Text-Derived Knowledge Helps Vision: A Simple Cross-modal Distillation for Video-based Action Anticipation

Sayontan Ghosh[1] **Tanvi Aggarwal**[1]

Minh Hoai[1] **Niranjan Balasubramanian**[1]

[1]**Stony Brook University**

{sagghosh, taggarwal, minhhoai, niranjan }@cs.stonybrook.edu

Abstract

Anticipating future actions in a video is useful for many autonomous and assistive technologies. Most prior action anticipation work treat this as a vision modality problem, where the models learn the task information primarily from the video features in the action anticipation datasets. However, knowledge about action sequences can also be obtained from external textual data. In this work, we show how knowledge in pretrained language models can be adapted and distilled into vision-based action anticipation models. We show that a simple distillation technique can achieve effective knowledge transfer and provide consistent gains on a strong vision model (Anticipative Vision Transformer) for two action anticipation datasets (3.5% relative gain on EGTEA-GAZE+ and 7.2% relative gain on EPIC-KITCHEN 55), giving a new state-of-the-art result[1].

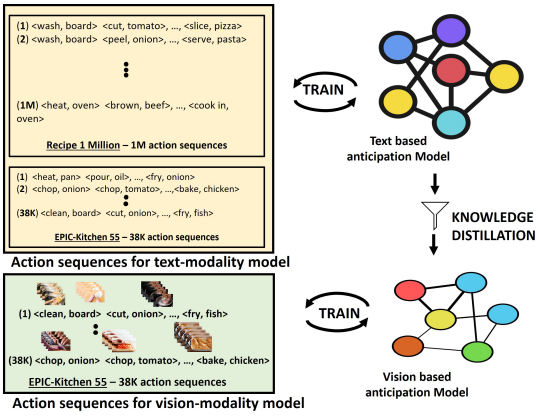


Figure 1: A model learning the action anticipation from only the vision modality (video frames) is essentially exposed to a very limited set of action sequences. Language models, which are pre-trained on large-scale text, can learn this distribution from the task, and a much larger domain-relevant text. We propose distilling this knowledge from text modality models to vision modality model for video action anticipation task.

1 Introduction

Anticipating future actions in the video of an unfolding scenario is an important capability for many applications in augmented reality (Salamin et al., 2006; Azuma, 2004), robotics (Duarte et al., 2018; Schydlo et al., 2018), and autonomous driving (Chaabane et al., 2020; Suzuki et al., 2018). Anticipating what actions will likely happen in a scenario, requires one to both recognize what has happened so far, and use anticipative general knowledge about how action sequences tend to play out. Most models for this task use a pre-trained video encoder to extract information about what has happened so far in the scenario, and use a text-based decoder to predict what action is likely to happen in the future (Carion et al., 2020; Dessalene et al., 2021; Liu et al., 2020; Sener et al., 2020).

However, when trained on the target video datasets, the generalization of the models depends on how well these video datasets cover the space of action sequence distributions. In other words, the knowledge that is learnt for predicting future actions is, in effect, limited to the information in the target video datasets, where obtaining large scale coverage of action sequences is difficult.

Knowledge about action sequences can also be obtained from text resources at scale. Language models, (e.g. BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b)), are typically pre-trained on large collections of unlabeled texts with billions of tokens, where they acquire a wide-variety of knowledge including large scale knowledge about common action sequences. For example, Table 1 illustrates how the pre-trained BERT is able to predict the next action in a sequence of actions extracted from a recipe video in terms of its verb and the object. Also, it is easier to collect a much larger collection of action sequences from text sources compared to video annotated with

[1]The models and code used are available at:https://github.com/StonyBrookNLP/action-anticipation-lmtovideo

Findings of the Association for Computational Linguistics: EACL 2023, pages 1882–1897
May 2-6, 2023 ©2023 Association for Computational Linguistics

Masked action sequence	BERT @top5
Clean the board → takeout pan → wash the onion → clean the fish → cut the onion → heat the pan → pour oil in pan → **[MASK]** the fish.	fry, cook, boil, wash, clean
Clean the board → takeout pan → wash the onion → clean the fish → cut the onion → heat the pan → pour oil in pan → fry **[MASK]**.	pan, fish, chicken, it, onion

Table 1: Given a sequence of actions extracted from a video, **BERT@top5** shows the top5 prediction made by a standard pre-trained BERT for the masked verb and object of the next action.

segments. As illustrated in Figure 1, EPIC55, a video dataset of about 800GB only has about $38K$ action sequences, whereas there are around $1M$ sequences in the text recipes dataset Recipe1M. Text modality models can thus be exposed to a much larger variety of action sequences compared to video-modality anticipation models. However, because the task is defined only over the video inputs there is a question of how one can transfer this knowledge.

In this work, we show that we can augment video-based anticipation models with this external text-derived knowledge. To this end, we propose a simple cross-modal distillation approach, where we distill the knowledge gained by a language model from the text modality of the data into a vision-modality model. We build a teacher using a pre-trained language model which already carries general knowledge about action sequences. We adapt this teacher to the action sequences in the video domain by fine-tuning them for the action anticipation task. Then, we train a vision-modality [2] student, which is now tasked with both predicting the target action label as well as matching the output probability distribution of the teacher.

There are two aspects of language models that can be adjusted further for improved distillation. First, while they may contain knowledge about a broad range of action sequences, we can focus them towards specific action sequences in the target dataset. Second, the text modality teacher can be further improved by pretraining on domain-relevant texts (e.g. cooking recipes), to further adapt it to the action sequences in the task domain.

[2]The task requires the anticipation model to make inference based on the vision modality (video frames) of the video

Our empirical evaluation shows that this cross-modal training yields consistent improvements over a state-of-the-art Anticipative Vision Transformer model (Girdhar and Grauman, 2021) on two egocentric action anticipation datasets in the cooking domain. Adapting the teacher to the task domain by pretraining on domain relevant texts yields further gains and the gains are stable for different language models. Interestingly, our analysis shows that the language model based teacher can provide gains even when it is not necessarily better than the vision student, suggesting that distillation benefits can also come from the complementary of knowledge, as in the case of the text modality.

