# Knowledge Distillation for Multimodal Egocentric Action Recognition Robust to Missing Modalities

Maria Santos-Villafranca<sup>\*1</sup> Dustin Carrión-Ojeda<sup>\*2</sup> Alejandro Perez-Yus<sup>1</sup> Jesus Bermudez-Cameo<sup>1</sup> Jose J. Guerrero<sup>1</sup> Simone Schaub-Meyer<sup>2,3</sup>

<sup>1</sup>I3A - University of Zaragoza <sup>2</sup>Department of Computer Science, TU Darmstadt <sup>3</sup>hessian.AI

#### Abstract

Action recognition is an essential task in egocentric vision due to its wide range of applications across many fields. While deep learning methods have been proposed to address this task, most rely on a single modality, typically video. However, including additional modalities may improve the robustness of the approaches to common issues in egocentric videos, such as blurriness and occlusions. Recent efforts in multimodal egocentric action recognition often assume the availability of all modalities, leading to failures or performance drops when any modality is missing. To address this, we introduce an efficient multimodal knowledge distillation approach for egocentric **a**ction **r**ecognition that is robust to missing modalities (KARMMA) while still benefiting when multiple modalities are available. Our method focuses on resource-efficient development by leveraging pretrained models as unimodal feature extractors in our teacher model, which distills knowledge into a much smaller and faster student model. Experiments on the Epic-Kitchens and Something-Something datasets demonstrate that our student model effectively handles missing modalities while reducing its accuracy drop in this scenario.

# 1. Introduction

Egocentric vision aims to capture and interpret the world from a first-person perspective. Recently, there has been an increasing interest in this topic due to the advances in wearable technology and growing interest in AR/VR. This area has a wide range of applications, including assistive devices that support individuals with disabilities to facilitate their daily activities [25], as well as human-computer interaction [3], surveillance [2], and VR gaming [6].

\*Equal contribution.

dustin.carrion@tu-darmstadt.de



Figure 1: Overview of our proposed multimodal-to-multimodal distillation pipeline for egocentric action recognition. Both teacher and student models receive the same action sequence, but the student operates under real-world conditions where some modalities may be missing (*e.g.*, due to sensor failures). As a result, our student model is a smaller and faster version of the teacher and also robust to missing modalities.

The introduction of new egocentric datasets [7, 14-16, 27, 33, 47] has enabled research on several tasks using egocentric videos, such as action recognition [12, 40, 43], object recognition [1, 20, 57], localization [10, 29, 52], and others [24, 37, 46]. However, the egocentric perspective presents more challenges than the exocentric (third-person) view due to the increased motion blur and object occlusions, often caused by the movements of the person recording the video. To address these challenges in egocentric action recognition, recent approaches have suggested leveraging multiple modalities [12, 23, 26, 41, 43, 44]. For instance, Li et al. [26] proposes a set of egocentric features and combines them with motion and object features. Other works [12, 43] explore methods for fusing the features from all available modalities. However, these methods typically assume the availability of all modalities during inference.

Corresponding authors: m.santos@unizar.es,

Project website: https://visinf.github.io/KARMMA

This assumption is not always feasible in real-world scenarios due to privacy concerns or sensor failures, such as video camera malfunctions [58]. Moreover, research has shown that multimodal architectures experience significant performance drops when the most informative modality for a specific task is unavailable [34], *e.g.*, when the video is missing for egocentric action recognition.

In this work, we propose a method that leverages all available modalities while being robust to missing ones. To achieve this, we introduce a strategy to compensate for missing modalities, ensuring our approach maintains accuracy even if the most informative modality (video) is completely unavailable. However, one of the main drawbacks of multimodal models is their high computational cost. To overcome this, we propose a novel multimodal-to-multimodal distillation strategy, illustrated in Fig. 1, that allows us to use a student model that is small, fast, and robust to missing modalities instead of a large teacher model. Additionally, our teacher model leverages the availability of pre-trained unimodal feature extractors, facilitating the creation of multimodal models through modality fusion.

**Contributions.** (1) To our knowledge, we propose the first multimodal-to-multimodal distillation approach for egocentric action recognition. (2) Our distillation process incorporates a novel technique for handling missing modalities, resulting in a robust student model that maintains competitive accuracy even when some modalities are completely unavailable during inference. At the same time, it benefits from the availability of all modalities. (3) Our teacher model fuses features from frozen pre-trained unimodal feature extractors, eliminating the need to train them, allowing easy integration of newer feature extractors. (4) We introduce a fusion block that uses a novel parameter-free token reduction strategy, thereby reducing its computational cost.

# 2. Related work

Egocentric action recognition has been addressed mostly using video data [50, 51, 55], as it is the most informative modality for this task. While video alone provides substantial information, several works have shown that incorporating additional modalities such as Inertial Measurement Units (IMUs) [53, 62], gaze [36], or audio [23], can improve recognition accuracy, particularly when video alone is insufficient to identify actions. However, most existing multimodal research has focused only on bimodal approaches. Recently, the availability of modern datasets [7, 14-16, 27, 33, 47] that include multiple modalities has encouraged the development of multimodal approaches beyond two modalities for egocentric action recognition. For instance, Gong et al. [12] studied the generalization to new modalities and the impact of missing ones, while Dong et al. [8] explored domain generalization under missing modality conditions. Additionally,

Radevski et al. [43] and Hatano et al. [17] leveraged multiple modalities to distill the knowledge from a multimodal teacher into a unimodal student to improve its accuracy. In contrast to previous works, our approach focuses on distilling knowledge from a large multimodal teacher into a small *multimodal* student capable of handling missing modalities.

Modality fusion is often performed using crossattention [41, 61] due to the success of transformer architectures. However, cross-attention primarily fuses pairs of modalities, limiting its applicability. Another common strategy for fusing modality pairs is the usage of contrastive losses [18, 28, 41]. In this direction, Lin et al. [28] introduced a contrastive loss specifically designed for aligning egocentric video and language data. Alternatively, some methods [43, 59] pre-train expert models for each modality and then fuse their features using model-agnostic techniques, enabling fusion across more than two modalities. However, all these approaches assume the availability of all modalities at inference time. To address this limitation, Liu et al. [31] proposed a data-dependent self-attention-based fusion strategy adaptable to any number of input modalities. Similarly, Gong et al. [12] introduced a transformer-based fusion block with modality dropout during training to improve robustness to missing modalities and evaluate its zero-shot generalization capabilities to new ones. Moreover, Ramazanova et al. [45] proposed adding a learnable modality token that is only used when a modality is missing. Our transformer-based approach differs from these works by combining modality dropout with two types of learnable tokens that remain active at all times.

Token reduction mechanisms have been crucial in developing self-attention-based methods, as their computational cost increases quadratically with the number of input tokens [54]. This high cost becomes a significant problem when designing fusion blocks since the number of tokens increases significantly with the number of modalities. To address this, Fayyaz et al. [9] proposed a parameter-free module that prunes the attention matrix within self-attention layers. Similarly, Shang et al. [48] introduced a token-pruning strategy for large multimodal models using the Interquartile Range (IQR) scoring function [5] to cluster similar tokens and sample from the remaining ones. Instead of pruning, Bolya et al. [4] combined redundant tokens based on the cosine similarity between their keys, while Marin et al. [35] treated tokens as discrete samples of a continuous signal, selecting a subset that best approximates it. In this work, we reduce the number of tokens per modality using a simple yet effective parameter-free approach that averages contiguous tokens to reduce the computational requirements of our method.

