LanViKD: Cross-Modal Language-Vision Knowledge Distillation for Egocentric Action Recognition

Yizheng Sun[1,*], Hao Li[1], ChengHua Lin[1] and Riza Batista-Navarro[1]

[1]*The University of Manchester*

Abstract

Understanding human actions through the analysis of egocentric videos is a desirable capability of intelligent agents, and is a research area that has gained popularity recently. Thus far, most approaches to egocentric (video) action recognition (EAR), i.e., the task of classifying a given video clip according to a predefined set of natural-language descriptions (actions), represent the target action classes (label) using one-hot encoding, thus ignoring any relationships or similarities between some of the actions. The goal of this work is to augment the generalisation capability of vision models through leveraging the pre-existing knowledge encoded within pre-trained language models. Specifically, we propose a language-vision knowledge distillation framework to distil a pre-trained language model's knowledge about actions (expressed in text) into a vision model. Instead of using the one-hot encoding representation of a label, we employ the probability distribution across all action classes—given by a language model—as a teaching signal. Our experiments demonstrate that our framework obtains improved performance and generalisation capability on EAR based on the EPIC-Kitchens, Something-Something V2 and Something-Else benchmarks.

Keywords

Egocentric Action Recognition, Language-Vision Multi-modality, Knowledge Distillation

1. Introduction

Egocentric vision is a subfield of computer vision that analyses first-person viewpoint vision data captured by a wearable camera. Núñez-Marcos et al. [1] highlights that compared with third-person view (exocentric) videos, egocentric videos usually involve rich hand-object interactions. Our framework leverages the observation that different egocentric actions often involve the same objects (e.g., both "Taking cutting board" and "Cutting onion" involve a cutting board) and captures such correlations using pre-trained large language models.

Early work has demonstrated that, in addition to the RGB modality, leveraging multiple modalities such as audio, optical flow, and the bounding box and category of an object help improve a model's capability to understand egocentric videos [2, 3, 4]. Such efforts have explored the potential of multi-modal knowledge distillation, where the teacher and student models receive different input modalities [5, 6, 7, 8]. Their results show that using the teacher's knowledge from certain modalities for training improves the student's performance on a different

HAII5.0: Embracing Human-Aware AI in Industry 5.0, at ECAI 2024, 19 October 2024, Santiago de Compostela, Spain.

*Corresponding author.

yizheng.sun@manchester.ac.uk (Y. Sun)

0009-0004-2600-1236 (Y. Sun); 0000-0002-9923-4346 (H. Li); 0000-0003-3454-2468 (C. Lin); 0000-0001-6693-7531 (R. Batista-Navarro)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

modality during inference. It is, however, unrealistic to assume that multiple modalities are always available. In contrast, the language modality is usually available because most existing EAR datasets are annotated according to target actions expressed in natural language [9, 10, 11]. Additionally, the rapid growth and impressive performance of pre-trained Language Models (LMs) on natural language processing (NLP) and computer vision (CV) tasks have been notable [12, 13, 14, 15]. Pre-trained LMs bring broader knowledge of human actions, that can support the language modality.

Extensive research has delved into exploring the potential of learning vision representations through supervision embedded in natural language [16, 17, 18, 19]. Consequently, it is natural to investigate whether LMs can be employed for video action recognition. Siddharth et al. [20] utilised language models to generate textual descriptions of videos, enabling their vision model to comprehend and identify actions more effectively through textual cues. Sun et al. [21] jointly trained video and language modalities, enabling tasks like action recognition to benefit from textual context. While previous studies demonstrated the advantages of integrating the language modality into video learning, they typically fuse video and language modalities together instead of utilising a pre-trained language model's latent knowledge directly. Several considerations drive the advancement of leveraging pre-existing knowledge in modelling. Firstly, language models (LMs) have showcased exceptional capabilities in few-shot and zero-shot transfer learning [22]. Consequently, LMs can be employed effectively with relatively small datasets, as their objective is solely to assist the existing LMs during inference. Secondly, methods based on LMs for video need little or even no training. Through plug-in modules, they can be utilised in a convenient manner [23]. In this study, we take a different route and propose a cross-modal language-vision knowledge distillation framework for EAR.

Figure 1 depicts our framework. The conventional training approach employs one-hot encoding to represent target actions, treating "Taking cutting board" and "Cutting onion" as distinct target classes. Consequently, a vision model perceives these two videos as unrelated due to the lack of consideration for the correlation between the action classes in the one-hot encoding scheme. However, this perspective fails to reflect the inherent relationships within the video data, leading to a lack of generalisation. This is different from a human standpoint, as humans would recognise that both videos share relevant visual features associated with the cutting board object. Conversely, a language model perceives textual action labels such as "Taking cutting board" and "Cutting onion" as relevant, given their shared usage of the word "cut", which better aligns with the video content. To address this discrepancy, our framework leverages a language model as the teacher to capture and incorporate this contextual relevance information into the EAR training process to help improve vision models' general understanding of videos. Furthermore, our framework also follows a multi-task learning approach for capturing correlations between the vision and language representations. We demonstrate that utilising a pre-trained language model as teacher can improve a vision model's performance and generalisation capability on the EAR task.