In summary we make the following contributions: (i) We show that a simple distillation scheme can effectively transfer text-derived knowledge about action sequences (i.e. knowledge external to the video datasets) to a vision-based action anticipation model. (ii) We show that text-derived knowledge about actions sequences contain complementary information that is useful for the anticipation task, especially for the case where the action label space is large. (iii) Using a strong action anticipation model as a student, we achieve new state-of-the-art results on two benchmark datasets.

2 Related Work

There has been a wide range of solutions for action anticipation ranging from hierarchical representaions (Lan et al., 2014), unsupervised representation learning (Vondrick et al., 2016), to encoder-decoder frameworks that decode future actions at different time scales (Furnari and Farinella, 2019), and transformers trained on multiple auxiliary tasks (Girdhar and Grauman, 2021). However, these only use the vision modality features of the observed video to train the model for the anticipation task. Our work aims to distill text-derived knowledge to improve action anticipation. Here we relate our work to others that have made use of (i) textual knowledge for related tasks, (ii) general knowledge distillation, and (iii) multimodal models which also allow for integration of information from different modalities.

Textual Knowledge for Action Anticipation: Other works have also shown the utility of modeling text-modality. Sener and Yao (2019) transfer knowledge in a text-to-text encoder-decoder to a video-to-text encoder-decoder, by substituting the text encoder with the video encoder. However, this

relies on projecting the image and text features in a shared space, which requires lots of properly aligned text and its corresponding image. Camporese et al. (2021) model label semantics with a hand engineered deterministic label prior based on the global co-occurrence statistics of the action labels from the overall training data, which can be ineffective in case the underlying joint action distribution is complex. In contrast, our work proposes a different approach to leverage the text in the training data by using language models to learn the complex underlying distribution of action sequences in the video and then distill this knowledge into a vision model to improve their performance.

Cross-modal Knowledge Distillation: Thoker and Gall (2019) propose learning from RGB videos to recognize actions for another modality. Others have used cross-modal distillation for video retrieval tasks (Hu et al., 2020; Chen et al., 2020) and for text-to-speech (Wang et al., 2020). Most relevant to ours is a recent system that improves language understanding of text models by transferring the knowledge of a multi-modal teacher trained on a video-text dataset, into a student language model with a text dataset (Tang et al., 2021) . In contrast, our proposed method for action anticipation transfers knowledge gained by a text-based teacher model into a vision-based student model.

Mutlimodal Models: Due to the recent prevalence of multimodal data and applications (Lin et al., 2014; Sharma et al., 2018; Antol et al., 2015; Krishna et al., 2017; Ordonez et al., 2011; Abu Farha et al., 2018; Talmor et al., 2021; Afouras et al., 2018), there has been plethora of recent work on multimodal transformers. One commonly used approach used to train these models is to learn a cross-modal representation in a shared space. Examples include learning to align text-image pairs for cross-modal retrieval (Radford et al., 2021; Wehrmann et al., 2020), grounded image representations (Liu et al., 2019a), and grounded text representations (Tan and Bansal, 2020; Li et al., 2019). Hu and Singh (2021) extend the idea for multi-task settings with multiple language-vision based tasks. Tsimpoukelli et al. (2021) adapt a vision model to a frozen large LM to transfer its few-shot capability to a multimodal setting (vision and language). However these methods rely on large-scale image-text aligned datasets for the training the model, which may not always be available, for e.g. EGTEA-GAZE+ video dataset has only 10.3K labelled action sequences. In contrast our distillation approach does not require any image-text alignment for the anticipation task.

3 Language-to-vision knowledge distillation for action anticipation

The action anticipation task asks to predict the class label of a future action based on information from an observed video sequence. In this task setting, the model has access to both, video and annotated action segments (action text) during the train time, but needs to make the inference only using the video sequence. The input to the prediction model is a sequence of video frames up until time step t: $\mathbf{X} = (X_1, X_2, \ldots, X_t)$, and the desired output of the model is the class label Y of the action at time $t + \tau$, where τ is the anticipation time.

To learn an anticipation model, we assume there is training data of the following form: $\mathcal{D} = \{(\mathbf{X}^i, \mathbf{L}^i, Y^i)\}_{i=1}^n$, where $\mathbf{X}^i = (X_1^i, \ldots, X_{t^i}^i)$ is the ith training video sequence, Y^i is the class label of the future action at time $t^i + \tau$, and $\mathbf{L}^i = (L_1^i, \ldots, L_{k^i}^i)$ is the sequence of action label of the action segments in the video sequence $\mathbf{X}^i$. Each human action can span multiple time steps, so so the number of actions k^i might be different from the number of video frames t^i.

Our task is to learn a model g that can predict the future action label based on the vison modality of the video sequence $\mathbf{X}^i$ only. A common approach is to optimize cross entropy loss $\mathcal{L}$ between the model's predicted label $g(\mathbf{X}^i)$ and the ground truth label Y^i of each training instances, i.e., to minimize: $\sum_i \mathcal{L}(g(\mathbf{X}^i), Y^i)$. Although the sequence of action labels $\mathbf{L}^i$ is available in the training data, the semantics associated with these labels is not properly used by the existing methods for training the anticipation model.

Here we propose to learn a text-based anticipation model g_{text} and use it to supervise the training of the vision-based anticipation model g. This training approach utilizes the knowledge from the text domain, which is easier to learn than the vision-based knowledge, given the abundance of event sequences described in text corpora. Hereafter, we will refer to the language-based model as the teacher, and the vision-based model as the student.

