Supplementary material

The Supplementary material is organized as follows:

- Extended discussion and details regarding the datasets we use (§A).
- Additional details about each of the modalities, as well as the modality-specific models we use (§B).
- An ablation study, alike the one in Table 5 (main paper), conducted on the Something-Something and Something-Else datasets (§C).
- Details on approaches which leverage multimodal data during training (§D).
- Analysis of the performance of our approach on seen and unseen environments (§E).
- Extended per-class performance breakdown for Epic-Kitchens and Something-Something (§F).
- Learning curves on the Epic-Kitchens and Something-Something datasets (§G).
- Additional qualitative examples on the Epic-Kitchens and the Something-Something datasets (§H).

A. Datasets

A.1. Epic-Kitchens

EPIC-Kitchens is a large-scale benchmark dataset consisting of 700 videos recorded by 32 participants [13, 14], totalling 100 hours of egocentric videos capturing daily activities in kitchen environments. In our experiments, we use the annotations for the action recognition task. It features compositional actions which can be broken down into the noun (the active object participating in the action, e.g. "carrot", "pan", etc.) and the verb (the activity itself, e.g. "cutting", "washing"). In total, there are 300 noun and 97 verb categories, while the training and validation set contain $\sim$ 68k and $\sim$ 10k videos respectively. The modalities we use in our experiments include the RGB frames, the audio, and the optical flow. We extract the audio directly from the mp4 files, and use the optical flow as released by the dataset authors [14].

A.1.1 Epic-Kitchens Unseen Participants

This particular subset of the Epic-Kitchens validation split contains 1065 action sequences from two participants which were not observed in the training dataset (i.e. videos recorded by them are not included in the training data). We use this data split to more explicitly gauge the compositional generalization performance of the models. Namely, standard RGB models tend to pick up undesirable biases to discriminate between different actions, i.e. objects or environment cues unrelated to the action [33]. Using the Epic-Kitchens Unseen Participants split, we verify the extent to which students distilled from multimodal teachers are robust w.r.t. this type of distribution shift.

A.2. Something-Something

The Something-Something V2 [21] dataset consists of (mainly) egocentric videos of people performing 174 unique object-agnostic actions with their hands, e.g. "pushing [something] left", "taking [something] out of [something]". While the dataset also contains exocentric videos, all the sequences describe similar hand-object interactions, with the main distinction being the orientation of the hands. Notice that the action classes do not account for specific objects, but rather, only for the activity itself. Therefore, on Something-Something, there is increased focus on capturing temporal relationships that characterize the actions. The training and validation set contain $\sim$ 169k and $\sim$ 26k videos respectively. Furthermore, to deal with the environment bias (models relying on unrelated environmental cues to discriminate between the actions), videos recorded by the same participant can be in either the training or validation set. Nevertheless, the objects the participants interact with – even though unrelated to the action label – can appear in both the training and the testing data, indicating that models that observe the videos through the RGB modality can overfit on the objects' appearance.

A.3. Something-Else

In Something-Something, the objects present in the scene (where the action takes place) may appear in both the training and the testing data. The goal of Something-Else [33] is to deal with the issue of models exploiting visual cues related to objects' appearance. Materzynska *et al*. [33], propose a data split according to the objects' distribution at training and test time. The data is divided such that the models encounter distinct objects during training and testing. The training and validation set contain $\sim$ 55k and $\sim$ 58k videos respectively, with 174 action categories. This data split is explicitly aimed at testing the compositional generalization of the models. Furthermore, Materzynska *et al*. [33], show that a standard RGB-based model [8] exhibits significantly lower performance on the Something-Else split compared to the standard Something-Something split. To improve the generalization ability of standard RGB-based models, the work of Materzynska *et al*. [33], as well as subsequent works [25, 43], propose using object detections as input to the model [44], as they are agnostic to the appearance of individual objects.