Contributions *(i)* We provide a cross-modal language-vision knowledge distillation framework for EAR. Our framework is highly flexible, and is not constrained in terms of the vision and language models involved. *(ii)* We demonstrate through experiments that a pre-trained

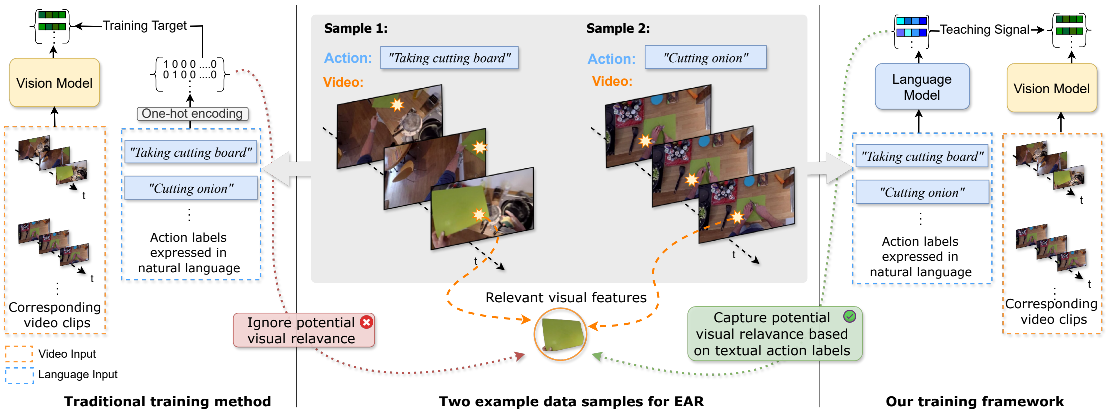

Training Target
Vision Model
1 0 0 0 0
0 1 0 0 0
One-hot encoding
"Taking cutting board"
"Cutting onion"
Action labels expressed in natural language
Corresponding video clips
Video Input
Language Input
Traditional training method
Sample 1:
Action: "Taking cutting board"
Video:
Sample 2:
Action: "Cutting onion"
Video:
Relevant visual features
Ignore potential visual relavance
Capture potential visual relavance based on textual action labels
Two example data samples for EAR
Teaching Signal
Language Model
Vision Model
"Taking cutting board"
"Cutting onion"
Action labels expressed in natural language
Corresponding video clips
Our training framework

Figure 1: In EAR data, samples include action labels described in natural language along with corresponding video clips. These video clips often exhibit relevant visual features corresponding to different action labels. However, traditional training methods commonly utilise one-hot encoding for action labels, which does not adequately capture this correlation and lacks generalisation. In contrast, our framework applies a language model on textual action labels to better understand the relationships among them, thereby aligning more closely with the inherent information in video data.

language model's pre-existing knowledge is beneficial for a vision model's understanding of egocentric vision. *(iii)* Our experiments show that our framework's performance in terms of accuracy improves upon a baseline approach by up to 2.6%. This superior performance is achieved without adding any additional computation for inference.

2. Related Work

Natural language supervision for vision learning focusses on learning visual representations from semantic information contained in natural language. Various methods have been introduced to learn visual presentations from text paired with images [16, 18, 24, 19]. Notably, a close work to ours is that of Gomez-Bigorda et al. [17], which projects given textual information into topic classes using Latent Dirichlet Allocation (LDA). They then use the probability distribution of topic classes as a supervisory signal to train a CNN with cross-entropy loss. In our case, we use pre-trained language models to generate the probability distribution and employ standard practice in knowledge distillation to train a transformer-based vision model. Furthermore, most of the aforementioned work are for pre-training visual representations, while our framework is directly applied to downstream tasks such as egocentric action recognition.

Multi-modal knowledge distillation. In the context of multi-modal knowledge distillation, several methods have been introduced in a cross-modal fashion [5, 6], where a student and a teacher receive a different modality, respectively. Alternatively, some efforts explored the distillation of knowledge between more than two modalities [7, 25, 8, 26], which have utilised vision and audio-based data such as raw RGB, optical flow and sound waves, etc. In contrast, we focus on knowledge distillation from a teacher model receiving language modality to a student model receiving RGB modality. Compared with vision and audio-based modalities, the strength of using language as a teaching modality comes from modern pre-trained language models, whose pre-existing knowledge contain strong generalisation and understanding capability.

Egocentric action recognition (EAR). One line of work has focussed on model architecture design to model the interplay between spatial and temporal information within RGB video frames [27, 28, 29]. Concurrently, another strand of research demonstrated that using object bounding boxes and categories to model hand-object interaction significantly improves EAR performance [30, 4]. Recent work showed that utilising multiple modalities demonstrates promising performance [2, 3, 8]. They utilised vision and audio-based modalities and have used a shared model architecture for different modalities. Notably, the language modality poses unique challenges due to its distinct data format, making direct application of existing methods impractical. Thus, we propose a novel framework aimed at harnessing the language modality specifically for EAR tasks.

Multi-task learning was originally introduced by Caruana [31], where a shared model generates output predictions for multiple tasks on the same input. Recent research highlighted the strong performance of multi-task learning in computer vision tasks [32, 33, 34]. In our study, we extend this concept to our knowledge distillation framework by incorporating a regression head. This head projects vision latent representations from a student onto pre-trained language latent representations provided by a teacher.

3. Methodology

This section provides a formal definition of the EAR task and delineates the procedural aspects of our framework, which we refer to as LanViKD. Figure 2 presents an overview of the architecture of LanViKD, which is comprised of two primary stages: Stage 1 entails the preparation of a language model designated as the teacher model, while Stage 2 involves performing cross-modal knowledge distillation.