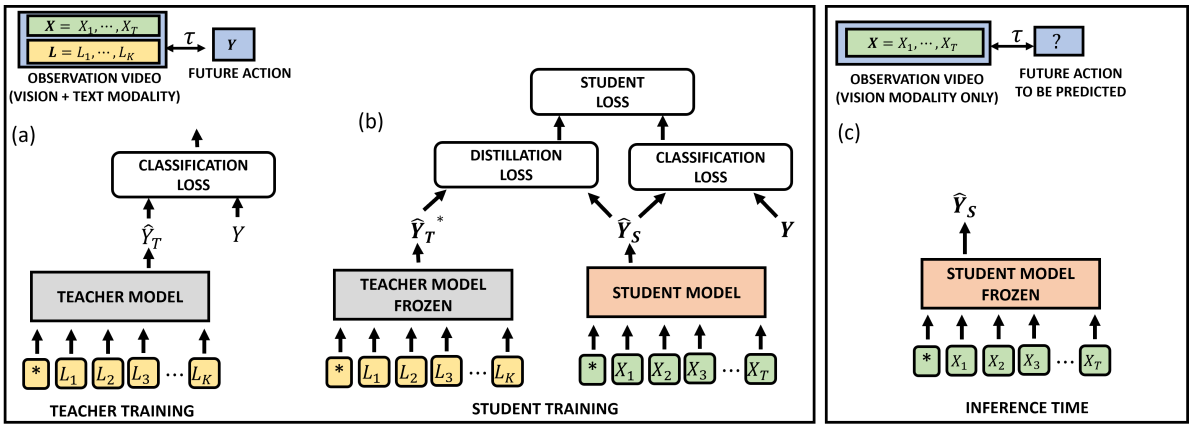


Figure 2: **METHOD OVERVIEW: Training**- The observation video has two set of features, a sequence of T image frames $\mathbf{X}$, and a sequence action labels (`e.g. cut-onion, peel-onion etc.`) $\mathbf{L}$ corresponding to the K action segments in $\mathbf{X}$. $(\mathbf{a})$ We train a teacher model to predict Y using the text features $\mathbf{L}$. Then we distill the knowledge gained by the teacher on text features into the student model that operates on vision modality $\mathbf{X}$. For this, $(\mathbf{b})$, we train a student model on the vision modality feature $\mathbf{X}$ while using the corresponding prediction from the teacher model as a label prior. **Inference**- During the inference or test time, the trained student model is used to predict the future action using only the vision modality of the observed video.

3.1 Overview

The overview of our proposed method is shown in Figure 2. We augment vision-based anticipation models (students) with knowledge distilled from text-based models (teachers) that have access to knowledge from large scale action sequences. To this end we fine-tune a pre-trained language model on the action sequences in the training data. However, unlike the student, the teacher gets to see the action labels of the input video segment to make its predictions (Figure 2a). Then, we train a vision-based student that learns from the text-based teacher (Figure 2b).

The teacher in our setting is built using a pre-trained language model g_{txt} that has access to broad knowledge about action sequences. We fine-tune it on the target dataset as follows. For each instance, the teacher is given the textual action sequence $\mathbf{L}^i$ in the input video as the input (or conditioning context), which then predicts the anticipated future action $\hat{Y^i}_{txt}$. The teacher is trained to minimize the loss defined over the predicted and true labels, i.e., to minimize: $\sum_i \mathcal{L}_{txt}(g_{txt}(\mathbf{L}^i), Y^i)$, where $\mathcal{L}$ denotes the cross-entropy loss and $\hat{Y^i}_{txt} = g_{txt}(\mathbf{L}^i)$ is the output of the text-based teacher model.

We then freeze the *teacher*, and train a vision-based *student* model g that predicts the future action using the vision features $\mathbf{X}^i$. The student is trained to minimize the loss $\mathcal{L}_S(Y^i, \hat{Y}^i, \hat{Y}^i_{txt})$ such that it's output probability distribution $\hat{Y}^i = g(\mathbf{X}^i)$ matches that of the teacher's output $\hat{Y^i}_{txt}$, in addition to matching the true label Y^i.

3.2 Teacher

The input to the teacher is a sequence of action phrases $\mathbf{L} = (L_1, \cdots, L_k)$ that denotes the sequence of actions observed in the input video segment. The teacher first uses a standard language model ϕ_{LM} to produce a vector f_{txt}, of the input sequence $\mathbf{L}$. In transformer-based language models, a special token (e.g. `[CLS]` in `BERT`) is prepended to the input sequence. The output contextual representation of this special token is used as the final representation of the entire input sequence.

The teacher uses this f_{txt} vector to predict the output labels using the standard linear transformation $(\mathbf{W}, \mathrm{b})$ followed by a softmax layer. In addition we also train the teacher to predict the main verb Y_{vb} and the object Y_{ob} of the action Y. These are predicted using separate linear transformations $(\mathbf{W_v}, \mathbf{b_v})$ and $(\mathbf{W_o}, \mathbf{b_o})$, followed by softmax.

The full set of predictions for input $\mathbf{L} = (L_1, \cdots, L_k)$ is obtained as:

$$
\begin{aligned}
f_{txt} &= \phi_{LM}(L_1, ..., L_k) \\
\hat{Y}_{txt} &= \mathbf{softmax}(\mathbf{W} f_{txt} + \mathbf{b}) \\
\hat{Y}_{ob} &= \mathbf{softmax}(\mathbf{W_o} f_{txt} + \mathbf{b_o}) \\
\hat{Y}_{vb} &= \mathbf{softmax}(\mathbf{W_v} f_{txt} + \mathbf{b_v})
\end{aligned}
$$

To fine-tune the teacher model, we minimize the weighted sum of the cross-entropy loss between

the predicted triplet of action, verb and noun and their corresponding ground truth values.

$$\mathcal{L}_{txt}(Y, (\hat{Y}_{txt}, \hat{Y}_{ob}, \hat{Y}_{vb})) = \lambda \mathcal{L}(Y, \hat{Y}_{txt}) + \lambda_o \mathcal{L}(Y_{ob}, \hat{Y}_{ob}) + \lambda_v \mathcal{L}(Y_{vb}, \hat{Y}_{vb}) \quad (1)$$

3.2.1 Adapting Teacher using Domain Relevant Texts

Pre-trained LMs have been shown to contain a wide variety of knowledge, which we hope to distill into the vision student model. However, there are two aspects about LMs which limit their applicability. First, LMs are general purpose models that cover many domains, but the target video datasets cover specific domains. For example, many action anticipation datasets are built for the cooking domain. Second, unlike fluent texts that LM's are trained on, the action sequences in the videos are annotated using simpler verb/object constructions. Adapting LMs to these differences can benefit the knowledge distillation. To this end, we make use of domain-relevant texts (for e.g. the recipes in `Recipe1M` (Marin et al., 2019) dataset). The recipes are natural language instructions. To mimic the target sequences in the video datasets, we convert these into simpler verb-object constructs, and then use the standard Masked Language Modeling training task. Thus, this allows us to not only make use of generic knowledge about action sequences but also adapt the text-derived knowledge to the target domain.