B. Data Modalities & Models

B.1. RGB frames (RGB)

To encode the RGB frames we follow the standard setting of [31] for all datasets – Epic-Kitchens (including the Unseen split), Something-Something (including the Something-Else compositional generalization split). Unless stated otherwise, we use the Swin-T [31] model to process the RGB frames. During training, we resize the image such that the shorter dimension (typically the height) is set to a value randomly chosen from the interval $[224, 320]$, and subsequently select a random 224×224 crop. Additionally, we adopt random horizontal flips with probability 50% (only for Epic-Kitchens), and color jittering. During inference, we resize the image such that the shorter dimension (typically the height) is set to 224, and then select a 224×224 central crop for each frame.

In the case of the R3D [27] models, where we focus on testing our approach on computationally cheaper and faster architectures and settings, we keep the same train and inference setup, with the exception of the final crop size which we reduce to 112, as per Kataoka *et al*. [27].

B.2. Optical Flow (OF)

We process the optical flow frames in the same fashion as the RGB frames and use the same vision backbone (Swin-T) for both Epic-Kitchens (including the Unseen split) and Something-Something (including the Something-Else compositional generalization split). We use the two components of the velocity as the first two channels of the input, and in order to maintain the same architecture, we append an additional channel where we set each pixel intensity to 0.0, effectively expanding the number of input channels to 3. We use the same data augmentations as with the RGB model, with the exception of color jitter.

B.3. Audio (A)

On Epic-Kitchens, we first convert the 24000Hz stereo audio to 16000Hz monoaural audio. We compute the mel-spectrograms of audio segments using 1024 FFT bins and 128 mel filter banks. We use the Hann window with length of 160, with an 80 sample overlap between successive windows. We square the magnitude after computing the FFT, and thus obtain the signal power at each frequency bin for each time-step. For the audio segments of 1.116 with the sample frequency of 16000, we thus obtain spectrograms with 128 frequency bins and 224 timesteps.

During training, as data augmentation, we perform random time and frequency masking of the spectrograms, as per the work of [38]. In time masking, with a probability of 50%, we randomly chose the number of masked time-steps T_n from the range $[0, 80]$, and the starting time-step from the range $[0, 224 - T_n)$, such that, for all the frequency bins, the range of time-steps $[T_s, T_s + T_n)$ is masked by setting the power value in the spectrogram to 0. In frequency masking, with a probability of 50%, we randomly chose the number of masked frequency bins F_n from the range $[0, 80]$, and the starting frequency bin F_s from the range of $[0, 128 - F_n)$, such that, for all the timesteps, the range of bins $[F_s, F_s + F_n)$ is masked by setting the power value in the spectrogram to 0. Afterwards, we resize the spectrogram height to a value randomly chosen from the interval $[224, 320]$, and finally select a random 224×224 crop.

During inference, we do not perform time and frequency masking, we simply resize the height of the spectrogram to 224 and select a 224×224 central crop for each frame.

We use the obtained spectrogram repeated 3 times to construct a 3-channel input for the Swin-T backbone. Despite the simple setup, our audio-specific model performs on-par with more sophisticated state-of-the-art audio models [48] on Epic-Kitchens.

B.4. Object Detections (OBJ)

When pre-processing the Object Detections (on Something-Something and Something-Else) we closely follow the setup of [43]. We represent each video frame with only its object detections – bounding boxes & object categories. We use the object detections released from Herzig *et al*. [25] for Something-Something and Materzynska *et al*. [33] for Something-Else, which had been obtained using a Faster R-CNN [44], trained as per the setting of [46]. We use the STLT (Spatial-Temporal Layout Transformer) model to encode the object detections, while following the settings and the implementation of [43].

In the STLT model, one Transformer model [53] encodes the spatial relations between the objects in each frame independently, while another Transformer encodes the temporal relations given the embedding of each frame (output of the Spatial-Transformer).

