Revisiting the "Video" in Video-Language Understanding

Shyamal Buch[1], Cristóbal Eyzaguirre[1], Adrien Gaidon[2], Jiajun Wu[1], Li Fei-Fei[1], Juan Carlos Niebles[1]
[1]Stanford University, [2]Toyota Research Institute
{shyamal, ceyzagui, jiajunwu, feifeili, jniebles}@cs.stanford.edu, adrien.gaidon@tri.global

Abstract

What makes a video task uniquely suited for videos, beyond what can be understood from a single image? Building on recent progress in self-supervised image-language models, we revisit this question in the context of video and language tasks. We propose the atemporal probe (ATP), a new model for video-language analysis which provides a stronger bound on the baseline accuracy of multimodal models constrained by image-level understanding. By applying this model to standard discriminative video and language tasks, such as video question answering and text-to-video retrieval, we characterize the limitations and potential of current video-language benchmarks. We find that understanding of event temporality is often not necessary to achieve strong or state-of-the-art performance, even compared with recent large-scale video-language models and in contexts intended to benchmark deeper video-level understanding. We also demonstrate how ATP can improve both video-language dataset and model design. We describe a technique for leveraging ATP to better disentangle dataset subsets with a higher concentration of temporally challenging data, improving benchmarking efficacy for causal and temporal understanding. Further, we show that effectively integrating ATP into full video-level temporal models can improve efficiency and state-of-the-art accuracy.[1]

1. Introduction

Videos offer the promise of understanding not only what can be discerned from a single image (e.g. scenes, people, and objects), but also multi-frame event temporality, causality, and dynamics (Figure 1(a)). Correspondingly, there lies a central question at the heart of video research: *What makes a video task uniquely suited for videos, beyond what can be understood from a single image?*

As a field, video analysis has considered this question deeply in the context of action classification in videos [3, 17, 43, 50]. The emergence of strong convolutional models for image classification [15] enabled researchers to better characterize the limits of single-frame understanding for recognizing actions [17, 50]. A key finding from this analysis was that, in many standard video datasets [24, 47] at the time, temporal understanding was simply not required

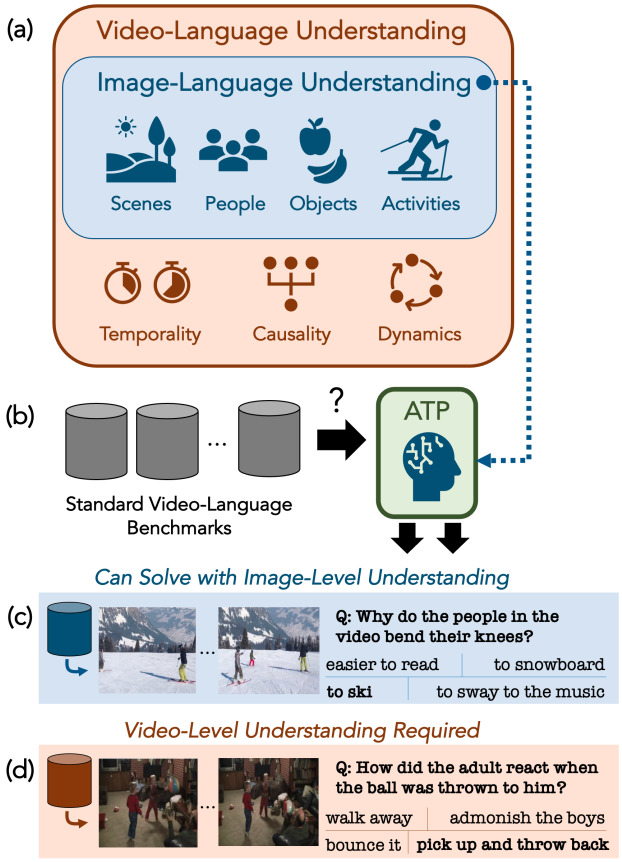

Figure 1. **(a)** The promise of *videos* lies in the potential to go beyond *image-level* understanding (scenes, people, etc.) to capture event temporality, causality, and dynamics. **(b)** In this work, we propose an atemporal probe (ATP) model to *revisit the video* in standard benchmarks [29, 53, 55] for video question answering and text-to-video retrieval, offering a stronger image-centric baseline and analytical tool. For example, ATP finds non-trivial subsets of "causal" questions that can be answered with **(c)** only image-level understanding, rather than **(d)** full video-level understanding.

[1]Project website: https://stanfordvl.github.io/atp-revisit-video-lang/

to perform well on these benchmarks. For example, recognizing static scene context like the presence of a pool was sufficient to recognize the "diving" activity from a single frame [31, 50]. The impact of such analysis was tremendous: later datasets were designed to capture a richer distribution of temporal understanding [6, 13, 46] with better disentanglement of such cues [33], and model designs evolved further to better capture the now necessary dynamics to address these improved tasks [9–11, 30, 51].

