A Appendix Introduction

The following sections present supporting material for the main paper. We first list all our notations for easy referencing in Appendix A.1. We then present proofs for the mathematical results in the paper in Appendix B. Next, we describe the experimental setup in Appendix C and go on to provide an runtime analysis in Appendix D and additional results on TSN in Appendix E.

A.1 Notation Reference

Variable	Definition	Description
X	$(x_i)_{i=1}^n$	An ordered sequence comprised of n elements. In our experiments x_i is a frame and X is a video.
x_i	$x_i \in X$	An element of our sequence X.
X'	$X' \subseteq X$	A subsequence of X, at times this is constrained to be a proper subset ($X' \subset X$). Check surrounding context for constraints.
$f(X)$	-	A model that operates on a sequence X and produces a vector of class scores.
$f_c(X)$	-	The score for class c produced when the model f is evaluated on the sequence X.
$f_c(\emptyset)$	-	The score for class c when there is no input, we choose the empirical class distribution over the training set to represent this.
$f_{gt}(X)$	-	The ground truth class score.
$f_{pt}(X)$	-	The predicted class score.
$f_{cc}(X)$	$f_{gt}(X) - fpt(X)$	The difference between ground truth and predicted class score.
$f^s(X)$	$f^s : \mathbb{R}^{s \times D} \to \mathbb{R}^C$	A single-scale model mapping from a sequence of length s where each element has dimension D to a set of class scores.
$f^{\text{ms}}(X)$	$\mathbb{E}_s \left[\mathbb{E}_{X'\|s}[f^s(X')]\right]$	A multi-scale model built from a set of single scale models.
$\Delta_i^c(X')$	$f_c(X' \cup \{x_i\}) - f_c(X')$	The *marginal contribution* of x_i on the subsequence X' with the condition that $x_i \notin X'$.
ϕ_i^c	$\sum_{X' \subseteq X \setminus \{x_i\}} w(X')\Delta_i^c(X')$	The Element Shapley Value for x_i with respect to class c.
$w(X')$	$\frac{(\|X\|-\|X'\|-1)!\|X'\|!}{\|X\|!}$	The weighting factor used in the Element Shapley Value definition.
$\hat{\phi}_i^c$	-	Approximated Element Shapley Value computed via Algorithm 1.
δ_i	$\phi_i^{gt} - \phi_i^{pt}$	The difference in Element Shapley Value computed w.r.t the ground truth and predicted class.

B Proofs

B.1 Shapley Value Expectation Forms

Single Expectation Form. The Shapley value for an element x_i from a sequence X can be interpreted as the expected marginal contribution of x_i on a random coalition $X' \subseteq X \setminus \{x_i\}$, where X' maintains the order of elements in X.

Proof. The Shapley value is originally defined [1] as

$$\phi_i = \sum_{X' \subseteq X \setminus \{x_i\}} \frac{(|X| - |X'| - 1)!|X'|!}{|X'|!} [f(X' \cup \{x_i\}) - f(X)] \,. \tag{11}$$

We define a random variable X' whose probability mass function (pmf) is

$$p(X') = \frac{(|X| - |X'| - 1)!|X'|!}{|X|!} \,. \tag{12}$$

$p(X')$ is a valid pmf, given all values are non-negative and its sum is equal to one by direct proof.

$$\sum_{X' \subseteq X \setminus \{x_i\}} p(X') = \sum_{X' \subseteq X \setminus \{x_i\}} \frac{(|X| - |X'| - 1)!|X'|!}{|X|!} \tag{13}$$

$$= \sum_{s=0}^{|X|-1} \sum_{\substack{X' \subseteq X \setminus \{x_i\} \\ |X'|=s}} \frac{(|X| - |X'| - 1)!|X'|!}{|X|!} \tag{14}$$

$$= \sum_{s=0}^{|X|-1} \frac{(|X| - s - 1)!s!}{|X|!} \sum_{\substack{X' \subseteq X \setminus \{x_i\} \\ |X'|=s}} 1 \tag{15}$$