**Knowledge distillation** is a commonly used technique for transferring knowledge from a larger teacher model to a

smaller student model [19]. More recent approaches have explored dynamically changing the teacher-student roles during training [11, 60]. In multimodal learning, Radevski et al. [43] and Hatano et al. [17] used knowledge distillation to enhance the accuracy of a unimodal student by transferring knowledge from a multimodal teacher. Similarly, Wei et al. [56] proposed a multimodal-to-multimodal distillation framework with modality dropout for classification and segmentation tasks. Inspired by these works, we introduce a novel multimodal-to-multimodal distillation framework for egocentric action recognition, which not only incorporates modality dropout but also proposes an additional strategy to further enhance robustness to missing modalities.

# 3. KARMMA

This work focuses on the multimodal egocentric action recognition task, which aims to identify and classify human actions by analyzing data from multiple egocentric modalities. Fig. 1 provides an overview of our proposed KAR-MMA framework, which distills knowledge from a large teacher model into a smaller student model.

# 3.1. Problem definition

Given an egocentric dataset with M available modalities, let  $\mathbf{x}_i^j \in \mathbb{R}^{T \times D_1^j \times D_2^j \cdots \times D_n^j}$  represent the  $i^{\text{th}}$  action sequence for the  $j^{\text{th}}$  modality, where T is the number of time-steps, and  $D_1^j \times D_2^j \cdots \times D_n^j$  define the n specific data dimensions for the  $j^{\text{th}}$  modality. The goal is to develop a model f that processes all egocentric action sequences  $\mathbf{X}_i = \{\mathbf{x}_i^1, \mathbf{x}_i^2, \dots, \mathbf{x}_i^M\}$  and predicts the corresponding action  $\mathbf{y}_i$  from a set of possible actions, formulated as  $\hat{\mathbf{y}}_i = \operatorname{argmax} \sigma (f(\mathbf{X}_i))$ , where  $\sigma$  denotes the softmax operator.

It is important to mention that the representation of actions varies across datasets. For example, in the Something-Something dataset [14], an action is a single output,  $y_i = {\text{phrase}}$ . In contrast, in the Epic-Kitchens dataset [7], an action consists of multiple outputs,  $y_i = {\text{noun, verb}}$ .

### 3.2. Teacher and student models

The teacher and student models share the same overall architecture, as shown in Fig. 1. This architecture consists of three main components: feature extractors, a fusion block, and a multi-head MLP. However, the student uses smaller feature extractors and a more compact fusion block, reducing memory requirements and increasing inference speed. Additionally, a key advantage of our *multimodal* student is its ability to perform inference on any combination of the trained modalities without requiring separate models for each combination.

**Feature extractors (FEs).** As described in Sec. 3.1, the input data comprises M modalities. To reduce the training

cost of the teacher, we use *M* unimodal, frozen pre-trained FEs and train only the remaining components. This approach facilitates the updating of FEs as newer models become available. For each modality, we select a well-known FE available in various sizes, allowing the teacher to use a larger version while the student uses a smaller one. If available, the student FEs use pre-trained weights, which are then fine-tuned.

**Fusion block (FB)** is a crucial component of our framework, designed to fuse the features from all M modalities. Based on previous works [12, 31, 34, 38, 49], we designed a transformer-based FB capable of handling an arbitrary number of input tokens and modalities. As shown in Fig. 2, our FB begins with an embedding layer that linearly projects modality-specific tokens into a uniform dimension. These tokens, along with a learnable CLS token, are then processed by l transformer layers, where l is larger for the teacher than for the student. The output of our FB consists of M + 1tokens: one averaged token per modality and the CLS token, which aggregates information from all modalities. The CLS token is particularly important as it prevents reliance solely on per-modality averages by incorporating a compact representation of all modalities.

**Multi-head MLP (MH-MLP).** The final stage of our framework processes the M + 1 tokens produced by the FB using a MH-MLP. The final output of the model is obtained by averaging the M + 1 predictions of the MH-MLP. Our MH-MLP consists of a shared fully connected layer followed by multiple independent classification heads, where the number of heads corresponds to the number of outputs required to describe an action. As explained in Sec. 3.1., actions consists of single or multiple outputs.

#### 3.3. Knowledge distillation

As shown in Fig. 2, our distillation pipeline consists of two stages: first, we train our multimodal teacher, and second, we freeze it and distill its knowledge into our student (both models are detailed in Sec. 3.2.). Since multimodal egocentric action recognition is a classification task, we train the teacher using the cross-entropy loss

$$\mathcal{L}_{\rm CE} = -\frac{1}{N} \sum_{i=1}^{N} \mathbf{y}_i \cdot \log \hat{\mathbf{y}}_i^T, \tag{1}$$

where N is the number of action sequences in the training dataset,  $\mathbf{y}_i$  represents the ground-truth labels, and  $\hat{\mathbf{y}}_i^T$  denotes the predicted actions of the teacher.

Additionally, inspired by previous works [12, 44] demonstrating the benefits of feature alignment in multimodal settings, we include an alignment loss to enforce the fusion block to project all modalities into a common feature space. Given that video is the most informative modality for egocentric action recognition, we aligned other modalities to it



Figure 2: **KARMMA framework.** The left diagram illustrates the first stage of our knowledge distillation pipeline, where the teacher is trained. The right diagram shows the second stage, where the student learns from the frozen, previously trained teacher. During this stage, the student uses modality dropout and a strategy for handling missing modalities (as described in Sec. 3.4.) to improve robustness in incomplete input scenarios.

using the InfoNCE loss [39]:

$$\mathcal{L}_{\text{InfoNCE}} = \frac{1}{M-1} \sum_{m \neq \text{video}} \text{CE}\left(\frac{\sin(\mathbf{z}_{\text{video}}, \mathbf{z}_{\text{m}})}{\tau}, \widetilde{\mathbf{y}}\right), \quad (2)$$

where CE and  $sim(\cdot, \cdot)$  represents the cross-entropy and cosine similarity operators, respectively,  $\mathbf{z}_m$  is the output of the fusion block for modality  $m, \tau$  is a temperature parameter, and  $\tilde{\mathbf{y}} = \{1, 2, \dots, b\}$  represents a constructed set of target labels for a batch of b samples.

Our teacher is trained using a weighted combination of the cross-entropy and alignment losses

$$\mathcal{L}_{\mathrm{T}} = \alpha \, \mathcal{L}_{\mathrm{CE}} + \beta \, \mathcal{L}_{\mathrm{InfoNCE}} \,, \tag{3}$$

where  $\alpha$  and  $\beta$  balance the contribution of these losses.

Once trained, the teacher is frozen, and its knowledge is distilled into the student. We apply Kullback-Leibler (KL) divergence to align the class probability distributions from both models after their multi-head MLPs:

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{N} \sum_{i=1}^{N} \bar{\mathbf{y}}_{i}^{\mathrm{T}} \cdot \left( \log(\bar{\mathbf{y}}_{i}^{\mathrm{T}}) - \log(\bar{\mathbf{y}}_{i}^{\mathrm{S}}) \right), \qquad (4)$$

where  $\bar{\mathbf{y}}_{i}^{T} = \sigma\left(f^{T}\left(\mathbf{X}_{i}\right)\right)$  and  $\bar{\mathbf{y}}_{i}^{S} = \sigma\left(f^{S}\left(\mathbf{X}_{i}\right)\right)$  are the class probability distributions from the teacher and student models, respectively.