C. Loss Term Weighting - Something-Something & Something-Else

In the vein of the ablation study reported in Table 5 (main paper), we conduct experiments on the Something-Something and the Something-Else datasets. Namely, the main findings from Table 5 suggest that (i) training with the ground-truth labels cross-entropy loss, in conjunction with the multimodal knowledge distillation loss, overcomes the issue of inferior modality-specific teachers, and (ii) weighting the teachers in the ensemble (such that their each individual cross-entropy loss on a holdout set of $Z = 1000$ samples are minimized) improves the performance further. In Table 6, however, we observe that in the case of Something-Something and Something-Else, the addition of the loss term featuring ground-truth labels has a small effect on

			Something-Something		Something-Else	
Objective	λ	γ	Action@1	Action@5	Action@1	Action@5
$\mathcal{L}_{\text{CE}}$	0.0	—	60.3	86.4	51.8	79.5
$\mathcal{L}_{\text{KL}}$	1.0	30.0	63.0	88.9	59.1	86.1
$\mathcal{L}_{\text{CE}} \wedge \mathcal{L}_{\text{KL}}$	0.8	30.0	63.1	88.3	59.3	86.3

Table 6: Ablation study on Something-Something and Something-Else; λ: Distillation and Cross-Entropy loss balancing term; γ: Temperature of the Ensemble Teacher Weighting.

Dataset	Objective	λ	γ	Noun@1	Verb@1	Action@1
Epic Kitchens **Regular**	$\mathcal{L}_{\text{CE}}$	0.0	—	52.0	61.7	38.3
	$\mathcal{L}_{\text{CE}} \wedge \mathcal{L}_{\text{KL}}$	0.8	1.0	53.5	65.4	41.2
Epic Kitchens **Unseen**	$\mathcal{L}_{\text{CE}}$	0.0	—	38.3	51.7	25.4
	$\mathcal{L}_{\text{CE}} \wedge \mathcal{L}_{\text{KL}}$	0.8	1.0	42.5	54.7	30.4
Epic Kitchens **Seen**	$\mathcal{L}_{\text{CE}}$	0.0	—	53.7	63.0	39.9
	$\mathcal{L}_{\text{CE}} \wedge \mathcal{L}_{\text{KL}}$	0.8	1.0	54.8	66.7	42.5

Table 7: Epic-Kitchens dataset (regular validation set, environments **Unseen** during training, environments **Seen** during training).

the performance of the student. Namely, as discussed in the main paper, on both the Something-Something and the Something-Else datasets, all modality-specific models perform well, and are complementary to each other, therefore, there is a lesser need for joint training using the ground truth labels and the distillation loss.

D. Details on action recognition models trained on multimodal data

Multiple works explore a similar setting, i.e. using multiple modalities for training while performing inference using only RGB frames. Some of the most prominent works are ModDrop [37], DMCL [17, 18], and Omnivore [20].

ModDrop. Neverova *et al*. [37] propose a method where a multimodal model is made robust to missing modalities during inference by randomly dropping out modalities during training. Namely, the model is trained such that it might observe all modalities, a partial set of modalities or only a single modality during training. This makes the model recognize cues, generally multimodal, from RGB data, and is therefore superior to an RGB model.

DMCL. Garcia *et al*. [17, 18] propose a four-step mulitmodal distillation framework which is tested on non-egocentric data. They train a model on multimodal inputs, where for each training video-action sample, the teacher network is established as the model which exhibits the lowest cross-entropy w.r.t. ground truth action, and the remaining models are the students. Then, the student models are trained on the soft teacher labels. On the other hand, our method is simple – standard knowledge distillation – and flexible – other models can easily be added to the ensemble and the student model can be retrained while keeping the existing models fixed.

Omnivore. [20] To the best of our knowledge, Omnivore is the latest and best performing method that uses multimodal data during training, while using only unimodal data during inference. Compared to multimodal distillation, Omnivore can perform inference using a single set of weights across all different modalities it was trained on. In particular, Girdhar *et al*. [20] use multimodal data during pre-training, while for the downstream task, the model is directly fine-tuned on the RGB frames. The resulting model – pre-trained on omnivorous data – is superior. In our work, to establish an Omnivore baseline parallel to multimodal distillation, we perform training on the multimodal data the downstream task features.