$$= \sum_{s=0}^{|X|-1} \frac{1}{|X|} \binom{|X| - 1}{s}^{-1} \sum_{\substack{X' \subseteq X \setminus \{x_i\} \\ |X'|=s}} 1 \tag{16}$$

$$= \frac{1}{|X|} \sum_{s=0}^{|X|-1} \binom{|X| - 1}{s}^{-1} \sum_{\substack{X' \subseteq X \setminus \{x_i\} \\ |X'|=s}} 1 \tag{17}$$

$$= \frac{1}{|X|} \sum_{s=0}^{|X|-1} \binom{|X| - 1}{s}^{-1} \binom{|X| - 1}{s} \tag{18}$$

$$= \frac{1}{|X|} \sum_{s=0}^{|X|-1} 1 \tag{19}$$

$$= 1 \,. \tag{20}$$

Consequently, we can see that

$$\phi_i = \mathop{\mathbb{E}}_{X'} \left[f(X' \cup \{x_i\}) - f(X') \right] \tag{21}$$

Conditional Expectation Form. The Shapley value can also be formulated as the expectation over the conditional expectation of the marginal contribution of an element on a coalition of size s where all values of s are equally probable.

$$\mathop{\mathbb{E}}_{X'}\left[f(X' \cup \{x_i\}) - f(X')\right] = \mathop{\mathbb{E}}_{s}\left[\mathop{\mathbb{E}}_{X'|s}\left[f(X' \cup \{x_i\}) - f(X')\right]\right] \tag{22}$$

This is a consequence of the law of total expectation, as we split the sample space into $|X|$ non-empty and non-overlapping partitions, each containing subsets of size $s \in \{0\,..\,|X|-1\}$.

Proof. We show the equivalence by direct proof. Since there are $\binom{|X|-1}{s}$ instances of X' of size s the conditional probability of X' given s is

$$p(X'|s) = \binom{|X|-1}{s}^{-1} \tag{23}$$

Expanding out the RHS of Eq. (22) we get

$$\sum_{s=0}^{|X|-1} p(s) \sum_{\substack{X' \subseteq X \setminus \{x_i\} \\ |X'|=s}} p(X'|s)[f(X' \cup \{x_i\}) - f(X')] \tag{24}$$

$$= \sum_{s=0}^{|X|-1} \frac{1}{|X|} \sum_{\substack{X' \subseteq X \setminus \{x_i\} \\ |X'|=s}} \binom{|X|-1}{s}^{-1} [f(X' \cup \{x_i\}) - f(X')] \tag{25}$$

$$= \sum_{X' \subseteq X \setminus \{x_i\}} \frac{1}{|X|} \binom{|X|-1}{|X'|}^{-1} [f(X' \cup \{x_i\}) - f(X')] \tag{26}$$

$$= \mathop{\mathbb{E}}_{X'}\left[f(X' \cup \{x_i\}) - f(X')\right] \tag{27}$$

B.2 Recursive Definition of Variable-Length Input Model

We define our multi-scale model f (note this is the same as f^{ms}, but we drop the superscript in this proof for notational simplicity) as a combination of the results from a set of single scale models $\{f^s\}_{s=1}^{n_{\max}}$

$$f(X) = \mathop{\mathbb{E}}_{s}\left[\mathop{\mathbb{E}}_{X'|s}\left[f^s(X')\right]\right]. \tag{8 revisited}$$

To improve efficiency when computing Shapley values, it is desirable to formulate this in a recursive fashion, which we denote $\check{f}$. This enables the computation of $f(X)$ in terms of the expected result of $\check{f}(\check{X})$ where $\check{X}$ is a random variable over subsequences of X with one element less (all equally probable).

$$\check{f}(X) = \begin{cases} f^1(X) & |X| = 1 \\ |X|^{-1}\left[f^{|X|}(X) + (|X|-1)\,\mathbb{E}_{\check{X}}\left[\check{f}(\check{X})\right]\right] & |X| \leq n_{\max} \\ \mathbb{E}_{\check{X}}\left[\check{f}(\check{X})\right] & |X| > n_{\max} \end{cases} \tag{28}$$

Proof. We prove the equivalence between f and $\check{f}$ by induction on $|X|$.