3.3 Student

The student is trained to take the video frames in the video segment $\mathbf{X} = (X_1, ..., X_t)$ as input and predict the future action Y as output. Though the applicability of the proposed distillation method is not restricted to any particular class of student model, we use the recent state-of-the-art Anticipative Vision Transformer (`AVT`) (Girdhar and Grauman, 2021) as our student model. In `AVT`, the video to action prediction is done in two stages, first a *backbone* network $\mathcal{B}$ generates the feature representation of the individual frames in $\mathbf{X}$ in a non-contextual manner.

$$z_1, \ldots, z_t = \mathcal{B}(X_1), \ldots, \mathcal{B}(X_t)$$

This is then followed by a transformer based decoder *head* network $\mathcal{D}$, that generates the contextual representation of the frames by transforming the frame features z_i's in an autoregressive manner.

$$f_{v_1}, ..., f_{v_t} = \mathcal{D}(z_1, ..., z_t)$$
$$\hat{Y}_{v_j} = \mathbf{softmax}(\mathbf{W}_s f_{v_j} + \mathbf{b}_s) \quad \forall j \in \{1, \cdots, t\}$$
$$\hat{Y} = \hat{Y}_{v_t}$$

The feature representations from the *head* network f_{v_j}'s are then used to make predictions for the anticipated action $\hat{Y}_{v_j}$ at time unit j. The anticipated action $\hat{Y}$ for the input video $\mathbf{X}$ is simply the predicted label at time unit t i.e. $\hat{Y}_{v_t}$. During training the model is also supervised for two other auxiliary tasks namely future feature prediction and intermediate action prediction (see (Girdhar and Grauman, 2021) for details). We denote this combined training loss function as $\mathcal{L}_{AVT}$.

For the teacher to student distillation, we want the AVT's output distribution over action classes $\hat{Y}$ to match the teacher's distribution $\hat{Y}_{txt}$. To this end, we minimize the KL divergence between the teacher prediction $\hat{Y}^{\gamma}_{txt}$ and student predictions $\hat{Y^{\gamma}}$, after smoothing the distributions using a temperature parameter γ, following the standard distillation technique (Hinton et al., 2015).

$$\mathcal{L}_S = \mathcal{L}_{AVT} + \lambda_S \cdot \mathcal{D}_{KL}(\hat{Y}^{\gamma}_{txt} \parallel \hat{Y^{\gamma}}) \quad (2)$$

Dataset	Segments	Classes	τ
`Epic 55`	28.6K + 9 K	2,513	1.0 sec
`EGTEA-Gaze+`	7.3K + 3K	106	0.5 sec

Table 2: **Datasets** on which the proposed method is benchmarked. **Segments** are the number of action segments in the train + test set, **Classes** are the number of action classes in the dataset, τ is the anticipation time.

4 Experimental Setup

4.1 Datasets

1. Anticipation Datasets We evaluate the proposed method on two different datasets that are summarised in Table. 2. Both the datasets, `Epic-Kitchen 55` (Damen et al., 2018) and `EGTEA-GAZE+` (Li et al., 2018), are egocentric (first-person) videos of people cooking some recipe. Note the proposed method is broadly applicable to other types of dataset as long as the input video segments in the training set contain action sequence annotations. For the `Epic-Kitchen 55` dataset, we use the standard train-test split followed in Furnari and Farinella (2019). For the `EGTEA-GAZE+` dataset,

we report performance on the first of the three train-test splits following previous work by Girdhar and Grauman (2021).

2. Domain-Relevant Dataset The teacher can be improved further by adapting its language model (LM) to domain relevant texts. To test the effectiveness of this, we use the `Recipe1M` dataset (Marin et al., 2019) to pre-train the LM. The `Recipe1M` dataset contains one million recipes along with associated images (which are not used in this work). The instructions in a recipe can be seen as a sequence of cooking actions to be performed.

4.2 Performance Metrics

For the `EGTEA-Gaze+`, we report the performance on `top1 accuracy` (**`Acc@1`**) and `class mean recall` (**`Rec@1`**)-mean recall of the individual classes, as reported by Girdhar and Grauman (2021). For the `Epic-Kitchen 55` dataset, there are a set of action classes that occur only in the train set but not in the test set and vice versa. Existing anticipation methods, including our proposed work does not support zero-shot learning. Therefore `top5 many-shot class-mean recall` (**`MS-Rec@5`**)-mean top5 recall of the classes in the many-shot-classes, as mentioned in Furnari et al. (2018), is our primary metric for model evaluation.

4.3 Implementation Details

1. Teacher Training: The teacher model is a classification layer on top a pre-trained language model. For the main set of experiments we used **`AlBERT`** (Lan et al., 2019) as the base language model. Our choice here is motivated by two main factors: (i) the pre-training task for **`AlBERT`** focuses on modeling the inter-sentence coherence which is important when modeling the sequence of disparate action phrases (ii) it enables faster training of deeper models. For the `EGTEA-GAZE+` dataset, we trained the model for 4 epochs by minimizing the weighted cross-entropy loss (inversely weighted by the relative class frequency) due to the high degree of class imbalance in the dataset ($\sim 1:24$). For the `EPIC-Kitchen-55` dataset, the model was trained for 8 epochs using regular cross-entropy loss instead of weighted cross-entropy as a lot of classes in the test label set are not present in the train-set, and vice versa.

The classification head is a single linear layer ($\mathbf{W} \cdot x + \mathbf{b}$) that projects the feature representation of the input action sequence into the label space of the target dataset. For optimizing on both the datasets, we used the AdamW ((Loshchilov and Hutter, 2017)) optimizer, with a learning rate of 10^{-5} and weight decay of 10^{-7}. The context window for the Epic-Kitchen was set to 5 action segments whereas for the `EGTEA-GAZE+` it was set to 15 action segments. The teacher training was performed on two Nvidia RTX Titan-X GPUs. The teacher training for the `EGTEA-GAZE+` takes about 2-4 hours depending on the LM base whereas the `EPIC-Kitchen-55` takes about 3-5 hours to train.

2. Teacher Pre-training: We first parse each instruction in the `Recipe1M` dataset into a sequence event tuples of the form (subject, verb, object) using an open information extraction system (Stanovsky et al., 2018) made available by AllenNLP (Gardner et al., 2018). To match the action label structure we see in the video datasets, we represent each instruction using the sequence of action, i.e. <verb, object> part of the event. The actions in the action sequence are sorted by the discourse order of their corresponding verb in the instruction. The language model is pretrained on these (verb, object) sequences using the standard masked language modeling objective (Devlin et al., 2019), where some token in the sequence is masked at random and the model is tasked with predicting the masked token.

