This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Exploring Missing Modality in Multimodal Egocentric Datasets

Merey Ramazanova Alejandro Pardo Humam Alwassel
Bernard Ghanem
Center of Excellence in Generative AI, KAUST, Saudi Arabia
firstname.lastname@kaust.edu.sa

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.

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 settings.

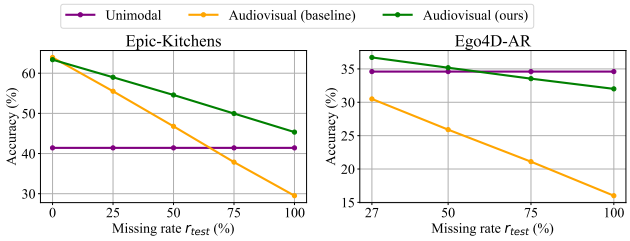


Figure 1. Multimodal models are typically trained on modal-complete data. However, these models (orange) fail when encountering modal-incomplete inputs at test time. Our proposed adaptation strategy for handling missing modalities (green) significantly improves performance across datasets. Notably, when all test inputs are modal-incomplete ($r_{test} = 100\%$), our approach surpasses unimodal performance (violet) by 5 points in Epic-Kitchens and doubles 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, 36, 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, 35]. 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 setups, no works have diagnosed how recent transformer-based ap-

proaches perform when modalities are missing for action recognition.

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 twofold: **(1)** We present a comprehensive study of the challenge of missing modalities in egocentric action recognition, exploring datasets with varying degrees of modal incompleteness and evaluating the influence of the fusion layer. **(2)** We introduce the Missing Modality Token (MMT) as a novel solution to handle missing modalities during both training and testing. In conjunction with a training strategy termed *random-replace*, our approach is extensively evaluated and demonstrates significant improvements over existing baselines. This work provides valuable insights and lays the groundwork for developing robust multimodal backbones.

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*. [35] 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*. [36] 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 (see supplementary materials), 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 classification [1]. 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, 36, 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]

[1] Code and data are not available

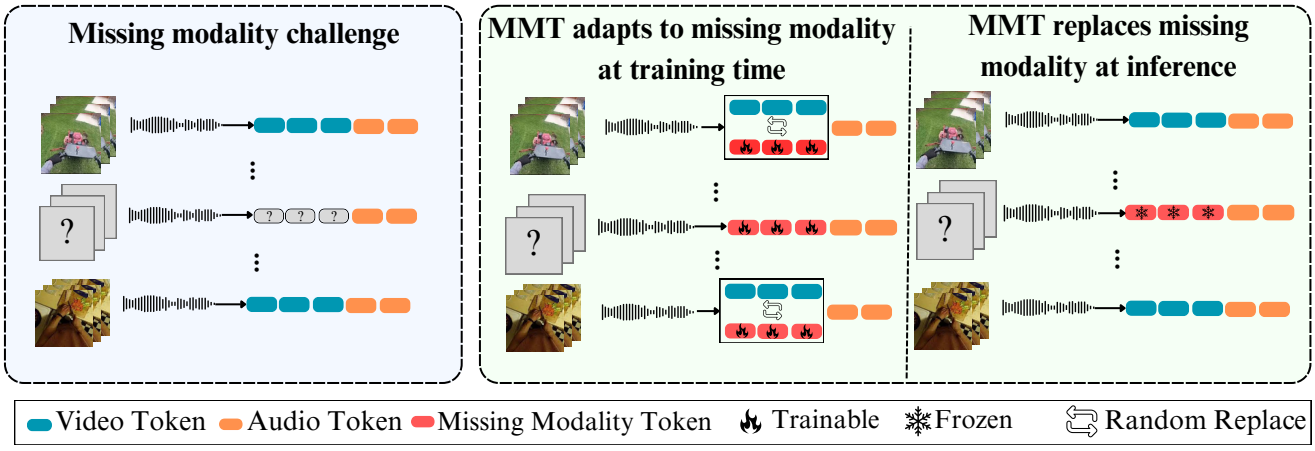


Figure 2. **Learning and Predicting with Missing Modalities. Left:** Given modal-incomplete data, it is still unclear how to effectively train and predict with a multimodal model (we present some naive baseline methods in Sec. 3.3). **Right:** To address this issue, we introduce a Missing Modality Token (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.

