Do Video-Language Foundation Models have a Sense of Time?

Piyush Bagad
University of Amsterdam

Makarand Tapaswi
IIIT Hyderabad

Cees G. M. Snoek
University of Amsterdam

Abstract

Modelling and understanding time remains a challenge in contemporary video understanding models. Time also appears in language through temporal relations. Video-language models can benefit from having a sense of time, especially since language provides an interface for generalization. In this paper, we consider a specific aspect of temporal understanding: consistency of time order as elicited by before/after relations. We construct a simple synthetic dataset to measure such temporal understanding in video-language models and find that six existing models struggle to understand even such simple relations. We then posit whether it is feasible to equip these foundation models with temporal awareness without re-training them from scratch. Towards this, we propose a temporal adaptation recipe on top of one such model, VideoCLIP, based on post-pretraining on a small amount of video-text data. Our work serves as a first step towards probing and instilling a sense of time in existing video-language models without needing data- and compute-intense training from scratch. Project page: https://bpiyush.github.io/testoftime-website.[1]

1 Introduction

Self-supervised pretraining on multimodal web corpora tied with powerful architectures (Vaswani et al. (2017)) has led to foundation models (Bommasani et al. (2021)) for images (Radford et al. (2021); Ramesh et al. (2021)) and videos (Xu et al. (2021); Alayrac et al. (2022)). These models have enabled remarkable improvements on a plethora of downstream tasks, particularly, video-language tasks such as retrieval and question-answering. Given the cost and difficulty of video annotations, even for a small amount of downstream data, such foundation models are emerging as the de-facto backbone for zero-shot visual understanding tasks (Xu et al. (2021); Zeng et al. (2022); Alayrac et al. (2022)). However, it remains unclear if these video-language models capture essential properties of a video beyond what can be learned from static images, most notably: *time*.

Many before us have shown that existing video-language models (Xu et al. (2021); Bain et al. (2021a); Luo et al. (2021)) can achieve impressive performance on several video benchmarks without reliably encoding time (Buch et al. (2022); Lei et al. (2022); Li et al. (2022)). Buch et al. (2022) show that a model that uses a single (carefully selected) frame often outperforms recent video-language models on standard video benchmarks such as MSR-VTT (Xu et al. (2016)). Lei et al. (2022) report similar findings with a single-frame pretraining approach. These findings hint at a lack of time awareness in video models. However, it remains unclear if these findings are caused by the lack of time in video models or whether the benchmarks themselves do not mandate time awareness. Furthermore, there is no clear definition of what it means for a model to be time aware. In this paper we strive to shed light on all these factors of time awareness in video-language models.

As a first step, we consider a simple notion of understanding time through temporal relations such as *before* and *after* (Allen (1984)). Consider a pair of visual events, E_1, E_2 that occur in that order. This order can also be captured in language by describing the video as "E_1 `before` E_2". If we swap the order of events in the text, then the swapped sentence that is *consistent* with the time order in the video should be assigned a higher similarity score than the original sentence that is *inconsistent*. Thus, the ***first question*** we ask in Section 2: do the representations learnt by video-language foundation models encode this sense of time? To reliably attribute lack of time awareness

[1] Original version of the paper (Bagad et al. (2023)) to appear at CVPR 2023.

to models and not existing benchmarks, we design our own synthetic dataset to probe models. We create video-language pairs that show a sequence of two events. Then, we alter the order of events either in the text or the video and check if models can connect (reversed) video with (reversed) language consistently. We find that existing video-language models indeed struggle to associate the time order across video and language.

In light of these findings, the ***second question*** we ask in Section 3 is: can we adapt a video-language model, without expensive re-training from scratch, to instill this sense of time? Towards this, we take inspiration from the literature on understanding time in natural language, where there has been much work on developing time aware language models (Dhingra et al. (2022); Han et al. (2021; 2020); Zhou et al. (2020; 2021)). To instill time awareness, we propose TACT: **T**emporal **A**daptation by **C**onsistent **T**ime-ordering based on two key components: (i) we artificially create samples that provide temporal signal, for example, by flipping the order of events in the video or the text, (ii) we introduce a modified contrastive loss to learn time order consistency based on these samples. Instead of training from scratch, we adapt an existing model, VideoCLIP (Liang et al. (2022)), using the paradigm of *post-pretraining* (Luo et al. (2021)) on a *small* amount of video-text data. We demonstrate the effectiveness of TACT on four diverse datasets in Section 3.