For pre-training, the language models were trained on the `Recipe1M` dataset for 200K steps with a batch size of 16. The optimizer used was AdamW (Loshchilov and Hutter, 2017), with a learning rate of 10^{-5} and weight decay of 10^{-7}. LM pre-training was performed on a single Nvidia A100 GPU with the training time varying from 12 hrs for the smallest model (**`DistilBERT`**) to 24 hrs for **`BERT, RoBERTa`**, and **`AlBERT`**.

3. Student Training: For the student training, all the hyperparameters and initial conditions (parameter initialization) are exactly identical to the ones used to train the **`AVT`** (Girdhar and Grauman, 2021) baseline model. So any change in the performance from the baseline is the result of adding the knowledge distillation. The distillation loss coefficient λ_S, for the `EGTEA-GAZE+` dataset was set to 150, and 20 for `EPIC-Kitchen 55`.

4. Top-K logit distillation: The label space of `EPIC-Kitchen 55` has $2,513$ classes, out of which only 31% of the classes in the training data are present in the test data. This leads to the teacher

model assign relatively low probability values to many classes, which may not be reliable signals for distillation. Therefore, instead of matching the probability distribution over all the action classes, we only match the relative probability distribution of the top-50 classes with the highest teacher probabilities. For this, we consider the classes corresponding to the top-50 logits from the teacher prediction, normalize them, and only minimized the KL-Divergence between them and their corresponding logits of the student prediction. The student training was performed on either Nvidia Tesla V100 GPU and the training time was ~ 24 hrs for EGTEA-GAZE+ and ~ 6 hrs for the EPIC-55 dataset.

5 Results and Analysis

We present the results of text to video knowledge distillation on the **AVT** (Girdhar and Grauman, 2021) model as the student. AVT is the state-of-the-art model for action anticipation on the EGTEA-GAZE+ and EPIC-Kitchen 55 datasets on all performance metrics.

For each of these datasets, we consider the AVT variants with the best performance as our baseline and student model. For the EGTEA-GAZE+ dataset, we consider **AVT-h + AVT-b** in (Girdhar and Grauman, 2021) as our baseline model. Similarly for the EPIC-Kitchen 55 dataset, we consider, **AVT-h + irCSN152** in (Girdhar and Grauman, 2021) as our baseline model. [3] Throughout this section, we refer to **AVT-h + AVT-b** and **AVT-h + irCSN152** as **AVT-1** and **AVT-2** respectively. The baseline models distilled with LM based teacher is denoted as **AVT-1(or 2) + LM Distillation** and in case teacher LM is pre-trained on the recipe domain text, the resulting model is referred to as **AVT-1(or 2) + RcpLM Distillation**. We tried to reproduce the **AVT** model to use as our student and obtain stronger results than the published version (see Table 3), on all but one metric. We use this stronger implementation as our baseline and our student model.

5.1 Does Knowledge distillation from Language Models help ?

Table 3 shows the result of training the state-of-the-art baseline model **AVT**, with and without the text to vision knowledge distillation, for the EGTEA-GAZE+ and EPIC-Kitchen 55 dataset. We can observe that applying text to vision knowledge distillation to the **AVT** leads to performance gains on both the datasets. For EGTEA-GAZE+, adding knowledge distillation leads to 2.1% and 2% relative percentage improvement over **AVT-1** on the **Acc@1** and **Rec@1** metrics respectively. For the EPIC-Kitchen 55 dataset, knowledge distillation leads to a relative performance gain of 3.5% over **AVT-2** on **MS-Rec@5** metric.

5.2 Does domain-adaptive pre-training of teacher improves the task performance ?

To analyze the effect domain adaptive pre-training on the task, we pre-train the teacher LM on the Recipe1M dataset through the MLM task. The pre-trained model was then finetuned on the task-specific video dataset for the anticipation task. As seen in Table 3, the performance gain of the teacher directly translates to the performance gain of the student. For EGTEA-GAZE+ dataset, pre-training teacher leads to 3.9% and 3.4% relative improvement over the **AVT-1** on **Acc@1** and **Rec@1** metric compared to 2.1% and 2% relative improvement when not pretraining the teacher. For the EPIC-Kitchen 55 dataset, teacher pre-training leads to a relative improvement of 7.2% on **MS-Rec@5** metric compared to only 3.5% when not pre-training the teacher.

Model	EGTEA-GAZE+		EPIC-55
	Acc@1	Rec@1	MS-Rec@5
AVT (Girdhar and Grauman, 2021) **-published**	43.0	35.5	13.6
AVT (Girdhar and Grauman, 2021) **-reproduced**	43.52	34.87	15.25
AVT + LM Distillation	44.41	35.54	15.79
AVT + RcpLM Distillation	**45.2**	**36.1**	**16.36**

Table 3: **Effect of knowledge distillation:** Distilling knowledge from teacher (**ALBERT** LM) trained on text-modality of the video data, into vision based student model leads to student performance gain. Pre-training the teacher on domain relevant text before task specific finetuning leads to further performance improvement.♣

5.3 How sensitive is the distillation to the choice of Language Model ?

In order to analyze the sensitivity of the distillation scheme towards the choice of the language model,

[3] ♣**AVT** variants used for the EGTEA-GAZE+ and EPIC-Kitchen 55 baselines are **AVT-1** and **AVT-2**.