Similar to the teacher, the student is trained with a weighted combination of the cross-entropy and distillation losses

$$\mathcal{L}_{\rm S} = \gamma \, \mathcal{L}_{\rm CE} + \lambda \, \mathcal{L}_{\rm KL} \,, \tag{5}$$

where  $\gamma$  and  $\lambda$  are the respective weighting parameters. Note that, unlike the teacher, which uses frozen feature extractors, all components of the student are trained. Therefore, our proposed multimodal-to-multimodal distillation pipeline aims to create a student with improved accuracy, which requires less memory and is faster than the teacher.



Figure 3: **Missing modality strategy.** To handle missing modalities, the embedding layer projects the tokens from the feature extractor when available. Then, it adds the learned modality token  $\mathbf{t}^m$  to all the projected tokens. Finally, a learned token  $\mathbf{t}^m_i$  is added to each individual token. The difference between  $\mathbf{t}^m$  and  $\mathbf{t}^m_i$  is that  $\mathbf{t}^m$  is learned per modality, whereas  $\mathbf{t}^m_i$  is learned per token.

#### 3.4. Missing modality robustness

Previous works [12, 45, 56] have demonstrated that using modality dropout during training improves robustness to missing modalities. Modality dropout works similarly to standard neuron dropout but operates at modality level; *i.e.*, each modality is randomly dropped with probability p, ensuring at least one remains. As shown in Fig. 2, we apply this technique exclusively to the student during distillation, allowing it to learn from the teacher how to compensate for missing modalities.

To further enhance the robustness of the student model to missing modalities, we propose a simple yet effective strategy illustrated in Fig. 3. Our strategy uses two types of learnable tokens:

- 1. Modality-specific token  $(\check{t}^m)$ : A single token learned per modality, helping to differentiate modalities and acting similarly to positional encodings.
- 2. Token-specific tokens  $(\dot{\mathbf{t}}_i^m)$ : A set of k tokens learned within each modality, enabling better adaptation when

#### a modality is missing.

Assuming that the feature extractor for modality m produces k tokens for a given input  $\mathbf{x}_i^m$ , the output of the embedding layer  $\mathbf{o}_i^m$  is expressed as

$$\mathbf{o}_{i}^{m} = \begin{cases} \left\{ \mathbf{t}_{i}^{m} + \breve{\mathbf{t}}^{m} + \dot{\mathbf{t}}_{i}^{m} \right\}_{i=1}^{k} & \text{if } m \text{ is available} \\ \left\{ \breve{\mathbf{t}}^{m} + \dot{\mathbf{t}}_{i}^{m} \right\}_{i=1}^{k} & \text{otherwise} \end{cases}, \quad (6)$$

where  $\mathbf{t}_i^m$  represents the *i*<sup>th</sup> projected token.

Unlike previous methods that either replace missing tokens with zeros [56] or replicate a single learned modality token [45], our approach provides richer information through the combination of both types of learnable tokens. Moreover, it ensures that even when one or more modalities are missing, the number of input tokens for the transformer layers of the fusion block remains the same, facilitating the effective handling of missing modalities.

#### **3.5.** Token reduction

Since the memory requirements of our fusion block increase with each additional modality, we propose a parameter-free token reduction technique ( $\Theta$ -Average) to reduce these requirements. Our proposed technique is based on a predefined threshold,  $\Theta$ . If the feature extractor of a specific modality produces more than  $\Theta$  tokens ( $k > \Theta$ ), the tokens are divided into  $\Theta$  groups, each containing  $\left\lfloor \frac{k}{\Theta} \right\rfloor$  tokens. If  $k \mod \Theta \neq 0$ , the last group includes the remaining tokens. Then, each group is averaged, resulting in exactly  $\Theta$  tokens that are fed to the fusion block. On the other hand, if  $k \leq \Theta$ , all tokens per modality,  $\Theta$ -Average reduces the computational cost of our fusion block. It is important to mention that this technique is applied to the output of the feature extractors of both the teacher and student models.

### 4. Experiments

**Datasets.** We evaluate our method on two widely used datasets for egocentric action recognition: Epic-Kitchens-100[7] and Something-Something (V2) [14]. Epic-Kitchens includes 300 nouns and 97 verbs used to describe actions, where each action consists of a combination of a verb and a noun, *e.g.*, "pick up + knife." For this dataset, we use three modalities: video (V), audio (A), and optical flow (F). On the other hand, Something-Something comprises data from 174 object-agnostic actions, *e.g.*, "moving [something] up." We also use three modalities for this dataset: video (V), optical flow (F), and object detection annotations (D).

As in most previous works, we report results on the validation sets of both datasets, since the test sets are not publicly available. However, we found that the validation set of Epic-Kitchens contained four nouns absent from the training set. To ensure a fair evaluation, we removed all samples containing those nouns and used this pruned validation set for our evaluations. Since computing results using the validation set as both a validation and test set may not reflect the generalization capability of our method, we created a custom split of the Epic-Kitchens training set, named Epic-Kitchens\*, allocating 90% for training and 10% for validation, and kept the pruned validation split for testing.<sup>1</sup>

**Implementation details.** For the teacher, we use Swin-B [30], pre-trained on Kinetics-400 [22], as the feature extractor for V, A, and F. For D, we use STLT [42] with 12 layers, pre-trained on Action Genome [21]. The fusion block has an embedding dimension of 768, with 2 transformer layers (l = 2), 8 attention heads, and 30% of attention dropout. We apply our  $\Theta$ -Average token reduction (see Sec. 3.5.) with  $\Theta = 300$ . Additionally, for both datasets, we train the teacher for 100 epochs using a batch size of 32, weight decay of 0.05, and gradient clipping at 1.0. Training is performed with the AdamW optimizer [32] using an initial learning rate of  $1e^{-5}$ , linearly increased to  $5e^{-4}$  during the first 10 epochs, followed by cosine decay to the initial value. For the loss function (Eq. 3), we set  $\alpha = 0.7$ ,  $\beta = 0.3$ , and  $\tau = 0.1$ .

For the student, we maintain most configurations from the teacher. Therefore, we just described the configurations that change. The student uses Swin-T [30], pre-trained on Kinetics-400 [22], as the feature extractor for V and F. For A and D, it uses AST-T [13] and STLT [42] with 9 layers, respectively, both models without pre-trained weights due to unavailability. The fusion block has an embedding dimension of 384, with l = 1, no attention dropout, and 50% of modality dropout. Training is performed with a batch size of 6, weight decay of 0.01, gradient clipping at 2.0, and a maximum learning rate of  $1e^{-4}$ . For the loss function (Eq. 5), we set  $\gamma = 0.7$  and  $\lambda = 0.3$ .

#### 4.1. Analysis of our KARMMA framework

To evaluate the effectiveness of our proposed framework, Tab. 1 presents the results of our teacher (KARMMA<sub>T</sub>) and student (KARMMA<sub>S</sub>) models, described in Sec. 3.2.. KARMMA<sub>S</sub> includes multimodal-tomultimodal knowledge distillation (see Sec. 3.3.), modality dropout (see Sec. 3.4.), and our strategy for handling missing modalities (see Sec. 3.4.). We refer to these modifications as "KARMMA enhancements." To evaluate their impact, we compare KARMMA<sub>S</sub> to two baselines. The first, Baseline, shares the same architecture as KARMMA<sub>S</sub> but is trained end-to-end with cross-entropy, without any of the KARMMA enhancements. The second, Baseline w/ $\delta$  incorporates modality dropout and our proposed strategy for missing modality robustness.

The results of Tab. 1 show that when all modalities are available (V+F+[A/D]), both baselines and the student out-

<sup>&</sup>lt;sup>1</sup>The Epic-Kitchens\* split will be integrated into our codebase.

