

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

SUPPLEMENTARY MATERIAL

Piyush Bagad
University of Amsterdam

Makarand Tapaswi
IIIT Hyderabad

Cees G.M. Snoek
University of Amsterdam

bpiyush.github.io/testoftime-website

As part of the supplementary material, we describe pre-processing 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 ζ_i, ζ_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 τ 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 τ . For example, if τ is before, then $u_{ij} := [v_i; v_j]$ as shown in first of the two stitched clips. If τ 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, Δ_{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.

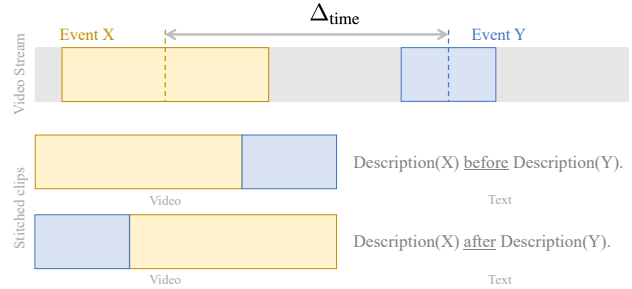


Figure 1. Illustration of clip stitching. We consider two non-overlapping events in a video and stitch them with temporal relations - before and after. Δ_{time} denotes the time difference between midpoints of the two events.

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 Δ_{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 upto ${}^{10}C_2=45$ 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

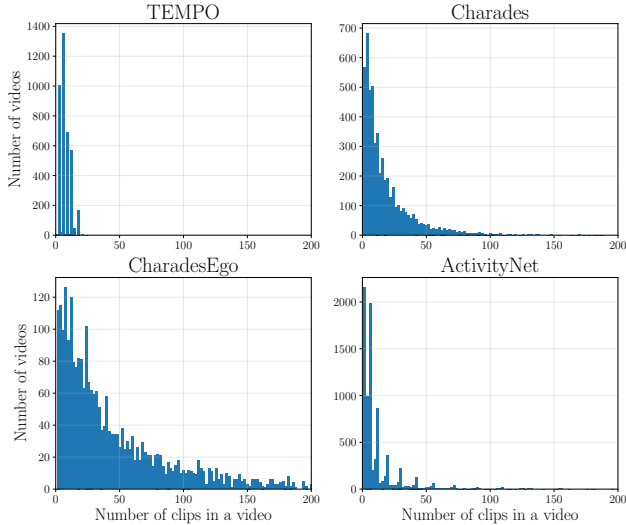


Figure 2. Number of clips in a video. The number of clips per video is lower in TEMPO and ActivityNet as compared to Charades and Charades-Ego.

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

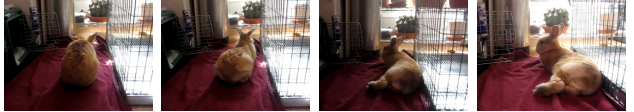
Analyzing more pretrained models. We present preliminary experimental results for other pretrained models on the Charades dataset in Tab. 1. The other models perform (slightly) better than random, but are not as promising as VideoCLIP. We observe similar trends in performance on the TEMPO dataset. We 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

Model	Temporal receptive field	Pre-training strategy	Visual backbone	Encoder	A_{time}
Frozen [1]	4	Contrastive	TimeSformer	Multimodal	53.0
VindLU [2]	8	Autoencoding	ViT+Temp. attn.	Multimodal	54.1
CLIP4Clip [3]	12	Contrastive	ViT+Temp. attn.	Two-tower	57.5
VideoCLIP [4]	32	Contrastive	BERT on S3D	Two-tower	77.0

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.

A rabbit lays down on its stomach before bunny lying on it’s side



Little girl eats from cup after the child walks downhill



(a) TEMPO

A woman is standing in a room holding a hula hoop before she begins to use the hula hoop

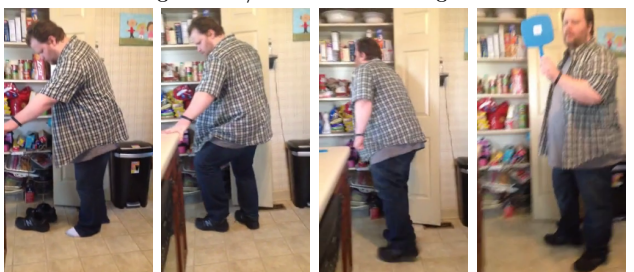


The team shakes hands with the opposing team after a team groups together holding a trophy



(b) ActivityNet

Putting on shoe/shoes before holding a mirror



(c) Charades

Taking a broom from somewhere before holding a dish



(d) Charades-Ego

Figure 3. Examples from datasets used for temporal adaptation. The first two frames are linearly spaced from the first event while the last two from the second event. Notice how there is a significant change in visual context between the two events in ActivityNet in contrast to other datasets. Best viewed on a screen.

of TACT is α_{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 α_{same} controlling the trade-off between spatial and temporal understanding.

