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

Piyush Bagad
University of Amsterdam

Makarand Tapaswi
IIIT Hyderabad

Cees G.M. Snoek
University of Amsterdam

bpiyush.github.io/testoftime-website

Abstract

Modelling and understanding time remains a challenge in contemporary video understanding models. With language emerging as a key driver towards powerful generalization, it is imperative for foundational video-language models to have a sense of time. In this paper, we consider a specific aspect of temporal understanding: consistency of time order as elicited by before/after relations. We establish that seven existing video-language models struggle to understand even such simple temporal relations. We then question whether it is feasible to equip these foundational 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. We conduct a zero-shot evaluation of the adapted models on six datasets for three downstream tasks which require varying degrees of time awareness. We observe encouraging performance gains especially when the task needs higher time awareness. Our work serves as a first step towards probing and instilling a sense of time in existing video-language models without the need for data and compute-intense training from scratch.

A. Dog runs away *before* it brings a ball to the man

B. The dog brings a a ball to the man *before* it runs away

C. The baby eats food *after* it looks into the camera

D. The baby looks into the camera *after* it eats food

Figure 1. Can you match the correct video-text pairs? Understanding the time order of events across video and language is necessary to be able to solve this task. See footnote on next page for answers.

1. Introduction

Self-supervised pretraining at scale on multimodal web corpora tied with powerful architectures [107] has led to foundational models [12] for images [2, 49, 59, 83, 84] and videos [2, 6, 26, 109, 119, 126]. These models have enabled remarkable improvements on a plethora of downstream video-language tasks such as video-text retrieval, video question-answering, and action recognition. Given the cost and difficulty of video annotations, even for a small amount of downstream data, such foundational models are emerging as the de-facto backbone for zero-shot [119, 122, 127] and few-shot generalization [2]. 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 [6, 57, 66, 119] can achieve impressive performance on several video benchmarks [22, 41, 120] without reliably encoding time [13, 56, 59]. For example, Buch *et al*. [13] show that a model that uses a single (carefully selected) frame often outperforms recent video-language models [57, 119] on standard video benchmarks such as MSR-VTT [120]. Lei *et al*. [56] 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, indeed, 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, *i.e*., understanding temporal relations such as *before* and *after* [4]. Consider the task presented in Fig. 1. A time invariant model shall be able to associate (A) with (1) or (2) and (B) with (3) or (4) based on static frames alone. But to distinguish between (1) and (2), one needs to be able to understand time order and connect it across video and language[1]. Thus, the ***first question*** we ask in Section 3: do the representations learnt by foundational video-language models encode this sense of time? To reliably attribute lack of time awareness to models and not existing benchmarks, we design our own synthetic dataset to probe models for this sense of time. 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 the order in video and language. 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 4 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 literature on understanding time in natural language, where there has been much work on developing time aware language models [20, 36, 37, 130, 131]. Our objective is to instill time awareness in a video-language model without having to pretrain from scratch. To do that, 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, and (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 video-language model, VideoCLIP [61], using the paradigm of *post-pretraining* on a *small* amount of video-text data [66, 121]. We demonstrate the effectiveness of TACT in connecting the time order in video and language on four diverse real datasets in Section 5.

Finally, in line with the original motive of video-language models for zero-shot generalization, we evaluate in Section 6 our TACT-adapted model for three sets of tasks on six downstream datasets which require a varying degree of time awareness. On tasks that need higher time awareness, with the appropriate choice of adaptation dataset, TACT outperforms a strong baseline that is based on post-pretraining on canonical clip-text pairs without consideration of time-order.

In summary, our contributions are: (i) We show that existing video-language models struggle to associate time order in video and language through controlled experiments on synthetic data and several evaluations on real datasets. (ii) Based on VideoCLIP [119], we propose TACT, a method for temporal adaptation using this time order consistency without having to pretrain from scratch. (iii) We demonstrate improved zero-shot generalizability of TACT-adapted models on tasks that require higher time awareness.

[1] Answers: (A)-(2), (B)-(1), (C)-(4), (D)-(3).

2. Background and Related Work

We briefly discuss recent advances in video-language models followed by their consideration of time.

Foundational video-language models. Large-scale datasets, self-supervision, and the advent of Transformers [107] have led to the emergence of powerful encoders for images [21,39,103], videos [5,11,24,104,117], language models [19,64,69,86] and even universal encoders [32,46]. These encoders form the basis for several vision-language foundational models. Popular image-language models such as CLIP [83] and ALIGN [49] are trained on massive datasets by using web images and alt-text. Similarly, video-language models are catching up and can be categorised into two broad directions: (i) adapting image-language models for videos [8, 23, 50, 51, 63, 66, 71, 110, 112, 121], and (ii) pure video-based models that are learned using large video-text datasets [3,7,27–29,31,58,62,65,67,68,95,119]. Recently, a new paradigm of *post-pretraining* has emerged where an existing image- or video-language model goes through another stage of self-supervised pretraining on a *small* amount of video data before it is evaluated on downstream tasks [66, 121]. This is promising as it circumvents the prohibitive cost of pretraining on large datasets from scratch. In [66], the post-pretraining uses time-invariant mean-pooling, while [121] strives to bridge the domain gap between image captions and video subtitles. In contrast, our proposed temporal adaptation involves post-pretraining of VideoCLIP [119] with a small amount of data that instills the model to learn the time-order of events in a video.