To train a single model (single set of weights) using multimodal data, Girdhar *et al*. [20] propose two strategies to sample the batches: (i) Batches contain data of mixed modalities – heterogenous batches, or (ii) each batch is unimodal – homogenous – with a randomly chosen modality. In our work, we found that (i) yields a model with performance similar to the simply training the model on RGB frames, and therefore, we opted for (ii).

E. Performance Breakdown on Seen and Unseen Participants

To assess whether our approach yields improvements in terms of generalizing to new visual environments, we isolate the videos of participants in the Epic-Kitchens validation dataset (which we dub as Epic-Kitchens Regular in Table 7 that are not included in the Epic-Kitchens Unseen split (Section A.1.1). We dub this dataset split as Epic-Kitchens Seen, as it consists videos of participants that have also been featured in the training data (and thus show visual environments that have already been observed during training). In Table 7, on the Epic-Kitchens dataset, we contrast the results on the Unseen split with those on the Seen split. We notice that while the performance improves on the Seen split, the gain of our approach is even larger on the Unseen split (in line with optical flow and audio being more invariant w.r.t. environment appearance). Consequently, we observe a smaller performance drop going from Seen to Unseen dataset splits in the case of our approach, suggesting a lower degree of overfitting to the appearance of objects and environments.

F. Per-Class Performance Breakdown

In Figure 4 in the main paper, we provided a per-class performance breakdown of the 20 most frequent actions for Epic-Kitchens and Something-Something. We also present the per-class performance breakdown for the teacher the top-20 most frequent action classes in Figure 8. On these actions, the teacher ensemble performance generally follows a similar trend as the student, shown in Figure 4, albeit

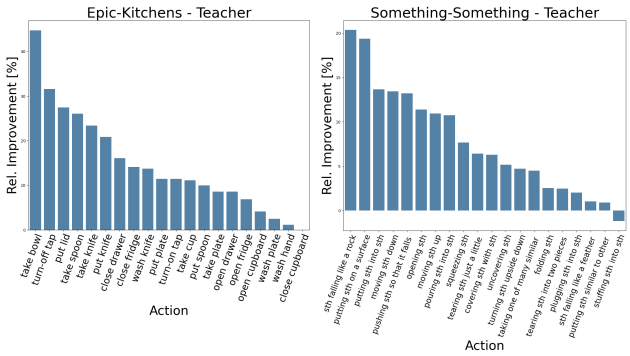

Figure 8: Per-class improvement of the teacher ensemble over the RGB baseline on the top-20 most frequent actions across datasets.

achieving higher improvement with respect to the baseline, as expected from its higher overall performance.

Furthermore, in Figure 9 and Figure 10, we provide an extended per-class performance breakdown for the student on the 100 most frequent actions. On both datasets, we observe that the student is superior to the model trained on the ground truth labels on the majority of action classes.

Finally, we also provide an extended per-class performance breakdown for the teacher on the 100 most frequent actions. On both datasets, we observe that the teacher achieves slightly higher overall performance than the student, achieving positive relative improvement on a larger number of action categories.

G. Learning curves

To ensure better reproducibility, we also report learning curves on the Epic-Kitchens and Something-Something dataset. Namely, in Figure 13 we report the top-1 validation set accuracy measured at the end of each epoch on the y-axis, and the number of executed training epochs on the x-axis. We observe that for both datasets, the distilled student converges to a model which generalizes better than training on the ground truth labels alone.

H. Qualitative Examples

We report additional qualitative examples in Figure 14 supplementing the results of Figure 7 in the main paper.