3.1. Egocentric Action Recognition Formulation

Following Radevski et al. [8], we formally define the EAR task as follows. An RGB video clip is in the format of $c \in \mathbb{R}^{T \times C \times W \times H}$, where T is the number of sampled RGB frames, and C, W and H represent the number of channels, height and width. An egocentric action recognition dataset $\mathbb{D} = \{(c_1, w_1, y_1), ..., (c_N, w_N, y_N)\}$ contains N video clips c_i, together with textual narrations w_i describing actions in the clips, and one-hot encoding $y_i \in \mathbb{R}^C_+$ of the narrations. The goal of EAR is to predict $\hat{y}_i \in \mathbb{R}_+$ as the action class for a given video clip c_i, or alternatively, $(\hat{v}, \hat{n}) \in \mathbb{R}^2_+$ as the verb and noun constituting the action in a video. The traditional training target for EAR is the one-hot encoding of actions expressed in text [35, 29, 36]. However, as shown in Figure 1 some action classes such "taking cutting board" and "cut carrot" share common features with respect to the "cutting board" object in their corresponding RGB video frames. One-hot encoding ignores the this relationship between different action classes. The goal of our work

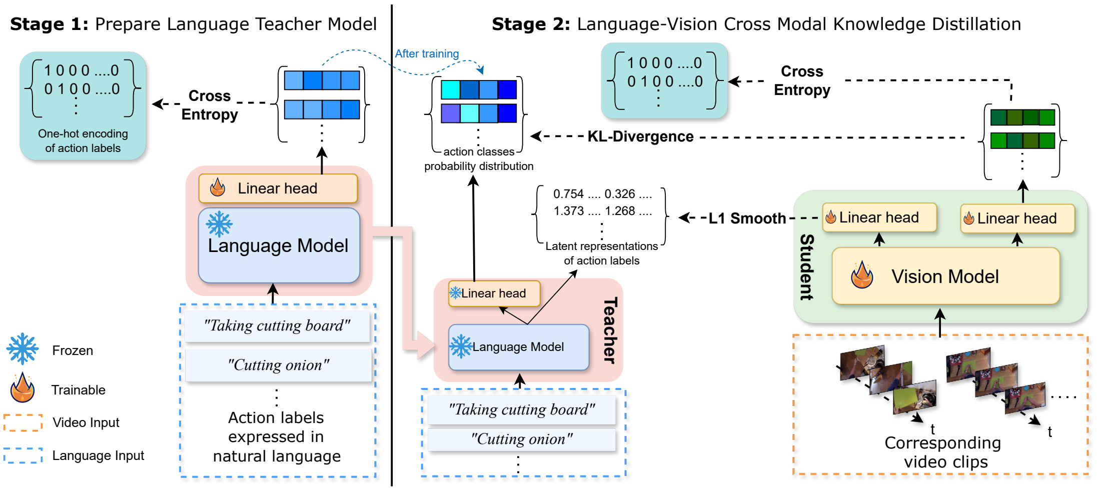


Figure 2: Overview of our LanViKD framework architecture. LanViKD consists of two main stages. During the first stage, we prepare a pre-trained language model to serve as the teacher by training a linear projection head atop it. In the second stage, we employ a vision model as the student and perform knowledge distillation.

is to utilise this relationship information for EAR training by distilling the knowledge of a pre-trained language model into an RGB video model.

3.2. Language Teacher Model Preparation

As shown in Figure 2, given an EAR dataset $\mathbb{D} = \{(c_1, w_1, y_1), ..., (c_N, w_N, y_N)\}$, we employ a pre-trained language model capable of processing sequences of text tokens to generate latent representations. Subsequently, we freeze the parameters of the language model and proceed to train a linear projection layer (or two separate linear projections in scenarios involving verb-noun compositional actions) atop the language model. This trained projection layer is tasked with classifying a textual action description w_i into its corresponding one-hot encoding index y_i (or verb and noun indices, as previously specified). Following training, the linear projection facilitates the generation of a soft probability distribution across all action classes given a textual action description as input. This soft distribution contains valuable semantic information, differing from conventional one-hot encoding. For instance, consider the actions "taking cutting board", which is associated with the noun label "cutting board" encoded as 1, and "cut carrot", labelled with the noun "carrot" encoded as 2. When inputting "taking cutting board" into the language model for noun index classification, it assigns the highest probability to 1 while also allocating a considerable probability to 2. This is due to the shared term "cut" in both textual actions, despite their distinct noun classes. Moreover, this semantic relationship is echoed in the video data, wherein both actions involve the object "cutting board". While one-hot indices categorise these videos into separate, unrelated classes, the probability distribution reflects their semantic connection, aligning more closely with the visual modality.

3.3. Cross-modal Language-Vision knowledge distillation

Once the language teacher model is prepared, we opt for a vision model to serve as the student model, taking RGB video frames as its input. Similar to the teacher model, we apply linear projection(s) atop the student model. The parameters of the teacher model are then fixed, and we proceed with knowledge distillation, as originally proposed by Hinton et al. [37].