Base case: When $|X| = 1$, observe that $p(s = 1) = 1$ and the sample space $\Omega(X'|s = 1) = \{X\}$ therefore $f(X) = f^1(X)$ from Eq. (8).

Inductive step: We split the inductive step into two parts, one where $|X| \leq n_{\max}$ and one where $|X| > n_{\max}$. For both parts of the proof, we start by assuming $f(\check{X}) = \check{f}(\check{X})$, which we have proven for the base case $s = 1$. For $|X| \leq n_{\max}$, we start from $\check{f}(X)$, expanding out the definition according to clause 2 in Eq. (28)

$$\check{f}(X) = |X|^{-1} \left[f^{|X|}(X) + (|X| - 1) \mathop{\mathbb{E}}_{\check{X}} \left[\check{f}(\check{X}) \right] \right] . \tag{29}$$

Our strategy will be to expand the expectation in Eq. (29), and show that substituting the expanded form back into Eq. (29) recovers Eq. (8). Focusing on the expectation and substituting our assumption, $f(\check{X}) = \check{f}(\check{X})$, yields

$$\mathop{\mathbb{E}}_{\check{X}} \left[\check{f}(\check{X}) \right] = \mathop{\mathbb{E}}_{\check{X}} \left[f(\check{X}) \right] \tag{30}$$

$$= \mathop{\mathbb{E}}_{\check{X}} \left[\mathop{\mathbb{E}}_{s} \left[\mathop{\mathbb{E}}_{\check{X}'|s} \left[f^s(X') \right] \right] \right] \tag{31}$$

It is important to note the definitions of the random variables here: $\check{X}' \subseteq \check{X}$ and $s \in \{1 \,..\, |\check{X}|\}$. We then expand out the expectations

$$\mathop{\mathbb{E}}_{\check{X}} \left[\check{f}(\check{X}) \right] = \sum_{\substack{\check{X} \subset X \\ |\check{X}| = |X| - 1}} \frac{1}{|X|} \sum_{s=1}^{|\check{X}|} \frac{1}{|\check{X}|} \sum_{\substack{\check{X}' \subseteq \check{X} \\ |\check{X}'| = s}} \binom{|\check{X}|}{s}^{-1} f^s(\check{X}') , \tag{32}$$

reordering the summations, and replacing $|\check{X}|$ with $|X| - 1$ yields

$$\mathop{\mathbb{E}}_{\check{X}} \left[\check{f}(\check{X}) \right] = \frac{1}{|X|} \sum_{s=1}^{|X|-1} \frac{1}{|X| - 1} \sum_{\substack{\check{X} \subset X \\ |\check{X}| = |X| - 1}} \sum_{\substack{\check{X}' \subseteq \check{X} \\ |\check{X}'| = s}} \binom{|X| - 1}{s}^{-1} f^s(\check{X}') . \tag{33}$$

As $\{\check{X} \subset X : |\check{X}| = |X| - 1\}$ is all subsequences of X excluding a single element, we can substitute this set with $\{x_i \in X\}$ and replace $\check{X}$ with $X \setminus \{x_i\}$,

$$\mathop{\mathbb{E}}_{\check{X}} \left[\check{f}(\check{X}) \right] = \frac{1}{|X|} \sum_{s=1}^{|X|-1} \frac{1}{|X| - 1} \sum_{x_i \in X} \sum_{\substack{X' \subseteq X \setminus \{x_i\} \\ |X'| = s}} \binom{|X| - 1}{s}^{-1} f^s(X') . \tag{34}$$

Next, we show that

$$\frac{1}{|X|} \sum_{x_i \in X} \sum_{\substack{X' \subseteq X \setminus \{x_i\} \\ |X'| = s}} \binom{|X| - 1}{s}^{-1} = \sum_{\substack{X' \subseteq X \\ |X'| = s}} \binom{|X|}{s}^{-1} \tag{35}$$