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 outlines our approach to addressing the missing modality problem. Specifically, we discuss the evaluation scenarios (3.1), multimodal design and fusion (3.2), naive baselines (3.3), and our proposed method (3.4).

3.1. Problem statement, setup, and evaluation

Given the training and testing multimodal samples, we denote the missing modality rates as r_{train} and r_{test}, computed by dividing the number of modal-incomplete samples by the total number of samples, assuming that only one modality is missing per sample. Our work considers two training scenarios: a **modal-incomplete training set** ($r_{train} \neq 0\%$), and a **modal-complete training set** ($r_{train} = 0\%$). To assess the model's performance under varying degrees of missing data, we create several test set variants by progressively removing modality information until $r_{test} = 100\%$.

Following previous works [35, 36] when using fully modal-complete datasets ($r_{train} = 0\%$ and $r_{test} = 0\%$), we assume that the modality with the best unimodal performance (*e.g*., audio for Epic-Sounds) is made incomplete at test time. Unlike prior research [30, 35, 36], we also validate our adaptation strategies on datasets with naturally incomplete modalities in both training and testing splits. We conduct our evaluations on egocentric action recognition datasets using the two commonly available modalities: visual (RGB frames) and audio.

3.2. Efficient and effective multimodal fusion

We address the challenge of missing modalities by employing advanced multimodal fusion strategies. While previous transformer-based methods predominantly used simple fusion techniques—such as early or mid-fusion with cross-modal self-attention, where tokens are concatenated at the fusion layer—these approaches encounter scalability issues for videos due to the quadratic complexity inherent in the attention mechanism [30, 37].

To enhance both efficiency and effectiveness, we adopt MBT [37], tthe current state-of-the-art in audiovisual fusion. MBT utilizes a bottleneck transformer that streamlines cross-modal interactions by introducing a small set of learnable "bottleneck" tokens. This design allows each modality to perform self-attention with minimal computational overhead, making it especially well-suited for information-dense modalities such as video. The original MBT model set the fusion layer at 8 for optimal downstream performance.

Fusion Layer. A critical aspect of multimodal fusion is the design of the fusion layer, denoted as L_f, where cross-

modal interactions occur. We evaluate the original bottleneck model on modal-complete audiovisual inputs and train variants with different fusion layers. Fig. 5 presents results (marked as *baseline*) for Epic-Kitchens and Epic-Sounds.

Our analysis reveals that when all modalities are present ($r_{test} = 0$), performance remains largely unaffected by the choice of fusion layer. However, under missing modality conditions, the fusion layer significantly impacts performance. For instance, in Epic-Sounds at $r_{test} = 50\%$, test accuracy improves from 35% with $L_f = 0$ to 42% with $L_f = 11$. Interestingly, earlier fusion is more effective in Epic-Kitchens, while later fusion yields better results in Epic-Sounds. This aligns with prior observations [36] that the optimal fusion strategy is dataset-dependent.

While this finding may seem intuitive, it poses a practical challenge—when model performance is highly sensitive to the fusion layer, identifying the best setting can be computationally expensive. To address this, we analyze the effect of fusion layers under missing modality conditions and present further results in the supplementary material.

3.3. Intuitive baselines

Designing effective solutions for missing modalities is challenging (Fig. 2). However, simple and intuitive adaptations can be applied at test time. We introduce two *training-free baselines* and one training-time strategy to handle missing modalities.

Passing Missing Inputs as Zero Tensors. A straightforward approach is to replace missing modality inputs with tensors filled with zeros. This method, widely used in prior work [1, 30, 40], is popular for its simplicity and ease of implementation. Unless stated otherwise, we adopt this as our primary baseline, referring to it as *baseline*.

Passing Only Complete Inputs. Since transformers can handle sequences of varying lengths, we can omit tokens corresponding to missing modalities and provide only the available modality tokens. While intuitive, this approach becomes impractical for batched inference when some inputs within a batch have missing modalities.