Meanwhile, the recent advent of self-supervised image-language models [20, 41] with competitive performance to standard image-classification models [7, 15] means that we have a unique opportunity to reconsider this fundamental question in the context of standard discriminative *video-language* tasks, such as video question answering [29, 53, 55] and video-language retrieval [16, 23, 55]. In particular, we can now build *beyond* prior (video-only) analysis work, largely constrained to recognition settings of limited atomic actions in relatively short clips, towards more complex (temporal, causal) event understanding in longer-horizon, multimodal settings where the expressivity of natural language can potentially describe a richer event space.

The primary motivation of our work is to analyze these existing video-language benchmarks by *revisiting the video*, and derive insights that can help guide the further development of the field. Our driving question is, to what extent can image-level understanding obtained from a single frame (well-chosen, without temporal context) address the current landscape of video-language tasks? To accomplish this, we make the following key contributions:

First, we introduce the atemporal probe (ATP) model to provide a stronger bound on the capabilities of image-level understanding in video-language settings than traditional random frame and mean pooling baselines [50]. Here, we leverage a *frozen* self-supervised image-language model (e.g. CLIP [40]) to extract a set of image and language representations: our ATP model must then learn to select a *single* frozen representation corresponding to a single frame, and forward that to the downstream video-language task. Critically, our framework is constrained to *not* be capable of reasoning temporally, and its output is ultimately bottlenecked by what a frozen image-language model can discern from an individual, decontextualized video frame.

Second, we apply ATP to analyze a wide range of video-language datasets, focusing primarily on video question answering with extensions to text-to-video retrieval (per Figure 1(b)). To our surprise, we find that many standard and recent benchmarks can be potentially well-addressed with single-frame image understanding. In particular, while this was not our primary aim, we find that our *learned* ATP model is able to outperform recent state-of-the-art video-language models on standard vision-language benchmarks [16, 23, 29, 53, 55], despite its substantial bottleneck constraints on model capacity, capability, and inputs. We find that even recent benchmarks that explicitly design for temporal and causal understanding (e.g., [53]), can have a non-trivial subset of questions answerable by simple single-frame event recognition. As shown in Figure 1(c), while the question asking "why" an event occurred suggests causal understanding may be needed, our ATP model shows that in practice simple scene and object recognition can ascertain the correct answer from a single chosen frame.

Finally, we examine how ATP and the insights it provides can help with improving both dataset and video-level temporal modeling designs. As a case study, we closely examine the NExT-QA benchmark [53]. We find that ATP is able to better identify collections of "causal" and "temporal" questions that *cannot* be well-addressed with single-frame understanding. In Figure 1(d), ATP struggles to answer this question since it necessitates multi-event reasoning across time. By improving the disentanglement of video- and image-level understanding in the benchmark data, we can better understand the progress of state-of-the-art video techniques leveraging motion features and event reasoning architectures over image-centric models, a result that is not as apparent in the original setting. We further validate our analysis by training a temporal video-level model on top of our ATP selectors, achieving a new state-of-the-art for this benchmark with improved efficiency. Taken together, our analysis suggests key avenues by which our ATP technique can guide continued development of video-language datasets and models in future work.

2. Background and Related Work

Our work is related to many different areas of vision and vision-language research, including video-specific and image-specific settings. In this section, we discuss the key relevant areas of prior work that motivate our contributions.

Video-language understanding (tasks). Understanding events in their multimodal vision-language context is a long-standing challenge for the computer vision community. Standard video-language tasks include both discriminative tasks, such as video question answering [12, 19, 26, 27, 29, 53, 55, 57, 58], text-to-video/moment retrieval [16, 23, 42, 55, 61], and generative tasks, such as video captioning [4, 23] and open-ended VQA [53, 55]. In context, we choose a representative subset of these video-language benchmarks well-suited to studying event temporality and causality. In particular, we choose to focus on *discriminative* tasks, since automatic metrics (without human-in-the-loop) for generative tasks with causal descriptions remains an open research challenge [38]. Furthermore, many video-language tasks involve heavy reasoning over auxiliary text inputs, such as scripts [8, 58]. These exciting directions are complementary to our goal: we focus instead on revisiting event temporality in the real-world videos themselves.

Video-language understanding (approaches). Standard approaches for addressing these tasks [21, 23, 28, 34, 48, 54] often operate on a combination of image-derived appearance [7, 15] and video-derived motion features [3, 35, 39, 45] as input to an architecture [49, 60] that combines information across the temporal dimension for the final task. While these models are traditionally quite heavy, employing dense features extracted from many frames, recent work [25] has suggested that enabling end-to-end training through sparsity can improve accuracy. Our proposed approach aims to complement these prior lines of work by taking a different approach: instead of focusing explicitly on improving state-of-the-art accuracies, we impose strong learnability and representation constraints to better *analyze* the degree to which full video-level understanding is truly necessitated by current benchmarks, to help guide future model and dataset designs for capturing deeper event understanding.

Temporality in videos (action recognition). Action and event recognition are fundamental tasks for video understanding, and the subject of recurring deep analysis regarding the role of temporality in action classification [3, 17, 43, 50, 60], with profound downstream impacts on dataset [6, 13, 46] and subsequent model designs [9–11, 30, 51]. We draw inspiration from this foundational prior work, while also aiming to broaden analysis beyond characterizing limited sets of atomic actions towards longer-horizon temporal and causal event understanding, which multimodal video-language contexts have the potential to better capture [53].