Model	EGTEA-GAZE+		EPIC-55
	Acc@1	Rec@1	MS-Rec@5
AVT (Girdhar and Grauman, 2021)	43.52	34.87	15.25
+ Rcp-ALBERT Distillation	45.2	36.1	**16.36**
+ Rcp-BERT Distillation	44.81	35.57	15.98
+ Rcp-RoBERTa Distillation	**45.5**	**36.53**	15.97
+ Rcp-ELECTRA Distillation	45.2	35.58	15.34
+ Rcp-DistillBERT Distillation	44.86	35.64	16.23

Table 4: **Effect of the choice of teacher LM** on the distillation performance. Each of the pre-trained LMs that we tested as a teacher, showed performance gain over the baseline **AVT** model for both the datasets.♣

Model	EGTEA-GAZE+		EPIC-55
	Acc@1	Rec@1	MS-Rec@5
AVT (Girdhar and Grauman, 2021)	**43.52**	**34.87**	15.25
Rcp-ALBERT Teacher	21.66	22.63	21.78
Rcp-BERT Teacher	22.05	23.39	21.43
Rcp-RoBERTa Teacher	19.98	21.58	**22.41**
Rcp-ELECTRA Teacher	21.46	23.71	15.19
Rcp-DistillBERT Teacher	21.86	22.58	21.56

Table 5: **Teacher performance** on the anticipation task. For the EGTEA-GAZE+ dataset, the teacher performance is much lower than the video only **AVT** model, where as for the EPIC-Kitchen 55 dataset, the teacher performance is much better than the video-only **AVT** model.♣

we also trained multiple teachers with different pre-trained LMs. The result of using different teachers for the anticipation task is specified in Table 4. From the table, we can observe that all the teacher distilled models perform better than the baseline **AVT** on all the metrics for both the datasets. This indicates that the text modality has some information that complementary to the vision modality that if properly exploited can lead to improved performance for the anticipation task.

5.4 Should the teacher be always better than the student ?

To understand the impact of the quality of the teachers, we measured the performance of the teacher models by themselves on the anticipation task as show in Table 5. For the EPIC-Kitchen 55 dataset the teacher performance is much better than the video-only baseline, whereas, for the EGTEA-GAZE+ dataset, the baseline vision model's performance is much better than any of the teachers. Despite this, the performance gain due to distillation is greater for the EGTEA-GAZE+ dataset compared to the EPIC-Kitchen 55 dataset, as seen in Table 3. This suggests that what matters more for distillation in this case is the complementary information gained from the text modality that is not already present in the vision modality.

6 Conclusions

Action anticipation is a challenging problem that requires training large capacity video models. In this work, we showed how the textual modality of the input videos, which is often ignored in training, can be leveraged to improve the performance of the video models. In particular, we can exploit the large scale knowledge acquired by pre-trained language models to build a text-modality teacher that can provide useful complementary information about the action sequences to a vision modality student. This cross-modal distillation strategy yields consistent gains achieving new state-of-the-art results on multiple datasets. Last, the gap between the performance of the teacher and the student models for domains with large label space suggests that there is still room for improvement with better distillation techniques.

7 Limitations

Real life scenarios have a large space of human actions which cannot be exhaustively covered by manually annotated training data. As such it is important to have models with zero-shot anticipation capabilities to predict unseen actions. This work did not explore zero-shot settings but we believe text-to-video distillation holds promise given the recent successes of language models in zero-shot tasks.

In this work we have shown the capability of text based language models for action anticipation, especially when the action space is very large and sparse. Though this work is intended to be a proof of concept for leveraging text based model for im-

proving video based action anticipation, there is still a large performance gap between the a text based language model and vision modality model. This performance gap indicates fruitful research avenues in text to vision knowledge distillation for action anticipation task.

8 Ethical Considerations

Anticipation future action based on videos is an important for many applications such as assistive technologies, augmented reality etc. Our work demonstrates that knowledge derived from text sources can be used to further improve the performance of video based action anticipation model. Even though our proposed work is able to improve the current state-of-art numbers on the standard benchmark datasets, the absolute performance is still low, especially in the case where the action space is very large. As such we would recommend to carefully analyze the cost of erroneous prediction before deploying the system for real world application.

Since the proposed method involves distilling the knowledge gained by pre-trained language model from text sources into a vision based model for action anticipation, this can also transfer the biases that these languages models can learn from the training text. As such data on which these text-based teacher models are trained should be analyzed for potential biases before deploying the proposed system for actual application. Analysis of bias propagation during knowledge distillation and devising bias reduction techniques are some potential extension of this work that we are highly interested in.

9 Acknowledgement

This material is based on research that is supported in part by the Air Force Research Laboratory (AFRL), DARPA, for the KAIROS program under agreement number FA8750-19-2-1003 and in part by the National Science Foundation under the award IIS #2007290.

References

Yazan Abu Farha, Alexander Richard, and Juergen Gall. 2018. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5343–5352.

Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Ronald Azuma. 2004. Overview of augmented reality. In *ACM SIGGRAPH 2004 Course Notes*, pages 26–es.

Guglielmo Camporese, Pasquale Coscia, Antonino Furnari, Giovanni Maria Farinella, and Lamberto Ballan. 2021. Knowledge distillation for action anticipation via label smoothing. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3312–3319. IEEE.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Mohamed Chaabane, Ameni Trabelsi, Nathaniel Blanchard, and Ross Beveridge. 2020. Looking ahead: Anticipating pedestrians crossing with future frames prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2297–2306.

Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12655–12663.

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

Eadom Dessalene, Chinmaya Devaraj, Michael Maynord, Cornelia Fermuller, and Yiannis Aloimonos. 2021. Forecasting action through contact representations from first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nuno Ferreira Duarte, Mirko Raković, Jovica Tasevski, Moreno Ignazio Coco, Aude Billard, and José Santos-Victor. 2018. Action anticipation: Reading the intentions of humans and robots. *IEEE Robotics and Automation Letters*, 3(4):4132–4139.

Antonino Furnari, Sebastiano Battiato, and Giovanni Maria Farinella. 2018. Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0.

Antonino Furnari and Giovanni Maria Farinella. 2019. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6252–6261.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Rohit Girdhar and Kristen Grauman. 2021. Anticipative video transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13505–13515.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Hengtong Hu, Lingxi Xie, Richang Hong, and Qi Tian. 2020. Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3123–3132.

Ronghang Hu and Amanpreet Singh. 2021. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1439–1449.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. 2014. A hierarchical representation for future action prediction. In *European Conference on Computer Vision*, pages 689–704. Springer.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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

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

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

Fenglin Liu, Yuanxin Liu, Xuancheng Ren, Xiaodong He, and Xu Sun. 2019a. Aligning visual regions and textual concepts for semantic-grounded image representations. *Advances in Neural Information Processing Systems*, 32.