Time in vision. Time separates videos from static images or an unordered set of frames. While modeling time remains a challenge, it also presents a natural source of supervision that has been exploited for self-supervised learning. For example, as a proxy signal by posing pretext tasks involving spatio-temporal jigsaw [1, 44, 53], video speed [10, 17, 48, 94, 111, 124], arrow of time [78, 80, 114], frame/clip ordering [25, 70, 90, 97, 118], video continuity [61], or tracking [45, 108, 113]. Several works have also used contrastive learning to obtain spatio-temporal representations by (i) contrasting temporally augmented versions of a clip [47, 77, 81], or (ii) encouraging consistency between local and global temporal contexts [9, 18, 85, 123]. Nevertheless, it remains unclear if the learnt representations actually encode time reliably. Time-aware features have also been explored for specific downstream tasks such as action recognition [30, 100, 101]. There has been some very recent work on evaluating self-supervised video representations [87, 99] on their temporal recognition ability instead of only relying on time as a guidance for training.

In the same spirit, a related direction pursues evaluation and benchmarking of time awareness in video datasets [88], models [13, 14, 30, 56, 89, 125] or both [43, 92]. Huang *et* *al*. [43] measure the effect of motion on temporal action recognition to find that only a subset of classes in UCF-101 and Kinetics-400 require motion information. Ghodrati *et al*. [30] propose new tasks to evaluate temporal asymmetry, continuity and causality in video models. Our work derives inspiration from these but applies more generally to video-language models as language provides a basis for open-world generalization.

Time in language. Time has also been extensively studied in the natural language literature. Early works identified temporal structures in language such as temporal prepositions and quantifiers [4, 79]. More recent literature focuses on tasks such as extracting temporal relations [35, 72–74], as well as temporal reasoning [36, 37, 82, 130, 131]. For example, Han *et al*. [36, 37] and Zhou *et al*. [131] pretrain language models specifically to focus on understanding temporal relations such as before, after, during, *etc*. The emergence of large language models has also spurred an increased interest in developing benchmarks to test for time awareness in these models [20, 75, 76, 102, 106, 129]. For example, Ning *et al*. [75] propose a new benchmark of reading comprehension with questions involving before/after relations. Since temporal relations in language are grounded in the video, we draw inspiration from [36,37,131] and aim to instill time awareness in video-language models.

Time in video-language models appears implicitly in tasks like video-text alignment [38, 98] and temporal grounding [42, 60]. In this work, we focus on self-supervised video-language models that can generalize to a variety of tasks rather than models designed for a specific task, *e.g*., temporal grounding. Some recent works have shown the under-utilisation of time in classic video-text benchmarks such as MSR-VTT [120], YouCook [132], ActivityNet [22], and DiDeMo [41]. For example, [13, 56, 57] discover that on several benchmarks, using only one or few frames or clips achieves competitive performance. Adaptations of the popular CLIP architecture for videos (*e.g*., CLIP4Clip [66]) show that weighted pooling of frames already achieves impressive performance on retrieval benchmarks.

These raise some key questions: do existing video-language models truly understand time in the sense of correctly associating order of events in language and video? If not, can we adapt them to instill time awareness? Our work addresses these questions. There has been some work in using time-order across video and language as a source of self-supervision. Specifically, concurrent to our work, both Sun *et al*. [96] and Cao *et al*. [15] propose fine-grained temporal alignment between video and text as the pretraining objective. Different from these works, we consider the notion of time-order and we aim to adapt a given video-language model using *post-pretraining*, which circumvents the need for a new round of compute-intense pretraining.

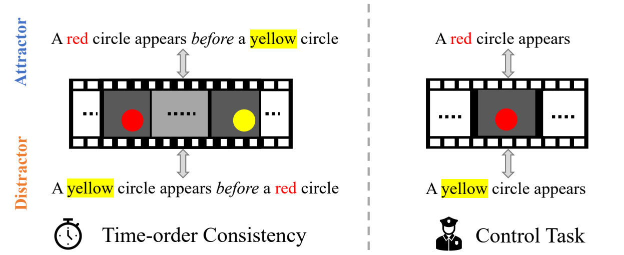

Figure 2. Overview of the proposed task to evaluate time-order consistency across synthetic video-language pairs having before/after relations. We also define a control task to check if the synthetic videos are considered out-of-distribution by the model.

3. 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 of 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. 2 (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*. To rule out the possibility that synthetic videos are out-of-distribution, we also perform the same experiment with canonical clips with a single event and the text describes that same event as shown in Fig. 2 (right). We refer to this as the *control task*.