Image-language understanding. The advent of new self-supervised vision-language models trained at scale [20, 40], where models learn a joint embedding space for vision [7, 15] and language [5, 32] *without* explicit low-level labels, has proven disruptive for image and image-language understanding tasks [1, 40, 44]. We leverage these models, both vision and language components, as foundations for our analytical technique to better characterize the extent to which image-language understanding can address current video-language tasks. Our work is *complementary* to prior image-language analytical work [14] which revealed unintended language bias: we aim to characterize the extent of unintended video-specific biases in this multimodal setting.

Efficient image-centric video modeling. Finally, we note that aspects of our technical approach draw inspiration from efficient image-centric video modeling literature, which aim to improve efficiency and for tasks like action recognition [52] and localization [56] by learning how to selectively process a sparse number of frames from the input video.

3. Technical Approach

In this section, we describe our technical approach for our atemporal probe (ATP), a new modeling tool for characterizing the boundary of image-constrained understanding in the context of standard discriminative video-language tasks.

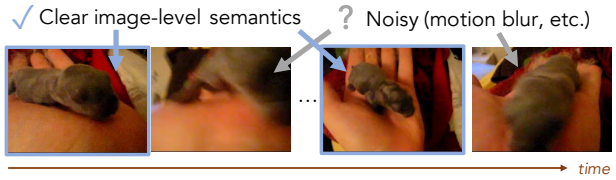

Figure 2. **Motivating a stronger image-centric baseline**. Videos are noisy, correlated collections of frames [31]: while some frames have clear image-level semantics (*above:* a small puppy dog in a human hand), a significant fraction of frames can contain camera motion blur, difficult perspectives, and uninformative frames. Standard atemporal techniques, such as evaluating image-level models on a random frame or mean pooling, may be susceptible to such noise, and thus not necessarily represent a true bound on image-level semantics understanding in video-language contexts. This motivates our atemporal probe (ATP) model (Sec. 3).

3.1. Preliminaries: Video-Language Tasks

We first briefly introduce the notation and discriminative video-language tasks we consider in this work, namely video question answering and text-to-video retrieval:

Video question answering. Our primary analysis setting is on video question answering: given a paired collection of videos C_V, and language questions and answers $C_L = \{C_Q, C_A\}$, the goal is for each (video, question) pairing (V, Q) to provide the correct answer in A.

Video-language retrieval. We also examine video-language retrieval, to assess the generality of our approach. In text-to-video retrieval, the objective is complementary: given a paired collection of videos C_V and language descriptions C_L, the goal is to use the language L to retrieve the specific video V that it originally corresponded with.

We note that in both settings, there exist video V and language L $(= (Q, A))$ inputs common to each task. While our work ultimately analyzes performance on these downstream tasks with respect to their inputs and metrics, our core goal for this work is to provide an improved analytical tool for characterizing specific instantiations of these tasks.

3.2. Motivating a Stronger Image-Centric Baseline

Traditionally, video models and benchmarks establish their efficacy over image-level understanding by reporting results with a model based on a single (center-most, randomly, etc.) chosen video frame [50]. Because videos can be considered noisy collections of frames, such baselines may not truly represent the bounds of what image-constrained understanding can achieve in video-language contexts (Figure 2). In particular, we seek to answer the question: if we can select a "good frame" from the video and only derive our understanding from that one frame, what video-language tasks are we capable of performing?