Related work. We focus on video-language models that can be categorized into (i) image-language models adapted for videos (*e.g.*Luo et al. (2021)), (ii) video-language models trained by contrastive objective (*e.g.*, Xu et al. (2021)), and (iii) video-language models trained by masking objective (*e.g.*, Ge et al. (2022)). Recently, *post-pretraining* approaches (Luo et al. (2021)) have been proposed, *e.g.*, to adapt image-language models for video-language tasks by adding another stage of self-supervised training before using models on downstream tasks. This is attractive as it circumvents the cost of large-scale pretraining. We follow this line of work and aim to instil time awareness in pretrained video-language models.

Understanding of time has been studied independently in videos and language. For videos, aspects of time such as direction(Wei et al. (2018)), frame/clip ordering (Xu et al. (2019)), continuity (Liang et al. (2022)), *etc.*have been used as a self-supervision signal to learn time aware representations. However, it remains unclear if such representations actually encode time reliably. In a similar spirit, a related direction pursues evaluation and benchmarking of time awareness in video datasets, models or both (Huang et al. (2018)). Modelling time in natural language text has also been thoroughly investigated (Zhou et al. (2021; 2020)). Our work derives inspiration from these but applies more generally to video-language models. As far as we know, we are the first to shed light on time awareness in a cross-model (video-language) setting. Note that, in this study, we do not consider supervised video-language models trained on a particular task (*e.g.*, temporal grounding).

2 Do Video-Language Models Sense Time?

Probing video-language models for temporal understanding is an open direction of research. In this work, we consider a specific sense of temporal understanding: consistency in the order of events in a video with the associated textual description. For example, consider a text description: `A red circle appears before a yellow circle`. This imposes an order constraint on the video stream to have the event `red circle appears` happen before the event `yellow circle appears`. Can existing video-language models connect time-order in text with that in video? To answer this, we design an experiment with synthetic data.

Synthetic dataset. We construct simple videos that comprise a pair of events such as the ones mentioned above. We generate $N{=}180$ video-language pairs as a combination of $C{=}6$ colors, $S{=}3$ shapes, and $|\tau|{=}2$ temporal relations: *before* and *after*. The corresponding caption describes the order of events connected with a *before/after* temporal relation. We call this caption as an *attractor* since it is consistent with the time-ordering in the video. Likewise, we construct a *distractor* where we flip the order of event descriptions while retaining the temporal relation. An example pair is illustrated in Fig. 1 (left). Ideally, a time aware video-language model should be able to associate the video with the temporally consistent text, or vice versa. We refer to this task as *time-order consistency check*. In order to rule out the possibility that synthetic videos are out-of-distribution, we also perform the same experiment with canonical clips where a video displays a single event and the text describes that same event as shown in Fig. 1 (right). We refer to this as the *control task*.

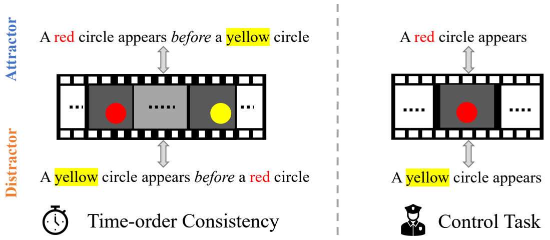

Figure 1: Proposed task to evaluate time-order consistency across synthetic video-language pairs having before/after relations. We also define a control task to check if the synthetic videos are considered out-of-distribution by the model.

Paradigm	**Method**	**Video-to-Text**		**Text-to-Video**	
		👮	⏱	👮	⏱
Chance	-	50.0	50.0	50.0	50.0
Image-Language adapted to video	CLIP4Clip	49.4	51.1	50.0	49.4
	CLIP2Video	100.0	47.8	97.8	52.3
	CenterCLIP	91.7	46.1	97.2	51.1
Video-Language Contrastive	VideoCLIP	87.1	51.1	66.7	48.3
	Frozen in Time	97.8	49.4	100.0	50.6
Video-Language Masking	BridgeFormer	100.0	51.1	97.2	42.2

Table 1: Results on synthetic control (👮) and time-order consistency (⏱) task as described in Fig. 1. Existing models show random performance on our time-order task.

