A. View Routing and Matching

In the main paper (Sec. 3.2), we illustrate the concept of SVT using a single global view passed through the teacher, which generates target for all the other views passed through the student model. However, in practice, multiple global views are all passed through the teacher model, and we separately map each student view (global and local) to the multiple teacher targets. In the case of two global views, $g1$ ($T = 8$) and $g2$ ($T = 16$), we obtain two targets, $\tilde{f}_{gt}^{(1)}$ and $\tilde{f}_{gt}^{(2)}$. Both these global views are also passed through the student model to obtain $\tilde{f}_{gs}^{(1)}$ and $\tilde{f}_{gs}^{(2)}$. We map $\tilde{f}_{gs}^{(1)}$ to $\tilde{f}_{gt}^{(2)}$ and $\tilde{f}_{gs}^{(2)}$ to $\tilde{f}_{gt}^{(1)}$. The local views passed through the student that generates $\tilde{f}_{ls}^{(1)} \dots \tilde{f}_{gs}^{(8)}$ which are separately mapped to both teacher targets, $\tilde{f}_{gt}^{(1)}$ and $\tilde{f}_{gt}^{(2)}$. Our proposed loss is applied over each mapped student-teacher feature pair.

B. Comparison to Supervised Training

In SVT, we use a standard ViT backbone with split attention across space and time dimensions similar to [9]. We compare SVT with supervised pre-training based initialization for Kinetics-400 training reported in [9]. For fairness, our comparison with [9] includes the highest reported input resolution used in their work since the SVT uses slow-fast inference. These results are presented in Table I.

Table I. Comparison of SVT with supervised pretraining methods containing similar backbone (ViT-B) on Kinetics-400. For each different pre-training strategy, we finetune on Kinetics-400 and report accuracy (top-1) on Kinetics-400 validation set.

Pretrain Dataset	Supervision	Accuracy
Random-init	-	64.8 [9]
ImageNet-1K	✓	75.8 [9]
ImageNet-21K	✓	79.7 [9]
ImageNet-1K	✗	69.9
Kinetics-400	✗	78.1

C. Dataset Description

We use the Kinetics-400 [14] training set for the SVT self-supervised training and its validation set for evaluation of learned self-supervised representations. Kinetics-400 is a large-scale dataset containing 240k training videos and 20k validation videos belonging to 400 different action classes. On average, these videos are of duration around 10 seconds, with 25 frames per second (*i.e.*, around 250 frames per video). Interestingly, most classes of this dataset are considered to be separable with appearance information alone [92]. In addition to Kinetics-400, we evaluate our approach on three downstream datasets, UCF-101 [69], HMBD-51 [49], and Something-Something v2 (SSv2) [33]. UCF-101 and HMBD-51 are small-scale datasets each containing 13k videos (9.5k/3.7k train/test) belonging to 101 classes and 5k (3.5k/1.5k train/test) videos belonging to 51 classes respectively, while SSv2 is a large-scale dataset heavily focused on motion with 168k training and 24k validation videos belonging to 174 action classes. Unlike UCF101 and HMDB51 which contain action classes similar to Kinetics-400, SSv2 contains very different actions involving complex human object interactions, such as 'Moving something up' or 'Pushing something from left to right'.

D. Future Directions

As discussed in the main paper, the key limitation of SVT is being constrained to operating within a single modality input (RGB video). We hope to explore how SVT to can improved to utilize alternate modalities (Optical Flow, Audio) for better self supervision in future work.

In this work, we focus on evaluating the effectiveness of our proposed cross-view and motion correspondences (that compose the core of SVT) in relation to ViT backbones. The question of applicability of our proposed approach under convolutional neural network (CNN) settings remains unexplored. However, we highlight that the main components (temporal attention, dynamic input sizes, and slow-fast inference) of our proposed SVT are designed to leverage some unique characteristics of ViTs, which could not be directly implemented with a CNN backbone. On the other hand, we believe that self-distillation and view matching, also core to SVT, can be applied to CNNs and is an interesting future direction.

Another key issue is the significant drop in performance (top-1 accuracy) for linear evaluation in large-scale datasets (Kinetics-400 and SSv2). Particularly in SSv2, our features perform poorly in the linear evaluation setting (in comparison to fine-tune setting). A key reason for this could be the significant domain difference between Kinetics-400 and SSv2 (as opposed to UCF-101 and HMDB-51 which contain videos and classes similar to Kinetics-400). The self-supervised training phase of SVT uses Kinetics-400 only, and the SSv2 experiments use that representation for linear evaluation. An interesting future direction we hope to explore is self-supervised training using the SSv2 dataset itself, which could potentially reveal more interesting insights on representations learned by SVT.

E. Attention Visualization

Following the approach in [13], we visualize the attention of our classification token (feature vector) towards each spatiotemporal patch token within the final encoder block of SVT for two randomly selected videos. As illustrated in

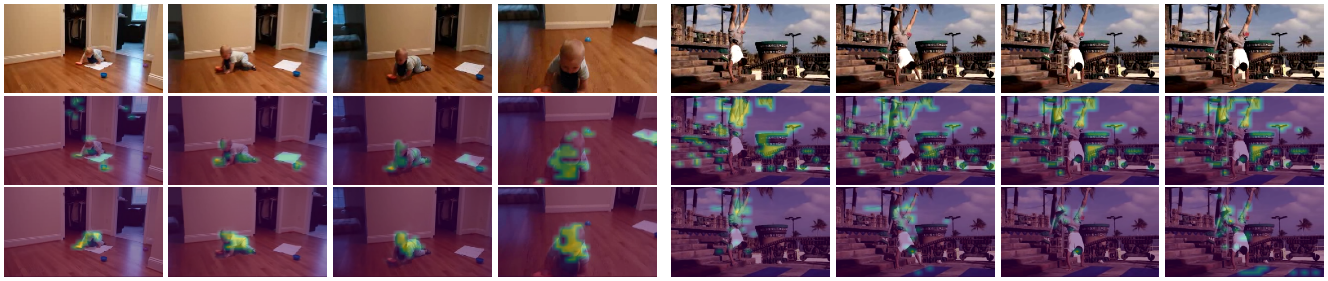

Figure I. **Attention Visualization:** We uniformly sample four frames from two videos (cols 1-4 and 5-8 respectively) and visualize the attention from the classification token of self-supervised vision transformer DINO [13] (second row) and our SVT (last row). Observe how DINO attention is scattered around multiple objects, while SVT is focused on *'crawling baby'* and *'person walking on hands'* across frames which are the salient objects for these action. This highlights how SVT learns to pay attention to the motion within a video.

Fig. I, SVT attends to the regions of motion in these videos, even in the case of highly detailed backgrounds (right). Attention to the salient moving object in each case qualitatively demonstrates how our proposed cross-view and motion correspondences learn spatiotemporally invariant representations.