To get from the LHS to the RHS, we need to consider how many times we will repeatedly count X' due to the outer summation. If we consider a fixed X', then there will be $|X| - |X'|$ times where it will be drawn in the inner summation. Hence we can replace the outer summation with the number of occurrences

$$\frac{1}{|X|} \sum_{x_i \in X} \sum_{\substack{X' \subseteq X \setminus \{x_i\} \\ |X'|=s}} \binom{|X|-1}{s}^{-1} \tag{36}$$

$$= \frac{1}{|X|} \sum_{\substack{X' \subseteq X \setminus \{x_i\} \\ |X'|=s}} (|X| - s) \binom{|X|-1}{s}^{-1} \tag{37}$$

$$= \sum_{\substack{X' \subseteq X \setminus \{x_i\} \\ |X'|=s}} \frac{(|X|-s)}{|X|} \frac{(|X|-1-s)!s!}{(|X|-1)!} \tag{38}$$

$$= \sum_{\substack{X' \subseteq X \setminus \{x_i\} \\ |X'|=s}} \frac{(|X|-s)!s!}{(|X|)!} \tag{39}$$

$$= \sum_{\substack{X' \subseteq X \setminus \{x_i\} \\ |X'|=s}} \binom{|X|}{s}^{-1} . \tag{40}$$

Having proven Eq. (35), we can simplify Eq. (34) to

$$\mathop{\mathbb{E}}_{\check{X}} \left[\check{f}(\check{X}) \right] = \frac{1}{|X|-1} \sum_{s=1}^{|X|-1} \sum_{\substack{X' \subseteq X \\ |X'|=s}} \binom{|X|}{s}^{-1} f^s(X') \,. \tag{41}$$

Substituting this form back into Eq. (29) we get

$$\check{f}(X) = |X|^{-1} \left[\sum_{s=1}^{|X|-1} \sum_{\substack{X' \subseteq X \\ |X'|=s}} \left[\binom{|X|}{s}^{-1} f^s(X') \right] + f^{|X|}(X) \right] . \tag{42}$$

Now $f^{|X|}(X)$ can be merged into the summation over s by increasing its bound from $|X|-1$ to $|X|$

$$\check{f}(X) = \sum_{s=1}^{|X|} \frac{1}{|X|} \sum_{\substack{X' \subseteq X \\ |X'|=s}} \binom{|X|}{s}^{-1} f^s(X') \,. \tag{43}$$

Which is the expansion of the expectations of $f(X)$. The proof for the equivalence between $\check{f}(X)$ and $f(X)$ for $|X| > n_{\max}$ is very similar to the above and has been omitted for brevity.

B.3 Linearity of Shapley Values

The Shapley value $\phi_i^{w_1c_1+w_2c_2}$ for a model $f_{w_1c_1+w_2c_2}(X) = w_1 f_{c_1}(X) + w_2 f_{c_2}(X)$ where $w_1, w_2 \in \mathbb{R}$ is $w_1\phi_i^{c_1} + w_2\phi_i^{c_2}$.

Proof.

$$\phi^{w_1c_1+w_2c_2} = \sum_{X' \subseteq X \setminus \{x_i\}} w(X') f_{w_1c_1+w_2c_2}(X') \tag{44}$$

$$= \sum_{X' \subseteq X \setminus \{x_i\}} w(X') \left[w_1 f_{c_1}(X') + w_2 f_{c_2}(X') \right] \tag{45}$$

$$= w_1 \sum_{X' \subseteq X \setminus \{x_i\}} w(X') f_{c_1}(X') + w_2 \sum_{X' \subseteq X \setminus \{x_i\}} w(X') f_{c_2}(X') \tag{46}$$

$$= w_1\phi_i^{c_1} + w_2\phi_i^{c_2} \tag{47}$$

C Detailed Experimental Setup

TRN. We extract 256D features for every frame in Something-something v2 using the publicly available 8 frame multi-scale TRN[1] (this uses a BN-Inception [2] backbone). We then train a set of MLP classifiers $\{f^s\}_{s=1}^{n_{\max}=8}$ with one hidden layer of 256 units, dropout of 0.1 on the input features, and ReLU activation. Each MLP f^s takes in the concatenation of s frame features and produces class scores. This is the same as the multi-scale variant of TRN, however, rather than training the classifiers jointly, we train them separately and then combine their results at inference time through Eq. (8). We tried training these classifiers jointly, producing a final class score by Eq. (8), however the individual classifiers perform very poorly when tested in isolation, and the overall performance was improved by training them separately. We train for 30 epochs with a batch-size of 512 and learning rate of 1e-3 using Adam to optimise parameters.