Intuitively, settings where only scene-level descriptions

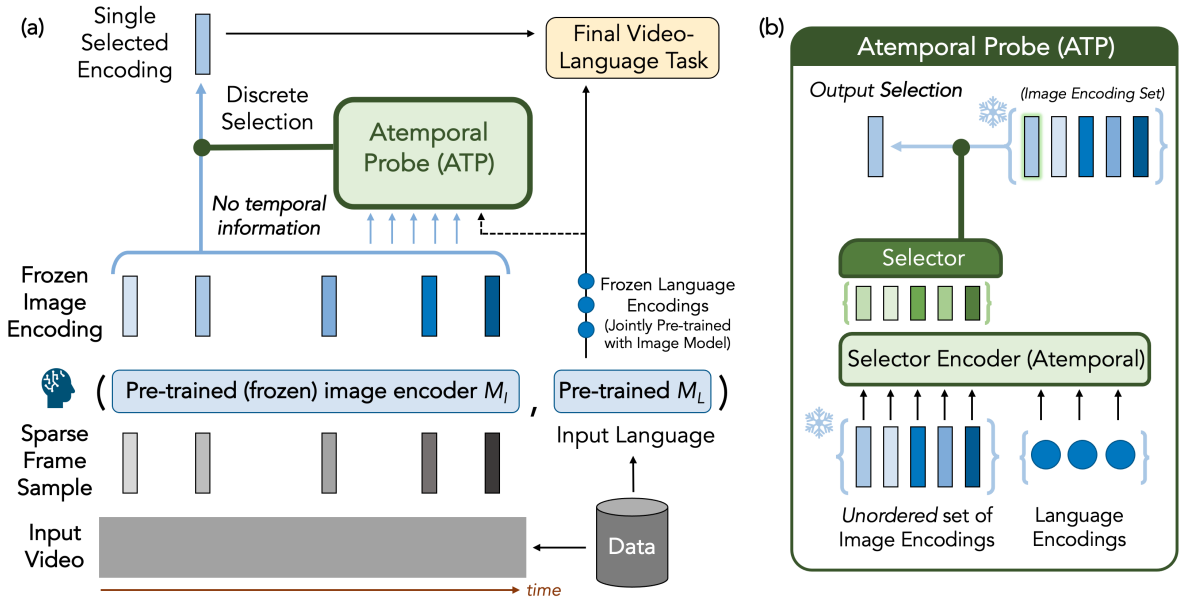

Figure 3. **Atemporal Probe (ATP).** We propose ATP: a new, stronger baseline for characterizing the degree to which video-language tasks can be addressed exclusively with vision-language understanding derived from image-only settings (i.e. jointly learned pre-trained encoders for image M_I and language M_L). **(a)** In the broader context of a video-language task, such as video question answering, our ATP model must learn to select a *single* (frozen, image-derived) embedding that can provide as strong a signal as possible for the final task. **(b)** Zooming in, we emphasize that our ATP model does not use any temporal information as part of this selection, operating on an unordered (shuffled) set of frame-level embeddings (without temporal positional encodings) with permutation invariant self-attention operations. Furthermore, the learnable atemporal selector encoder remains low capacity. Please see Section 3.3 for additional details.

are being assessed should likely be addressable from a single frame, as should simple event recognition (per prior analysis in the domain of action recognition, Section 2). However, by the same intuition, questions/tasks that attempt to fully assess deeper event dynamics, causal, or temporal understanding should in principle be *unanswerable* from a single frame alone, requiring reasoning over multiple events which are not necessarily co-located in time. A compelling baseline that effectively bounds image-level understanding can thus potentially help distinguish between these settings.

3.3. Atemporal Probe (ATP) Model

Overview. With the motivating insight above, we propose an atemporal probe (ATP) model: a new, stronger analytical approach for characterizing the degree to which video-language tasks can be addressed exclusively with vision-language representations derived from image-only settings. The ATP model (Figure 3) is tasked with finding a single (frozen, image-derived) embedding from the video and forwarding this to the downstream video-language task. Our ATP model does *not* use any temporal information to perform this selection, processing unordered frame embeddings with permutation invariant self-attention operations (without any sequence positional information). Further, we ensure that the learnable portion of ATP remains low capacity, with only a few, small layers and number of heads.

ATP (Context). We illustrate an overview of our ATP model in the larger video-language task context in Figure 3(a). For each video $V \in C_V$, we draw a random sparse (shuffled) subset of frames $F = \{v_1, \ldots, v_n\} \in V$, where usually $n << |V|$, the length of the video. We also take as input to our task a pretrained, self-supervised image-language model $M = \{M_I, M_L\}$, which consists of two components M_I and M_L for the vision and language components, respectively. These are used to encode all video V and language L inputs to the original video-language task.

We proceed to encode each of the frames with the pre-trained vision encoder $M_I(F) = \{x_1, \ldots, x_n\}$ to get vision embeddings x_i corresponding to each frame v_i. Intuitively, because our encoder is completely frozen and never updated, x_i is a representation of what an *image-constrained* visual encoder can discern; no additional information of the broader video is encoded here. Furthermore, our model treats the set $\{x_1, \ldots, x_n\}$ as an *unordered* set, without any temporal positional information.

Now, ATP can be properly formulated as:

$$ATP : \{x_1, \ldots, x_n\} \mapsto x_i, \tag{1}$$

where the goal is to select a single representation $x_i \in \{x_1, \ldots, x_n\}$ to pass to the final video-language task. Depending on the original video-language task formulation, ATP can take additional language inputs $M_L(L)$ (e.g. the

encoded question for video question answering; Sec. 4.1).

ATP (Selection). In Figure 3(b), we illustrate a more detailed view of the ATP selection operation. Given the inputs provided by the frozen pre-trained image and language encoders, the ATP model must now perform *embedding selection*, passing one of these input visual embeddings, unmodified, to the downstream video-language task. To accomplish this, ATP first encodes the (unordered, shuffled) input image encoding sequence $\{x_1, \ldots, x_n\}$ with a learnable selector encoder E_s as follows:

$$E_s(\{x_1, \ldots, x_n\}; M_L(L)) \mapsto \{s_1, \ldots, s_n\}, \quad (2)$$

where $\{s_1, \ldots, s_n\}$ correspond to the original $\{x_1, \ldots, x_n\}$ and are only used for selection. We instantiate E_s in our work as low-capacity transformer [49], with 3 or fewer layers and heads: we choose a self-attention architecture here because it is permutation invariant. Because our original embedding sequence $\{x_1, \ldots, x_n\}$ is unordered, and we provide no positional encodings (only learnable modality encodings [25] to differentiate vision from language inputs), this operation is thus strictly *atemporal*.[2] These encodings $\{s_1, \ldots, s_n\}$ are input to a final multilayer perceptron (MLP) to obtain logits for the final selection operation:

$$MLP(\{s_1, \ldots, s_n\}) \mapsto g \in \mathbb{R}^n. \quad (3)$$

Our final selection operation is discrete: ATP must select a single embedding x_i. To ensure learnability, we consider two versions of our selector S *during training*, both operating on the logits g: the first is to employ a straight-through Gumbel-Softmax estimator [18], the second is to apply softmax and ensuring entropy decreases over time [9]. In either case, at final test-time inference, the operation is made fully discrete ($S(g) \mapsto x_i$); see supplement for details.