Choice of models. We consider recent video-language models, broadly categorized into three groups: (i) image-language models like CLIP [83] that are adapted to videos [23,66,128], (ii) pure video-language models trained on a contrastive learning objective [6,16,119], and (iii) pure video-language models trained on a masking objective [29].

Findings. We evaluate video-to-text and text-to-video retrieval on both time-order consistency and control tasks.

Paradigm	Method	Video-to-Text		Text-to-Video	
Chance	-	50.0	50.0	50.0	50.0
Image-Language adapted to video	CLIP4Clip [66]	49.4	51.1	50.0	49.4
	CLIP2Video [23]	100.0	47.8	97.8	52.3
	CenterCLIP [128]	91.7	46.1	97.2	51.1
Video-Language Contrastive	VideoCLIP [119]	87.1	51.1	66.7	48.3
	Frozen in Time [6]	97.8	49.4	100.0	50.6
	VindLU [16]	98.0	52.0	100.0	51.1
Video-Language Masking	BridgeFormer [29]	100.0	51.1	97.2	42.2

Table 1. Results on synthetic control ( ) and time-order consistency ( ) task as described in Fig. 2. Existing video-language models show random performance on our time-order task.

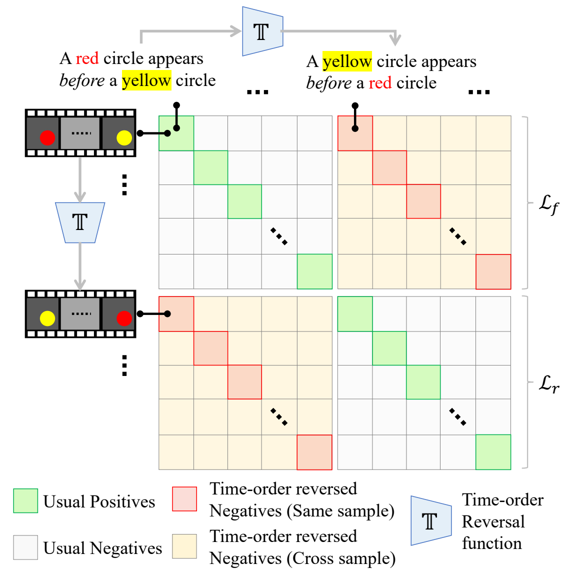

Figure 3. Overview of TACT. Along with the usual contrastive loss, where negatives come from other samples in the batch, we make use of time-order reversal within the same sample and cross samples to generate additional negatives for both video and text. We also extend the contrastive loss to time-order reversed video and text corresponding to reverse consistency $\mathcal{L}_r$.

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. Note that while synthetic data allows for controlled experiments, we also expand this evaluation to real video datasets in the following section.

4. Adaptation by Consistent Time-Ordering

We describe a post-pretraining recipe to instill a video-language model with a sense of time. 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; and (ii) we introduce a modified contrastive loss to learn temporal consistency based on these samples. We start by defining the notation and then describe the key components of our adaptation 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, $\tau \in \{before, after\}$. 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 classic contrastive learning paradigm for video-language models aligns components of a video clip v_i with its 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 or the scene. This is evident from simple bag-of-word-like methods that are shown to work well for contrastive learning, both on the visual (*e.g.*, CLIP4Clip [66]) and textual (*e.g.*, MIL-NCE [67]) modalities. We hypothesize that unless there are negatives in a contrastive setup that contain the same scenes and objects, models need not 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:

$$\mathbb{T}(u) = \mathbb{T}([v_i; v_j]) := [v_j; v_i], \quad \text{and} \tag{1}$$
$$\mathbb{T}(t) = \mathbb{T}([\zeta_i; \tau; \zeta_j]) := [\zeta_j; \tau; \zeta_i]\,. \tag{2}$$

An illustration of $\mathbb{T}$ is shown in Fig. 3. 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 pre-trained 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 $\mathbf{z}_t := g_\phi(t) \in \mathbb{R}^d$. Our goal is to adapt $\Theta = \{\theta, \phi\}$ via post-pretraining such that the resulting model is time aware while maintaining its original performance on tasks such as retrieval. As we aim to use a small amount of data, we only update some parameters of the model (Θ), such as the last few layers.

We now introduce our recipe for temporal adaptation based on the InfoNCE loss [105] 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.

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

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

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

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:

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

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. (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,

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

Here, the TNCE term operates on time-reversed clips and $\mathcal{C}^{\text{time}}$ contrasts $(\mathbb{T}(u), \mathbb{T}(t))$ with un-reversed text clips in the batch $(\mathbb{T}(u), t)$. The overall loss function is defined as:

$$\mathcal{L} = \mathcal{L}_f + \beta\mathcal{L}_r. \quad (7)$$

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 α_{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. 3 (best seen in colour). $\mathbb{T}$ 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 on stitched pairs. While the green diagonal terms are positive pairs, the red diagonal terms are the strongest drivers for instilling temporal understanding in the model.