TSN. We train TSN [3] with a ResNet-50 backbone [4]. We introduce a bottleneck FC layer mapping from the 2048D features from the GAP layer to 256D, we then add a FC layer atop of this for classification. The Element Shapley Values for TSN can be obtained since the model is inherently capable of operating on variable length sequences. This means we can directly evaluate $f_c(X')$ for all subsequences X'. When scaling up to a large number of frames, we use the approximation technique as specified in Algorithm 1 but lines 7-12 are replaced with

$$\mathcal{F}_j^s \leftarrow f_c(\mathcal{X}_j^s)\,. \tag{48}$$

This adaptation to the algorithm is the same for any other model already supporting a variable length input.

The accuracy of both models on the validation set is shown in Fig. 10 across a range of different sequence lengths where frames are uniformly sampled from the video.

[1] `https://github.com/zhoubolei/TRN-pytorch`

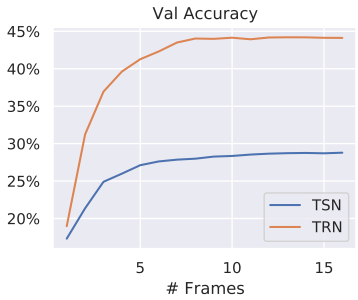

Fig. 10. TRN and TSN validation accuracy by number of frames input to the model.

D Computational Cost

The computational cost of the Element Shapley Value without approximation is $\mathcal{O}(2^{|X|})$ in the number of forward passes. When we apply our approximation method presented in Algorithm 1 we reduce that complexity to $\mathcal{O}(mi|X|)$ where m is the maximum number of sequences sampled per scale and i is the number of iterations of the approximation. Practically speaking, we can compute Element Shapley Values for both TSN and TRN exactly in reasonable time (under 10s/example) thanks to an optimised batched GPU implementation for up to 16 frames. Beyond this, the exponential scaling results in prohibitively long runtimes. Our approximation enables us to scale to more than 16 frames. We present a runtime analysis on a single NVIDIA 1080Ti in Appendix D.

Table 1. Runtime analysis of exact and approximate ESV on sequences of varying length. When applying our approximation algorithm (Algorithm 1) we have to choose m (max number of subsequences/scale) and i (number of approximation iterations).

Model	No. of frames	Configuration	Time (s)
TRN	8	Exact	$.080 \pm 0.0001$
	16	Exact	6.79 ± 0.01
	16	m = 1024, i = 1	$.122 \pm 0.005$
TSN	8	Exact	$.0043 \pm 0.0007$
	16	Exact	$.035 \pm 0.0001$
	20	Exact	$.614 \pm 0.006$
	20	m = 1024, i = 1	$.116 \pm 0.0005$

E Additional Analysis for Shapley Values using TSN

In this section we present analogous results for TSN, compared to the TRN counterparts in the main paper. Figure 11a reports the average marginal contribution at each scale (cf. Fig. 6a). The trends seen here are much the same as for TRN: as the scale increases, the average marginal contribution decreases. Figure 11b shows the number of supporting frames vs. number of frames input to the model (cf. Fig. 6b). Whilst the trend is similar to that for TRN, the proportion of supporting frames for TSN is much lower. Figure 11c shows the results of the frame shifting experiment for TSN (cf. Figure 6c) also demonstrating temporal smoothness.