Modality Dropping. As a training-time strategy, we can randomly drop modalities during training as a form of augmentation. This helps the model learn robustness to missing modalities without requiring architectural modifications. Similar to the training-free baseline, when a modality is dropped during training, it is replaced with a zero tensor.

3.4. Our approach to dealing with missing modalities

We propose a simple and generalizable approach for handling missing modalities. Instead of filling missing inputs with zero tensors or discarding them, we introduce a learnable Missing Modality Token (MMT). MMT serves as a "template" for missing inputs, learning to represent them based on the non-missing modality.

Fig. 2 (right) illustrates MMT in action. When a sample lacks a modality, MMT—a shared parameter—is repeated to match the missing tokens and passed to the transformer along with the available modality. Each token retains distinct positional embeddings, akin to the masked token in [18]. This allows MMT to generalize across modal-incomplete samples by leveraging non-missing tokens.

However, if MMT is trained solely on naturally missing samples, it never directly observes the inputs it aims to mimic. Moreover, in datasets where only a fraction of samples are incomplete (*e.g.*, $r_{train} = 10\%$), MMT has limited training exposure. To address this, we introduce the *random-replace* strategy:

This increases training opportunities for MMT and enables the model to see both complete and incomplete versions of the same sample, reinforcing cross-modal dependencies.

However, setting p too high may lead to excessive information loss, especially if a critical modality is replaced (Sec. 4.4). Thus, MMT is trained in two ways: (1) Using naturally missing modality samples. (2) Applying random replacement on modal-complete samples. Below, we detail how these strategies are applied across different dataset conditions.

Modal-Incomplete Training Set. Given that r_{train} of the training samples are modal-incomplete and $100\% - r_{train}$ are modal-complete, MMT can be trained in two ways:

(1) *Learning from Modal-Incomplete Only*: MMT replaces missing modality inputs only for modal-incomplete samples. Modal-complete samples remain unchanged (*i.e.*, $p = 0$).

(2) *Learning from Both Modal-Complete and Modal-Incomplete*: In addition to (1), we apply *random-replace* with $p > 0$ to expose MMT to artificially missing inputs.

Modal-Complete Training set. Since all training samples are modal-complete, MMT is trained exclusively using *random-replace* ($p > 0$).

Inference. Regardless of the training strategy, MMT replaces missing input tokens 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. Then, we compare our method to the baselines we proposed and [30] in Sec. 4.8. Additionally, we study the effect of fusion layers L_f in supplementary materials.

4.1. Datasets

We use **Ego4D** [16] videos to pre-train the MBT backbone. Due to privacy regulations, only 2.5K of the 3.7K video hours retain original audio. From these, we extract 450K 10-second modal-complete clips. For downstream tasks, we use the following egocentric action recognition benchmarks:

Epic-Kitchens-100 [5]: 90K variable-length clips spanning 100 hours, labeled with 300 nouns and 97 verbs. We use two output heads to jointly predict verb and noun classes. All videos have complete audio-visual streams ($r_{train} = r_{test} = 0\%$).

Epic-Sounds [22]: 79K clips from the same 100 hours as Epic-Kitchens, annotated with 44 unique sound labels, independent of noun-verb labels.

Ego4D-AR: A dataset derived from Ego4D's Short-Term Action Anticipation task [16]. We use *time-to-contact* timestamps to trim clips and assign action labels. It contains 142K noun-verb clips, with 128 noun and 81 verb classes. Verbs are highly imbalanced, so we apply class-weighted loss. Unlike Epic-Kitchens, Ego4D-AR naturally has missing modalities, with audio present in only 71% of training and 73% of test clips ($r_{train} = 29\%$, $r_{test} = 27\%$).

For clarity, we report verb accuracy for Epic-Kitchens, class accuracy for Epic-Sounds, and noun accuracy for Ego4D-AR, with additional metrics in 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 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 $L_f = 8$.

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. When comparing the modality dropping baseline with MMT, we train it with the same drop probability.