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 2. 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.

5. Experiments: TACT Ablations

Base model. We demonstrate the effectiveness of TACT as an adaptation recipe on top of VideoCLIP [119] owing to its simple architecture, contrastive objective, and use of pre-computed S3D [116] features that enable faster experimentation and allow encoding a long temporal context ($\sim$32 secs). We initialize Θ from the model pretrained on HowTo100M [68] 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 [7] or HowTo100M [68]. 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*** [40]: the subset of stitched diverse third-person videos from DiDeMo [41] with text descriptions for fixed 5s segments that contain before/after relations; (ii) ***ActivityNet Captions*** [55]: a dense video captioning dataset with human-centric actions; (iii) ***Charades*** [93]: a scripted indoor daily human activities video dataset; and (iv) ***Charades-Ego*** [91]: 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. 2).

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:

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

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 [54] 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 adaptation run takes about 1-3 hours.

Evaluation on the test set. Results in Tab. 3 show that TACT* with optimal loss coefficients outperforms TACT† (all 0 loss coefficients) and the zero-shot baseline (no post-pretraining similar to the synthetic data experiment), both on the retrieval and time-order consistency tasks. This indicates the robustness of the adaptation.

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 $\text{GeometricMean}(R@1, \max(A_{\text{time}} - 50, 0))$ on the validation sets. The above metric ensures the geometric mean is not overpowered by R@1. 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 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. 2) additional positives introduced by setting $\beta{=}1$ are not necessary and in fact, hurt performance. Finally, we note that carefully setting the value of α_{same} provides a convenient trade-off between spatial and temporal understanding. Please refer to the supplement for detailed experiments.

Dataset	Method	R@1↑	MedR ↓	A_{time}↑
		Retrieval		**Time-order**
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 3. Results for the best TACT model on test sets. TACT* has optimal loss coefficients and TACT† is a baseline with all coefficients 0. 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 $\beta{=}1$ which helps retrieval with a small amount of data.

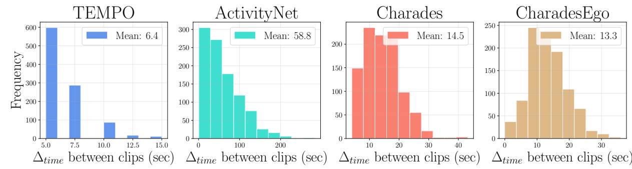

Figure 4. Time-distance between stitched clips in datasets for temporal adaptation (Δ_{time}). TEMPO has stitched clips close to each other while those in Charades-Ego are farthest apart suggesting a correlation between Δ_{time} and difficulty of temporal adaptation.

What makes temporal adaptation hard? We observe a large gap in A_{time} between TEMPO and ActivityNet. We hypothesize that the distance (in seconds) between the two clips (Δ_{time}) in a stitched clip is strongly correlated with the difficulty of adaptation. Intuitively, it is easier to infer the time-order consistency for a stitched clip u with text t that has distant constituent clips v_i, v_j since the objects and scene could be vastly different. In contrast, it is harder to discern the correct time-order when the constituent clips are closer in time. Fig. 4 shows distribution of Δ_{time} for each dataset. Indeed, Δ_{time} in ActivityNet (58.8s) is much higher than that in TEMPO (6.4s) making the task harder on TEMPO. To further test our hypothesis, we conduct a controlled experiment where we gradually vary the distribution of Δ_{time} for Charades-Ego to match it to that of TEMPO. We find a strong correlation (ρ=0.92) between Δ_{time} and hardness of adaptation. More details are in the supplement.

6. Experiments: Downstream Evaluation

The goal of video-language foundation models is to generalize in a zero- or few-shot manner to a diverse range of downstream tasks. We evaluate TACT models on three sets of downstream tasks that need low-to-high time awareness.

Baseline: Standard post-pretraining. Comparing our temporally adapted models with pretrained VideoCLIP is not fair since adapted models see data beyond the pretraining phase. In addition to the zero-shot comparison, we compare against a baseline model that is trained for standard video-text retrieval on the same datasets as temporal adaptation. Instead of using stitched clips, we use simple canonical pairs, *i.e.*, (v_i, ζ_i) instead of (u_{ij}, t_{ij}).

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.