Training. ATP is trained within the context of the overall video-language task framework, where the groundtruth answer or retrieval supervises the task loss, and gradients are backpropagated into the learnable ATP parameters. We reiterate that no modifications are made to the frozen image-language encodings, and the final video-language task is performed directly on these frozen representations without any additional downstream learnable parameters. For both tasks, we optimize for the groundtruth similarity between the vision and language encodings. For video question answering, we consider a cross entropy loss over the answer set [53], and for retrieval our loss is based on the standard InfoNCE contrastive loss [40]; see supplement for details.

3.4. Improving Temporal Modeling with ATP

In the final part of our experiments (Section 4), we additionally consider how our learned ATP embedding selector models (in Section 3.3) can improve downstream temporal models (Figure 4). Intuitively, ATP learns to be an effective (language-conditional) event recognizer; building on this intuition, we propose a straightforward model that partitions the original video V into k partitions $V^{(1)}, \ldots, V^{(k)}$ and runs (a learned, *now frozen*) ATP model on each partition to obtain selected candidate embeddings $x_i^{(1)}, \ldots, x_j^{(k)}$ for the k partitions. These per-partition outputs are then useful candidates for a separate, final learnable model T that performs temporal reasoning and outputs a video-level embedding for the final video-language task. In Section 4.3 experiments, this *downstream* temporal model is a distinct transformer model, equipped to perform video-level reasoning *on top of* ATP's output selections (for details, see supplement).

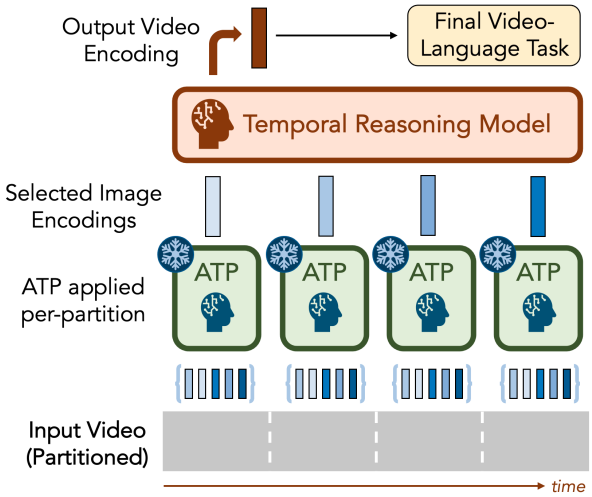

Figure 4. **Improving temporal modeling with ATP.** In our Section 4.3 case study, we make use of learned single-embedding ATP selectors to improve temporal modeling. Intuitively, ATP learns to surface frames rich in single event-level information. Building upon this, we propose a simple approach to partition the original video and run (a now *frozen*) ATP on each part. These per-partition selection outputs are then useful candidates for a separate downstream learnable model to perform temporal reasoning and output a video-level embedding for the final video-language task.

4. Experiments

4.1. Benchmark and Implementation Details

Benchmarks. We consider three representative benchmarks for video question answering: NExT-QA [53], VALUE-How2QA [28, 29], and MSR-VTT-MC [55]. We also examine the generality of our ATP model for text-to-video retrieval on DiDeMo [16], MSR-VTT [55], and ActivityNet [23]. For each benchmark, we follow standard protocols outlined by prior work [25, 29, 53, 54] for dataset processing, metrics, and settings; see supplement for details and analysis. We choose these benchmarks specifically

[2] We include detailed experimental analysis and discussion of ATP atemporality (including *relative* vs. *absolute* encoder designs) in the supplement.

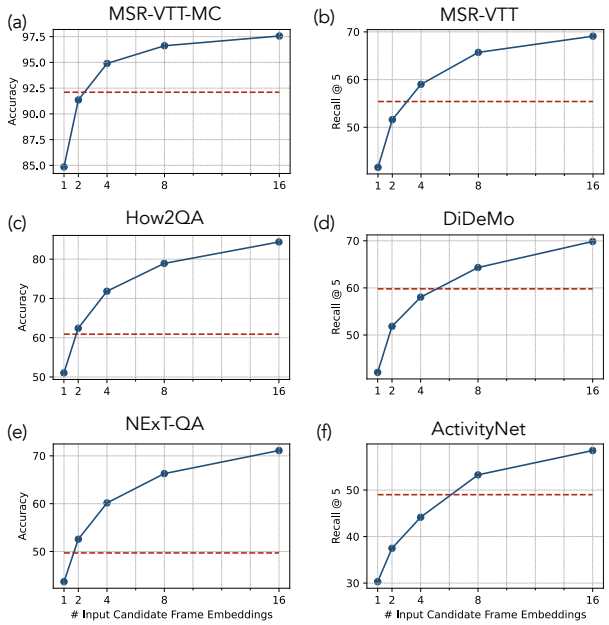

Figure 5. **Oracle upper bound analysis.** As a preliminary step, we analyze the performance upper bound of ATP under oracle conditions (with respect to the downstream task). Recall and accuracy (y-axis) averaged over multiple random samples of n frames (x-axis). We observe that the upper bounds are competitive with state-of-the-art video models even when choosing one embedding from relatively few-frame samples. Dashed reference lines are state-of-the-art models (**(a,b)** [54], **(c)** [29], **(d)** [2], **(e)** [53], **(f)** [25]).