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

Tab. 1 presents the unimodal and multimodal performance across downstream datasets. As in the original MBT, multimodal models are trained exclusively on fully modal-complete samples. These models serve as our *baseline* across all datasets. In scenarios where r_{test} is small (i.e., test data is nearly modal-complete), the adapted model should ideally maintain performance close to the multimodal baseline, ensuring that *multimodal reasoning capabilities are preserved.*

Due to the annotation strategy, video is the dominant modality in Epic-Kitchens and Ego4D-AR, while audio is more informative in Epic-Sounds. Consequently, as outlined 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 in Sec .4.6, where either modality may be absent). Additionally, we report unimodal performance for the non-missing modality (e.g., video in Epic-Sounds) since *adapted models should converge toward this performance at higher values of* r_{test}.

Epic-Sounds, having fewer training samples and a more balanced unimodal performance across modalities, is analyzed more extensively in Sec .4.5 and Sec .4.6.

4.4. Results with MMT.

We apply MMT as described in Sec. 3.4. Fig. 3 illustrates how performance varies with different modality drop probabilities p using the *random-replace* and compares these results to both unimodal (purple) and baseline multimodal (orange) performance across the datasets. For Epic-Kitchens, we experiment with $p \in \{12.5\%, 25\%, 50\%\}$. In Ego4D-AR, where the training set naturally has $r_{train} = 29\%$ missing modalities, we consider $p \in \{0\%, 25\%, 50\%, 75\%\}$. In Epic-Sounds, we observe that higher values of p lead to better performance, so we employ $p \in \{30\%, 60\%, 90\%\}$.

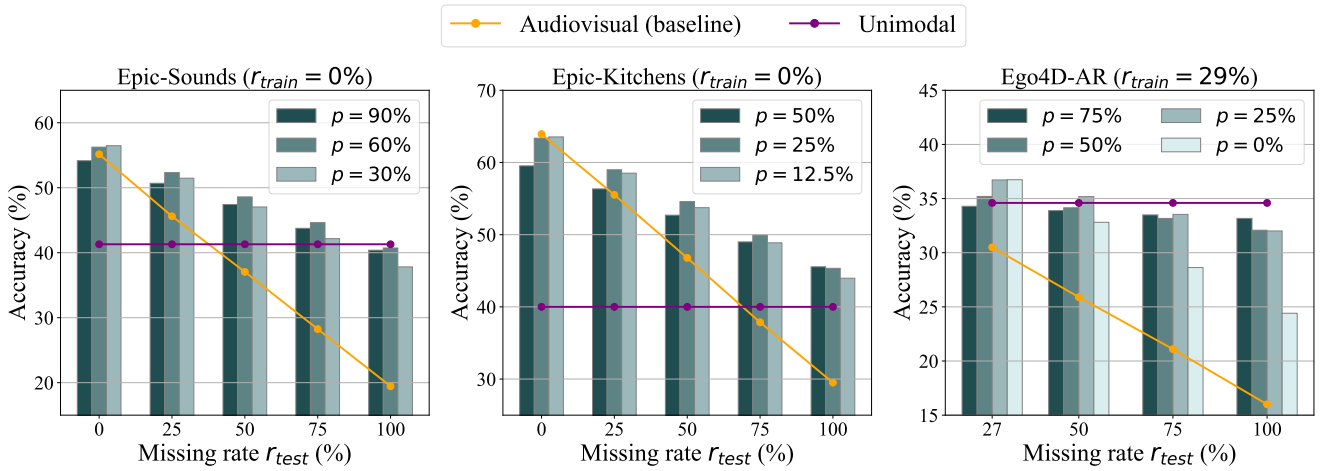


Figure 3. **Modality drop probability p *vs*. accuracy for Epic-Sounds, Epic-Kitchens and Ego4D-AR at various r_{test} levels.** Each dataset's training missing modality rate r_{train} is also indicated. In all datasets, our method significantly outperforms the baseline (orange). The optimal p varies among datasets, likely due to different missing modality, r_{train}, and individual modality performance (see Tab. 1).