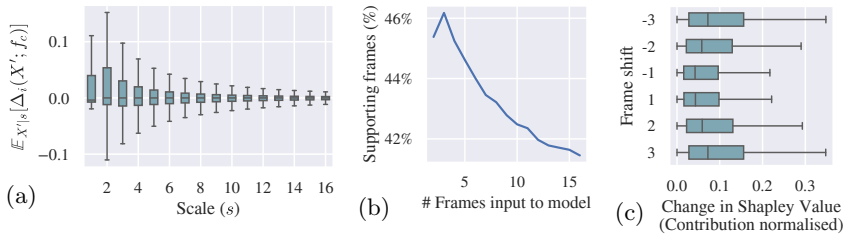

Fig. 11. TSN: (a) Box-plot of marginal contributions at each scale. (b) Increasing the number of frames fed to the model decreases the percentage of supporting frames. (c) Change in ESV when compared to neighbouring frames. Larger shifts result in larger differences in ESV indicating temporal smoothness. (cf. Fig. 6 for TRN).

We conduct the same experiment on frame discarding for TSN as shown for TRN in Fig. 3, presenting results in Fig. 12. Since both GC and IG produce similar attribution values to ESV for TSN, we see less of a performance gap when discarding frames by the attribution ranks in descending order. However, ESV still produces a larger performance gain than GC and IG when removing frames in ascending order of attributions.

Finally, in Figs. 13 and 14, we present heatmaps showing the percentage of videos where $\phi_i^c > \phi_i^c$ for both TSN and TRN. Figure 13 shows classes where the distributions are similar and Figure 14 where they are different.

References

1. Shapley, L.S.: A Value for n-Person Games. In: Contributions to the Theory of Games (AM-28), Volume II. Volume 2. Princeton University Press, Princeton (1953)
2. Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: International Conference on Machine Learning (ICML). (2015) 448–456

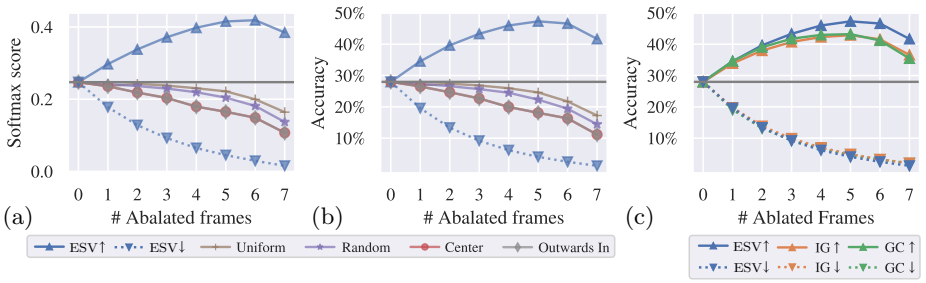

Fig. 12. TSN class score and accuracy after iteratively discarding frames in order of their attribution rank (ascending ▲ vs descending ▼). We compare our method (ESV) to baselines (a,b) and two alternate attribution methods in (c): GradCam (GC) and Integrated Gradients (IG), keeping figures (b) and (c) separate for legibility. (cf. Fig. 3 for TRN).

3. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: Proceedings of the European Conference on Computer Vision (ECCV), Springer (2016) 20–36
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 770–778

