Test of Time: Instilling Video-Language Models with a Sense of Time

SUPPLEMENTARY MATERIAL

Piyush Bagad
University of Amsterdam

Makarand Tapaswi
IIIT Hyderabad

Ceës G.M. Snoek
University of Amsterdam

bpiyush.github.io/testoftime-website

As part of the supplementary material, we describe preprocessing steps as well as some qualitative examples from the datasets in Appendix A. In Appendix B, we present additional ablations on what makes temporal adaptation hard. This expands on the last paragraph of Sec. 5 of the main paper. Finally, in Appendix C, we conduct a qualitative analysis to verify if the model has indeed learnt to connect the time order.

A. Datasets and Pre-processing

We sketch out the procedure we use for stitching two clips within a video.

Clip stitching. Consider a video containing two events (clips) \( v_i, v_j \) with associated captions \( \zeta_i, \zeta_j \) as shown in Fig. 1. We assume these are non-overlapping (in time). We stitch the text descriptions to construct a new caption \( t_{ij} := [\zeta_i; \tau; \zeta_j] \). Since \( \tau \) can be either before or after, we end up with two newly constructed sentences. Corresponding to each of these new sentences, we also stitch the video events to construct a stitched video. Note that the order of stitching video events depends on the value of \( \tau \). For example, if \( \tau \) is before, then \( u_{ij} := [v_i; v_j] \) as shown in the first of the two stitched clips. If \( \tau \) is after, then \( u_{ij} := [v_j; v_i] \) as shown in the second of the two stitched clips.

From each stitched clip in Fig. 1, we construct negatives for the contrastive loss by reversing the time order in either video or text. This step happens on-the-fly during loss computation, and hence, we do not show it here. For a given dataset, we can either use all possible tuples of non-overlapping events to create such stitched clips or sample from all possible tuples. Since the TEMPO dataset already comes with stitched event descriptions (based on DiDeMo), we directly use its subset which has before/after relations in the text. For all the other datasets, we apply the stitching process as described. Recall, \( \Delta_{time} \) is the time distance between the two events, and plays a key role in deciding the difficulty of temporal adaptation, as observed empirically.

Next, we describe dataset properties and show some qualitative examples after the clip stitching step.

Adaptation datasets. To gain a sense of the diversity in the datasets we consider for adaptation, we present examples of stitched clips from these datasets in Fig. 3. Since TEMPO has short adjacent clips, the context remains almost the same, we think this is important to instill a sense of time in models. In contrast, for ActivityNet, since the stitched events are far apart, the context changes make it easy to infer which event description goes with which part of the video, or the time order of events. In this regard, Charades and Charades-Ego are similar to TEMPO. Quantitatively, this change in context is captured by \( \Delta_{time} \) which is lowest for TEMPO (mean 6.8s), followed by Charades-Ego (13.3s), Charades (14.5s) and ActivityNet (58.8s).

Distribution of number of clips in a video. A single video with 10 non-overlapping individual event clips can make up to \( 10C_2 = 45 \) stitched clips. We plot the number of clips per video against the number of videos in a given dataset in Fig. 2. A single video with \( > 30 \) stitched clips is rare in TEMPO and ActivityNet while much more frequent in Charades and Charades-Ego. Overall, the number of clips per video is lower in TEMPO and ActivityNet as compared to Charades and Charades-Ego.

Downstream datasets. In Fig. 4, we also show some examples from some downstream datasets (tasks) that need
higher time awareness since they typically involve multiple temporally linked events (e.g., walk and eat in Fig. 4(b)). On these datasets, we perform zero-shot evaluation of temporally adapted models in Sec. 6 of the main paper.

B. Experiments

Analyzing more pretrained models. We present preliminary experimental results for other pretrained models on the Charades dataset in Tab. 1. The other models perform (slightly) better than random, but are not as promising as VideoCLIP. We observe similar trends in performance on the TEMPO dataset. We hypothesize that VideoCLIP’s larger temporal receptive field and contrastive pre-training objective similar to TACT helps it achieve superior performance. This merits further investigation into how various factors (as tabulated in Tab. 1) influence temporal adaptation.