Overall, the results emphasize the need to select a sufficiently high p so that the model can adequately adapt to missing modalities, while avoiding excessive information loss. For instance, in Epic-Sounds, the model trained with $p = 30\%$ achieves 42.7% accuracy at $r_{test} = 75\%$, whereas increasing p to 60% improves performance 44.6%. Similarly, in Ego4D-AR, training with naturally modal-incomplete samples only ($p = 0\%$) results in 28.6% accuracy at $r_{test} = 75\%$ but increasing p to 25% boosts accuracy to 33.5%. Conversely, setting p too high can lead the model to over-rely on the modality that is always present: for example, the model trained with $p = 75\%$ in Ego4D-AR achieves higher accuracy at $r_{test} = 100\%$ but drops to unimodal video performance at $r_{test} = 27\%$, suggesting it has effectively ignored the audio modality.

We find that the models trained with $p = 60\%$ for Epic-Sounds, $p = 25\%$ for Epic-Kitchens, and $p = 25\%$ for Ego4D-AR deliver the best overall performance on their respective datasets. In missing modality scenarios, these models *substantially outperform the baselines*. For example, at $r_{test} = 50\%$, the adapted models achieve improvements of 11.5 points in Epic-Sounds, 7.8 points in Epic-Kitchens, and 9.3 points in Ego4D-AR compared to the baseline. Notably, for Epic-Sounds and Epic-Kitchens, the MMT-trained models *match or exceed unimodal performance* even when all modalities but one are missing ($r_{test} = 100\%$). At the same time, they retain the baseline multimodal performance when $r_{test} = 0\%$, demonstrating that *the adaptation strategy does not compromise* the model's overall capabilities.

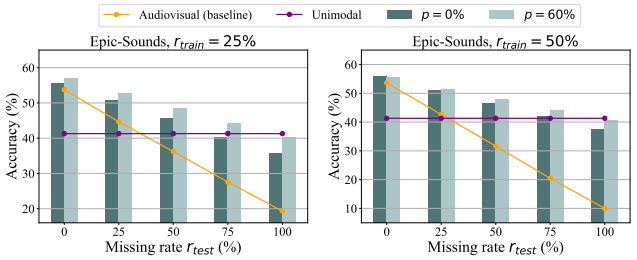
In Ego4D-AR, the baseline model fails to reach unimodal accuracy even at $r_{test} = 27\%$, primarily because $r_{train} = 29\%$ of the training samples were filtered out, resulting in information loss. This makes the MMT-trained models a significantly better choice, as they better leverage the available data in modal-incomplete datasets.

In Epic-Kitchens, the model achieves better results with a lower drop probability of $p = 25\%$ compared to Epic-Sounds. We attribute this to the fact that the video modality, which is missing in Epic-Kitchens, generates nearly four times as many tokens as audio. Consequently, replacing video tokens results in a higher degree of information loss, necessitating a smaller p to preserve critical content.

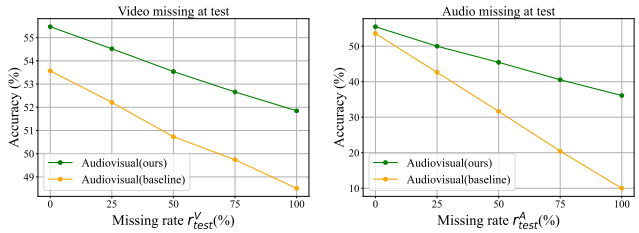
4.5. Datasets with modal-incomplete training data.

In Sec. 4.4, Ego4D-AR is the only naturally modal-incomplete dataset, , prompting us to investigate the generalization of our method across datasets with varying r_{train}. In cases where r_{train} is even higher—for instance, when filtering out noisy or corrupted data leads to significant information loss—MMT enables us to learn from all instances rather than discarding valuable samples.

Fig. 4a presents results on our modal-incomplete version of Epic-Sounds with $r_{train} = 25\%$ and $r_{train} = 50\%$ (see Supplementary for details on the dataset creation). As observed in Ego4D-AR, the baseline model performs suboptimally under high r_{train}. However, applying the *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 $r_{train} = 50\%$. Moreover, training MMT on modal-complete samples ($p = 60\%$) yields a substantial performance boost compared to models trained solely on modal-incomplete samples ($p = 0\%$).



(a) **Results with modal-incomplete training data.** Since Epic-Sounds does not naturally contain missing modalities in the training data, we manually remove audio from (**left**) $r_{train} = 25\%$ and (**right**) $r_{train} = 50\%$ of the training samples.