Low —— Time awareness ——→ High

Adaptation		I. Text-to-Video Retrieval						II. Temporal VQA		III. Action-to-Video Retrieval		
Dataset	Method	MSR-VTT			YouCookII			Next-QA (ATP)	AGQA	SSv2	SSv2 (events)	Temporal
		$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	Accuracy	Accuracy	mAP	mAP	mAP
-	Chance	0.1	0.5	1.0	0.1	0.5	1.0	20.0	50.0	0.6	2.0	2.0
-	None	10.6	23.4	29.9	18.2	45.5	59.9	23.4	49.9	3.4	6.4	13.0
TEMPO	Baseline	12.0	29.3	37.3	21.5	48.2	61.8	25.0	50.8	3.9	6.8	15.9
	TACT	12.8	26.5	35.7	20.4	45.1	58.7	27.6	57.1	4.2	7.7	16.1
ActivityNet	Baseline	15.7	34.4	44.9	15.6	38.8	51.4	23.7	50.7	3.7	7.0	16.0
	TACT	13.8	29.6	39.6	16.0	36.9	49.8	25.4	50.3	3.5	7.2	16.2
Charades	Baseline	12.3	25.8	33.6	21.5	48.6	61.7	26.0	50.5	4.1	7.1	13.7
	TACT	11.7	25.2	32.4	22.4	49.1	62.4	25.2	54.8	4.3	7.8	14.1
Charades-Ego	Baseline	13.1	27.5	34.5	19.4	47.1	60.8	24.3	49.7	4.0	6.9	14.7
	TACT	11.1	24.6	30.6	21.9	48.2	61.9	25.6	58.4	3.9	6.9	13.5

Table 5. Results on downstream zero-shot evaluation with tasks requiring increasing time awareness from I to III. None corresponds to direct evaluation of the VideoCLIP model on the downstream dataset. Green denotes an improvement for the TACT adapted model w.r.t. the baseline, red denotes a deterioration. As we move from tasks that need low to high time awareness, the effectiveness of TACT increases. See Sec. 6 for a more detailed discussion. The table is best viewed on screen in colour.

Evaluating TACT adapted models on synthetic data. On the video-to-text variant, TACT adapted on TEMPO achieves 64.4%, ActivityNet 52.5%, Charades 65.0%, Charades-Ego 85.6%. This is usually higher than the performance that non-adapted models achieve in Tab. 1. This highlights that TACT models learn useful signals to match time-order in video and language.

I. Text-to-video retrieval. We consider two widely used benchmarks: ***MSR-VTT*** [120] and ***YouCookII*** [132] and adopt standard retrieval metrics. Recent work has identified a bias for spatial understanding in these datasets [8, 13, 43, 56, 59, 66]. Thus, we consider this class of tasks as requiring low time awareness. As shown in Tab. 5 set I, on MSR-VTT [120], we observe that TACT is worse (marked in red) or at par with the baseline across datasets. This aligns well with findings in [13, 56] that these benchmarks do not need time awareness. On YouCookII [132], TACT models based on Charades(-Ego) outperform the baseline (marked in green). We believe this is a consequence of a lower domain shift between YouCookII and Charades.

II. Temporal video QA. Next, we use subsets of recently released multiple-choice video question answering benchmarks: ***Next-QA*** [115] and ***AGQA*** [34]. The idea is to check if we can probe models for temporal understanding by asking questions with temporal language. Buch *et al*. [13] identify a subset of Next-QA, dubbed as `ATP-hard`[2], with a higher concentration of temporally challenging data. For AGQA, we pick a subset of ∼6k questions that explicitly have a question with before/after relation. We consider

[2] Available here: github.com/StanfordVL/atp-video-language

these benchmarks as requiring a moderate-high level of time awareness and AGQA in particular is also close to our adaptation task. We use accuracy as the standard metric. We observe (see Tab. 5 set II) that indeed TACT almost always outperforms (marked in green) baselines on both Next-QA and AGQA. TEMPO-adapted TACT seems to generalize particularly well on both benchmarks. Likewise, Charades-adapted TACT does well on AGQA since AGQA is also based on the Charades videos accounting for reduced domain shift. We affirm that temporal adaptation is useful, especially when the downstream tasks require it.

III. Action-to-video retrieval. Finally, we consider action recognition benchmarks such as ***Something-Something*** (SSv2) [33] and ***Temporal*** [88]. SSv2 was designed to capture richer temporal information [33, 56] . We follow Lei *et al*. [56], who propose the *template-retrieval* task that encourages temporal modelling and use their evaluation split[3] containing C=174 actions and K=12 videos per class. Interestingly, different actions in SSv2 require differing levels of time awareness. We create a subset ***SSv2 (events)*** with C_{events}=49 actions that have at least two verbs in the label as the occurrence of multiple verbs is indicative of multiple events occurring in sequence. Finally, we also evaluate against the Temporal benchmark [88], a combination of 50 action classes from SSv2 [33] and Kinetics-400 [52] for which temporal information is deemed to be essential for recognition. Similar to text-to-video retrieval, we use the action class as a text query and obtain a ranking over all videos. Different from the retrieval setup, since a single query has multiple correct answers (up to K=12 videos), we report mAP as the metric for these benchmarks. This task set needs high time awareness. Furthermore, unlike QA tasks in II, there is a *shift* in several (uncontrolled) factors as we move from temporal adaptation task to these tasks. From Tab. 5, we observe that TEMPO- and Charades-adapted models generalize well across set III benchmarks. ActivityNet-adapted TACT underperforms on SSv2 but outperforms the baseline on strongly temporal actions in SSv2 (events) and Temporal. Finally, TACT adapted on Charades-Ego is at-par or slightly worse than the baseline on SSv2 variants, and also on Temporal, perhaps due to the shift from egocentric to third-person videos. Overall, despite SSv2 and Temporal requiring high time awareness, TACT models show promising zero-shot generalization with the right choice of the adaptation dataset.