Spatial vs. temporal understanding. An interesting facet of TACT is $\alpha_{\text{same}}$ which controls how well a model adapts to temporal tasks. We highlight this on the TEMPO dataset in Tab. 2, where, $\alpha_{\text{same}}=0$ results in $A_{\text{time}} \sim 50\%$ while $\alpha_{\text{same}}=1$ improves performance. Further investigation on downstream tasks shows that adaptation with $\alpha_{\text{same}}=1$ does not perform well on MSR-VTT (a non-temporal benchmark) but shows consistent improvements on AGQA (a temporal benchmark),and the trade-off between spatial- and temporal-understanding. This hints at $\alpha_{\text{same}}$ controlling the trade-off between spatial and temporal understanding.

What makes temporal adaptation difficult? To recall, we define $\Delta_{\text{time}}$ as the time-distance (in seconds) between the midpoints of the two clips in a stitched pair. We hy-

<table>
<thead>
<tr>
<th>Model</th>
<th>Temporal receptive field</th>
<th>Pre-training strategy</th>
<th>Visual backbone</th>
<th>Encoder</th>
<th>$A_{\text{time}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frozen [1]</td>
<td>4</td>
<td>Contrastive</td>
<td>TimeSFormer</td>
<td>Multimodal</td>
<td>53.0</td>
</tr>
<tr>
<td>VideoCLIP [4]</td>
<td>32</td>
<td>Contrastive</td>
<td>BERT on S3D</td>
<td>Two-tower</td>
<td>77.0</td>
</tr>
</tbody>
</table>

Table 1. Adaptation results for more pre-trained models on Charades. Models with smaller temporal receptive field perform worse in comparison to VideoCLIP. The temporal receptive field is reported in terms of the number of input frames. Systematically understanding the influence of various factors on making models time-aware by post-pretraining makes for interesting future work.
Table 2. Impact of $\alpha_{\text{same}}$ on spatial- vs. temporal understanding. Gray denotes better performance for $\alpha_{\text{same}}=0$ or 1. While $\alpha_{\text{same}}=1$ drives temporal understanding, it comes at a cost of retrieval performance on MSR-VTT [5]. This hints at $\alpha_{\text{same}}$ controlling the trade-off between spatial- and temporal-understanding.

C. Qualitative Analysis

To get an intuitive sense of whether a TACT model understands time order of events, we perform a qualitative analysis on the model trained on TEMPO. Our demo interface looks like the one shown in Fig. 6. First, a user uploads a video and adds text descriptions for two events within the video. These descriptions are then connected via a temporal relation such as before or after. We also experiment with

---

**Figure 4.** Examples from datasets used for downstream evaluation. These tasks demand time awareness since it is often not possible to infer the action from a single frame.
First, we consider samples from the TEMPO validation set and show their results in Fig. 7. Notably, for some examples, it connects time order for before relations but not the other two. We suspect this is because a majority (~60%) of the TEMPO dataset has descriptions involving before. Note that TEMPO already comes with temporal captions of which we pick subset of before/after relations. Second, we also consider samples from other datasets which the model has never seen. To our surprise, albeit qualitatively, the model does generalize well to such examples as shown in Fig. 8.

These results reinforce the promise of our method and also raise the possibility of extending this work to consider more general temporal relations. Having said that, we reiterate that these are qualitative examples and should be treated as such.

References


Test of Time: Instilling Video-Language Models with a Sense of Time

Rank sentences based on their relevance to a video

The child runs into the room before he sits near the gifts

he sits near the gifts before the child runs into the room

The child runs into the room before he sits near the gifts

he sits near the gifts before the child runs into the room

Figure 6. Interface of our demo for qualitative analysis. The user uploads a video and is asked to describe two events in the video. These event descriptions are then connected via one of the three temporal relations shown at the bottom left. We construct one sentence that is consistent with the time order of events in the video and another that is not. The output on the right shows the ranking of the constructed sentences in terms of cosine similarity with the video representation. Higher score for correct matching indicated by a longer orange bar.
Figure 7. Qualitative examples from TEMPO validation set. We evaluate similarity of a given video with sentences with different temporal order with the usual temporal connectors (before/after). Green bordered boxes indicate correct predictions (consistent time order between video and language) while red denote mispredictions. For some examples, e.g., in the bottom example, the model gets predictions incorrect particularly for relations other than before. Furthermore, we also try a new temporal connector First, ..., then, ... and observe that the model qualitatively generalizes to that as well.

Figure 8. Qualitative results on samples not from TEMPO. We see that despite not having seen these examples, the model still connects the time order across video and language correctly.