(b) **Multiple modalities missing.** 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 use Epic-Sounds with $r^A_{train} = 25\%, r^V_{train} = 25\%$.

Figure 4. Comparison of performance under different training and testing conditions with incomplete modalities.

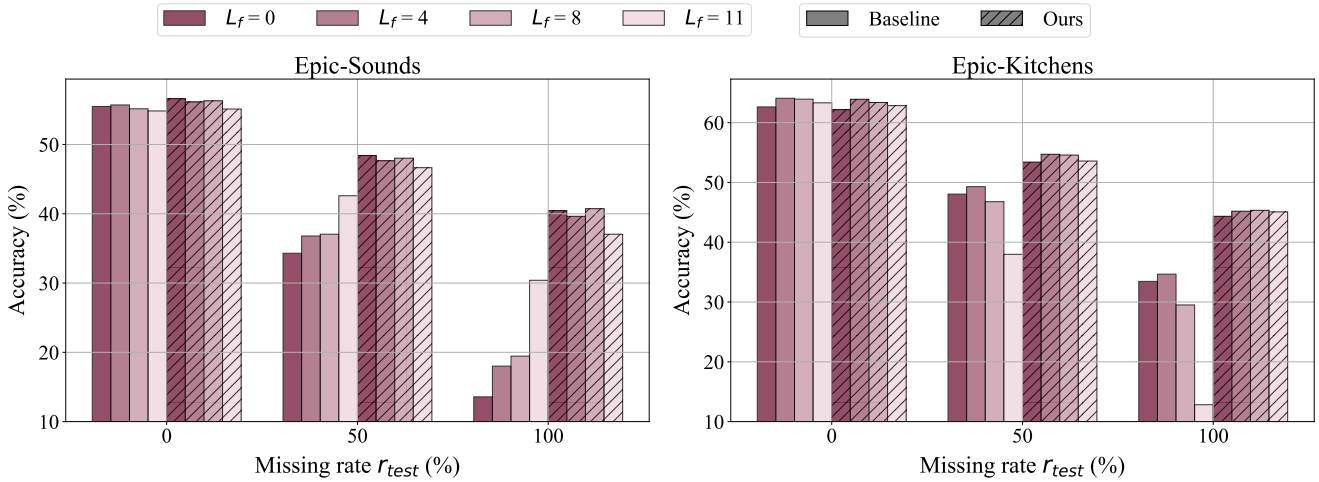


Figure 5. **Fusion Layer L_f *vs*. accuracy in the models trained with no adaptation strategy (Baseline) and trained with MMT (Ours).** For training with MMT, we use *random-replace* strategy with $p = 60\%$ for Epic-Sounds and $p = 25\%$ for Epic-Kitchens. This strategy makes MBT more robust to missing modalities across all L_f and significantly reduces the negative effect of missing modalities.

4.6. Both modalities missing

In our setup, we train a single Missing Modality Token (MMT) under the assumption that only one modality is missing at a time. However, in real-world scenarios, either modality could be absent for each sample. Conventional multimodal baselines handle this by filtering out all modal-incomplete samples, reducing the training set to a small subset where both modalities are present—potentially degrading model performance.

Extending our approach to handle both missing modalities is straightforward: we introduce two MMTs, one for each modality. To simulate this scenario, we modify Epic-Sounds by removing audio from $r^A_{train} = 25\%$ of training samples and video from another $r^V_{train} = 25\%$. For simplicity, we set $p = 0\%$ for both MMTs when training.

For the baseline, we exclude all modal-incomplete samples, effectively discarding half of the training data. Fig. 4b compares our model's performance to the baseline for missing video (left) and missing audio (right). As shown, training with MMT significantly improves model performance.

4.7. The effect of the fusion layer.

As discussed in Sec. 3.2, the fusion layer significantly impacts the performance of the bottleneck model, particularly when test inputs are modal-incomplete. In Fig .5, we analyze whether training with MMT improves robustness to missing modalities across different fusion layers and whether models trained with MMT exhibit the same sensitivity to the fusion layer as baseline models.