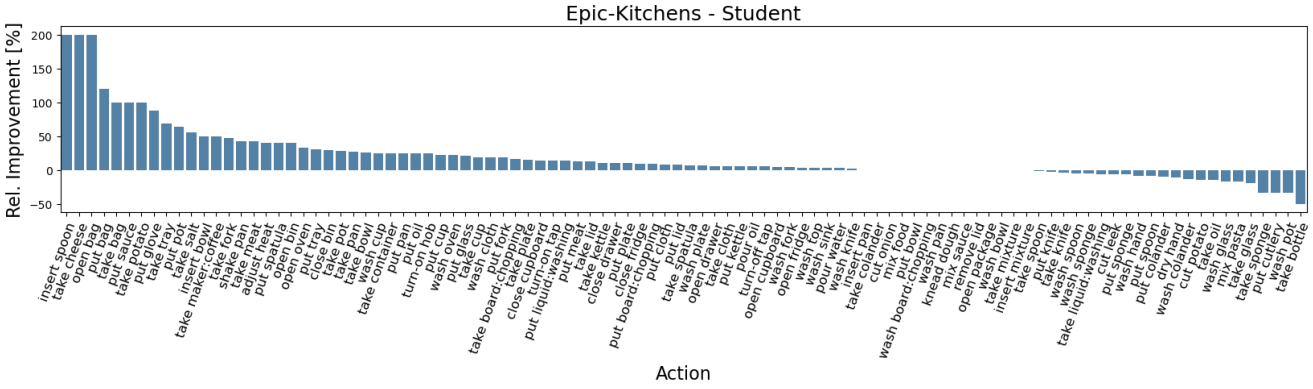

Figure 9: Per-class performance change between the student and the RGB baseline on the Epic-Kitchens regular split.

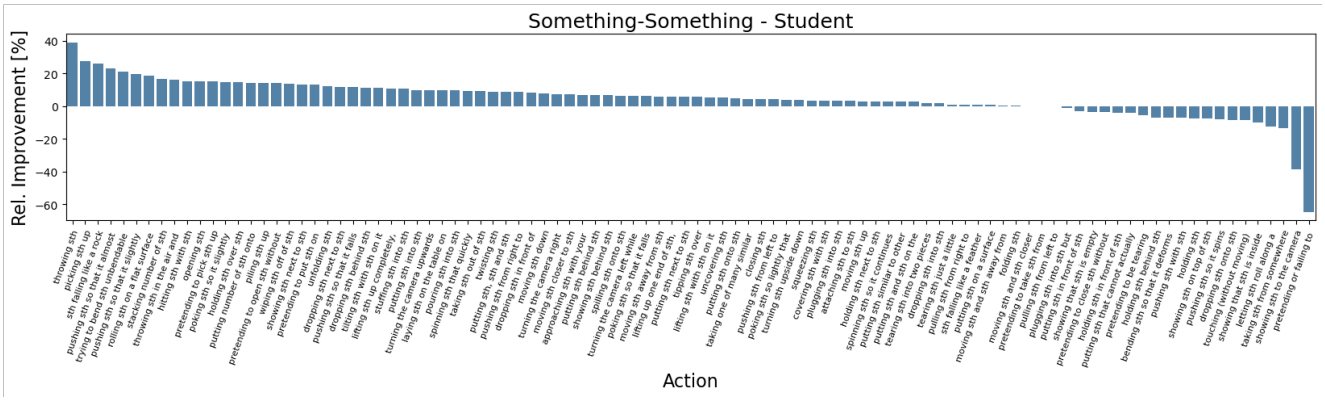

Figure 10: Per-class performance change between the student and the RGB baseline on the Something-Something dataset.