Training Objective. As described above, given a dataset $\mathbb{D} = \{(c_1, w_1, y_1), ..., (c_N, w_N, y_N)\}$, the teacher model takes w_i (action expressed in text) as input and predicts the class probability distribution $\hat{y}_i^t = [y_{i,1}^t, ..., y_{i,C}^t]$. Similarly, the student model takes c_i (RGB video frames) as input and predicts $\hat{y}_i^s = [y_{i,1}^s, ..., y_{i,C}^s]$. We minimise the KL-divergence between $\hat{y}_i^t$ and $\hat{y}_i^s$ as $\mathscr{L}_{KL} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i^t \cdot (log\hat{y}_i^t - log\hat{y}_i^s)$. Following standard practice [37, 8], we use a temperature parameter τ to control the entropy of class probabilities predicted by the teacher $\hat{y}_i^t = \sigma(\hat{y}_i^t/\tau)$ and the student $\hat{y}_i^s = \sigma(\hat{y}_i^s/\tau)$, where σ is the softmax operator. We then scale the KL-divergence loss according to the temperature parameter $\mathscr{L}_{KL} = \mathscr{L}_{KL} \cdot \tau^2$. Additionally, we also minimise the standard cross-entropy objective of class probabilities predicted by the student $\mathscr{L}_{CE} = \frac{1}{N}\sum_{i=1}^{N} y_i \cdot log\sigma(\hat{y}_i^s)$. In the case of compositional actions containing verbs and nouns, the training objective becomes the sum of corresponding loss terms with respect to the verb and the noun, where $\mathscr{L}_{KL} = \frac{1}{2}(\mathscr{L}_{KL}^n + \mathscr{L}_{KL}^v)$ and $\mathscr{L}_{CE} = \frac{1}{2}(\mathscr{L}_{CE}^n + \mathscr{L}_{CE}^v)$.

Furthermore, we apply a multi-task learning approach in LanViKD by adding an extra linear projection layer on top of the student model to generate o_i^s. We take the output from the last hidden layer from the teacher h_i^t, which is the latent representation of the input text given by the original pre-trained language model. We minimise the smooth L1 objective $\mathscr{L}_{L1}$ to regress o_i^s towards h_i^t.

$$\mathscr{L}_{L1} = \begin{cases} 0.5(o_i^s - h_i^t)^2/\beta & \text{if } |o_i^s - h_i^t| < \beta \\ |o_i^s - h_i^t| - 0.5 * \beta & \text{otherwise} \end{cases}$$

Where β determines the threshold for switching between L1 and L2 loss, with a value of 1 used in our experiments. We compute the final loss as $\mathscr{L} = \lambda \cdot \mathscr{L}_{KL} + (1 - \lambda) \cdot \mathscr{L}_{CE} + \mu \cdot \mathscr{L}_{L1}$. We note that the weights sum of $\mathscr{L}_{KL}$ and $\mathscr{L}_{CE}$ is 1 because they are based on the same output linear layer. Instead, we use a separate loss weight for $\mathscr{L}_{L1}$ because it is based on the linear layer of a separate task. During the inference process, it is important to note that the language teacher model is dispensable. The student vision model operates solely on RGB video frames as its input.

4. Experimental Setup

In our experiments, our primary objective is to assess the potential benefits of integrating knowledge from a language model into a vision model for the EAR task. Specifically, we ask the following questions: *(i)* What is the performance of utilising LanViKD on regular EAR data samples, i.e. training and testing samples containing overlapping environments and/or objects. *(ii)* To what extent can a student model, trained using LanViKD, effectively generalise to unseen environments and/or objects not encountered during training? *(iii)* How does the incorporation of a language model's teaching signal alongside the standard one-hot target affect the training of a student model, and what is the optimal balance between the two? *(iiii)* How does using the language modality compare to using the audio modality in cross-modal knowledge distillation with the RGB modality? We choose to compare language with audio because it is unlike optical flow and object bounding box/category which need to be computed using external algorithms or models for RGB data [38, 39]; audio and language are both raw data sources that are readily available in EAR datasets.

To address questions *(i)* and *(ii)*, we conduct experiments across various datasets, encompassing those with overlapping environments and objects for both training and validation, as well as those featuring unseen or under-represented elements during validation. For question (*iii*), we perform experiments with different λ settings, regulating the ratio of the language model's teaching signal to the traditional one-hot target within the training objective. To address question (*iv*), we compare our findings with those of Radevski et al. [8], who conducted similar knowledge distillation from audio modality to RGB video modality.

Datasets. Our experiments are conducted on three datasets: Epic-Kitchens-100 [9], Something-Something V2 [10] and Something-Else [40].

Epic-Kitchens-100 (EK-100) is a large-scale dataset of egocentric videos. It contains 100 hours of non-scripted videos recorded by 37 participants in kitchen environments. The actions depicted in the videos include narrations in the form of English phrases. The training targets are verbs and nouns expressing the actions (e.g. "cutting onion" is an action narration, whose training targets are "cut" and "onion"). There are 300 unique noun classes and 97 unique verb classes. An action is considered to be correctly predicted if both the verb and the noun are correct.

The Something-Something V2 (SSV2) dataset is a large collection of (mostly egocentric) videos that show people performing 174 pre-defined basic actions with everyday objects (e.g. putting something on a surface, moving something up) [10]. Notably, videos in SSV2 initially feature annotations with specific object names, which are then replaced with the word "something" for training targets (e.g., "putting box on a surface" becomes "putting something on a surface").

Something-Else (SthElse) is an alternative data re-split of the original SSV2 [40]. SthElse splits SSV2 in such a way that the training and validation sets contain distinct objects. Therefore, SthElse focusses on using unseen objects during training to measure the generalisation capability of a model.

In a similar vein, we also incorporate the EK-100 Unseen and Tail split. The unseen split is a subset of the EK-100 validation set, which contains videos that are recorded by two participants who did not appear in the training set. The unseen split is specifically designed to measure the ability of models on unseen environments during training. The tail split is a subset containing action classes that have little training samples. Notably, the EK-100 regular split encompasses all samples excluding the unseen split.

Language Backbone. In this study, our language model of choice is MiniLM, featuring 12 layers and a hidden size of 384 [41]. The rationale behind choosing MiniLM stems from its compact architecture and computational efficiency. Despite its smaller size, MiniLM maintains competitive performance over its teacher model, UniLM [42]. For the EK-100 dataset, we utilised the original textual action annotations, consisting of English phrases describing actions, as input to MiniLM. Similarly, for the SthElse dataset, we employed the original annotations, which include object names as inputs to MiniLM.

