— Supplementary Material —
GliTr: Glimpse Transformers with Spatiotemporal Consistency for Online Action Prediction

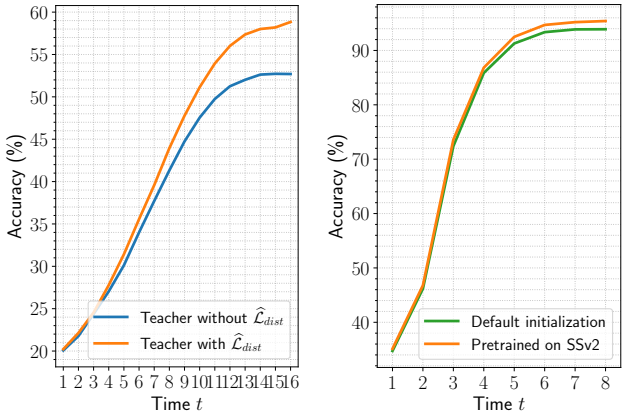
Samrudhdhi B Rangrej[1] Kevin J Liang[2] Tal Hassner[2] James J Clark[1]
[1]McGill University [2]Meta AI
samrudhdhi.rangrej@mail.mcgill.ca

1. Additional Results

Ablation on $\widetilde{\mathcal{L}}_{dist}$. We distill VideoMAE [1] (a transformers-based offline action recognition model) to our teacher model on SSv2 dataset. To do so, we minimize $\widetilde{\mathcal{L}}_{dist}$ *i.e.* KL-divergence between the class distributions predicted by our teacher model and VideoMAE based on complete video (equation 6 in main paper). To assess importance of this objective, we train our teacher model with and without $\widetilde{\mathcal{L}}_{dist}$ and display results in Figure 1(a). We observe improvement of approximately 6% in accuracy at $t = 16$ when $\widetilde{\mathcal{L}}_{dist}$ is included in the training objectives. Note, since a pretrained VideoMAE [1] in unavailable for Jester, we do not use $\widetilde{\mathcal{L}}_{dist}$ for training the teacher model on this dataset.

Ablation on Initialization Scheme. To improve the performance of the teacher model on the Jester dataset, we initialize its parameters using the parameters of the teacher model pretrained on SSv2 with a complete set of training objectives (equation 9 in the main paper), including $\widetilde{\mathcal{L}}_{dist}$. We compare the performance of the above model with the performance of the teacher initialized using default scheme *i.e.* $\mathcal{T}_f$ initialized using an open-source ViT-S model [2] pretrained on the ImageNet, and $\mathcal{T}_c$ and $\mathcal{T}_l$ initialized randomly. The result shown in Figure 1(b) indicates that once finetuned on Jester dataset, the teacher pretrained on SSv2 achieves higher performance than the teacher initialized with default scheme, especially for $t > 4$. Finally, at $t = 8$, the teacher with pretrained weights achieves nearly 1.5% higher accuracy.

Visualization. We show more visual results on example videos from SSv2 in Figure 2.



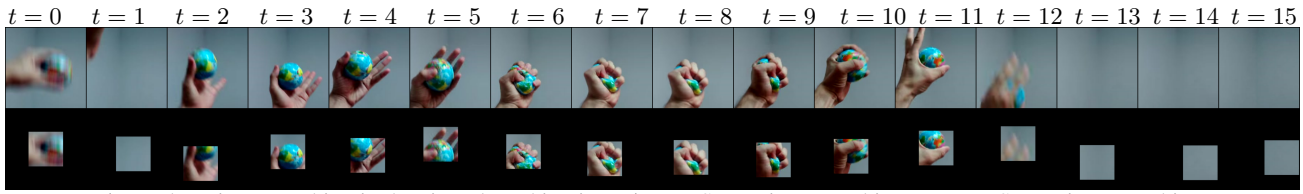
(a) (b)

Figure 1: **(a) Ablation on $\widetilde{\mathcal{L}}_{dist}$** objective for the teacher trained on SSv2 dataset. (b) **Ablation on initialization scheme** for the teacher trained on Jester dataset.

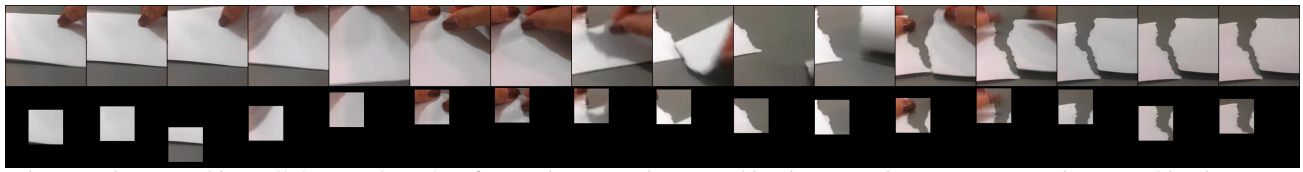
References

[1] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022.