Method	Inference Modalities	Epic -Kitchens	Epic -Kitchens*	Something -Something
KARMMA <sub>T</sub>	V+F+[A/D]	33.94	32.91	47.82
Baseline		40.00	39.26	57.31
Baseline w/ $\delta$	V+F+[A/D]	41.49	40.05	60.31
KARMMA <sub>S</sub>		42.55	40.44	63.19
Baseline		36.80	36.07	56.98
Baseline w/ $\delta$	V+F	39.69	39.00	60.38
KARMMA <sub>S</sub>		40.98	39.37	60.55
Baseline		37.48	36.60	40.03
Baseline w/ $\delta$	V+[A/D]	39.36	37.76	53.04
KARMMA <sub>S</sub>		40.90	39.49	59.57
Baseline		5.49	6.84	37.35
Baseline w/ $\delta$	F+[A/D]	27.09	27.21	<u>50.78</u>
KARMMA <sub>S</sub>		30.80	30.33	57.56
Baseline		32.10	32.50	39.91
Baseline w/ $\delta$	V	37.84	36.48	53.02
KARMMA <sub>S</sub>		39.35	38.28	55.39
Baseline		2.58	4.73	36.50
Baseline w/ $\delta$	F	25.64	26.01	50.97
KARMMA <sub>S</sub>		28.67	28.40	51.60
Baseline		2.05	2.25	0.07
Baseline w/ $\delta$	[A/D]	6.23	6.44	1.21
KARMMA <sub>S</sub>		7.58	6.73	35.75

Table 1: Analysis of KARMMA across different modality combinations. "KARMMA<sub>T</sub>" and "KARMMA<sub>S</sub>" refer to our proposed teacher and student models (see Sec. 3.2.). The "Baseline" model uses the same network architecture as "KARMMA<sub>S</sub>," but is trained end-to-end with cross-entropy, without the proposed KAR-MMA enhancements. "Baseline w/ $\delta$ " adds our proposed strategies for missing modality robustness (see Sec. 3.4.) to the "Baseline" model. "V," "F," "A," and "D" represent video, optical flow, audio, and object detection annotations, respectively. "[A/D]" indicates that either audio or object detection is used, depending on the dataset. The reported results are the action accuracies on both datasets (higher is better). Additionally, the "Epic-Kitchens\*" column shows results obtained using our custom Epic-Kitchens split (see "Datasets" in Sec. 4.). Gray indicates our final student. **Bold** and underline values indicate the best and second best results.

perform the teacher. Typically, in knowledge distillation, the teacher achieves better results than the student. However, in this work, the student benefits from full training, as described in Sec. 3.3.. While it uses smaller feature extractors, it trains them instead of freezing them like the teacher, allowing it to fine-tune its entire architecture for multimodal egocentric action recognition. Nevertheless, since the feature extractors of the teacher model are frozen, its fusion block is forced to learn the best strategy to fuse features from different modalities. Consequently, when the knowledge from the teacher is distilled to the student, it achieves absolute accuracy improvements of 2.55%, 1.18%, and 5.88% over the Baseline model on the Epic-Kitchens, Epic-Kitchens\*, and Something-Something datasets, respectively.

Additionally, Tab. 1 shows that the best results are ob-



Figure 4: **Impact of missing modalities at inference time.** The "Baseline" model uses the same network architecture as "KARMMA<sub>S</sub>," but it was trained end-to-end with cross-entropy, without the proposed KARMMA enhancements. "Baseline w/ $\delta$ " adds our proposed strategies for missing modality robustness (see Sec. 3.4.) to the "Baseline" model. "KARMMA<sub>S</sub>" refer to our proposed student model (see Sec. 3.2.).

tained when all modalities are available, indicating that our method effectively leverages multimodal information. The results also confirm that video is the most informative modality for egocentric action recognition, while audio and object detection annotations contribute the least. On Epic-Kitchens, switching from V+A to F+A causes a significant accuracy drop for all models. This effect is even more pronounced in unimodal settings, particularly when relying on [A/D], depending on the dataset. However, these drops are less severe when modality dropout and our missing modality handling strategy are applied, suggesting that they help prevent overfitting to the most informative modality.

Since KARMMA<sub>S</sub> consistently outperforms both baselines across all modality combinations, it demonstrates that while a model can be trained to handle missing modalities (Baseline w/  $\delta$ ), incorporating our knowledge distillation (KARMMA<sub>S</sub>) further enhances both accuracy and robustness to missing modalities. This enhancement is particularly evident on Something-Something, where KARMMA<sub>S</sub>

Method	Student Train. Mods.	Student Infer. Mods.	Flexible Mod. Inference	Teacher Train. Params. (M)	Student Train. Params. (M)	Action Acc. (%)	GFLOPs
Radevski et al. [43]	V	V	×	<u>84.65</u>	<b>28.21</b>	41.81	<b>175</b>
KARMMA <sub>V</sub>	V		×	11.25	29.64	40.46	<u>176</u>
KARMMA <sub>S</sub>	V+F+A	v	1	<b>11.25</b>	65.15	39.35	<u>176</u>
KARMMA <sub>[43]→S</sub>	V+F+A		1	<u>84.65</u>	65.21	39.40	<u>176</u>
KARMMA <sub>S</sub>	V+F+A	V+F+A	1	<b>11.25</b>	65.15	<u>42.55</u>	358
KARMMA <sub>[43]→S</sub>	V+F+A		1	<u>84.65</u>	65.21	<b>42.83</b>	358

Table 2: Comparison of KARMMA with the current SOTA multimodal-to-unimodal distillation approach for egocentric action recognition. All methods use knowledge distillation from a multimodal teacher trained with video (V), optical flow (F), and audio (A) modalities. "KARMMA<sub>V</sub>" represents a unimodal version of our multimodal "KARMMA<sub>S</sub>" student. "KARMMA<sub>[43]→S</sub>" represents our KARMMA<sub>S</sub> trained using knowledge distilled from the teacher of [43]. The reported results for Radevski et al. [43] were obtained using their provided checkpoint for their student model. Gray indicates our proposed multimodal-to-multimodal distillation framework with different teachers. Bold and <u>underline</u> values indicate the best and second best results.

achieves an absolute accuracy improvement of 34.54% over "Baseline w/  $\delta$ " when using only object detection annotations. For Epic-Kitchens, the improvement is smaller as audio is often less informative than object detection annotations, which provide some spatial and dimensional information about the objects involved in the action. Additionally, the results of the models trained with our custom Epic-Kitchens\* split (see Sec. 4., "Datasets") yield results comparable to those obtained from the original split, demonstrating the generalization of our KARMMA framework.

Although Tab. 1 demonstrates the ability of our student to handle missing modalities; it only evaluates the impact of missing entire modalities at inference time. However, real-world scenarios may involve dynamic missing patterns. Thus, Fig. 4 shows the evaluation of the two baselines and our student when the probability of missing modalities increases from 0% to 90% during inference. These results indicate that the Baseline model is highly sensitive to missing modalities, with just 10% missing data causing an absolute accuracy drop of approximately 5% on both datasets. In the extreme case of 90% missing modalities, accuracy drops by about 27% and 32% for Epic-Kitchens and Something-Something, respectively. However, Fig. 4 also demonstrates that our missing modality handling strategy enhances the robustness to missing modalities, as the accuracy drop of Baseline w/ $\delta$  is less severe compared to the one of Baseline. Moreover, our distillation pipeline consistently improves the accuracy of KARMMA<sub>S</sub>, achieving absolute accuracy improvements of approximately 3% and 13% over the Baseline w/ $\delta$  in the 90% missing modality scenario for Epic-Kitchens and Something-Something, respectively.