Vision Backbone. Following Radevski et al. [8], we chose the Swin Transformer Tiny version (Swin-T) model as the vision model in LanViKD. Each video clip is represented as a sequence of RGB frames, where each frame is represented by a $3 \times 224 \times 224$ tensor. Swin-T takes a video clip as input and produces a 768-dimension tensor as the latent representation of the video.

Implementation details. For teacher models, we train the linear head for 10 epochs across all datasets. As for the student models, we train them for 50 epochs on Epic-Kitchens, 40 epochs on SSV2 and 30 epochs on Something-Else. As per Radevski et al. [8], we employ the AdamW optimiser [43], setting the peak learning rate at $1e - 4$. Initially, the learning rate linearly increases for the first 3 epochs and then linear decreases to 0. A weight decay of $5e - 2$ is utilised, along with gradient clipping, limiting the maximum norm to 5. Across all experiments, τ remains fixed at a value of 3. For EK-100, during training, we select a random starting frame and sample 32 frames with a fixed stride of 2. In inference, frames are sampled in the same manner to cover the central section of the video. For SSV2 and SthElse, 16 frames are sampled to cover the entire video during both training and inference. Standard data augmentation techniques are applied to RGB frames, including random cropping, color jitter, and random horizontal flips (exclusive to EK-100). Consistency is maintained within each video clip by applying the same augmentation methods to every frame. A single temporal crop is employed for inference.

Direct Comparison. In the study by Radevski et al. [8], the Swin-T model was trained on the EK-100, SSV2 and SthElse datasets. A key distinction between their approach and ours is that while they incorporated multiple modalities, including RGB, optical flow, and audio, they did not include the language modality. In contrast, our work leverages only the language modality as the teacher modality. To ensure a direct and fair comparison, we adhered to the same experimental settings as Radevski et al. [8], including the use of the backbone model, data augmentation techniques, and frame sampling methods.

Evaluation Metrics. We calculate two widely used metrics, Accuracy@1 (ACC@1) and Accuracy@5 (ACC@5), on the test set, which play pivotal roles in assessing the effectiveness of such systems [44]. By measuring the correctness of predictions within the top-ranked recommendations, both ACC@1 and ACC@5 provide valuable insights into the system's ability to deliver relevant and satisfactory outcomes to users, where ACC@1 quantifies the proportion of correct predictions among the top-1 ranked results. It signifies whether the single highest-ranked item recommended by the system aligns with the user's preference. On the other hand, ACC@5 expands the assessment to the top-5 ranked results, thereby offering a broader evaluation of the system's performance.

Method	Teaching Parameters	EK-100 Regular			SSV2	
		Noun	Verb	Action	ACC@1	ACC@5
Baseline	$\lambda = 0, \mu = 0$	51.5	61.4	37.7	60.3	86.4
LanViKD	$\lambda = 0.4, \mu = 0$	$53.3_{+1.8}$	$\underline{63.0}_{+1.6}$	$\underline{39.7}_{+2.0}$	$58.4_{-1.9}$	$82.7_{-3.7}$
LanViKD	$\lambda = 0.4, \mu = 50$	$\underline{53.6}_{+2.1}$	$62.3_{+0.9}$	$39.6_{+1.9}$	$59.3_{-1.0}$	$83.0_{-3.4}$

Table 1
Performance on Standard Environments and Objects. Baseline model is a Swin-T [28] trained with RGB video frames on one-hot targets. λ is the loss weight for KL-divergence, which signifies the teaching signal received from the language model. μ is the loss weight for Smooth-L1. A value of 0 implies the absence of a regression head, whereas a value of 50 indicates the inclusion of the regression head, which is then scaled up accordingly.

Method	Teaching Parameters	EK-100 Unseen			EK-100 Tail			SthElse	
		Noun	Verb	Action	Noun	Verb	Action	ACC@1	ACC@5
Baseline	$\lambda = 0, \mu = 0$	37.8	50.0	25.9	30.9	38.4	21.4	51.8	79.5
LanViKD	$\lambda = 0.4, \mu = 0$	$39.3_{+1.5}$	$51.5_{+1.5}$	$26.7_{+0.8}$	$30.4_{-0.5}$	$37.8_{-0.6}$	$21.4_{+0.0}$	$\underline{54.4}_{+2.6}$	$79.7_{+0.2}$
LanViKD	$\lambda = 0.4, \mu = 50$	$\underline{39.7}_{+1.9}$	$\underline{51.9}_{+1.9}$	$\underline{27.2}_{+1.3}$	$\underline{30.9}_{+0.0}$	$36.0_{-2.4}$	$\underline{21.9}_{+0.5}$	$54.0_{+2.2}$	$\underline{79.8}_{+0.3}$