to provide a broad coverage of durations, source video domains (general activities, instructional, etc.), and designs.

Implementation. We implement our ATP model with a few-layer, low-capacity transformer [49] in PyTorch [36], and train all models using the Adam [22] optimizer. Main paper results here reported on ViT-B-32 (CLIP) inputs for consistency [7, 29, 41]. See supplement[3] for more.

4.2. Analyzing Video-Language with ATP

Preliminary (upper bound) analysis. As a preliminary step, we first examine the performance of ATP under oracle conditions (with respect to the downstream video-language task) to establish a kind of upper bound for ATP on the set of benchmarks. In this analysis, we sample n input frames from the video, varying n, and encode them with a pre-trained model. In this oracle setting *only*, ATP then selects a frozen embedding from this set that maximizes the downstream groundtruth accuracy on the video-language task. Note that in this analysis, the oracle empowered ATP is still bottlenecked by what the image-level representation is able to capture. We repeat this analysis for multiple samples (dependent on the video lengths), and report the average. As shown in Figure 5, we observe that upper bound accuracies are competitive with state-of-the-art video models, even with relatively small n sample sizes, suggesting the promise for analyzing these datasets with a learnable ATP.

[3]Please see project website for supplementary material and code release.

MSR-VTT-MC	Accuracy
ActBERT [62]	85.7
ClipBERT [25]	88.2
MERLOT [59]	90.9
VideoCLIP [54]	92.1
CLIP (single-frame)	84.8
Ours (ATP; $1 \leftarrow 4$)	91.4
Ours (ATP; $1 \leftarrow 8$)	**92.5**
Ours (ATP; $1 \leftarrow 16$)	**93.2**

Table 1. **VideoQA on MSR-VTT-MC.** We find that our learned ATP model significantly outperforms prior work, indicating that this dataset can be largely addressed with image-level understanding. ($1 \leftarrow n$ means 1 embedding chosen from n sampled.)

VALUE-How2QA	Accuracy
Random	25.0
HERO [28]	60.4
HERO+ [28]	60.9
CLIP (single-frame)	50.1
CLIP (mean pooling)	55.7
Ours (ATP)	**65.1**

Table 2. **VideoQA on VALUE-How2QA.** We observe strong performance over previous state-of-the-art baselines on instructional video data. HERO+ baseline here has the same preprocessing as our model, and all models leverage the same CLIP features (HERO baselines additionally leverage heavy motion features [11, 29]).

ATP analysis (video QA). We apply a learnable ATP model to analyze a suite of standard video-language benchmarks. We first center our analysis discussion on video question-answering (video QA) benchmarks, since we find these benchmarks provide strong potential for deep multi-event understanding. Per Section 4.1, we focus on three representative benchmarks for analysis: NExT-QA [53], VALUE-How2QA [28, 29], and MSR-VTT-MC [55]. In Tables 1, 2, and 3, we report the results for each benchmark.

On MSR-VTT-MC (Table 1), our learned ATP model outperforms recent state-of-the-art video-language models [25, 54, 59], when considering relatively few frames at inference and despite its substantial (single-frame) bottleneck constraints on model capacity, capability, and inputs. ATP also substantially improves over standard atemporal baselines, including random single-frame and mean-pooling baselines with CLIP [40], offering a stronger bound.

On VALUE-How2QA (Table 2), we find that our learned ATP model offers significantly stronger accuracies than prior state-of-the-art models. Note that the HERO baselines here also use the same input CLIP embeddings, and no auxiliary text inputs, for fair comparison. One takeaway finding from our analysis of this benchmark was that counting questions, often designed to track state over the course