Our findings show that introducing MMT enhances robustness to missing modalities across all fusion layers in both Epic-Sounds and Epic-Kitchens. Under extreme conditions ($r_{test} = 100\%$), the adapted models achieve $\sim 45\%$ accuracy in Epic-Kitchens, outperforming the unimodal au-

dio model, which reaches only 40%.

Additionally, adapted models do not display the same sensitivity to the fusion layer as the baseline. For instance, in Epic-Sounds, baseline models trained with $L_f = 11$ perform best, whereas this is not the case for adapted models—at $r_{test} = 100\%$, accuracy drops to 37% with $L_f = 11$, while remaining 40% across other L_f values.

Overall, adapted models perform consistently well across fusion layers, ensuring stable performance under modal-incomplete conditions. Our approach effectively mitigates a long-standing challenge [36]—selecting an optimal fusion layer when modalities are missing.

Table 2. We compare MMT with the baselines described in Sec. 3.3. We show the best result in **bold** and underline the runner-up. The training-free baselines are labeled as *test*. The missing modality representation is indicated as follows: *zeros* for passing zeros, *skip* for skipping the tokens of the missing input, or MMT. *Train-zeros* corresponds to the modality dropping baseline.

(a) Epic-Sounds

Missing rate	0%	25%	50%	75%	100%
Unimodal	41.4	41.4	41.4	41.4	41.4
Test-zeros	55.2	45.6	37.1	28.3	19.5
Test-skip	55.2	47.0	39.9	32.5	25.0
Train-zeros	**56.8**	52.1	47.6	43.2	38.8
Train-MMT (Ours)	56.3	**52.3**	**48.6**	**44.6**	**40.7**

(b) Epic-Kitchens

Missing rate	0%	25%	50%	75%	100%
Unimodal	40.0	40.0	40.0	40.0	40.0
Test-zeros	**63.9**	55.5	46.8	37.9	29.5
Test-skip	**63.9**	53.2	42.1	30.8	20.0
Train-zeros	61.5	57.5	53.5	49.3	45.2
Train-MMT (Ours)	63.4	**59.0**	**54.6**	**50.0**	**45.3**

(c) Ego4D-AR

r_{test}	27%	50%	75%	100%
Unimodal	34.6	34.6	34.6	34.6
Test-zeros	30.5	25.9	21.1	16.0
Test-skip	32.5	30.8	29.1	27.2
Train-zeros	36.2	34.8	33.4	**32.1**
Train-MMT (Ours)	**36.7**	**35.2**	**33.5**	32.0

4.8. MMT *vs.* other missing modality representations

In Tab. 2, we compare our *random-replace* against the baselines from Sec. 3.3, as well as unimodal performance.

The training-free baselines generally perform better than the unimodal models when the missing modality rate is moderate ($r_{test} \leq 50\%$) in Epic-Kitchens and Epic-Sounds. However, as the missing rate increases, the performance of these baselines drops below that of the unimodal models. This highlights the importance of explicitly adapting models to missing modalities during training.

Interestingly, the effectiveness of different baselines varies across datasets. In Epic-Kitchens, representing missing modalities by passing zero tensors yields better performance than skipping the missing modality tokens. In contrast, in Epic-Sounds and Ego4D-AR, skipping the missing modality tokens leads to better results than filling them with zeros. One possible reason for this difference is that video inputs generate a larger number of tokens compared to audio. As a result, completely removing video tokens may have a more detrimental effect than replacing them with zeros, whereas removing audio tokens may not lead to as much information loss.

The modality dropping baseline, which applies missing modality adaptation during training, consistently outperforms the training-free approaches by a significant margin. This is expected, as the modality dropping strategy directly trains the network to handle missing modalities by randomly omitting inputs during training. For fairness, we train this baseline with the same p used in our MMT model for the corresponding dataset.

Our proposed MMT approach further enhances robustness across all datasets, demonstrating the benefits of using a dedicated learnable missing modality token instead of relying on fixed inputs. By learning a suitable representation for the missing modality, MMT allows the model to better adapt to incomplete inputs. Importantly, this added flexibility does not introduce any tangible computational overhead compared to the modality dropping baseline, as both approaches employ a similar training framework.