Choice of models. We consider recent video-language models, broadly categorized into three groups: (i) image-language models like CLIP (Radford et al. (2021)) that are adapted to videos (Luo et al. (2021); Fang et al. (2021); Zhao et al. (2022)), (ii) pure video-language models trained on a contrastive learning objective (Xu et al. (2021); Bain et al. (2021a)), and (iii) pure video-language models trained on a masking objective (Ge et al. (2022)).

Findings. We evaluate video-to-text and text-to-video retrieval on both time-order consistency and control tasks. From Tab. 1, we observe that while most video-language models perform well on the control task, all of them struggle and perform on par with random chance on the temporal task. This gap in performance deserves attention given the importance of time in videos.

3 Instilling Video-Language Models with a Sense of Time

We describe a post-pretraining recipe for instilling this sense of time into a video-language model. We propose TACT: **T**emporal **A**daptation by **C**onsistency of **T**ime-order, that is based on two key components: (i) we artificially create samples that provide temporal signals, *e.g.*, by flipping the order of events; (ii) we introduce a modified contrastive loss to learn temporal consistency based on these samples. First, we define the notation and then describe the key components of our recipe.

Preliminaries. Let $\mathcal{V}$ be the space of video clips and $\mathcal{T}$ be the space of text clips. Consider two non-overlapping video clips $v_i, v_j \in \mathcal{V}$. Let $\zeta_i, \zeta_j \in \mathcal{T}$ be their respective captions. Let τ be a temporal relation, $\tau \in \{\textit{before}, \textit{after}\}$. Then, we denote a *stitched* and time-order consistent clip as (u_{ij}, t_{ij}), where $u_{ij} := [v_i; v_j]$, $t_{ij} := [\zeta_i; \tau; \zeta_j]$, and $[\cdot;\cdot]$ denotes concatenation. Note that depending on τ, the order of v_i and v_j may need to change in u_{ij}. For brevity, we drop the subscripts and refer to the stitched pair as (u, t) unless stated otherwise.

Time-order reversal. The classical contrastive learning paradigm for video-language models aligns components of a video clip v_i with it's text counterpart ζ_i and contrasts against other texts ζ_j that usually describe a completely different clip. This makes such models ignore the finer details of temporal understanding as it is easier to contrast the negatives by simply focusing on the objects/scene. This leads to bag-of-words like representations (Yuksekgonul et al. (2023)). We hypothesize that unless there are negatives in a contrastive setup that contain the same scenes and objects, models do not need to learn a sense of time. Thus, we present a simple strategy to generate negatives that force the learning process to focus on the temporal order. We define a time-order reversal function $\mathbb{T}$ that operates on the stitched video clip or text description and temporally swaps its constituents :

$$\mathbb{T}(u) = \mathbb{T}([v_i; v_j]) := [v_j; v_i], \qquad (1) \qquad\qquad \mathbb{T}(t) = \mathbb{T}([\zeta_i; \tau; \zeta_j]) := [\zeta_j; \tau; \zeta_i] \qquad (2)$$

An illustration of $\mathbb{T}$ is shown in Fig. 2. Note that $\mathbb{T}$ does not reverse the actual video, *i.e.*,, time does not flow backwards, but only changes the order in which events happen in the stitched clips. Our objective is to train a model that is able to distinguish between the original pair (u, t) and time-reversed versions $(u, \mathbb{T}(t))$, and $(\mathbb{T}(u), t)$.

Loss function. We assume access to an existing pretrained video-language model with a visual encoder f_θ and text encoder g_ϕ. We obtain the video encoding $\mathbf{z}_u := f_\theta(u) \in \mathbb{R}^d$ and the text encoding

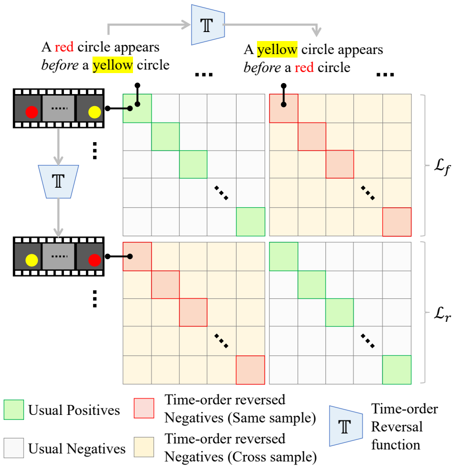

Figure 2: TACT overview. Along with the usual contrastive loss, we make use of time-order reversal within the same sample and cross samples to generate additional negatives for both video and text. We also extend the contrastive loss to time-order reversed video and text corresponding to reverse consistency $\mathcal{L}_r$.