7. Discussion and Conclusion

Generalization to other 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

[3] Available here: github.com/jayleicn/singularity

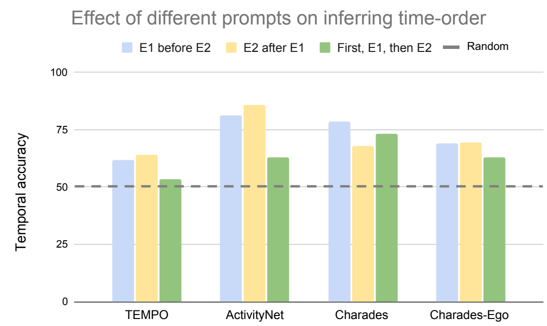

Figure 5. Models trained by TACT with before/after relations generalize to a new kind of prompt such as `First, .., then ..` indicating learning of the underlying true time-order of events.

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. 5, 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.

Limitations. While we present a promising way of instilling time in video-language models, our work is limited to the VideoCLIP [119] pretrained model. Our initial experiments with Frozen in Time [6] were not as promising, perhaps because it uses a much shorter temporal context (4 frames). Please see the supplement for results on more pretrained models. Furthermore, we consider a specific definition of time awareness derived from temporal relations like before/after. It is natural to ask if this can be extended to more general notions of temporality, *e.g*., as defined by Allen [4]. Finally, there can always be more downstream tasks considered such as (spatio-)temporal localization.

Conclusion. Given the essence of time in video-language models, we present a simple experiment based on synthetic data to test for time awareness. 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 [119], we present TACT, a recipe to instill this sense of time in video-language models. Finally, we analyze the zero-shot generalizability of TACT-adapted models to a diverse set of tasks. We hope that this work provokes further probing and instilling time awareness in video-language models, and also inspires other adaptations of foundational models to solve various challenging tasks.

Acknowledgements. We acknowledge support from ELLIS Amsterdam and the AMS Scholarship to Piyush. We thank Dr Dennis Koelma for their help with compute infrastructure and Dr Hazel Doughty for useful discussions.

References

[1] Unaiza Ahsan, Rishi Madhok, and Irfan Essa. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 179–189. IEEE, 2019. 2

[2] 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

[3] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelovi'c, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *ArXiv*, abs/2006.16228, 2020. 2

[4] James F. Allen. Towards a general theory of action and time. In *Artificial Intelligence*, 1984. 1, 3, 8

[5] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. *International Conference on Computer Vision (ICCV)*, pages 6816–6826, 2021. 2

[6] 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. 1, 3, 4, 8

[7] 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)*, 2021. 2, 5

[8] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. A clip-hitchhiker's guide to long video retrieval, 2022. 2, 7

[9] Nadine Behrmann, Mohsen Fayyaz, Juergen Gall, and Mehdi Noroozi. Long short view feature decomposition via contrastive video representation learning. *International Conference on Computer Vision (ICCV)*, pages 9224–9233, 2021. 2

[10] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9922–9931, 2020. 2

[11] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning (ICML)*, 2021. 2

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

[13] 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, 2, 3, 7

[14] Petr Byvshev, Pascal Mettes, and Yu Xiao. Are 3d convolutional networks inherently biased towards appearance? *Computer Vision and Image Understanding*, 220:103437, 2022. 2

[15] Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. Locvtp: Video-text pre-training for temporal localization. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[16] 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. 3, 4

[17] Hyeon Cho, Taehoon Kim, Hyung Jin Chang, and Wonjun Hwang. Self-supervised spatio-temporal representation learning using variable playback speed prediction. *IEEE Access*, 9:79562–79571, 2021. 2

[18] Ishan Rajendra Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding (CVIU)*, 219:103406, 2022. 2

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. 2

[20] 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, 3

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021. 2

[22] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015. 1, 3

[23] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 2, 3, 4

[24] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *International Conference on Computer Vision (ICCV)*, pages 6201–6210, 2019. 2

[25] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Conference on Computer

Vision and Pattern Recognition (CVPR), pages 3636–3645, 2017. 2

[26] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet : End-to-end video-language transformers with masked visual-token modeling. *ArXiv*, abs/2111.12681, 2021. 1

[27] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet : End-to-end video-language transformers with masked visual-token modeling. *ArXiv*, abs/2111.12681, 2021. 2

[28] Valentin Gabeur, Chen Sun, Alahari Karteek, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[29] 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)*, pages 16167–16176, June 2022. 2, 3, 4

[30] Amir Ghodrati, Efstratios Gavves, and Cees G. M. Snoek. Video time: Properties, encoders and evaluation. In *British Machine Vision Conference (BMVC)*, 2018. 2, 3

