representing the missing modality in brackets: *zeros* for the baseline of passing zeros, *skip* for the baseline skipping the tokens of the missing input, as mentioned in Sec. 3.3, or MMT.

Dataset	r_{test}	Unimodal	Baseline (zeros)	Baseline (skip)	Prompts [30]	Ours (MMT)
Epic-Sounds	0%	41.4	<u>55.2</u>	<u>55.2</u>	36.7	56.3
	25%	41.4	45.6	<u>47.0</u>	33.3	52.3
	50%	41.4	37.1	<u>39.9</u>	30.0	48.6
	75%	41.4	28.3	<u>32.5</u>	26.1	44.6
	100%	41.4	19.5	<u>25.0</u>	22.6	40.7
Epic-Kitchens	0%	40.0	63.9	63.9	32.5	63.4
	25%	40.0	<u>55.5</u>	53.2	31.1	59.0
	50%	40.0	<u>46.8</u>	42.1	30.2	54.6
	75%	40.0	<u>37.9</u>	30.8	29.2	50.0
	100%	40.0	<u>29.5</u>	20.0	28.2	45.3
Ego4D-AR	27%	34.6	30.5	<u>32.5</u>	18.4	36.7
	50%	34.6	25.9	<u>30.8</u>	16.4	35.2
	75%	34.6	21.1	<u>29.1</u>	13.9	33.5
	100%	34.6	16.0	<u>27.2</u>	11.5	32.0

We find that across all datasets, our MMT trained with *random-replace* either reach (in Epic-Sounds and Ego4D-AR) or exceed (in Epic-Kitchens) the unimodal performance in extreme $r_{test} = 100\%$. Interestingly, in Epic-Sounds, *random-replace* also regularizes the training and increases the $r_{test} = 0\%$ performance by 1.1 points.

5 Conclusion

We explore the missing modality problem in multimodal egocentric datasets. We suggest a simple yet effective method by learning the optimal token representation of the missing modality (MMT). Placing learnable tokens to represent

missing inputs provides an easy and intuitive way to train and test with modal-incomplete inputs. We propose strategy *random-replace* to learn MMT when training action recognition models and show how their performance brings us closer to robust and effective multimodal systems.

References

1. Azad, R., Khosravi, N., Dehghanmanshadi, M., Cohen-Adad, J., Merhof, D.: Medical image segmentation on mri images with missing modalities: a review (2022). URL: <https://arxiv.org/abs/2203.06217>, doi **10** **3**, **6**
2. Barrios, W., Soldan, M., Ceballos-Arroyo, A.M., Heilbron, F.C., Ghanem, B.: Localizing moments in long video via multimodal guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13667–13678 (2023) **1**
3. Colombo, P., Chapuis, E., Labeau, M., Clavel, C.: Improving multimodal fusion via mutual dependency maximisation. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 231–245. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). <https://doi.org/10.18653/v1/2021.emnlp-main.21>, <https://aclanthology.org/2021.emnlp-main.21> **2**, **3**
4. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 720–736 (2018) **4**
5. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. International Journal of Computer Vision pp. 1–23 (2022) **8**
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) **9**
7. Fabian Caba Heilbron, Victor Escorcia, B.G., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970 (2015) **4**
8. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019) **4**
9. Feichtenhofer, C., Li, Y., He, K., et al.: Masked autoencoders as spatiotemporal learners. Advances in neural information processing systems **35**, 35946–35958 (2022) **8**
10. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. pp. 214–229. Springer (2020) **4**
11. Garcia, N.C., Morerio, P., Murino, V.: Modality distillation with multiple stream networks for action recognition. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 103–118 (2018) **3**

12. Garcia, N.C., Morerio, P., Murino, V.: Learning with privileged information via adversarial discriminative modality distillation. *IEEE transactions on pattern analysis and machine intelligence* **42**(10), 2581–2593 (2019) [3](#)
13. Georgescu, M.I., Fonseca, E., Ionescu, R.T., Lucic, M., Schmid, C., Arnab, A.: Audiovisual masked autoencoders. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 16144–16154 (2023) [8](#)
14. Gong, X., Mohan, S., Dhingra, N., Bazin, J.C., Li, Y., Wang, Z., Ranjan, R.: Mmg-ego4d: Multimodal generalization in egocentric action recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6481–6491 (2023) [1](#), [2](#), [4](#)
15. Gong, Y., Chung, Y.A., Glass, J.: Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778* (2021) [9](#)
16. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., 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*. pp. 18995–19012 (2022) [1](#), [3](#), [4](#), [8](#)
17. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259* (2023) [1](#), [3](#)
18. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 16000–16009 (2022) [7](#), [8](#)
19. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781* (2019) [9](#)
20. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Computation* **9**(8), 1735–1780 (1997). <https://doi.org/10.1162/neco.1997.9.8.1735> [3](#)
21. Huang, P.Y., Sharma, V., Xu, H., Ryali, C., Fan, H., Li, Y., Li, S.W., Ghosh, G., Malik, J., Feichtenhofer, C.: Mavil: Masked audio-video learners. *arXiv preprint arXiv:2212.08071* (2022) [9](#)
22. Huh, J., Chalk, J., Kazakos, E., Damen, D., Zisserman, A.: Epic-sounds: A large-scale dataset of actions that sound. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 1–5. IEEE (2023) [4](#), [8](#)
23. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017) [4](#)
24. Kazakos, E., Huh, J., Nagrani, A., Zisserman, A., Damen, D.: With a little help from my temporal context: Multimodal egocentric action recognition. In: *British Machine Vision Conference (BMVC)* (2021) [1](#), [4](#)
25. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 5492–5501 (2019) [1](#), [4](#)
26. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Slow-fast auditory streams for audio recognition. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 855–859. IEEE (2021) [1](#), [4](#)

27. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. pp. 5583–5594. PMLR (2021) 4, 14
28. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**(11), 2278–2324 (1998). <https://doi.org/10.1109/5.726791> 3
29. Lee, H.C., Lin, C.Y., Hsu, P.C., Hsu, W.H.: 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). pp. 3956–3960. IEEE (2019) 2
30. Lee, Y.L., Tsai, Y.H., Chiu, W.C., Lee, C.Y.: Multimodal prompting with missing modalities for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14943–14952 (2023) 1, 2, 4, 5, 6, 8, 14
31. Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 11336–11344 (2020) 4
32. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019) 4
33. Lin, K.Q., Wang, J., Soldan, M., Wray, M., Yan, R., XU, E.Z., Gao, D., Tu, R.C., Zhao, W., Kong, W., et al.: Egocentric video-language pretraining. *Advances in Neural Information Processing Systems* **35**, 7575–7586 (2022) 1
34. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016) 9
35. Ma, M., Ren, J., Zhao, L., Testuggine, D., Peng, X.: Are multimodal transformers robust to missing modality? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18177–18186 (2022) 2, 3, 4, 5, 13
36. Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., Peng, X.: Smil: Multimodal learning with severely missing modality. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2302–2310 (2021) 2, 3, 5
37. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems* **34**, 14200–14213 (2021) 1, 2, 4, 5, 9
38. Neverova, N., Wolf, C., Taylor, G., Nebout, F.: Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**(8), 1692–1706 (2015) 2, 3
39. Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D., Le, Q.V.: SpecAugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779 (2019) 9
40. Parthasarathy, S., Sundaram, S.: Training strategies to handle missing modalities for audio-visual expression recognition. In: Companion Publication of the 2020 International Conference on Multimodal Interaction. pp. 400–404 (2020) 2, 3, 4, 6
41. Pramanick, S., Song, Y., Nag, S., Lin, K.Q., Shah, H., Shou, M.Z., Chellappa, R., Zhang, P.: EgoVLPv2: Egocentric video-language pre-training with fusion in the backbone. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5285–5297 (2023) 1
42. Radevski, G., Grujicic, D., Blaschko, M., Moens, M.F., Tuytelaars, T.: Multimodal distillation for egocentric action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5213–5224 (2023) 3

43. Ramazanova, M., Escorcia, V., Caba, F., Zhao, C., Ghanem, B.: Owl (observe, watch, listen): Audiovisual temporal context for localizing actions in egocentric videos. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4879–4889 (2023) [1](#), [4](#)
44. Tsai, Y.H.H., Liang, P.P., Zadeh, A., Morency, L.P., Salakhutdinov, R.: Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176* (2018) [2](#), [3](#)
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017) [3](#)
46. Wang, C., Niepert, M., Li, H.: Lrmm: Learning to recommend with missing modalities. *arXiv preprint arXiv:1808.06791* (2018) [2](#)
47. Wang, H., Mirmehdi, M., Damen, D., Perrett, T.: Centre stage: Centricity-based audio-visual temporal action detection. In: *The 1st Workshop in Video Understanding and its Applications (VUA 2023)* (2023) [1](#)
48. Woo, S., Lee, S., Park, Y., Nugroho, M.A., Kim, C.: Towards good practices for missing modality robust action recognition. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 37, pp. 2776–2784 (2023) [3](#)
49. Xiao, F., Lee, Y.J., Grauman, K., Malik, J., Feichtenhofer, C.: Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740* (2020) [4](#)
50. Zhao, J., Li, R., Jin, Q.: Missing modality imagination network for emotion recognition with uncertain missing modalities. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. pp. 2608–2618 (2021) [2](#), [3](#)