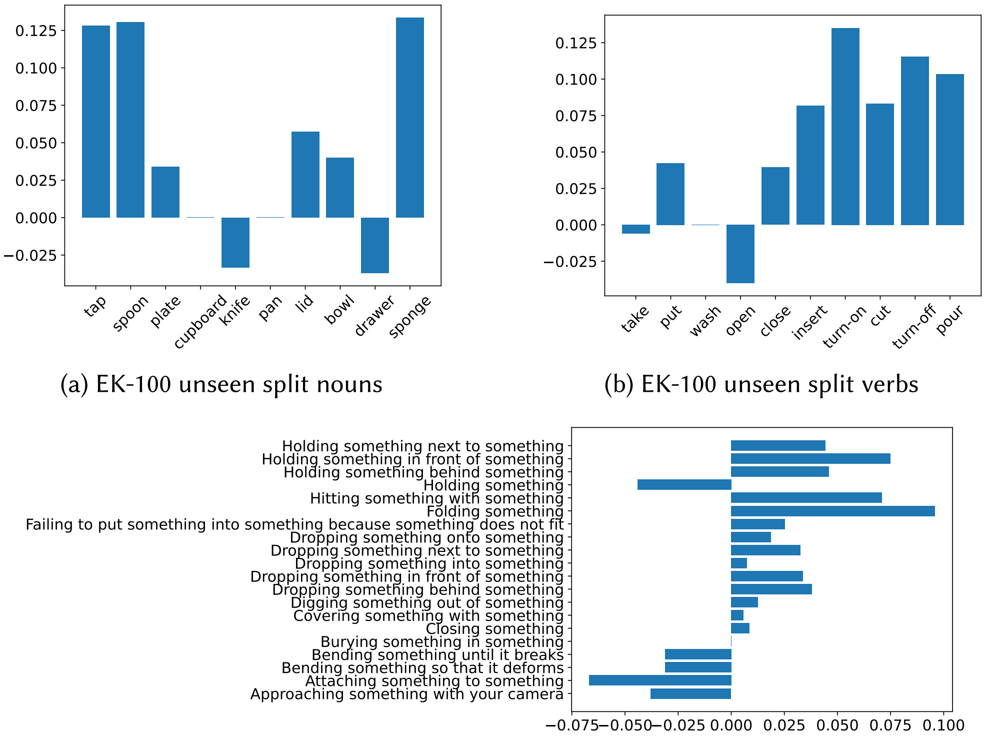
Table 2
Performance on Unrepresented and Unseen Environments & Objects: Baseline is a Swin-T model trained with RGB video frames on one-hot targets, whose results are reported by Radevski et al. [8].

5. Results and Analysis

Across all our experiments, we adopt baseline results derived from Radevski et al. [8], adhering to identical experimental settings. However, for the EK-100 dataset, since they did not include results for the EK-100 tail split, we replicated the baseline experiment to serve as our own baseline.

Performance on regular environments and objects. Table 1 shows the performance metrics obtained from experiments conducted on both the EK-100 regular split and SSV2 dataset. For EK-100, all results are based on ACC@1. For SSV2, we report both ACC@1 and ACC@5 accuracy.

We observe that incorporating knowledge distillation from a language model into a vision model generally enhances the performance of the vision model on the EK-100 regular split by up to 2%, while maintaining competitive results on the SSV2 dataset compared to the baselines. Specifically, in relation to the EK-100 dataset, integrating the regression head for LanViKD demonstrates superior performance in classifying nouns, whereas its removal results in improved classification of verbs. Furthermore, both scenarios show similar improvements in classifying actions, achieving approximately a 2% increase in ACC@1 over the baseline, which serves as the primary metric for EK-100. Conversely, for the SSV2 dataset, LanViKD's performance decreases by 1.9% compared to the baseline without the regression head. Moreover, incorporating the regression head yields performance that is competitive with the baseline.



(a) EK-100 unseen split nouns

(b) EK-100 unseen split verbs

(c) SthElse actions

Figure 3: Per-class ACC@1 improvement over baselines of the 10 most frequent nouns and verbs within the EK-100 unseen split dataset, as well as the 20 most frequent actions in SthElse.

Generalisation capability on unrepresented and unseen environments and unseen objects. Table 2 shows the performance on EK-100 unseen and tail splits, which contain unseen and unrepresented environments during training, respectively. It also shows the performance on SthElse, which contains videos involving objects that are unseen during training. These validation sets aim at evaluating a vision model's generalisation capability.

Our observations indicate that distilling knowledge from a language model into a vision model generally enhances the generalisation capability of the latter by up to 1.3% on the EK-100 unseen split and 2.6% on the SthElse dataset. Specifically, for the EK-100 unseen split, LanViKD outperforms the baseline across all three metrics (Noun, Verb, and Action) without the addition of the regression head. Furthermore, incorporating the regression head leads to an additional 1.3% improvement in performance specifically on the metrics for Action. For the EK-100 tail split, LanViKD demonstrates competitive results with the baseline when the regression head is absent. However, with the regression head, although LanViKD exhibits a slight performance decrease in the Verb metric compared to the baseline, it achieves a 0.5% enhancement in the primary metric, Action. Similarly, for the SthElse dataset, LanViKD surpasses the baseline by 2.6% in ACC@1 without the regression head. However, the addition of the regression head marginally diminishes performance by 0.4% compared to its absence. Moreover, Figure 3 shows per-class ACC@1 improvement in relation to the top 10 frequent nouns and verbs within the

Method	Teaching Parameters	EK-100 Regular			EK-100 Unseen			EK-100 Tail		
		Noun	Verb	Action	Noun	Verb	Action	Noun	Verb	Action
LanViKD	$\lambda = 0.4, \mu = 50$	53.6	62.3	39.6	39.7	51.9	27.2	30.9	36.0	21.9
LanViKD	$\lambda = 0.8, \mu = 50$	52.4	62.6	39.2	39.5	52.9	27.6	26.7	34.9	19.3

Table 3
The impact of varying teaching weights on EK-100. $\lambda = 0.4$ represents the weight assigned to KL-divergence loss. A higher value of λ indicates that the training objective for the student model prioritizes the teaching signals from the language model to a greater extent, while reducing emphasis on the one-hot targets.