[31] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. *ArXiv*, abs/2011.00597, 2020. 2

[32] Rohit Girdhar, Mannat Singh, Nikhil Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16081–16091, 2022. 2

[33] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *International Conference on Computer Vision (ICCV)*, pages 5842–5850, 2017. 8

[34] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 7

[35] Rujun Han, Qiang Ning, and Nanyun Peng. Joint event and temporal relation extraction with shared representations and structured prediction. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019. 3

[36] Rujun Han, Xiang Ren, and Nanyun Peng. Deer: A data efficient language model for event temporal reasoning. *ArXiv*, abs/2012.15283, 2020. 2, 3

[37] 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, 3

[38] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2896–2906, 2022. 3

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '16, pages 770–778, June 2016. 2

[40] 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)*, pages 1380–1390, 2018. 5

[41] 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)*, pages 5804–5813, 2017. 1, 3, 5

[42] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with temporal language. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 3

[43] 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)*, pages 7366–7375, 2018. 2, 3, 7

[44] Yuqi Huo, Mingyu Ding, Haoyu Lu, Zhiwu Lu, Tao Xiang, Ji-Rong Wen, Ziyuan Huang, Jianwen Jiang, Shiwei Zhang, Mingqian Tang, Songfang Huang, and Ping Luo. Self-supervised video representation learning with constrained spatiotemporal jigsaw. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021. 2

[45] A. Jabri, Andrew Owens, and Alexei A. Efros. Space-time correspondence as a contrastive random walk. *ArXiv*, abs/2006.14613, 2020. 2

[46] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning (ICML)*, 2021. 2

[47] S. Jenni and Hailin Jin. Time-equivariant contrastive video representation learning. *International Conference on Computer Vision (ICCV)*, pages 9950–9960, 2021. 2

[48] Simon Jenni, Givi Meishvili, and Paolo Favaro. Video representation learning by recognizing temporal transformations. In *European Conference on Computer Vision (ECCV)*, pages 425–442, 2020. 2

[49] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, 2021. 1, 2

[50] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. *ArXiv*, abs/2112.04478, 2021. 2

[51] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. *ArXiv*, abs/2112.04478, 2021. 2

[52] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 8

[53] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 8545–8552, 2019. 2

[54] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 6

[55] 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. 5

[56] Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning, 2022. 1, 2, 3, 7, 8

[57] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7327–7337, 2021. 1, 3

[58] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020. 2

[59] 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, 7

[60] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. Compositional temporal grounding with structured variational cross-graph correspondence learning. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3022–3031, 2022. 3

[61] 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

[62] Kevin Lin, Alex Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z. Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, Hongfa Wang, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. Egocentric video-language pretraining. *ArXiv*, abs/2206.01670, 2022. 2

[63] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Y. Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. *ArXiv*, abs/2208.03550, 2022. 2

[64] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019. 2

[65] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 2

[66] 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, 4, 7

[67] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9876–9886, 2020. 2, 4

[68] 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. 2, 5

[69] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations (ICLR)*, 2013. 2

[70] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision (ECCV)*, pages 527–544, 2016. 2

[71] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. *European Conference on Computer Vision (ECCV)*, 2022. 2

[72] Qiang Ning, Zhili Feng, and Dan Roth. A structured learning approach to temporal relation extraction. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2017. 3

[73] Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. Joint reasoning for temporal and causal relations. In *Association of Computational Linguistics (ACL)*, 2018. 3

[74] Qiang Ning, Sanjay Subramanian, and Dan Roth. An improved neural baseline for temporal relation extraction. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019. 3

[75] Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. Torque: A reading comprehension dataset of temporal ordering questions. *ArXiv*, abs/2005.00242, 2020. 3

[76] Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. Cogcomptime: A tool for understanding time in natural language. *ArXiv*, abs/1906.04940, 2018. 3

[77] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11200–11209, 2021. 2

[78] Lyndsey C. Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard

Scholkopf, and William T. Freeman. Seeing the arrow of time. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2043–2050, 2014. 2

[79] Ianthe Pratt and Nissim Francez. Temporal prepositions and temporal generalized quantifiers. *Linguistics and Philosophy*, 24:187–222, 2001. 3

[80] Will Price and Dima Damen. Retro-actions: Learning 'close' by time-reversing 'open' videos. *International Conference on Computer Vision Workshops (ICCVW)*, pages 1371–1380, 2019. 2

[81] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, H. Wang, Serge J. Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6960–6970, 2021. 2

[82] Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. Timedial: Temporal commonsense reasoning in dialog. In *Association of Computational Linguistics (ACL)*, 2021. 3

[83] 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, 2, 3

[84] 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

[85] Adrià Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Patraucean, Florent Altch'e, Michael Valko, Jean-Bastien Grill, Aäron van den Oord, and Andrew Zisserman. Broaden your views for self-supervised video learning. *International Conference on Computer Vision (ICCV)*, pages 1235–1245, 2021. 2