Finally, we observe that MMT trained with *random-replace* consistently achieves strong performance even when the missing modality rate is extreme ($r_{test} = 100\%$). In Epic-Sounds, MMT performs comparably to the unimodal model, achieving 40.7% accuracy versus 41.4% for the unimodal baseline. In Epic-Kitchens, MMT not only remains robust but even surpasses unimodal performance, reaching 45.3% accuracy compared to 40.0%. These results further reinforce the effectiveness of our approach in handling missing modalities in multimodal learning settings.

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.

Acknowledgements

The research reported in this publication was supported by funding from KAUST Center of Excellence on GenAI, under award number 5940.

References

[1] Reza Azad, Nika Khosravi, Mohammad Dehghanmanshadi, Julien Cohen-Adad, and Dorit Merhof. Medical image segmentation on mri images with missing modalities: a review (2022). *URL: https://arxiv. org/abs/2203.06217, doi*, 10. 2, 4

[2] Wayner Barrios, Mattia Soldan, Alberto Mario Ceballos-Arroyo, Fabian Caba Heilbron, and Bernard Ghanem. Localizing moments in long video via multimodal guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13667–13678, 2023. 1

[3] Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. Improving multimodal fusion via mutual dependency maximisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 231–245, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 1, 2

[4] 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. 2

[5] 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*, pages 1–23, 2022. 5

[6] 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. 5

[7] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 2

[8] 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. 2

[9] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022. 5

[10] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 214–229. Springer, 2020. 3

[11] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018. 2

[12] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Learning with privileged information via adversarial discriminative modality distillation. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2581–2593, 2019. 2

[13] Mariana-Iuliana Georgescu, Eduardo Fonseca, Radu Tudor Ionescu, Mario Lucic, Cordelia Schmid, and Anurag Arnab. Audiovisual masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16144–16154, 2023. 5

[14] Xinyu Gong, Sreyas Mohan, Naina Dhingra, Jean-Charles Bazin, Yilei Li, Zhangyang Wang, and Rakesh Ranjan. Mmg-ego4d: Multimodal generalization in egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6481–6491, 2023. 1, 2

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

[16] 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, 5

[17] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259*, 2023. 1, 2

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 4, 5

[19] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 5

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 2

[21] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Haoqi Fan, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, and Christoph Feichtenhofer. Mavil: Masked audio-video learners. *arXiv preprint arXiv:2212.08071*, 2022. 5

[22] Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. Epic-sounds: A large-scale dataset of actions that sound. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2, 5

[23] 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. 2

[24] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019. 1, 2

[25] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a little help from my temporal context: Multimodal egocentric action recognition. In *British Machine Vision Conference (BMVC)*, 2021. 2

[26] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 855–859. IEEE, 2021. 1, 3

[27] 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

[28] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[29] 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. 1

[30] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14943–14952, 2023. 1, 2, 3, 4

[31] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11336–11344, 2020. 3

[32] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

[33] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022. 1

[34] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[35] 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*, pages 2302–2310, 2021. 1, 2, 3

[36] 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. 1, 2, 3, 4, 8

[37] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021. 1, 2, 3, 5

[38] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2015. 1, 2

[39] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019. 5

[40] Srinivas Parthasarathy and Shiva Sundaram. Training strategies to handle missing modalities for audio-visual expression recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 400–404, 2020. 1, 2, 4

[41] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023. 1

[42] Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5213–5224, 2023. 2

[43] Merey Ramazanova, Victor Escorcia, Fabian Caba, Chen Zhao, and Bernard Ghanem. 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*, pages 4879–4889, 2023. 1, 2

[44] 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. 1, 2

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[46] Cheng Wang, Mathias Niepert, and Hui Li. Lrmm: Learning to recommend with missing modalities. *arXiv preprint arXiv:1808.06791*, 2018. 1

[47] Hanyuan Wang, Majid Mirmehdi, Dima Damen, and Toby Perrett. Centre stage: Centricity-based audio-visual temporal action detection. In *The 1st Workshop in Video Understanding and its Applications (VUA 2023)*, 2023. 1

[48] Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for

missing modality robust action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2776–2784, 2023. 2

[49] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 2

[50] Jinming Zhao, Ruichen Li, and Qin Jin. 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)*, pages 2608–2618, 2021. 1, 2