Miao Liu, Siyu Tang, Yin Li, and James M Rehg. 2020. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *European Conference on Computer Vision*, pages 704–721. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2019. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):187–203.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Patrick Salamin, Daniel Thalmann, and Frédéric Vexo. 2006. The benefits of third-person perspective in virtual and augmented reality? In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, VRST '06, page 27–30, New York, NY, USA. Association for Computing Machinery.

Paul Schydlo, Mirko Rakovic, Lorenzo Jamone, and José Santos-Victor. 2018. Anticipation in human-robot cooperation: A recurrent neural network approach for multiple action sequences prediction. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5909–5914. IEEE.

Fadime Sener, Dipika Singhania, and Angela Yao. 2020. Temporal aggregate representations for long-range video understanding. In *European Conference on Computer Vision*, pages 154–171. Springer.

Fadime Sener and Angela Yao. 2019. Zero-shot anticipation for instructional activities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 862–871.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Tomoyuki Suzuki, Hirokatsu Kataoka, Yoshimitsu Aoki, and Yutaka Satoh. 2018. Anticipating traffic accidents with adaptive loss and large-scale incident db. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3521–3529.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding via contextualized, visually-grounded supervision. In *EMNLP*.

Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. 2021. Vidlankd: Improving language understanding via video-distilled knowledge transfer. *Advances in Neural Information Processing Systems*, 34.

Fida Mohammad Thoker and Juergen Gall. 2019. Cross-modal knowledge distillation for action recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 6–10. IEEE.

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

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 98–106.

Disong Wang, Jianwei Yu, Xixin Wu, Songxiang Liu, Lifa Sun, Xunying Liu, and Helen Meng. 2020. End-to-end voice conversion via cross-modal knowledge distillation for dysarthric speech reconstruction. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7744–7748. IEEE.

Jonatas Wehrmann, Camila Kolling, and Rodrigo C Barros. 2020. Adaptive cross-modal embeddings for image-text alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12313–12320.

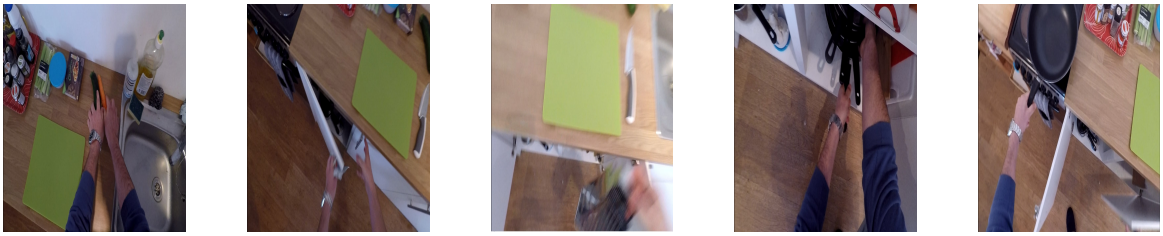
A Appendix

In this section, we present examples of model prediction for the video action anticipation task for the `EPIC-55` dataset. For each instance we show the top-5 predictions for (i) video-only model - **`AVT`** (ii) text-based teacher model - `LM-teacher` (**`Rcp-ALBERT`**) (iii) a `LM-teacher` distilled student video model - `AVT + LM teacher Distl` (**`AVT + Rcp-ALBERT Distillation`**). Note that the end-task setting is such that, the inference has to be done only from the video frames, as the text label for the action segment won't be available during the inference time.

Figure 3 and 4 shows example of cases where the base video only model makes incorrect prediction, where as the text-based teacher and the teacher-distilled video model makes correct predictions. Figure 5 and 6 shows example of cases where the base video only model makes incorrect prediction, the text-based teacher makes correct prediction, however the teacher-distilled video model makes incorrect predictions.

EXAMPLE 1

INPUT



put-down_vegetable open_door take_greater take_pan put-down_pan

TARGET: close_door

PREDICTIONS

AVT : [put-down_pan, take_pan, turn-on_hob, open_door, open_drawer]

LM-teacher : [open_door, close_door, put-down_pan, take_pan, turn-on_hob]

AVT + LM teacher Distl : [put-down_pan, take_pan, open_door, close_door, turn-on_hob]

EXAMPLE 2

INPUT

put-down_pan put_lid put-down_pan take_pan take_lid

TARGET: put_lid

PREDICTIONS

AVT : [put-down_pan, turn-on_hob, open_door, take_pan, close_door]

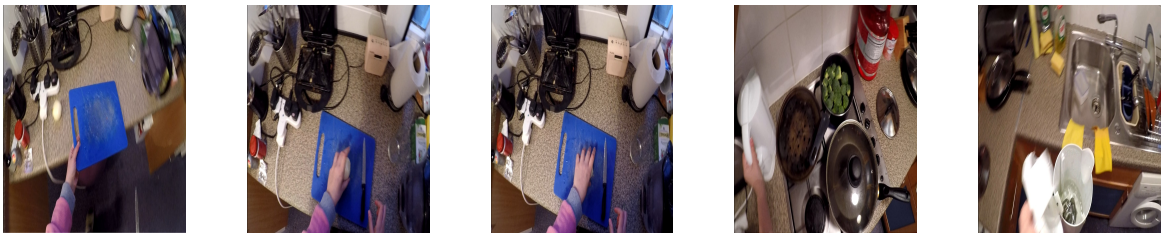
LM-teacher: [put-down_pan, open_door, take_pan, put_lid, wash_pan]

AVT + LM teacher Distl: [put-down_pan, turn-on_hob, open_door, take_pan, put_lid]

Figure 3: Example of instances where the base video-only model makes wrong prediction, whereas the text-based teacher and the teacher distilled video model makes correct prediction.