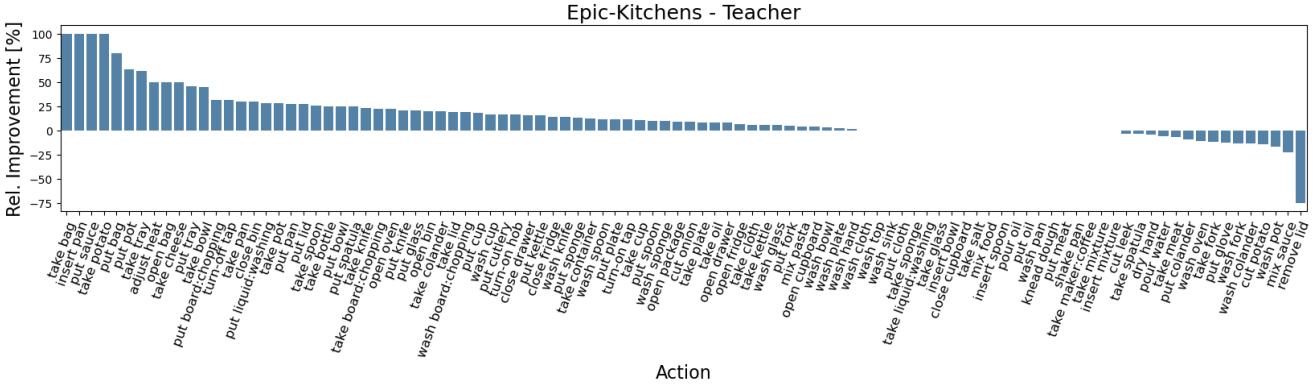

Figure 11: Per-class performance change between the teacher and the RGB baseline on the Epic-Kitchens regular split.

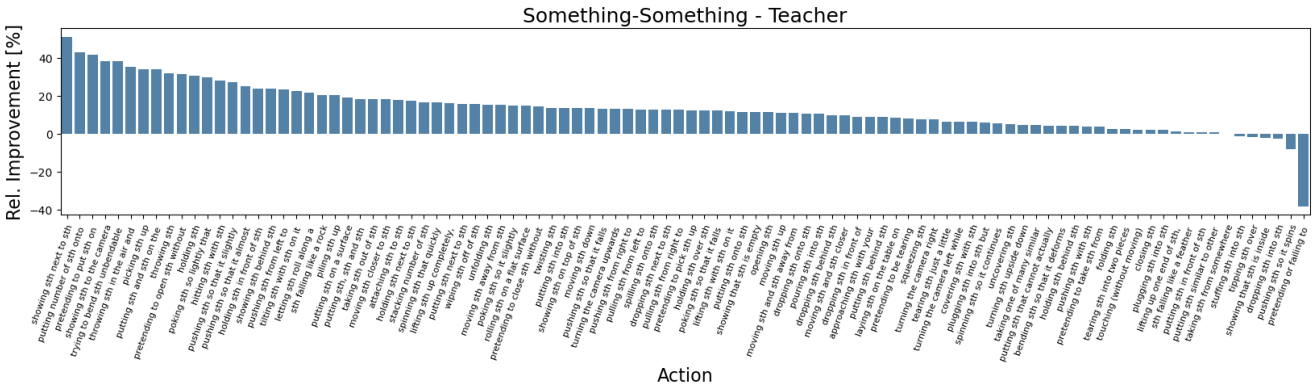

Figure 12: Per-class performance change between the teacher and the RGB baseline on the Something-Something dataset.