#### 4.2. Comparison with other methods

Since our method is inspired by Radevski et al. [43], which is the current state-of-the-art (SOTA) in multimodalto-unimodal distillation for egocentric action recognition, Tab. 2 presents a comparison between our approach and theirs on the Epic-Kitchens dataset. To ensure a fair comparison with their student, we also evaluate a unimodal version of our student (KARMMA<sub>V</sub>). Additionally, since Radevski et al. [43] use a fully trainable teacher, we leverage their teacher to train our multimodal student (KARMMA<sub>[43]→S</sub>).

The top portion of Tab. 2 shows that our KARMMA<sub>V</sub> achieves comparable accuracy to the student of Radevski et al. [43], despite using a teacher with approximately 7.5 times fewer trainable parameters. While our multimodal student (KARMMA<sub>S</sub>) performs slightly worse than theirs, the results of KARMMA<sub>[43]→S</sub> suggest that fully training our KARMMA<sub>T</sub> instead of freezing its feature extractors could further improve the accuracy of KARMMA<sub>S</sub>. These results also highlight the versatility of our approach, as it can also benefit from more powerful teachers if available.

Moreover, as shown in the bottom part of Tab. 2, a key advantage of our KARMMA<sub>S</sub> is its ability to perform inference with different modality combinations, unlike the unimodal student from Radevski et al. [43]. When leveraging all available modalities (V+F+A), KARMMA<sub>S</sub> surpasses their student accuracy. However, since KARMMA<sub>S</sub> is designed to handle and benefit from multiple modality combinations, its accuracy on video alone is slightly lower than that of its unimodal counterpart, which is optimized specifically for a single modality rather than for multiple modalities. All students have a similar computational cost, making our multimodal student just as efficient while offering the added flexibility of supporting multiple modalities. Therefore, our student can be used across various multimodal scenarios, benefiting from the availability of all modalities while maintaining robustness even if some modalities are sometimes missing.

### 4.3. Ablation studies

All ablation studies were conducted on Epic-Kitchens, with models trained for 50 epochs instead of 100. Additional ablations are included in the supplementary material.

Modality dropout. Tab. 3 demonstrates that applying

Method	Noun Acc. (%)	Verb Acc. (%)	Action Acc. (%)	
Baseline	50.08	64.74	38.94	
Baseline w/ Mod. dropout	51.30	<u>65.22</u>	<u>39.82</u>	
Baseline w/ $\delta$	<u>50.60</u>	65.61	39.87	

Table 3: Ablation study of our proposed strategies for improving robustness to missing modalities. The "Baseline" model uses the same network architecture as our KARMMA student, but it was trained end-to-end with cross-entropy, without the proposed KAR-MMA enhancements. "Baseline w/ Mod. dropout" and "Baseline w/  $\delta$ " indicate that the "Baseline" model incorporates either modality dropout alone or both modality dropout and our strategy for handling missing modalities (see Sec. 3.4.). Gray indicates the chosen strategy. **Bold** and <u>underline</u> values indicate the best and second best results.

Token Reduction	# of Tokens per Mod.	GFLOPs	Mem. Usage of FB (GB)	Action Acc. (%)
None	785	1165	6.52	31.19
Random [17]	<u>300</u>	1137	<u>1.97</u>	30.40
$\Theta$ -Average	500	1147	3.41	30.52
Θ-Average	<u>300</u>	<u>1137</u>	<u>1.97</u>	<u>31.11</u>
Θ-Average	100	1128	1.13	28.47

Table 4: Ablation study of our proposed token reduction strategy. "None" indicates that all tokens are used. "Random" corresponds to a baseline where random masking was applied to the tokens [17]. " $\Theta$ -Average" refers to our strategy described in Sec. 3.5.. The memory usage column reports the average memory consumption of the fusion block during training. The accuracy results were obtained using a bimodal version of our teacher model with video and audio modalities, following the implementation details in Sec. 4.. Gray indicates the chosen strategy. Bold and <u>underline</u> values indicate the best and second best results.

modality dropout improves the accuracy of the baseline. Additionally, incorporating our proposed strategy for enhancing robustness to missing modalities (see Sec. 3.4.) further improves verb and action accuracy.

**Token reduction.** Tab. 4 shows that our proposed token reduction strategy (see Sec. 3.5.) outperforms random token masking [17] while remaining a simple and parameter-free approach. Using  $\Theta = 300$  keeps fewer than half of the original 785 tokens while maintaining nearly the same action accuracy as using all tokens, providing the best accuracy-GFLOPs trade-off. Moreover, these results demonstrate that the main benefit of using a token reduction strategy is that it significantly reduces the memory consumption of our fusion block by processing fewer tokens.

**Resource efficiency.** Fig. 5 compares the memory usage and GFLOPs for the teacher and student during inference with different modality combinations. The results show that our KARMMA student reduces memory consumption by at least 50% compared to the teacher model while significantly



Figure 5: Ablation study on the resource efficiency of our student model. "KARMMA<sub>T</sub>" and "KARMMA<sub>S</sub>" refer to our proposed teacher and student models (see Sec. 3.2.). The top figure compares their memory consumption across multiple modality combinations at inference time, while the bottom figure compares their GFLOPs. "V," "F," and "A" denote video, optical flow, and audio, respectively. Some modality combinations are omitted as they yield identical results, *e.g.*, "F" matches "V" in memory and GFLOPs as they use the same feature extractor (see "Implementation details" in Sec. 4.).

decreasing GFLOPs, translating to faster inference times.

# 5. Conclusions

In this work, we introduced the first multimodal-tomultimodal approach for egocentric action recognition, resulting in a student model that leverages multiple modalities while remaining robust to missing ones. Unlike unimodal students, our model can perform inference with any combination of the trained modalities. Moreover, by using frozen pre-trained unimodal feature extractors for the teacher, our method reduces training costs and allows easy integration of newer models. Additionally, our parameter-free token reduction strategy improves computational efficiency without sacrificing accuracy. Our approach outperforms the state-ofthe-art multimodal-to-unimodal distillation method, achieving higher accuracy while being memory- and GFLOPsefficient. Furthermore, the flexibility and robustness of our method, make it well-suited for real-world scenarios, and future work could explore its generalization to exocentric datasets and additional modalities.

#### 6. Acknowledgments

This work was supported by projects PID2021-125209OB-I00 and TED2021-129410B-I00, (MCIN/AEI/10.13039/501100011033 and FEDER/UE and NextGenerationEU/PRTR), DGA 2022-2026 grant and Grant SMT Erasmus+, project 2022-1-ES01-KA131-HED-000065592 funded by Campus Iberus. The project has also been supported in part by hessian.AI through the Connectom Fund and by the State of Hesse through the cluster project "The Third Wave of Artificial Intelligence (3AI)."