Dataset	Method	Retrieval		Time-order
		R@1↑	MedR ↓	A_{time}↑
TEMPO	Zero-shot	3.7	49.0	48.1
	TACT†	7.7	13.0	46.5
	TACT*	**9.3**	**9.0**	**66.5**
ActivityNet	Zero-shot	1.1	44.0	49.6
	TACT†	**5.8**	**34.0**	59.7
	TACT*	**5.8**	35.0	**85.7**
Charades	Zero-shot	1.3	170.0	47.1
	TACT†	5.3	38.5	73.5
	TACT*	**5.7**	**35.0**	**77.0**
Charades-Ego	Zero-shot	1.6	64.0	53.7
	TACT†	6.4	35.0	60.1
	TACT*	**10.1**	**28.5**	**68.2**

Table 2: Results with TACT on test sets of various datasets. TACT* is the model with optimal loss coefficients and TACT† is a baseline with all coefficients 0 (usual contrastive learning without additional negatives/positives). On time order, TACT generalizes well with TACT* outperforming the baselines. On retrieval, for TEMPO and Charades-Ego, TACT* outperforms the baseline as their optimal models have β=1 which helps retrieval with a small amount of data.

$\mathbf{z}_t := g_\phi(t) \in \mathbb{R}^d$. Our goal is to adapt $\Theta = \{\theta, \phi\}$ via post-pretraining s.t. the resulting model is time aware while maintaining its original retrieval performance. As we aim to use a small dataset, we only update some parameters of the model (*e.g.*last few layers). We use the Info-NCE (van den Oord et al. (2018)) loss on a $2B \times 2B$-sized similarity matrix (B being batch size) instead of the usual $B \times B$ in contrastive loss (Fig. 2). Details of the loss computation are provided in Appendix A.

Experimental setup. We show the efficacy of TACT by adapting pretrained VideoCLIP model. We post-pretrain it on four diverse datasets: (i) ***TEMPO*** (Hendricks et al. (2018)): with text descriptions for fixed 5s segments ; (ii) ***ActivityNet Captions*** (Krishna et al. (2017)): a dense video captioning dataset with human-centric actions; (iii) ***Charades*** (Sigurdsson et al. (2016)): a scripted indoor daily human activities video dataset; and (iv) ***Charades-Ego*** (Sigurdsson et al. (2018)): similar to Charades, scripted human activities from the egocentric viewpoint. To construct stitched clips, we randomly sample any two non-overlapping clip-text pairs in the video. We evaluate adapted models on retrieval metrics such as $R@1$, median rank (MedR) and temporal accuracy A_{time} (% of video-text pairs with correctly identified time order). More details are provided in Appendix C.

Results. (i) **Synthetic data**: we first evaluate trained models on the same synthetic data. On the models trained on four datasets independently, we obtain 78.1%, 59.4%, 88.3% and 86.7% respectively. (ii) **Real data**: as shown in Tab. 2, on test sets of the real datasets, we obtain much superior performance in comparison to zero-shot or contrastive-only baseline. (iii) **Generalization to a new temporal prompt**: Having trained on sentences with before/after relations, we also test if adapted models generalize to new kinds of prompts such as `"First,`E_1`,then,`E_2`"`. On each of the four datasets, we obtain 53.2%, 62.9%, 73.1%, 62.9% accuracies respectively with this new prompt. More details are provided in Appendix D. This substantiates the learning of time order of events rather than merely learning the order of words in the sentence or learning about specific words.