Method	Teaching Parameters	SthElse	
		Acc@1	Acc@5
LanViKD	$\lambda = 0.4, \mu = 50$	54.0	79.8
LanViKD	$\lambda = 0.8, \mu = 50$	51.0	76.7

Table 4
Teacher's influence for SthElse. λ represents the weight assigned to KL-divergence loss.

Method	Training	Inference	EK-100 Regular			EK-100 Unseen		
			Noun	Verb	Action	Noun	Verb	Action
Baseline	RGB&Audio	RGB	51.5	62.4	37.9	41.8	51.8	27.5
LanViKD($\lambda = 0.4, \mu = 50$)	RGB&Language	RGB	$53.6_{+2.1}$	$62.3_{-0.1}$	$39.6_{+1.7}$	$39.7_{-2.1}$	$51.9_{+0.1}$	$27.2_{-0.3}$
LanViKD($\lambda = 0.8, \mu = 50$)	RGB&Language	RGB	$52.4_{+0.9}$	$62.6_{+0.2}$	$39.2_{+1.3}$	$39.5_{-2.3}$	$52.9_{+1.1}$	$27.6_{+0.1}$

Table 5
Comparison of utilising audio modality and language modality on EK-100 dataset. The baseline is introduced by Radevski et al. [8], which distils knowledge from audio modality to RGB modality during training. In contrast, our approach distils knowledge from language modality to RGB modality during training. Both approaches use only RGB video frames for inference.

EK-100 unseen split dataset, alongside the top 20 frequent actions identified in SthElse.

Teacher's influence on the student. To investigate the influence of the teacher language model on the student model's performance, we set the parameter λ to 0.4 and 0.8 for the EK-100 and SthElse datasets, respectively. Specifically, this adjustment increases the teaching signal's weight in the training objective from 40% to 80%, while maintaining the regression head.

Tables 3 and 4 present a comparative analysis of the model's performance with λ set at 0.4 and 0.8. The results indicate that increasing λ to 0.8 leads to a slight improvement on the unseen split of the EK-100 dataset. However, this increase is associated with a significant performance decline on the tail split of the EK-100 dataset and across the SthElse dataset.

Comparison with knowledge distillation on audio modality. We are interested in comparing the utilisation of audio modality for knowledge distillation, as opposed to optical flow (OF) and objects' bounding box and category (OBJ) modalities. Unlike OF and OBJ, which are derived from RGB modality through external algorithms or deep learning models [38, 39], audio and text modalities represent raw data from the datasets. This distinction is crucial, as the computation of OF and OBJ may introduce hidden external model knowledge into training, making it uncertain whether all the knowledge distilled into a student is solely from the teacher. In the study by Radevski et al. [8], they trained an audio model on EK-100 audio data alongside an RGB model. Subsequently, they combined these models as a teacher ensemble to train a Swin-T vision student model, which only received RGB video frames. Similarly, in our approach, we leverage knowledge from a language teacher alongside RGB video frames to train a vision student, also receiving only RGB frames; while Radevski et al. [8] utilised audio and RGB modalities for training, we employ language and RGB modalities. Both approaches exclusively use the RGB modality for inference.

It is important to note that the audio modality is exclusive to the EK-100 dataset. Table 5 presents a comparison between knowledge distillation using audio and RGB, and language and RGB modalities. Our findings indicate that training with language and RGB yields superior performance, surpassing training with audio and RGB by up to 1.7% on the EK-100 regular split, while also achieving competitive results on the unseen split.

6. Conclusion and Future Work

In this work, we propose a knowledge distillation framework, LanViKD, for language and vision (RGB) modalities. Our experiments demonstrate enhancement in performance compared to the baseline model, which is solely trained on one-hot labels utilising only the RGB modality. Additionally, we conduct a comparative analysis between the incorporation of audio modality and language modality for knowledge distillation. Our findings indicate the superiority of the language modality as a teacher for enhancing the learning of the vision-based student.

In our future work, we will investigate the integration of the language modality with additional modalities such as audio, depth, and thermography. We plan to find an approach for aligning multiple modalities and create a comprehensive teacher model with broader knowledge for knowledge distillation, potentially leading to further performance improvement.

Acknowledgments

We would like to acknowledge the use of the Computational Shared Facility at The University of Manchester. The computational resource used in this work is supported by the CSF (aka Danzek), which is a High Performance Computing (HPC) cluster at the University of Manchester, managed by IT Services for the use of University academics, post-doctoral assistants and post-graduates to conduct academic research.

References

[1] A. Núñez-Marcos, G. Azkune, I. Arganda-Carreras, Egocentric vision-based action recognition: A survey, Neurocomputing 472 (2022) 175–197. URL: https://doi.org/10.1016/j.neucom.2021.11.081. doi:10.1016/J.NEUCOM.2021.11.081.

[2] R. Girdhar, M. Singh, N. Ravi, L. van der Maaten, A. Joulin, I. Misra, Omnivore: A single model for many visual modalities, in: CVPR, IEEE, 2022, pp. 16081–16091.

[3] X. Xiong, A. Arnab, A. Nagrani, C. Schmid, M&m mix: A multimodal multiview transformer ensemble, CoRR abs/2206.09852 (2022).

[4] R. Herzig, E. Ben-Avraham, K. Mangalam, A. Bar, G. Chechik, A. Rohrbach, T. Darrell, A. Globerson, Object-region video transformers, in: CVPR, IEEE, 2022, pp. 3138–3149.

[5] S. Gupta, J. Hoffman, J. Malik, Cross modal distillation for supervision transfer, in: CVPR, IEEE Computer Society, 2016, pp. 2827–2836.