[86] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019. 2

[87] Madeline Chantry Schiappa, Yogesh Singh Rawat, and Mubarak Shah. Self-supervised learning for videos: A survey. *ArXiv*, abs/2207.00419, 2022. 2

[88] Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. *CoRR*, abs/1907.08340, 2019. 2, 8

[89] Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time? *International Conference on Computer Vision (ICCV)*, pages 9641–9649, 2021. 2

[90] Vivek Sharma, Makarand Tapaswi, and Rainer Stiefelhagen. Deep Multimodal Feature Encoding for Video Ordering. In *International Conference on Computer Vision Workshops (ICCVW)*, 2019. 2

[91] 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)*, pages 7396–7404, 2018. 5

[92] Gunnar A. Sigurdsson, Olga Russakovsky, and Abhinav Kumar Gupta. What actions are needed for understanding human actions in videos? *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2156–2165, 2017. 2

[93] 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)*, pages 510–526, Cham, 2016. 5

[94] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogério Schmidt Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10384–10394, 2021. 2

[95] Chen Sun, Austin Myers, Carl Vondrick, Kevin P. Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. *International Conference on Computer Vision (ICCV)*, pages 7463–7472, 2019. 2

[96] Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. Long-form video-language pretraining with multimodal temporal contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[97] Tomoyuki Suzuki, Takahiro Itazuri, Kensho Hara, and Hirokatsu Kataoka. Learning spatiotemporal 3d convolution with video order self-supervision. In *European Conference on Computer Vision Workshops (ECCV)*, pages 0–0, 2018. 2

[98] Makarand Tapaswi, Martin Bauml, and Rainer Stiefelhagen. Book2movie: Aligning video scenes with book chapters. In *CVPR*, June 2015. 3

[99] Fida Mohammad Thoker, Hazel Doughty, Piyush Bagad, and Cees G. M. Snoek. How severe is benchmark-sensitivity in video self-supervised learning? In *European Conference on Computer Vision (ECCV)*, 2022. 2

[100] Fida Mohammad Thoker, Hazel Doughty, and Cees Snoek. Tubelet-contrastive self-supervision for video-efficient generalization, 2023. 2

[101] Fida Mohammad Thoker, Hazel Doughty, and Cees G. M. Snoek. Skeleton-contrastive 3d action representation learning. *ACM MM*, 2021. 2

[102] Shivin Thukral, Kunal Kukreja, and Christian Kavouras. Probing language models for understanding of temporal expressions. In *BLACKBOXNLP*, 2021. 3

[103] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021. 2

[104] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6459, 2018. 2

[105] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 5

[106] Siddharth Vashishtha, Adam Poliak, Yash Kumar Lal, Benjamin Van Durme, and Aaron Steven White. Temporal reasoning in natural language inference. In *FINDINGS*, 2020. 3

[107] 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, 2

[108] Guangting Wang, Yizhou Zhou, Chong Luo, Wenxuan Xie, Wenjun Zeng, and Zhiwei Xiong. Unsupervised visual representation learning by tracking patches in video. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2563–2572, 2021. 2

[109] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. *ArXiv*, abs/2209.07526, 2022. 1

[110] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. *ArXiv*, abs/2209.07526, 2022. 2

[111] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *European Conference on Computer Vision (ECCV)*, pages 504–521, 2020. 2

[112] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14713–14723, 2022. 2

[113] X. Wang, A. Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2561–2571, 2019. 2

[114] 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)*, pages 8052–8060, 2018. 2

[115] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, 2021. 7

[116] 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)*, pages 318–335, Cham, 2018. 5

[117] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin P. Murphy. Rethinking spatiotemporal feature learning for video understanding. *ArXiv*, abs/1712.04851, 2017. 2

[118] 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)*, pages 10334–10343, 2019. 2

[119] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pretraining for zero-shot video-text understanding. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 6787–6800, 2021. 1, 2, 3, 4, 5, 8

[120] 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, 3, 7

[121] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Rui Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pretrained image-text model to video-language representation alignment. *ArXiv*, abs/2209.06430, 2022. 2

[122] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *arXiv preprint arXiv:2206.08155*, 2022. 1

[123] Xinyu Yang, Majid Mirmehdi, and Tilo Burghardt. Back to the future: Cycle encoding prediction for self-supervised contrastive video representation learning. *ArXiv*, abs/2010.07217, 2020. 2

[124] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6548–6557, 2020. 2

[125] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. *ArXiv*, abs/1910.01442, 2020. 2

[126] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel C. F. Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *ArXiv*, abs/2111.11432, 2021. 1

[127] 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

[128] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. Centerclip: Token clustering for efficient text-video retrieval. In *SIGIR*, 2022. 3, 4

[129] Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. *ArXiv*, abs/1909.03065, 2019. 3

[130] Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. Temporal common sense acquisition with minimal supervision. *ArXiv*, abs/2005.04304, 2020. 2, 3

[131] 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, 3

[132] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 7590–7598, 2018. 3, 7