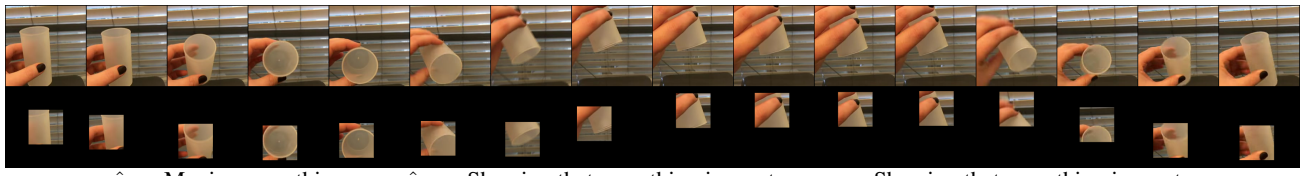
[2] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *International Conference on Learning Representations*, 2021.



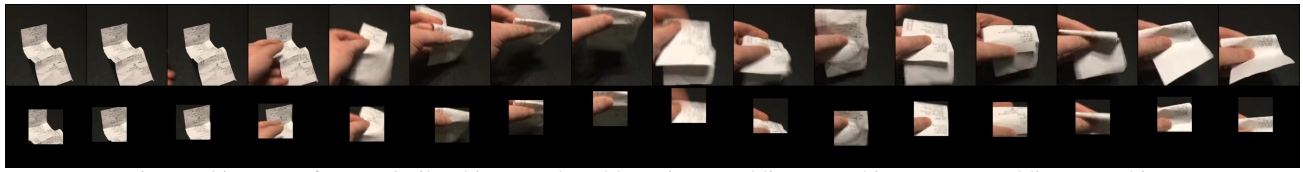
$\hat{y}_{1:9}$=Throwing something in the air and catching it; $\hat{y}_{10:16}$=Squeezing something; y_{GT}=Squeezing something



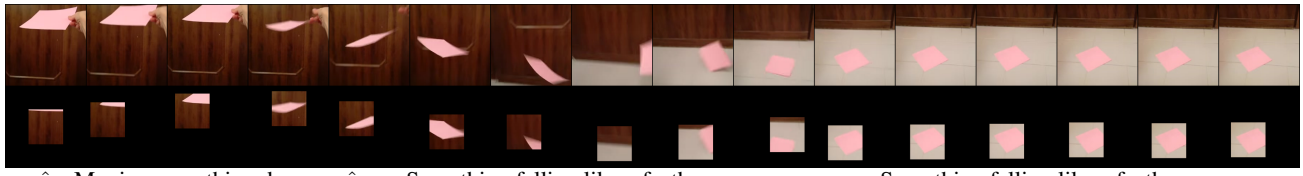
$\hat{y}_{1:4}$=Letting something roll down a slanted surface; $\hat{y}_{5:16}$=Tearing something into two pieces; y_{GT}=Tearing something into two pieces



$\hat{y}_{1:2}$=Moving something up; $\hat{y}_{3:16}$=Showing that something is empty; y_{GT}=Showing that something is empty

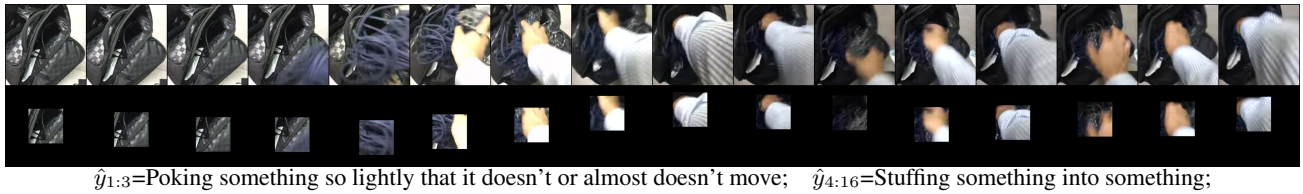


$\hat{y}_{1:4}$=Taking one of many similar things on the table; $\hat{y}_{5:16}$=Folding something; y_{GT}=Folding something

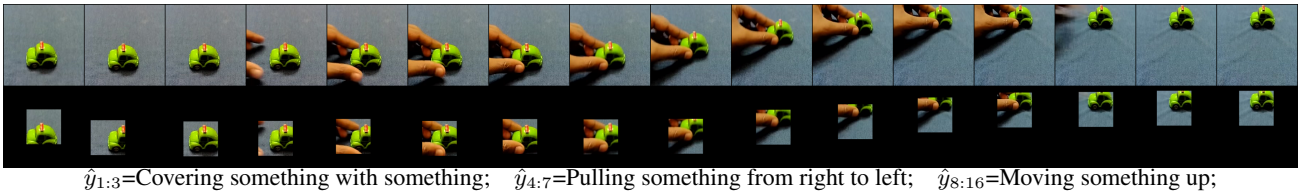


$\hat{y}_{1}$=Moving something down; $\hat{y}_{2:16}$=Something falling like a feather or paper; y_{GT}=Something falling like a feather or paper

$\hat{y}_{1:8}$=Folding something; $\hat{y}_{9:16}$=Unfolding something; y_{GT}=Unfolding something



$\hat{y}_{1:3}$=Poking something so lightly that it doesn't or almost doesn't move; $\hat{y}_{4:16}$=Stuffing something into something; y_{GT}=Stuffing something into something



$\hat{y}_{1:3}$=Covering something with something; $\hat{y}_{4:7}$=Pulling something from right to left; $\hat{y}_{8:16}$=Moving something up; y_{GT}=Moving something up

Figure 2: Visualization of glimpses (bottom rows) selected by GliTr on SSv2 dataset. Complete frames (top rows) are shown for reference.