Conclusion. Given the essence of time in video-language foundation models, we present a simple experiment based on synthetic data to test for time awareness in existing models. We find that existing models lack a sense of time defined in terms of consistency of order of events in video and language. To fill this gap, building upon VideoCLIP, we present TACT, a recipe to instill this sense of time in video-language models. As a result of such adaptation, we observe strong performance on the synthetic data and several real datasets. We hope that this work provokes further probing and instilling time awareness in video-language models; and also inspires other adaptations of foundation models to solve various challenging tasks.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022. 1

James F. Allen. Towards a general theory of action and time. In *Artificial Intelligence*, 1984. 1

Piyush Bagad, Makarand Tapaswi, and Cees G. M. Snoek. Test of Time: Instilling Video-Language Models with a Sense of Time. In *CVPR*, 2023. 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)*, 2021a. 1, 3

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *International Conference on Computer Vision (ICCV)*, 2021b. 9

Rishi Bommasani et al. On the opportunities and risks of foundation models. *ArXiv*, abs/2108.07258, 2021. 1

Shyamal Buch, Cristobal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "Video" in Video-Language Understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273, 2022. 2

Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 3

Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16167–16176, June 2022. 2, 3

Rujun Han, Xiang Ren, and Nanyun Peng. Deer: A data efficient language model for event temporal reasoning. *ArXiv*, abs/2012.15283, 2020. 2

Rujun Han, Xiang Ren, and Nanyun Peng. Econet: Effective continual pretraining of language models for event temporal reasoning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021. 2

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. *International Conference on Computer Vision (ICCV)*, pp. 5804–5813, 2017. 9

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1380–1390, 2018. doi: 10.18653/v1/D18-1168. URL https://aclanthology.org/D18-1168. 4, 9

De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7366–7375, 2018. doi: 10.1109/CVPR.2018.00769. 2

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL http://arxiv.org/abs/1412.6980. 10

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-Captioning Events in Videos. In *International Conference on Computer Vision (ICCV)*, 2017. 4, 9

Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning, 2022. URL https://arxiv.org/abs/2206.03428. 1

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, 2022. 1

Hanwen Liang, Niamul Quader, Zhixiang Chi, Lizhe Chen, Peng Dai, Juwei Lu, and Yang Wang. Self-supervised spatiotemporal representation learning by exploiting video continuity. *ArXiv*, abs/2112.05883, 2022. 2

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. 1, 2, 3

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *International Conference on Computer Vision (ICCV)*, 2019. 9

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 1, 3

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021. 1

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision (ECCV)*, pp. 510–526, Cham, 2016. 4, 9

Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7396–7404, 2018. doi: 10.1109/CVPR.2018.00772. 4, 9

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 4, 8

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. 1

Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8052–8060, 2018. 2

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *European Conference on Computer Vision (ECCV)*, pp. 318–335, Cham, 2018. 9

Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10334–10343, 2019. 2

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)*, pp. 6787–6800, 2021. doi: 10.18653/v1/2021.emnlp-main.544. URL https://aclanthology.org/2021.emnlp-main.544. 1, 2, 3, 9

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. 1

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX. 3

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv*, 2022. 1

Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. Centerclip: Token clustering for efficient text-video retrieval. In *SIGIR*, 2022. 3

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. Temporal common sense acquisition with minimal supervision. *ArXiv*, abs/2005.04304, 2020. 2

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. Temporal reasoning on implicit events from distant supervision. *ArXiv*, abs/2010.12753, 2021. 2

A Details of loss computation

We now introduce our recipe for temporal adaptation based on the InfoNCE loss van den Oord et al. (2018) to learn time-order sensitive video-text correspondence. For a positive (or time-order consistent) video-text pair (u, t), we first define a forward loss where the stitched pair is in its original time-order.

$$\mathcal{L}_f = \sum_{(u,t)\in\mathcal{B}} \left(\mathrm{TNCE}(\mathbf{z}_u, \mathbf{z}_t) + \mathrm{TNCE}(\mathbf{z}_t, \mathbf{z}_u)\right), \tag{3}$$

where TNCE is the Noise Contrastive Estimation (NCE) loss for temporal adaptation, defined as:

$$\mathrm{TNCE}(\mathbf{z}_u, \mathbf{z}_t) := -\log \frac{\exp(\mathbf{z}_u \cdot \mathbf{z}_t)}{\sum_{t'\in\mathcal{B}_t} \exp(\mathbf{z}_u \cdot \mathbf{z}_{t'}) + \mathcal{C}^{\text{time}}}, \tag{4}$$

where $\mathcal{B}$ is the batch of (u, t) pairs and $\mathcal{B}_t$ specifically refers to other stitched text captions in the batch. $\mathcal{C}^{\text{time}}$ accumulates negatives defined using time-order reversal as:

$$\mathcal{C}^{\text{time}} = \alpha_{\text{same}} \exp(\mathbf{z}_u \cdot \mathbf{z}_{\mathbb{T}(t)}) + \alpha_{\text{cross}} \sum_{t'\in\mathcal{B}_t\setminus\{t\}} \exp(\mathbf{z}_u \cdot \mathbf{z}_{\mathbb{T}(t')}), \tag{5}$$

where α_{same} controls the effect of contrasting text from the same sample but with reversed text time-order, *i.e.*,, $\mathbb{T}(t)$, and α_{cross} encourages the model to contrast between reversed versions of other text captions, *i.e.*,, $\mathbb{T}(t')$. Note that when both α_{same} and α_{cross} are 0, we revert back to the standard NCE formulation, albeit on stitched pairs. While Eq. equation 4 corresponds to the video-text loss $\mathrm{TNCE}(\mathbf{z}_u, \mathbf{z}_t)$, the text-video loss $\mathrm{TNCE}(\mathbf{z}_t, \mathbf{z}_u)$ is defined symmetrically.

Furthermore, we also apply a reverse loss $\mathcal{L}_r$ to bring time-order reversed versions of both the video and the text together. Note that as we consider (u, t) as a positive pair, $(\mathbb{T}(u), \mathbb{T}(t))$ also form a positive pair,

$$\mathcal{L}_r = \sum_{(\mathbb{T}(u),\mathbb{T}(t))\in\mathcal{B}} \left(\mathrm{TNCE}(\mathbf{z}_{\mathbb{T}(u)}, \mathbf{z}_{\mathbb{T}(t)}) + \mathrm{TNCE}(\mathbf{z}_{\mathbb{T}(t)}, \mathbf{z}_{\mathbb{T}(u)})\right). \tag{6}$$

Here, the TNCE term operates on time-reversed clips and $\mathcal{C}^{\text{time}}$ contrasts $(\mathbb{T}(u), \mathbb{T}(t))$ with unreversed text clips in the batch $(\mathbb{T}(u), t)$.

The overall loss function is defined as a combination,

$$\mathcal{L} = \mathcal{L}_f + \beta\mathcal{L}_r. \tag{7}$$

Depending on the choice of loss coefficients $\alpha_{\text{same}}, \alpha_{\text{cross}}, \beta \in \{0, 1\}$, we can vary properties of the adapted model. For example, setting $\alpha_{\text{same}}{=}1$ encourages high sensitivity to time-order reversal. As we will see empirically, the loss coefficients also provide the flexibility to adapt the model to various datasets.

We illustrate this temporal extension of the contrastive loss in Fig. 2 (best seen in color). $\mathbb{T}$ illustrates the time order reversal function. The top half corresponds to $\mathcal{L}_f$ while the bottom half visualizes $\mathcal{L}_r$. In particular, the top-left quadrant alone corresponds to the standard contrastive loss. While the green diagonal terms are positive pairs, the red diagonal terms are the strongest drivers for instilling temporal understanding in the model.

B Impact of loss coefficients

Choosing appropriate values for loss coefficients $\Theta_l{:=}\{\alpha_{\text{same}}, \alpha_{\text{cross}}, \beta\}$ allows the model to learn various aspects and adapt using different datasets. On each dataset, we vary $\Theta_l{\in}\{0, 1\}^3$ and find the best configuration based on the $\mathrm{GeometricMean}(R@1, \max(A_{\text{time}} - 50, 0))$ on the validation sets. The above metric ensures the geometric mean is not overpowered by A_{time}. The results are shown in Tab. 4.

As α_{same} is directly responsible for discriminating between original and time-reversed orders, unsurprisingly, setting it to 1 is necessary to achieve the best results on A_{time} for all the datasets. For TEMPO and Charades-Ego, using all loss components (all 1) provides the best results, whereas $\alpha_{\text{cross}}{=}1$ and $\beta{=}0$ achieves a better trade-off for ActivityNet and Charades. Choosing $\beta{=}1$ leads to an improvement in retrieval performance for TEMPO and Charades-Ego but leads to a decline