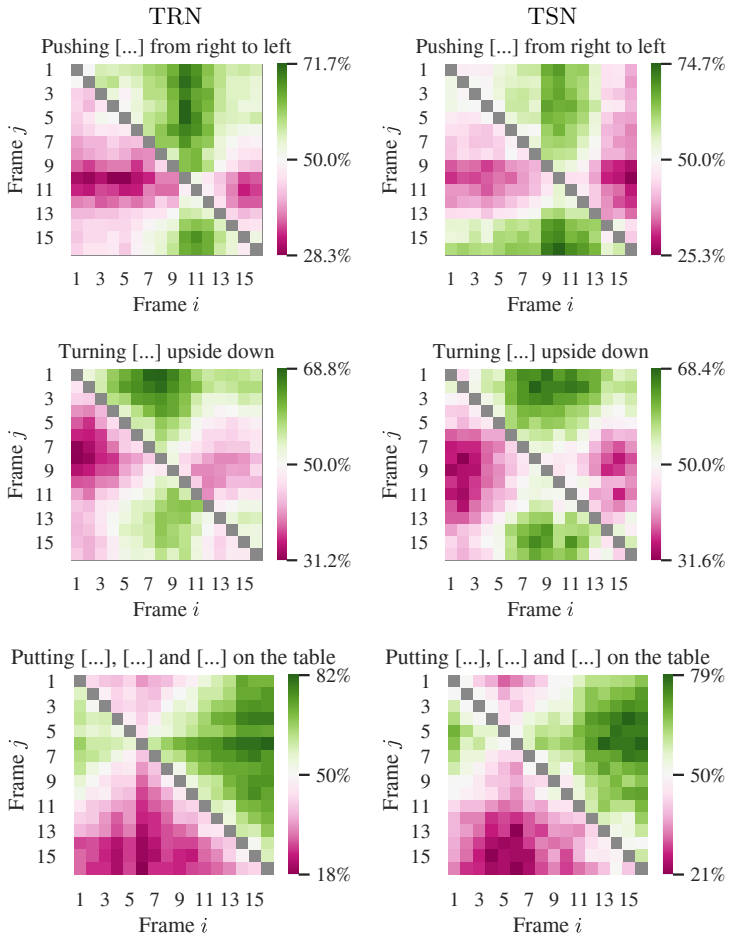

Fig. 13. Comparing the percentage of videos where $\phi_i^c > \phi_j^c$ for TRN (left) and TSN (right). These are classes where the models value frames from the same position similarly.

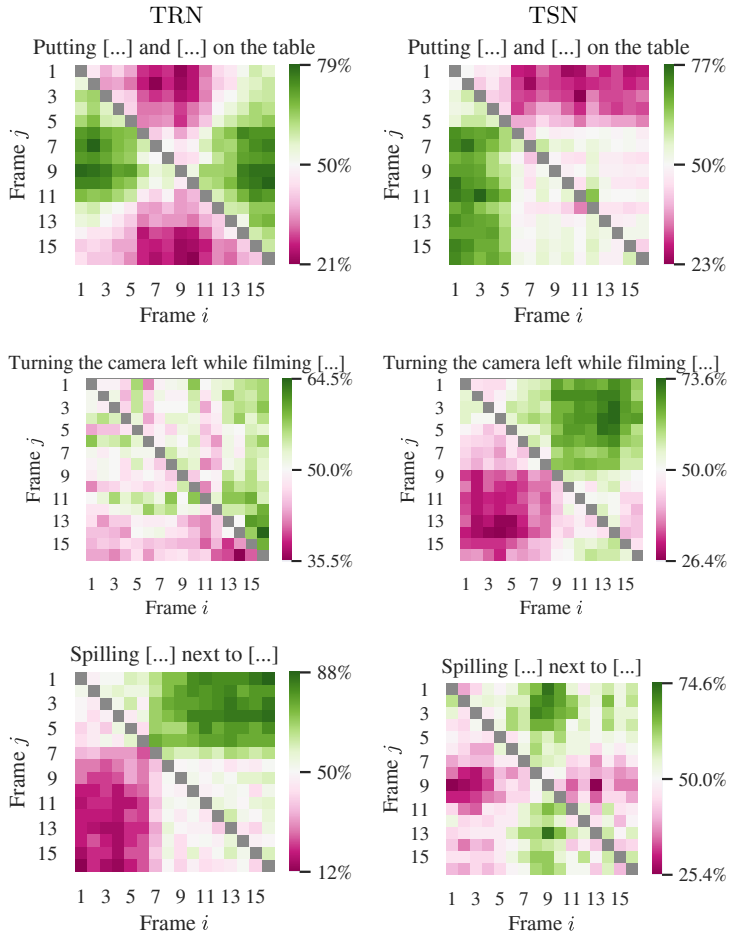

Fig. 14. Comparing the percentage of videos where $\phi_i^c > \phi_j^c$ for TRN (left) and TSN (right). These are classes where the models value frames from the same position differently.