# References

- Peri Akiva, Jing Huang, Kevin J. Liang, Rama Kovvuri, Xingyu Chen, Matt Feiszli, Kristin Dana, and Tal Hassner. Self-supervised object detection from egocentric videos. In *ICCV*, pages 5225–5237, 2023.
- [2] Shervin Ardeshir and Ali Borji. Integrating egocentric videos in top-view surveillance videos: Joint identification and temporal alignment. In ECCV, pages 300–317, 2018.
- [3] Md Mushfiqur Azam and Kevin Desai. A survey on 3D egocentric human pose estimation. In CVPR, pages 1643– 1654, 2024.
- [4] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *ICLR*, 2023.
- [5] Azzedine Boukerche, Lining Zheng, and Omar Alfandi. Outlier detection: Methods, models, and classification. ACM Comput. Surv., 53(3):1–37, 2020.
- [6] Jianchun Chen, Jian Wang, Yinda Zhang, Rohit Pandey, Thabo Beeler, Marc Habermann, and Christian Theobalt. EgoAvatar: Egocentric view-driven and photorealistic fullbody avatars. In SIGGRAPH, pages 1–11, 2024.
- [7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, et al. Scaling egocentric vision: The EPIC-KITCHENS dataset. In ECCV, pages 720–736, 2018.
- [8] Hao Dong, Ismail Nejjar, Han Sun, Eleni N. Chatzi, and Olga Fink. SimMMDG: A simple and effective framework for multi-modal domain generalization. In *NeurIPS\*2023*.
- [9] Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Jürgen Gall. Adaptive token sampling for efficient vision transformers. In ECCV, pages 396–414. Springer, 2022.
- [10] Antonino Furnari, Giovanni Maria Farinella, and Sebastiano Battiato. Temporal segmentation of egocentric videos to highlight personal locations of interest. In *ECCVW*, pages 474–489. Springer, ECCV.
- [11] Nuno Cruz Garcia, Sarah Adel Bargal, Vitaly Ablavsky, Pietro Morerio, Vittorio Murino, and Stan Sclaroff. Distillation multiple choice learning for multimodal action recognition. In WACV, pages 2754–2763, 2021.
- [12] 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 *CVPR*, pages 6481–6491, 2023.

- [13] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. In *Interspeech*, pages 571–575, 2021.
- [14] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, et al. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, pages 5842– 5850, 2017.
- [15] Kristen Grauman, Andrew Westbury, Eugene Byrne, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, 2022.
- [16] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, et al. Ego-Exo4D: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*, pages 19383–19400, 2024.
- [17] Masashi Hatano, Ryo Hachiuma, Ryo Fujii, and Hideo Saito. Multimodal cross-domain few-shot learning for egocentric action recognition. In *ECCV*, pages 182–199, 2024.
- [18] Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. Align and attend: Multimodal summarization with dual contrastive losses. In *CVPR*, pages 14867–14878, 2023.
- [19] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. arXiv:1503.02531 [stat.ML], 2015.
- [20] Mingzhen Huang, Xiaoxing Li, Jun Hu, Honghong Peng, and Siwei Lyu. Tracking multiple deformable objects in egocentric videos. In *CVPR*, pages 1461–1471, 2023.
- [21] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatiotemporal scene graphs. In CVPR, pages 10236–10247, 2020.
- [22] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. arXiv:1705.06950 [cs.CV], 2017.
- [23] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a little help from my temporal context: Multimodal egocentric action recognition. In *BMVC*, page 268, 2021.
- [24] Bolin Lai, Miao Liu, Fiona Ryan, and James M. Rehg. In the eye of transformer: Global-local correlation for egocentric gaze estimation. *Int. J. Comput. Vision*, 132(3):854–871, 2024.
- [25] Kyungjun Lee, Abhinav Shrivastava, and Hernisa Kacorri. Hand-priming in object localization for assistive egocentric vision. In WACV, pages 3411–3421, 2020.
- [26] Yin Li, Zhefan Ye, and James M. Rehg. Delving into egocentric actions. In CVPR, pages 287–295, 2015.
- [27] Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In ECCV, pages 639–655, 2018.
- [28] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, et al. Egocentric video-language pretraining. In *NeurIPS\*2022*.
- [29] Miao Liu, Lingni Ma, Kiran Somasundaram, Yin Li, Kristen Grauman, James M. Rehg, and Chao Li. Egocentric activity recognition and localization on a 3D map. In *ECCV*, pages 621–638. Springer, 2022.

- [30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.
- [31] Zecheng Liu, Jia Wei, Rui Li, and Jianlong Zhou. SFusion: Self-attention based N-to-one multimodal fusion block. In *MICCAI*, pages 159–169. Springer, 2023.
- [32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.
- [33] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, et al. Aria everyday activities dataset. arXiv:2402.13349 [cs.CV], 2024.
- [34] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *CVPR*, pages 18177–18186, 2022.
- [35] Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastegari, and Oncel Tuzel. Token pooling in vision transformers for image classification. In WACV, pages 12–21, 2023.
- [36] Kyle Min and Jason J. Corso. Integrating human gaze into attention for egocentric activity recognition. In WACV, pages 1069–1078, 2021.
- [37] Lorenzo Mur-Labadia, Jose J. Guerrero, and Ruben Martinez-Cantin. Multi-label affordance mapping from egocentric vision. In *ICCV*, pages 5238–5249, 2023.
- [38] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *NeurIPS\*2021*, pages 14200–14213.
- [39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv:1807.03748 [cs.LG], 2018.
- [40] Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, and Barbara Caputo. E<sup>2</sup>(GO)MOTION: Motion augmented event stream for egocentric action recognition. In *CVPR*, pages 19935– 19947, 2022.
- [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 *ICCV*, pages 5285–5297, 2023.
- [42] Gorjan Radevski, Marie-Francine Moens, and Tinne Tuytelaars. Revisiting spatio-temporal layouts for compositional action recognition. In *BMVC*, 2021.
- [43] Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *ICCV*, pages 5213– 5224, 2023.
- [44] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.
- [45] Merey Ramazanova, Alejandro Pardo, Humam Alwassel, and Bernard Ghanem. Exploring missing modality in multimodal egocentric datasets. arXiv:2401.11470 [cs.CV], 2024.
- [46] Fiona Ryan, Hao Jiang, Abhinav Shukla, James M. Rehg, and Vamsi Krishna Ithapu. Egocentric auditory attention localization in conversations. In *CVPR*, pages 14663–14674, 2023.

- [47] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *CVPR*, pages 21096–21106, 2022.
- [48] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLava-PruMerge: Adaptive token reduction for efficient large multimodal models. arXiv:2403.15388 [cs.CV], 2024.
- [49] Tim Siebert, Kai Norman Clasen, Mahdyar Ravanbakhsh, and Begüm Demir. Multi-modal fusion transformer for visual question answering in remote sensing. arXiv:2210.04510 [cs.CV], 2022.
- [50] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *CVPR*, pages 7396–7404, 2018.
- [51] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Learning to recognize actions on objects in egocentric video with attention dictionaries. *IEEE T. Pattern Anal. Mach. Intell.*, 2021.
- [52] Tamas Suveges and Stephen McKenna. Egomap: Hierarchical first-person semantic mapping. In *ICPR*, pages 348–363. Springer, 2021.
- [53] Shuhan Tan, Tushar Nagarajan, and Kristen Grauman. EgoDistill: Egocentric head motion distillation for efficient video understanding. In *NeurIPS\*2023*, pages 33485–33498.
- [54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS\*2017*, pages 5998–6008.
- [55] Huiyu Wang, Mitesh Kumar Singh, and Lorenzo Torresani. Ego-Only: Egocentric action detection without exocentric transferring. In *ICCV*, pages 5250–5261, 2023.
- [56] Shicai Wei, Chunbo Luo, and Yang Luo. MMANet: Marginaware distillation and modality-aware regularization for incomplete multimodal learning. In *CVPR*, pages 20039– 20049, 2023.
- [57] Jay Zhangjie Wu, David Junhao Zhang, Wynne Hsu, Mengmi Zhang, and Mike Zheng Shou. Label-efficient online continual object detection in streaming video. In *ICCV*, pages 19246–19255, 2023.
- [58] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. arXiv:2409.07825 [cs.CV], 2024.
- [59] Xuehan Xiong, Anurag Arnab, Arsha Nagrani, and Cordelia Schmid. M&M mix: A multimodal multiview transformer ensemble. arXiv:2206.09852 [cs.CV], 2022.
- [60] Lehan Yang and Kele Xu. Cross modality knowledge distillation for multi-modal aerial view object classification. In *CVPR*, pages 382–387, 2021.
- [61] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE T. Intell. Transp. Syst.*, 2023.
- [62] Mingfang Zhang, Yifei Huang, Ruicong Liu, and Yoichi Sato. Masked video and body-worn IMU autoencoder for egocentric action recognition. In *ECCV*, pages 312–330, 2024.