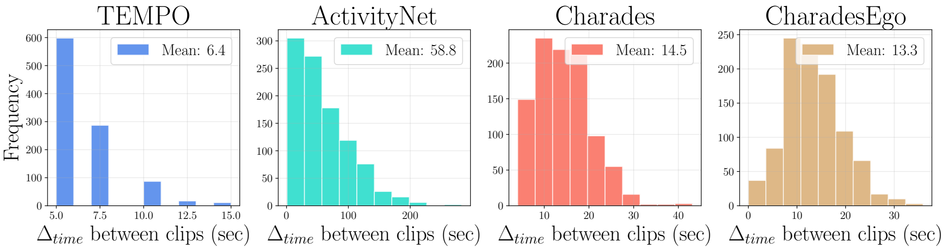

Figure 3: Time-distance between stitched clips in datasets for temporal adaptation (Δ_{time}). TEMPO has stitched clips close to each other while those in Charades-Ego are farthest apart suggesting a correlation between Δ_{time} and the difficulty of temporal adaptation.

Dataset	Train		Validation		Test		Ego	Length
	$N_{\mathcal{V}}$	$N_{\mathcal{C}}$	$N_{\mathcal{V}}$	$N_{\mathcal{C}}$	$N_{\mathcal{V}}$	$N_{\mathcal{C}}$		(s)
TEMPO	3,904	28,427	411	1,000	396	1,000	✗	30
ActivityNet	7,440	95,474	453	906	456	912	✗	120
Charades	5,262	99,928	500	1,000	500	1,000	✗	30
Charades-Ego	2,679	155,306	500	1,000	210	420	✓	31

Table 3: Statistics of datasets we consider for temporal adaptation. $N_{\mathcal{V}}$ is the number of unique videos and $N_{\mathcal{C}}$ is the number of stitched clips. Based on $N_{\mathcal{V}}$, TEMPO and Charades-Ego are smaller as compared to ActivityNet and Charades.

for ActivityNet and Charades. We attribute this to the number of unique videos in the train set for these datasets. As ActivityNet and Charades have more videos than TEMPO or Charades-Ego (see train $N_{\mathcal{V}}$ Tab. 3) additional positives introduced by setting $\beta{=}1$ are not necessary and in fact hurt performance.

C Details of TACT adaptation

Base model. We demonstrate the effectiveness of TACT as an adaptation recipe on top of VideoCLIP Xu et al. (2021) owing to its simple architecture, contrastive objective, and use of pre-computed S3D Xie et al. (2018) features that enable faster experimentation and allow encoding a long temporal context (∼32 secs). We initialize Θ from the model pretrained on HowTo100M Miech et al. (2019) and post-pretrain on multiple datasets.

Datasets. One of our key objectives is to post-pretrain on a small amount of data in contrast to massive pretraining datasets such as WebVid2M Bain et al. (2021b) or HowTo100M Miech et al. (2019). We consider dense video captioning datasets that offer sufficient diversity in terms of size, backgrounds, clip durations, viewpoints and activities. Specifically, we experiment with: (i) ***TEMPO*** Hendricks et al. (2018): the subset of stitched diverse third-person videos from DiDeMo Hendricks et al. (2017) with text descriptions for fixed 5s segments that contain before/after relations; (ii) ***ActivityNet Captions*** Krishna et al. (2017): a dense video captioning dataset with human-centric actions; (iii) ***Charades*** Sigurdsson et al. (2016): a scripted indoor daily human activities video dataset; and (iv) ***Charades-Ego*** Sigurdsson et al. (2018): similar to Charades, scripted human activities from the egocentric viewpoint. To construct stitched clips, we randomly sample any two non-overlapping clip-text pairs in the video. Since we require stitched clips instead of raw videos, we create new splits for each dataset (see Tab. 3). We will release all the splits publicly on our project page.

Evaluation metrics. We evaluate the post-pretrained model using two sets of metrics: (i) standard retrieval metrics, recall $R@1$, $R@5$, $R@10$ and median-rank evaluated on stitched video-text clips; and (ii) time-order consistency, *i.e.*,, the fraction of videos for which the model correctly associates text that is time order consistent with the video:

$$A_{\text{time}} := \frac{1}{|\mathcal{D}|} \sum_{(u,t)\in\mathcal{D}} \mathbb{I}[d(\mathbf{z}_u, \mathbf{z}_t) < d(\mathbf{z}_u, \mathbf{z}_{\mathbb{T}(t)})], \tag{8}$$

where (u, t) are time-order consistent pairs, $\mathcal{D}$ is the dataset, and $d(\cdot, \cdot)$ is a distance metric based on cosine similarity.