EXAMPLE 3

INPUT



put-down_board:cutting put_onion put_knife pick-up_kettle open_kettle

TARGET: fill_kettle

PREDICTIONS

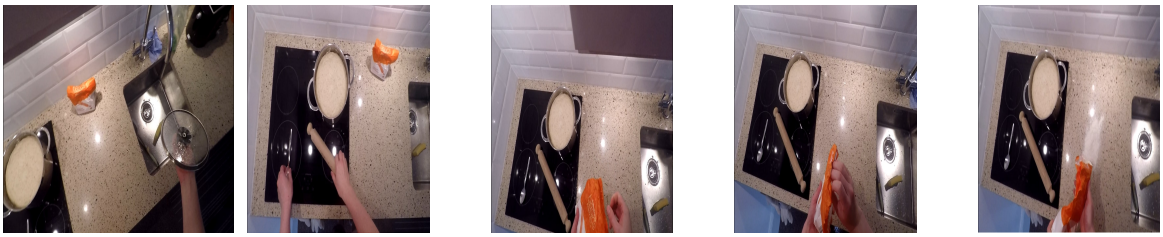
AVT: [open_door, turn-on_tap, pour_water, close_bin, open_tap]

LM-teacher: [pour_water, fill_kettle, put-down_kettle, open_kettle, close_kettle]

AVT + LM teacher Distl: [open_door, pour_water, close_bin, fill_kettle, turn-on_tap]

EXAMPLE 4

INPUT



put_lid move_spoon take_flour open_flour pour_flour

TARGET: put-down_flour

PREDICTIONS

AVT : [pour_flour, put-down_bag, mix_mixture, roll_dough, knead_dough]

LM-teacher : [put-down_flour, pour_flour, ' stir_flour, mix_mixture, check_flour]

AVT + LM teacher Distl: [pour_flour, roll_dough, mix_mixture, put-down_bag put-down_flour]

Figure 4: Example of instances where the base video-only model makes wrong prediction, whereas the text-based teacher and the teacher distilled video model makes correct prediction.

EXAMPLE 1

INPUT

open_fridge | take_carrot | open_drawer | close_fridge | putdown_vegetable

TARGET: open_door

PREDICTIONS

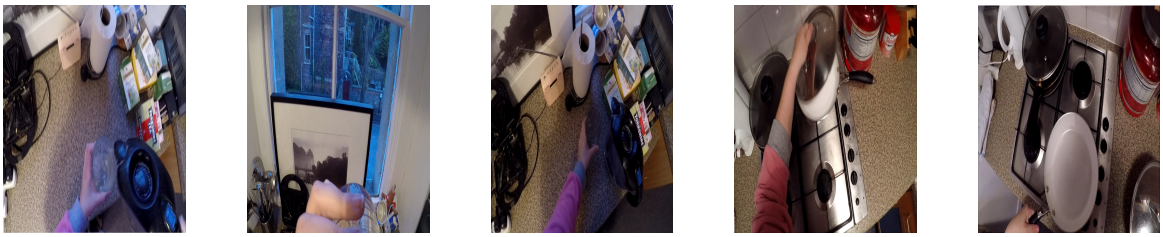
AVT: [close_door, close_fridge, put_container, open_drawer, take_knife]

LM-teacher: [close_fridge, open_drawer, open_door, close_door, take_sausage]

AVT + LM teacher Distl: [close_door, put_container, take_knife, open_drawer, take_container]

EXAMPLE 2

INPUT



put_filter:water | drink-from_cup | put_cup | take_lid | take_pan

TARGET: put-down_pan

PREDICTIONS

AVT : [put_lid, stir_pasta, put-down_spoon, change_temperature, stir_pan]

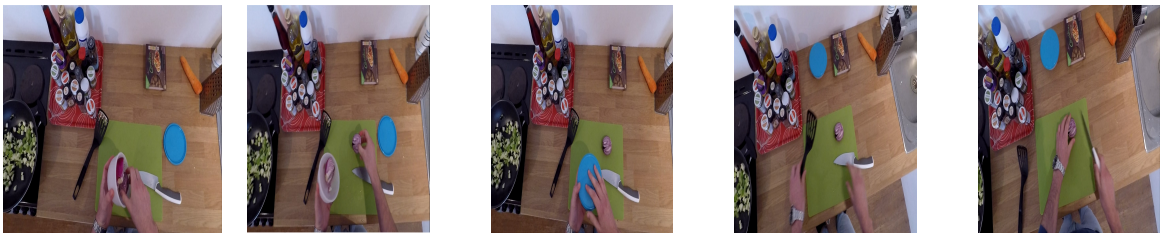
LM-teacher : [put-down_pan, wash_pan, open_door, take_pan, dry_saucepan]

AVT + LM teacher Distl: [put_lid, stir_pasta, put-down_spoon, change_temperature, open_door]

Figure 5: Example of instances where the base video-only model makes wrong prediction, the text-based teacher makes the correct prediction, however the teacher distilled video model makes incorrect prediction.

EXAMPLE 3

INPUT



take_onion | put-down_onion | close_container | take_spatula | take_knife

TARGET: cut_onion

PREDICTIONS

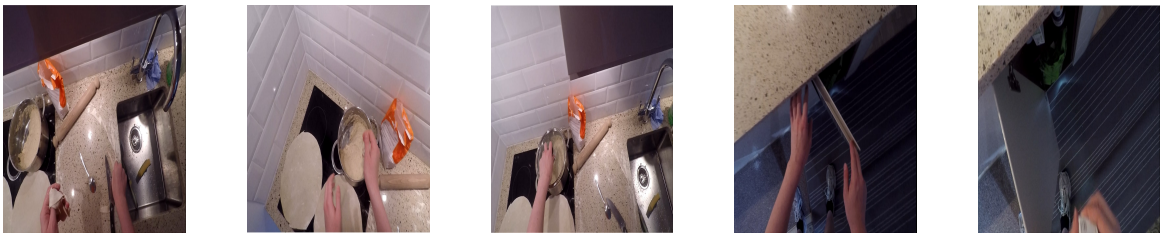
AVT: [put_container, take_knife, turn-on_tap, open_fridge, put-down_onion]

LM-teacher: [put-down_knife, cut_onion, mix_food, open_drawer, take_spoon]

AVT + LM teacher Distl: [put_container, take_knife, put-down_knife, take_container, open_fridge]

EXAMPLE 4

INPUT



take_dough | put_dough | put_lid | open_door | take_tomato

TARGET: put-down_tomato

PREDICTIONS

AVT : [open_fridge, open_door, turn-on_tap, open_drawer, rinse_hand]

LM-teacher : [put-down_tomato, close_door, take_tomato, take_plate, take_pan]

AVT + LM teacher Distl: [open_fridge, open_door, open_drawer, close_door, take_bowl]

Figure 6: Example of instances where the base video-only model makes wrong prediction, the text-based teacher makes the correct prediction, however the teacher distilled video model makes incorrect prediction.