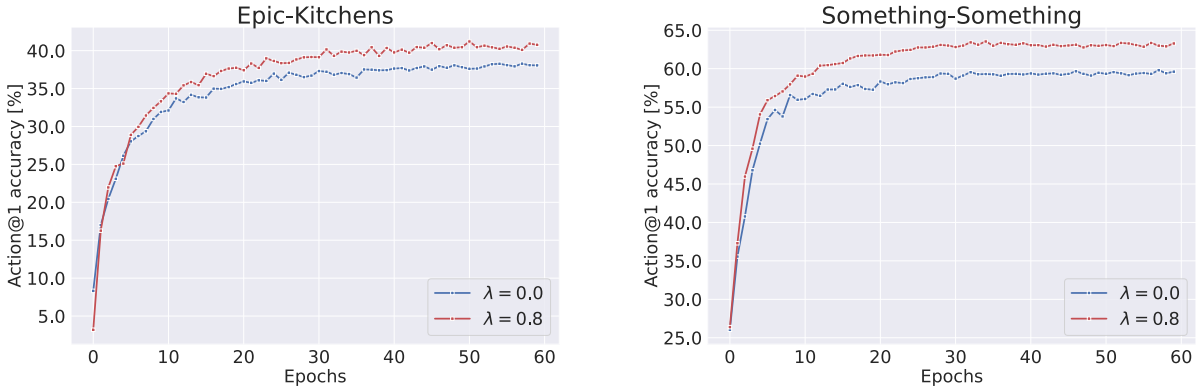

Figure 13: Learning curve on the Epic-Kitchens dataset (left) and the Something-Something dataset (right).

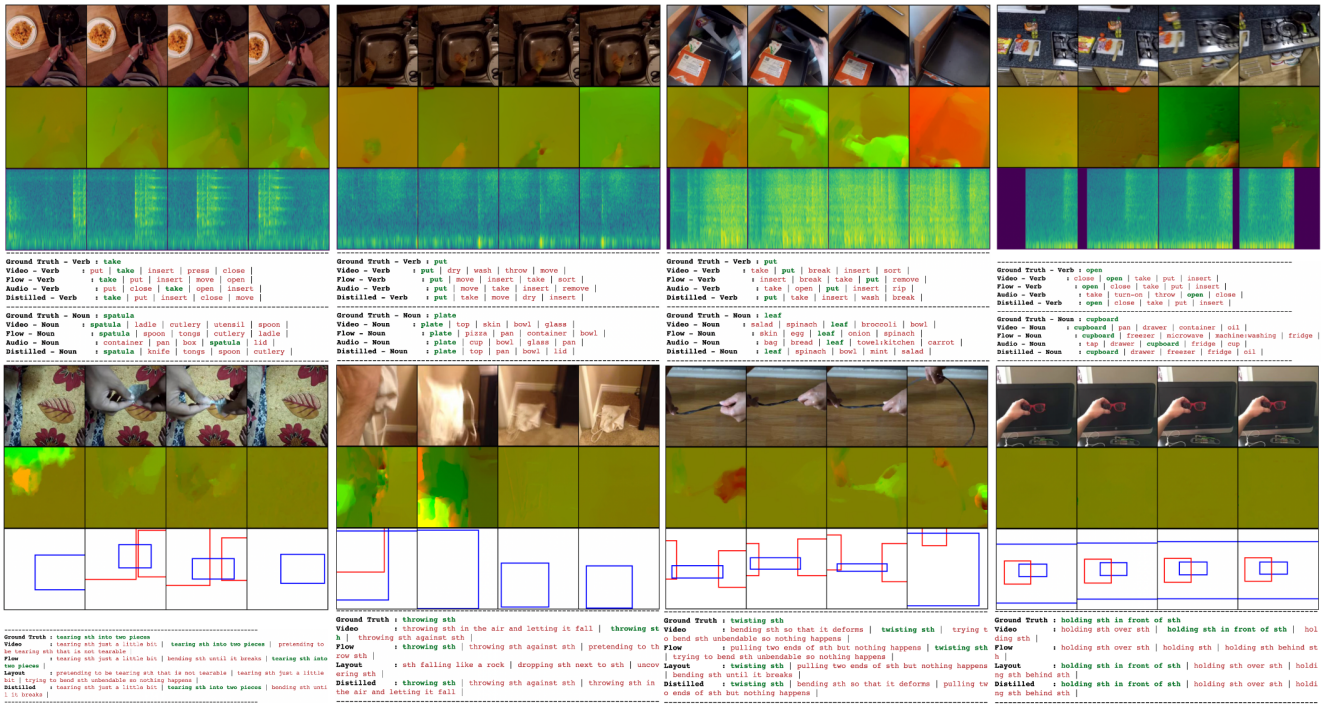

Figure 14: Qualitative evaluation for Epic-Kitchens (top) and Something-Something (bottom).