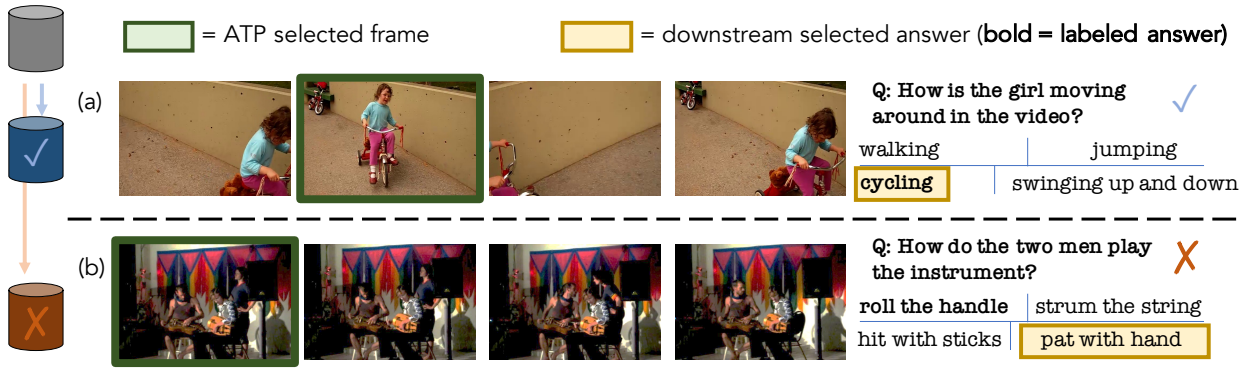

Figure 6. **ATP analysis (qualitative results).** We visualize example videos from the NExT-QA dataset [53], along with the selections ATP made from a random sparse sample of frames. Both questions shown here are examples of "causal-how" questions in the dataset (shown with the top-4 answer options, for clarity). **(a)** We find that our ATP model can select informative frames for the downstream Video QA task, when possible, and that many questions initially intended to assess causal or temporal understanding can be answered from single-frame semantics. **(b)** Conversely, for (video, question) inputs that necessitate a deeper *multi-frame* understanding of event relationships or dynamics, ATP's selected embedding is insufficient to answer the query. See Sec. 4.2 (additional visuals and datasets in supplement).

NExT-QA	Acc	Acc-D	Acc-T	Acc-C
Random	20.0	20.0	20.0	20.0
MAIN DATASET				(*Section 4.2*)
CLIP (single-frame)	43.7	53.1	39.0	43.8
HGA [21]	49.7	59.3	50.7	46.3
HGA [21] + CLIP [40]	50.4	59.3	52.1	46.8
Ours (ATP)	49.2	58.9	46.7	48.3
Ours (Temp[ATP])	**51.5**	65.0	49.3	48.6
Ours (Temp[ATP] + ATP)	**54.3**	66.8	50.2	53.1
ATP_{hard}-SUBSET				(*Section 4.3*)
Ours (ATP)	20.2	23.9	22.6	19.6
Ours (Temporal[ATP])	38.8	46.8	36.5	38.4
HGA [21]	44.1	51.2	45.3	43.3

Table 3. **VideoQA on NExT-QA.** We report accuracies on the overall main dataset and descriptive (D), temporal (T), and causal (C) splits. See Section 4.3 for details on the "Temp[ATP]" and "Temp[ATP] + ATP" models, and details on our ATP_{hard} subset.

	MSR-VTT		*DiDeMo*		*ActivityNet*	
	R@1	R@5	R@1	R@5	R@1	R@5
Support Set [37]	30.1	58.5	-	-	29.2	61.6
VideoCLIP [54]	30.9	55.4	16.6*	46.9*	-	-
ClipBERT [25]	22.0	46.8	20.4	48.0	21.3	49.0
CLIP (single-frame)	21.6	44.6	20.2	42.5	12.5	30.3
Ours (ATP)	27.8	49.8	26.1	50.5	17.7	41.8

Table 4. **Video-language (text-to-video) retrieval.** We show that our ATP analysis technique generalizes beyond video question answering settings. (* indicates zero-shot performance reported; see supplement for more complete prior work comparisons table.)

of a video, were in fact often addressable by a single well-chosen frame that showed sufficient number of the items.

Finally, on NExT-QA (Table 3), we find that even this recent benchmark, which is explicitly designed for temporal and causal understanding, can have a non-trivial subset of questions answerable by simple single-frame event recognition. In Figure 6, we show two different "causal-how" questions, which aim to assess both causality and dynamics. In the case of Figure 6(a) specifically, we observe that as long as the ATP model is able to select the informative frame with a clear depiction of the child on the cycle, the answer is readily apparent without deep video-level understanding. Quantitatively, our ATP model provides a stronger bound than standard image-level baselines; we also augment the the HGA baseline with CLIP features for fair comparison.

ATP analysis (retrieval). We also apply our learnable ATP model on standard retrieval benchmarks: DiDeMo [16], MSR-VTT [55], and ActivityNet [23]. In Table 4, we observe that our technique generalizes well to other discriminative video-language settings, establishing stronger bounds on image-centric performance and showing competitive accuracies with recent state-of-the-art methods. Our ATP model's performance on *paragraph* retrieval settings, like ActivityNet, highlights an area of improvement for image-bottleneck understanding: because paragraphs describe multiple dense events in long videos, it can be difficult to use a single frame embedding to capture this description well. We provide an extended discussion of prior work comparisons, limitations, and potential future directions for text-to-video retrieval as part of our supplement.

4.3. Improving Dataset and Model Design with ATP

Finally, we consider how to leverage our ATP model and its insights to improve both dataset and model design. For this section, we choose to focus on the NExT-QA benchmark [53] as a case study, since it is a key recent effort towards improving the field's focus on causal and temporal understanding in video-language tasks.