[6] Y. Aytar, C. Vondrick, A. Torralba, Soundnet: Learning sound representations from unlabeled video, in: NIPS, 2016, pp. 892–900.

[7] Z. Xue, S. Ren, Z. Gao, H. Zhao, Multimodal knowledge expansion, in: ICCV, IEEE, 2021, pp. 834–843.

[8] G. Radevski, D. Grujicic, M. B. Blaschko, M. Moens, T. Tuytelaars, Multimodal distillation for egocentric action recognition, in: ICCV, IEEE, 2023, pp. 5190–5201.

[9] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price, M. Wray, Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100, Int. J. Comput. Vis. 130 (2022) 33–55.

[10] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fründ, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, R. Memisevic, The ”something something” video database for learning and evaluating visual common sense, in: ICCV, IEEE Computer Society, 2017, pp. 5843–5851.

[11] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman, The kinetics human action video dataset, CoRR abs/1705.06950 (2017).

[12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 140:1–140:67.

[13] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT (1), Association for Computational Linguistics, 2019, pp. 4171–4186.

[14] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: ACL, Association for Computational Linguistics, 2020, pp. 7871–7880.

[15] W. Kim, B. Son, I. Kim, Vilt: Vision-and-language transformer without convolution or region supervision, in: ICML, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 5583–5594.

[16] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, C. P. Langlotz, Contrastive learning of medical visual representations from paired images and text, in: MLHC, volume 182 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 2–25.

[17] L. Gomez-Bigorda, Y. Patel, M. Rusiñol, D. Karatzas, C. V. Jawahar, Self-supervised learning of visual features through embedding images into text topic spaces, in: CVPR, IEEE Computer Society, 2017, pp. 2017–2026.

[18] A. Joulin, L. van der Maaten, A. Jabri, N. Vasilache, Learning visual features from large weakly supervised data, in: ECCV (7), volume 9911 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 67–84.

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: ICML, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763.

[20] N. Siddharth, A. Barbu, J. M. Siskind, Seeing what you're told: Sentence-guided activity recognition in video, in: CVPR, IEEE Computer Society, 2014, pp. 732–739.

[21] C. Sun, A. Myers, C. Vondrick, K. Murphy, C. Schmid, Videobert: A joint model for video and language representation learning, in: ICCV, IEEE, 2019, pp. 7463–7472.

[22] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: NeurIPS, 2020.

[23] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: Language models can teach themselves to use tools, in: NeurIPS, 2023.

[24] K. Desai, J. Johnson, Virtex: Learning visual representations from textual annotations, in: CVPR, Computer Vision Foundation / IEEE, 2021, pp. 11162–11173.

[25] N. C. Garcia, S. A. Bargal, V. Ablavsky, P. Morerio, V. Murino, S. Sclaroff, DMCL: distillation multiple choice learning for multimodal action recognition, CoRR abs/1912.10982 (2019).

[26] N. C. Garcia, P. Morerio, V. Murino, Modality distillation with multiple stream networks for action recognition, in: ECCV (8), volume 11212 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 106–121.

[27] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, C. Schmid, Vivit: A video vision transformer, in: ICCV, IEEE, 2021, pp. 6816–6826.

[28] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, H. Hu, Video swin transformer, in: CVPR, IEEE, 2022, pp. 3192–3201.

[29] D. Lee, J. Lee, J. Choi, CAST: cross-attention in space and time for video action recognition, in: NeurIPS, 2023.

[30] R. Yan, L. Xie, X. Shu, J. Tang, Interactive fusion of multi-level features for compositional activity recognition, CoRR abs/2012.05689 (2020).

[31] R. Caruana, Multitask learning, Mach. Learn. 28 (1997) 41–75.

[32] G. Ghiasi, B. Zoph, E. D. Cubuk, Q. V. Le, T. Lin, Multi-task self-training for learning general representations, in: ICCV, IEEE, 2021, pp. 8836–8845.

[33] K. Maninis, I. Radosavovic, I. Kokkinos, Attentive single-tasking of multiple tasks, in: CVPR, Computer Vision Foundation / IEEE, 2019, pp. 1851–1860.

[34] I. Misra, A. Shrivastava, A. Gupta, M. Hebert, Cross-stitch networks for multi-task learning, in: CVPR, IEEE Computer Society, 2016, pp. 3994–4003.

[35] F. Sener, D. Chatterjee, A. Yao, Technical report: Temporal aggregate representations, CoRR abs/2106.03152 (2021).

[36] D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M. Tan, M. Brown, B. Gong, Movinets: Mobile video networks for efficient video recognition, in: CVPR, Computer Vision Foundation / IEEE, 2021, pp. 16020–16030.

[37] G. E. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, CoRR

abs/1503.02531 (2015).

[38] B. D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: IJCAI, William Kaufmann, 1981, pp. 674–679.

[39] J. Redmon, S. K. Divvala, R. B. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: CVPR, IEEE Computer Society, 2016, pp. 779–788.

[40] J. Materzynska, T. Xiao, R. Herzig, H. Xu, X. Wang, T. Darrell, Something-else: Compositional action recognition with spatial-temporal interaction networks, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020). doi:10.1109/cvpr42600.2020.00113.

[41] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, in: NeurIPS, 2020.

[42] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, H. Hon, Unified language model pre-training for natural language understanding and generation, in: NeurIPS, 2019, pp. 13042–13054.

[43] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: ICLR (Poster), OpenReview.net, 2019.

[44] J. L. Favero, D. R. Ilgen, The effects of ratee prototypicality on rater observation and accuracy 1, Journal of Applied Social Psychology 19 (1989) 932–946.