# Supplementary material

Sec. A of this supplementary material explains our choice of the alignment loss used in our proposed teacher model. In Sec. B, we investigate the impact of various techniques for handling missing modalities. Sec. C evaluates the generalization capabilities of our student model to unseen environments. Finally, Sec. D discusses the limitations of our proposed multimodal knowledge distillation approach for egocentric action recognition that is robust to missing modalities (KARMMA). All results presented in this supplementary material were obtained using the pruned validation set of the Epic-Kitchens dataset [7], described in Sec. 4. of the main paper.

### A Alignment loss selection

As explained in Sec. 3.3. of the main paper, during the first stage of our proposed distillation pipeline, we train our multimodal teacher using a combination of cross-entropy loss and an alignment loss. The alignment loss enforces the fusion block to project the features from all modalities into a common feature space. While there are general-purpose alignment losses such as InfoNCE [39], recently Lin et al. [28] propose EgoNCE, loss specifically designed for aligning egocentric data. However, a common challenge when using alignment losses is that they often require large batch sizes to effectively align the modalities [28, 44].

Tab. A.1 compares the accuracy obtained when using InfoNCE [39] or EgoNCE [28] losses for aligning video and audio. The results indicate that while InfoNCE is a general-purpose alignment loss, it outperforms EgoNCE. The difference in accuracy between both losses could be attributed to the batch size since our teacher was trained with a batch size of 32, while Lin et al. [28] used a batch size of 512 for EgoNCE. In contrast, InfoNCE has been successfully applied with batch sizes ranging from 8 to 64 [39]. Given that InfoNCE improves the accuracy of our teacher model even with a small batch size (32), we use it in our proposed KARMMA teacher.

# **B** Techniques for handling missing modalities

As described in Sec. 3.4. of the main paper, our proposed KARMMA framework incorporates modality dropout during the distillation process to enhance the robustness of the student to missing modalities. When a modality is dropped, there are two possible approaches: ignoring the missing modality or applying a technique to handle it. Inspired by [12, 45], we propose five techniques for handling missing modalities.

Assuming that the feature extractor for modality m produces k tokens for a given input  $\mathbf{x}_i^m$ , which are then processed by the embedding layer that produces k projected

Alignment Loss	Noun Acc. (%)	Verb Acc. (%)	Action Acc. (%)
EgoNCE [28]	<u>43.48</u>	<u>56.37</u>	<u>29.88</u>
InfoNCE [39]	44.55	56.59	31.11

Table A.1: **Comparison of alignment losses.** The accuracy results were obtained using a bimodal version of our teacher model with video and audio modalities, following the implementation details in Sec. 4. of the main paper. All models were trained for 50 epochs. Gray indicates the chosen alignment loss. **Bold** and <u>underline</u> values indicate the best and second best results.

tokens  $\mathbf{o}_i^m = \{\mathbf{t}_i^m\}_{i=1}^k$ . Our proposed strategies are detailed below.

**Single learnable modality token.** A learnable token is introduced for each modality:

$$\mathbf{o}_{i}^{m} = \begin{cases} \left\{ \mathbf{t}_{i}^{m} + \breve{\mathbf{t}}^{m} \right\}_{i=1}^{k} & \text{if } m \text{ is available} \\ \left\{ \breve{\mathbf{t}}^{m} \right\} & \text{otherwise} \end{cases}, \quad (B.1)$$

where  $\check{\mathbf{t}}^m$  is the learned token representing the missing modality. This strategy results in k tokens when the modality is present but only one when it is missing.

**Replicated learnable modality token.** To maintain a consistent number of tokens, the learned modality token is replicated k times:

$$\mathbf{o}_{i}^{m} = \begin{cases} \left\{ \mathbf{t}_{i}^{m} + \breve{\mathbf{t}}^{m} \right\}_{i=1}^{k} & \text{if } m \text{ is available} \\ \left\{ \breve{\mathbf{t}}^{m} \right\}_{i=1}^{k}, & \text{otherwise} \end{cases}.$$
(B.2)

This technique ensures that the same number of tokens is used whether or not the modality is present.

Learnable tokens per missing modality. Instead of a single token, k learnable tokens are introduced:

$$\mathbf{o}_{i}^{m} = \begin{cases} \left\{ \mathbf{t}_{i}^{m} + \dot{\mathbf{t}}_{i}^{m} \right\}_{i=1}^{k} & \text{if } m \text{ is available} \\ \left\{ \dot{\mathbf{t}}_{i}^{m} \right\}_{i=1}^{k} & \text{otherwise} \end{cases}, \quad (B.3)$$

where  $\dot{\mathbf{t}}_i^m$  is the *i*<sup>th</sup> learned token for modality *m*. This strategy provides more information than replicating a single token while keeping a consistent number of output tokens.

**Learned tokens only for missing modalities.** This strategy is similar to the previous one, except that no additional tokens are learned when the modality is available:

$$\mathbf{o}_{i}^{m} = \begin{cases} \{\mathbf{t}_{i}^{m}\}_{i=1}^{k} & \text{if } m \text{ is available} \\ \left\{ \dot{\mathbf{t}}_{i}^{m} \right\}_{i=1}^{k} & \text{otherwise} \end{cases}.$$
 (B.4)

Method	Inference Modalities	Action Accuracy (%)	Ranking
Baseline		38.94	5
Baseline w/ mod. drop.		39.82	2
Baseline w/ mod. drop. and Eq. (B.1)	)	39.58	3
Baseline w/ mod. drop. and Eq. (B.2)	V+F+A	39.87	1
Baseline w/ mod. drop. and Eq. (B.3)	)	39.41	4
Baseline w/ mod. drop. and Eq. (B.4)	)	39.58	3
Baseline w/ mod. drop. and Eq. (B.5)	, )	39.87	1
Baseline		35.45	7
Baseline w/ mod_dron		38.79	1
Baseline w/ mod drop and Eq. (B.1)	)	37.77	5
Baseline w/ mod drop and Eq. $(B.2)$	, ) V+F	38.43	2
Baseline w/ mod. drop. and Eq. (B.2)	)	37.74	<u>=</u>
Baseline w/ mod. drop. and Eq. (B.4)	)	38.09	4
Baseline w/ mod. drop. and Eq. (B.4) Baseline w/ mod. drop. and Eq. (B.5)	)	38.17	3
Deedler	•	26.02	7
Baseline Descline m/ made data		30.02	1
Baseline W/ mod. drop.		37.73	Ĩ
Baseline w/ mod. drop. and Eq. (B.1)	)	37.18	5
Baseline w/ mod. drop. and Eq. (B.2)	) V+A	37.10	6
Baseline w/ mod. drop. and Eq. (B.3)	)	37.62	3
Baseline w/ mod. drop. and Eq. (B.4)	)	<u>37.64</u>	<u>2</u>
Baseline w/ mod. drop. and Eq. (B.5)	)	37.59	4
Baseline		5.28	7
Baseline w/ mod. drop.		24.91	4
Baseline w/ mod. drop. and Eq. (B.1)	)	24.83	5
Baseline w/ mod. drop. and Eq. (B.2)	) F+A	25.56	1
Baseline w/ mod. drop. and Eq. (B.3)	)	24.77	6
Baseline w/ mod. drop. and Eq. (B.4)	)	25.40	2
Baseline w/ mod. drop. and Eq. (B.5)	)	25.04	3
Baseline		30.70	7
Baseline w/ mod. drop.		36.63	1
Baseline w/ mod drop and Eq. (B.1)	)	35.23	6
Baseline w/ mod drop and Eq. $(B_2)$	v	35.41	5
Baseline w/ mod. drop. and Eq. (B.2)	) .	35.48	4
Baseline w/ mod. drop. and Eq. (B.4)	)	35.10	3
Baseline w/ mod. drop. and Eq. (B.5)	)	36.09	2
Basenne Winder urop: and Eq. (B.5)	,	<u>50.07</u>	-
Baseline		2.82	7
Baseline w/ mod. drop.		23.08	4
Baseline w/ mod. drop. and Eq. (B.1)	)	22.78	5
Baseline w/ mod. drop. and Eq. (B.2)	) F	23.33	2
Baseline w/ mod. drop. and Eq. (B.3)	)	22.76	6
Baseline w/ mod. drop. and Eq. (B.4)	)	24.00	1
Baseline w/ mod. drop. and Eq. (B.5)	)	23.18	3
Baseline		1.90	7
Baseline w/ mod. drop.		4.89	5
Baseline w/ mod. drop. and Eq. (B.1)	)	5.77	2
Baseline w/ mod. drop. and Eq. (B.2)	) A	5.61	3
Baseline w/ mod. drop. and Eq. (B.3)	)	4.84	6
Baseline w/ mod. drop. and Eq. (B.4)	)	5.42	4
Baseline w/ mod. drop. and Eq. (B.5)	)	6.11	1