Post-pretraining details. We freeze the word embeddings and layers 1 to 5 for both the video and text encoders in VideoCLIP. For adaptation, we use the Adam optimizer Kingma & Ba (2015) with learning rate $5e^{-6}$, batch size 32 trained on a single node with 4 GeForce GTX 1080 GPUs. On TEMPO, we train for 60 epochs while on the other datasets, we train for 10 epochs and pick the checkpoint that maximizes the geometric mean of $R@1$ and A_{time} on the respective validation set. A typical training run takes about 1-3 hours.

Loss coefficients			TEMPO			ActivityNet			Charades			Charades-Ego		
α_{same}	α_{cross}	β	R@1 ↑	MedR ↓	A_{time} ↑	R@1 ↑	MedR ↓	A_{time} ↑	R@1 ↑	MedR ↓	A_{time} ↑	R@1 ↑	MedR ↓	A_{time} ↑
	Chance		0.1	500.0	50.0	0.1	453.0	50.0	0.1	500.0	50.0	0.1	500.0	50.0
0	0	0	8.3	14.0	49.4	6.4	30.0	57.3	5.7	42.0	71.5	2.9	44.0	64.6
0	0	1	8.2	14.0	49.5	5.6	27.0	47.0	4.2	58.0	75.1	3.2	41.5	65.2
0	1	0	8.2	15.0	49.3	6.1	29.0	78.8	5.2	45.0	78.9	3.4	38.0	64.5
0	1	1	8.1	14.0	49.5	5.8	27.0	48.3	4.2	58.0	75.1	3.1	41.0	67.0
1	0	0	6.4	20.0	60.6	5.9	28.0	79.1	6.1	38.0	76.3	3.2	42.0	66.1
1	0	1	6.5	24.0	62.9	5.6	26.0	63.1	4.9	51.0	78.0	3.3	39.0	70.7
1	1	0	5.9	24.0	59.7	6.0	29.0	86.3	6.6	43.0	77.8	3.7	40.5	67.9
1	1	1	7.5	17.0	62.5	5.7	27.0	63.8	5.1	51.0	77.7	3.8	38.5	68.3

Table 4: Impact of loss coefficients for TACT post-pretraining on validation sets of various datasets. For each dataset, the corresponding *color-marked row denotes the best configuration* based on the geometric mean of $R@1$ and A_{time}. TACT is able to connect time-order in video and language while maintaining its retrieval capabilities across several datasets.

D GENERALIZATION TO NEW TEMPORAL PROMPTS

The time-order of events in language can be described using various sentence structures. Although we train video-language models using before/after relations, it is natural to ask if the model still correctly associates time-order for a different prompt such as `First,.., then,..`. To systematically test this, we gather event pairs E_1, E_2 (E_1 occurs before E_2 in the video) for each sample in the validation set and stitch them using three prompts as follows: (i) E_1 `before` E_2, (ii) E_2 `after` E_1, (iii) `First` E_1, `then` E_2. As shown in Fig. 4, TACT-adapted models generalize well to a new prompt (iii). This substantiates the learning of time-order of events rather than merely learning the order of words in the sentence.

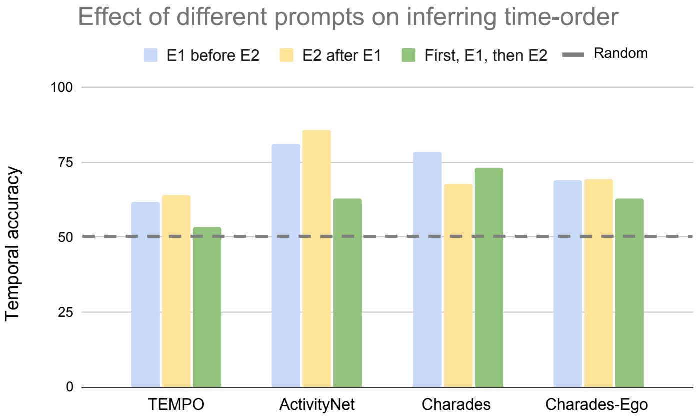

Figure 4: Models trained by TACT with before/after relations generalize to a new kind of prompt such as `First, .., then ..` indicating learning of the underlying true time-order of events.