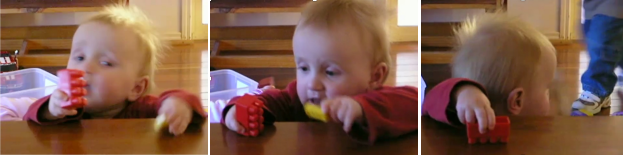
What makes temporal adaptation difficult? To recall, we define Δ_{time} as the time-distance (in seconds) between the midpoints of the two clips in a stitched pair. We hy-

Question: How did the boy react when he entered the room at the start?



Answer: Smile.

Question: Why does the baby turn around near the end of the video?



Answer: Sits down to play.

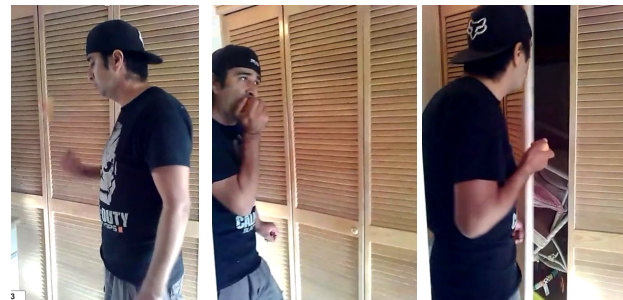
(a) Next-QA: Video question answering

Question: Did they reach for and grab a picture before or after putting a bag somewhere?



Answer: Before

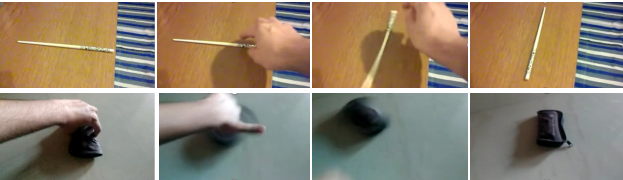
Question: Did they walk through a doorway before or after they eating the last thing they touched?



Answer: After

(b) AGQA: Video question answering

Template: Spinning [something] that quickly stops



(c) Something-Something: Template-based video retrieval

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.

Hyperparameters			Adaptation	Downstream		
α_{same}	α_{cross}	β	TEMPO $A_{\text{time}} \uparrow$	MSR-VTT $R@1 \uparrow$	MedR \downarrow	AGQA Accuracy \uparrow
0	0	0	49.4	15.0	20.0	50.5
0	0	1	49.5	14.2	20.0	49.9
0	1	0	49.3	14.4	19.0	50.2
0	1	1	49.5	15.1	19.0	50.2
1	0	0	60.6	11.7	27.0	56.6
1	0	1	62.9	9.4	36.0	58.3
1	1	0	59.7	9.1	37.0	56.9
1	1	1	62.5	12.8	27.0	57.1

Table 2. Impact of α_{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 α_{same} controlling the trade-off between spatial- and temporal-understanding.

pothesize that Δ_{time} is inversely related to the difficulty of temporal adaptation, *i.e.*, the larger Δ_{time} , the easier it is to distinguish between two stitched clips that have opposite time order. For example, consider ActivityNet examples in Fig. 3(b) where the visual context changes significantly making inference of the time order of events relatively easier.

We further test our hypothesis by sampling individual clips from the Charades-Ego dataset to match the Δ_{time} distribution of TEMPO. Concretely, assuming Δ_{time} for both these datasets follows a multinomial distribution, we construct a new distribution using a convex combination of the individual distributions where the mixing parameter $\lambda \in [0, 1]$ controls the extent to which we modify the distribution from TEMPO ($\lambda=0$) to Charades-Ego ($\lambda=1$). The resulting distributions are presented in Fig. 5 (left). With $\lambda=1$, we sample from the original Charades-Ego distribution and gradually move towards TEMPO as $\lambda \rightarrow 0$.

We then sample stitched clips according to this new distribution and post-pretrain temporal adaptation for varying values of λ . Note that for this experiment, we keep fixed $N_c=10,000$ for each λ . From Fig. 5 (right), we indeed find that as we move towards a more TEMPO-like distribution (shorter Δ_{time}), temporal accuracy deteriorates. The best fit also confirms that the distribution of Δ_{time} is strongly correlated ($\rho = -0.92$) with the difficulty of inferring time-order consistency.

