

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

Samrudhdi B Rangrej¹ Kevin J Liang² Tal Hassner² James J Clark¹
¹McGill University ²Meta AI
samrudhdi.rangrej@mail.mcgill.ca

1. Additional Results

Ablation on $\tilde{\mathcal{L}}_{dist}$. We distill VideoMAE [1] (a transformers-based offline action recognition model) to our teacher model on SSv2 dataset. To do so, we minimize $\tilde{\mathcal{L}}_{dist}$ *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 $\tilde{\mathcal{L}}_{dist}$ and display results in Figure 1(a). We observe improvement of approximately 6% in accuracy at $t = 16$ when $\tilde{\mathcal{L}}_{dist}$ is included in the training objectives. Note, since a pretrained VideoMAE [1] is unavailable for Jester, we do not use $\tilde{\mathcal{L}}_{dist}$ 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 $\tilde{\mathcal{L}}_{dist}$. 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.

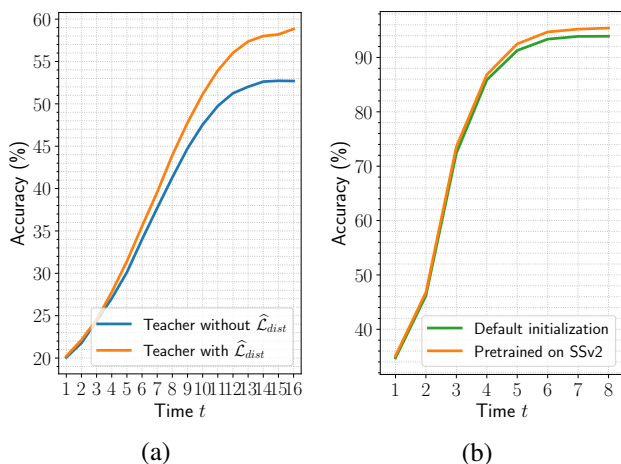


Figure 1: (a) Ablation on $\tilde{\mathcal{L}}_{dist}$ objective for the teacher trained on SSv2 dataset. (b) Ablation on initialization scheme for the teacher trained on Jester dataset.

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.

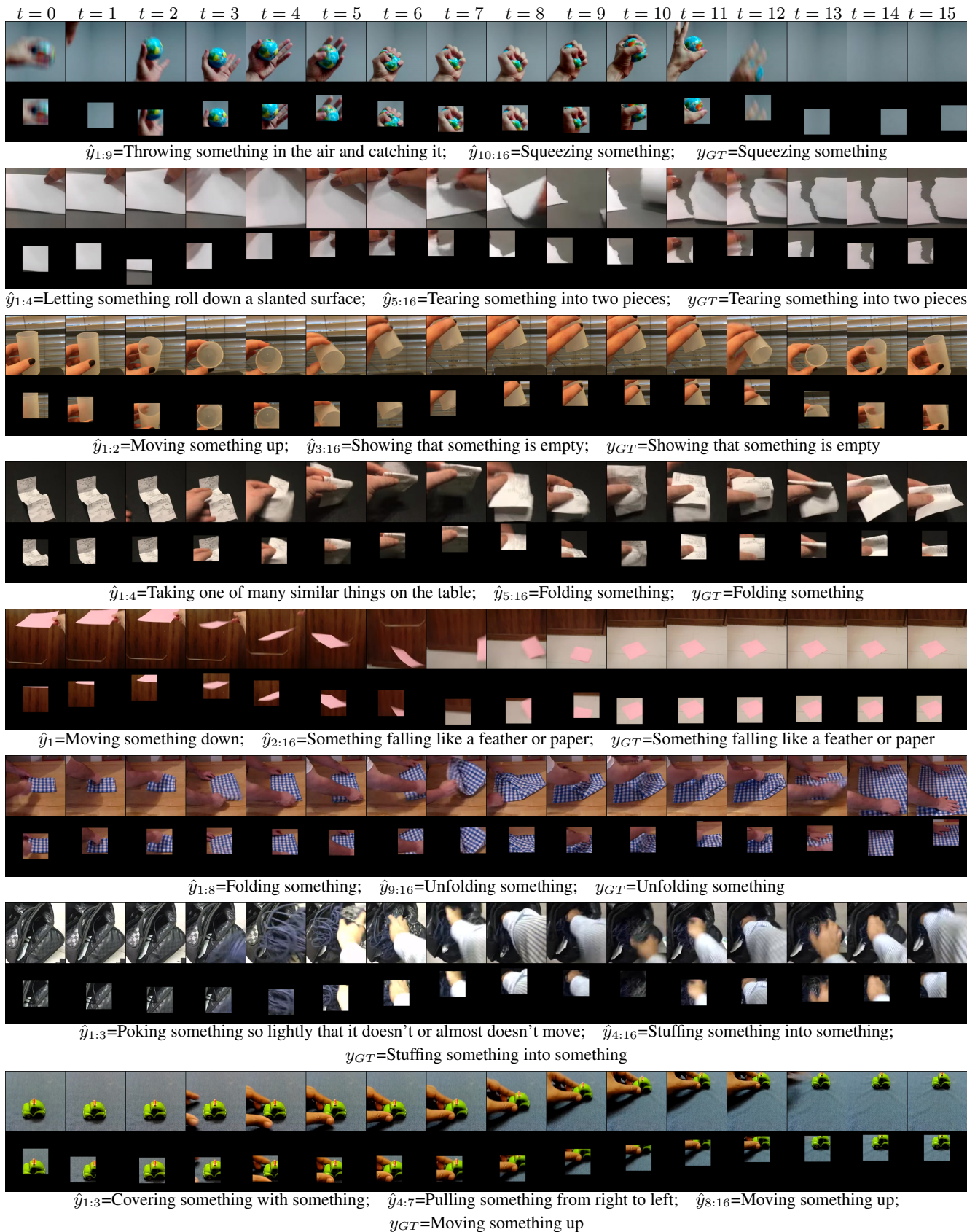


Figure 2: Visualization of glimpses (bottom rows) selected by GliTr on SSv2 dataset. Complete frames (top rows) are shown for reference.