Table B.1: Evaluation of techniques for handling missing modalities across different modality combinations. The "Baseline" model uses the same network architecture as our proposed student, but is trained end-to-end with cross-entropy, without the KARMMA enhancements. We then introduce 50% of modality dropout, described in Sec. 3.4. of the main paper, during the training of the Baseline and compare all techniques described in Sec. B. All models were trained for 50 epochs. Gray indicates the chosen technique. Bold and <u>underline</u> values indicate the best and second best results.

Here, the k learned tokens aim to best approximate the original modality tokens when the modality is missing.

**Combined learnable modality and token-specific tokens.** This final strategy combines both a learnable modality token

Method	Average Rank
Baseline	6.71
Baseline w/ mod. drop.	2.57
Baseline w/ mod. drop. and Eq. (B.1)	4.43
Baseline w/ mod. drop. and Eq. (B.2)	2.86
Baseline w/ mod. drop. and Eq. (B.3)	5.00
Baseline w/ mod. drop. and Eq. (B.4)	2.71
Baseline w/ mod. drop. and Eq. (B.5)	2.43

Table B.2: Average ranking of techniques for handling missing modalities. The methods are the same from Tab. B.1, with average ranking computed from the "Ranking" column of that table. Gray indicates the chosen technique. Bold and <u>underline</u> values indicate the best and second best results.

and k learnable tokens per token:

$$\mathbf{o}_{i}^{m} = \begin{cases} \left\{ \mathbf{t}_{i}^{m} + \mathbf{\check{t}}^{m} + \mathbf{\dot{t}}_{i}^{m} \right\}_{i=1}^{k} & \text{if } m \text{ is available} \\ \left\{ \mathbf{\check{t}}^{m} + \mathbf{\dot{t}}_{i}^{m} \right\}_{i=1}^{k} & \text{otherwise} \end{cases}.$$
(B.5)

This approach benefits from both strategies: the modality token provides general information about the missing modality, while the token-specific learnable tokens help preserve fine-grained details.

Tab. B.1 compares the action accuracy and ranking of not using modality dropout ("Baseline"), using modality dropout without any strategy to handle missing modalities ("Baseline w/ modality dropout"), and combine modality dropout with our five proposed strategies discussed above. The results demonstrate that incorporating modality dropout improves both robustness to missing modalities and accuracy. To facilitate the comparison of these methods, Tab. B.2 shows their average ranking. This average ranking indicate that combining a learnable modality token with learnable token-specific tokens (5<sup>th</sup> strategy) performs best across most modality combinations. Additionally, all the other evaluated techniques fail to outperform simple modality dropout alone, highlighting the importance of incorporating both types of learnable tokens. Consequently, we use the 5<sup>th</sup> strategy in our proposed KARMMA framework.

# C Adaptation to unseen environments

Egocentric data is often biased towards the individual capturing it. To account for this, the validation set of the Epic-Kitchens dataset [7] includes data captured by seen participants (who were present in the training set) and unseen participants (who were not included in the training set). Therefore, the "unseen" split serves as a benchmark for evaluating the ability of the models to adapt to distribution shifts. Tab. C.1 compares the accuracy of our KARMMA student model on the seen split, the unseen split, and the entire validation set (seen+unseen).

Validation Set Split	Noun Acc. (%)	Verb Acc. (%)	Action Acc. (%)	
Seen	55.05	68.52	43.73	
Unseen	45.07	58.59	33.05	
Seen+Unseen	<u>53.95</u>	<u>67.42</u>	42.55	

Table C.1: Analysis of the KARMMA student across different validation splits of the Epic-Kitchens dataset. The "Seen" and "Unseen" splits refer to portions of the full validation set ("Seen+Unseen"), where "Seen" contains participants also present in the training set, while "Unseen" consists of entirely new participants. A single KARMMA student model was trained for 100 epochs and evaluated on all splits. **Bold** and <u>underline</u> values indicate the best and second best results.

The results indicate that while our student model achieves high accuracy on the seen split, the noun, verb, and action accuracies drop by approximately 10% on the unseen split. This drop highlights the challenges of adapting to novel environments. However, when evaluated on the complete validation set, the decrease in accuracy is less pronounced, suggesting that our student has a reasonable degree of robustness to distribution shifts.

To further investigate the generalization capabilities of our method, we compare our KARMMA<sub>S</sub> with SimMMDG, a model proposed by Dong et al. [8], under their domain generalization setup for Epic-Kitchens with missing modalities. Fig. D.1 shows the absolute decrease in accuracy with respect to the full-modality model (which uses video, optical flow, and audio) across different modality combinations. These results show that while our KARMMA<sub>S</sub> is not explicitly designed for domain generalization, it achieves comparable performance to SimMMDG. Notably, our model demonstrates significantly higher robustness when audio is missing and video is present (V+F and V), further validating its effectiveness in handling missing modalities in real-world scenarios.

# **D** Limitations

The main limitation of our proposed KARMMA framework is that it assumes that all unimodal feature extractors are transformer-based models, as the fusion block is designed to process k tokens per modality, where k may vary across modalities. This constraint prevents the use of convolutional-based models as feature extractors, as their output format does not align with the expected input of the fusion block. Additionally, since our teacher relies on frozen unimodal feature extractors, its accuracy depends on the availability of suitable pre-trained models for each modality. Ideally, these feature extractors should be available in different sizes (*e.g.*, ViT-S and ViT-B), allowing the teacher to use the larger model while the student uses the smaller one, facilitating the distillation process. Finally, our distil-



Figure D.1: Comparison of robustness to missing modalities between KARMMA<sub>S</sub> and SimMMDG [8] under the domain generalization setup. The plot shows the absolute decrease in action accuracy (lower is better) relative to the full-modality model, which uses video (V), optical flow (F), and audio (A) for various modality combinations.

lation pipeline assumes the availability of all modalities for the teacher. However, as shown in Tab. B.1, models can still be trained end-to-end without this assumption.