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

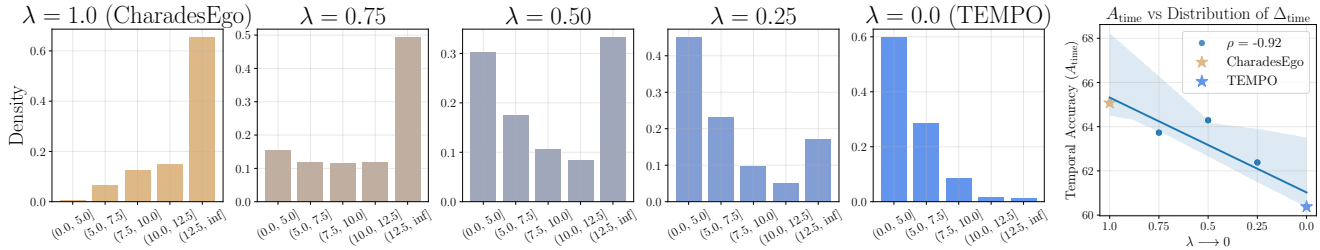


Figure 5. Impact of changing distribution of Δ_{time} , the time gap between two stitched clips. **Left:** We vary the distribution of Δ_{time} for Charades-Ego and make it similar to that of TEMPO as $\lambda \rightarrow 0$. Thus, crudely, as λ decreases, so does Δ_{time} . **Right:** A_{time} on Charades-Ego where the time difference between sampled clips is according to the distributions on the left. We observe that the accuracy deteriorates as the time-distance between a pair of clips decreases indicating a strong correlation between the distribution of Δ_{time} and difficulty of temporal adaptation.

a new temporal connector First, ..., then, to check if our model generalizes beyond before/after.

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 ($\sim 60\%$) 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


- [1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [2] Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. Vindlu: A recipe for effective video-and-language pretraining. *arXiv preprint arXiv:2212.05051*, 2022. 2
- [3] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 2
- [4] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 6787–6800, 2021. 2
- [5] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language.

In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

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

Rank sentences based on their relevance to a video

Video (stitched with two events)



Refresh video. Check this if you load a new video.

Write a description for event X (any event within the video)

The child runs into the room

Write a description for event Y (any event within the video)

he sits near the gifts

Choose a relation between the two events

before after First,... then...

Clear Submit

Constructed sentence 1

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

Constructed sentence 2

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

Ranking over sentences

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

The child runs into the room before he sits near the gifts	54%
he sits near the gifts before the child runs into the room	46%

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.

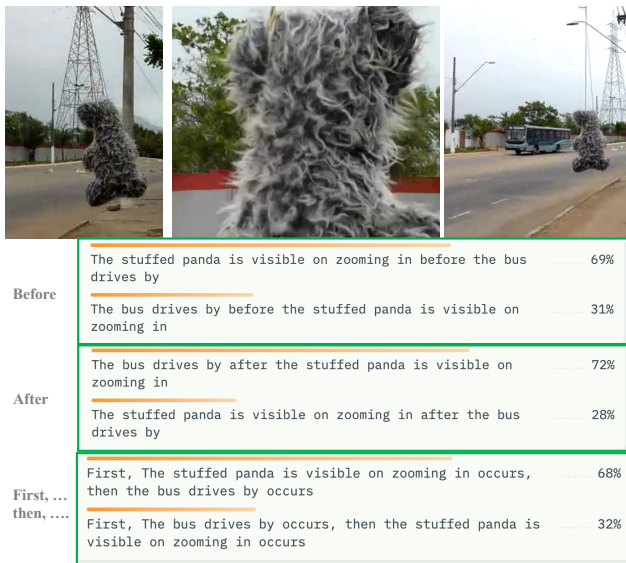
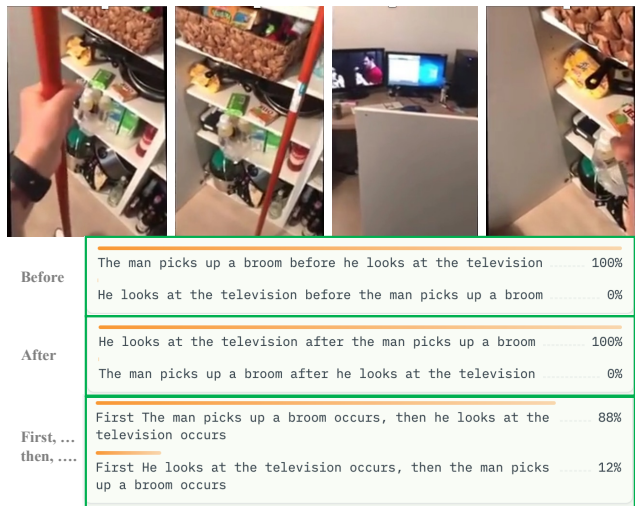


Figure 7. Qualitative examples from TEMPO validation set. We evaluate similarity of a given video with sentences with different temporal order with the usual temporal connectors (before/after). Green bordered boxes indicate correct predictions (consistent time order between video and language) while 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.



(a) Example from Charades-Ego



(b) Example from Next-QA

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.