Improving dataset design with ATP. From our initial analysis of the NExT-QA benchmark in Section 4.2, we found that ATP provides a surprising degree of accuracy on causal

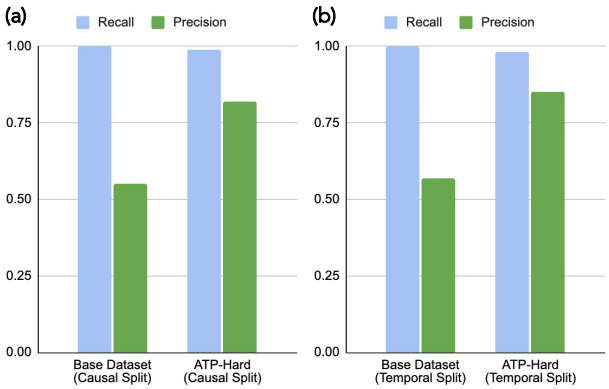

Figure 7. **Improving dataset design with ATP.** We analyze the original NExT-QA [53] benchmark, and find that our ATP models (denoted ATP$_{hard}$) are able to better disentangle input (video, question) pairs that *truly* necessitate video-level understanding compared with the original dataset, on both **(a)** casual and **(b)** temporal splits. This indicates promise for leveraging ATP in-the-loop for future dataset designs. See Section 4.3 for analysis details.

and temporal questions, despite its strong image-centric bottleneck. Because ATP provides a stronger bound on the capability of image-level understanding for these questions, it can help better disentangle questions that necessitate full video-level understanding (such questions will be largely unanswerable for the ATP model) from ones that do not.

We accomplish this by considering an ensemble of ATP models on the dataset, and leveraging their confidences and agreement to determine a subset of ATP$_{hard}$ questions. We determine any heuristics through k-fold cross validation on the *training* set. In parallel, we manually annotate a subset of the validation set for (video, question) pairs that necessitate true video-level understanding (see supplement for procedure details and limitations of our ATP technique). The results of our final analysis are shown in Figure 7. We find that our ATP based technique maintains the recall of the true video-level understanding questions on both the causal and temporal dataset splits, while substantially improving upon their precision (by filtering out "easy" questions).

Furthermore, we can also show how this ATP$_{hard}$ subset better benchmarks progress on video-level causal and temporal understanding (in Table 3) that may have been otherwise obscured. While ATP nearly matches the other models on the main dataset due to the inclusion of "easier" questions, this harder subset reveals a substantial gap relative to the state-of-the-art temporal reasoning model.

Together, these results suggest ATP in-the-loop can be an effective tool during future dataset design and creation.

Improving model design with ATP. As described in Section 3.4, we can leverage ATP to provide candidate frame embeddings for a downstream temporal model. As a first step towards improving temporal modeling (and efficiency), we introduce this model (denoted Temp[ATP] in Table 3) and benchmark it on the NExT-QA dataset. This model achieves a new state-of-the-art accuracy, outperforming the HGA (and HGA + CLIP) baselines on the main NExT-QA dataset, while operating at significantly reduced processing cost due to ATP (see supplement for efficiency discussion).

We make two additional observations: first, on the ATP$_{hard}$ subset, we find that this temporal model recovers much of the performance gap between ATP and the HGA model (we attribute the remaining gap to HGA's incorporation of additional motion features, which can aid in addressing some challenging dynamics questions), further verifying the potential dataset design contribution of ATP. Second, we observe that the aggregated confidence scores of the ATP ensemble provides a clear disentanglement signal on hard vs. easy problems, without access to groundtruth. Setting a heuristic threshold with k-fold cross validation on training, we use this signal to smartly ensemble ATP and Temp[ATP] further. For questions ATP can address, we do not need to consider additional (potentially noisy) frames, and we skip the full temporal model. For ones where ATP is less confident, the temporal model is run. This ensemble (denoted Temp[ATP] + ATP in Table 3) achieves a significant further accuracy-efficiency increase on NExT-QA.

5. Conclusion

In this work, we revisit a fundamental question of video understanding (*what makes a video task uniquely suited for videos, beyond what can be understood from a single image?*), building beyond prior analyses in action recognition towards video-language settings with more complex events. First, we propose an atemporal probe (ATP) model to provide a stronger bound on how much of video-language understanding can be addressed from image-language understanding only. Second, we use ATP to characterize both the limitations and potential of current video-language benchmarks for video question answering and video-language retrieval. Surprisingly, we find that single frame understanding can often achieve strong performance, even in settings intended for complex multi-frame event understanding and compared with recent large-scale video models. Third, we show how ATP can be leveraged to improve designs for both video-language datasets (disentangling unintentional atemporal biases) and video-level models (improving efficiency and accuracy). Going forward, we envision ATP as joining a broader, standard toolkit for video-language researchers and practitioners, revealing insights into complementary, *video-specific* sources of bias in multimodal video settings.

Acknowledgements. This work is supported by Toyota Research Institute, Stanford Institute for Human-Centered AI, Samsung, Salesforce, Office of Naval Research (N00014-19-1-2477), and NDSEG Fellowship (S.B.). Full ack. and discussion of limitations + broader impacts in supplement.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 3

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. *arXiv preprint arXiv:2104.00650*, 2021. 6

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1, 3

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv*, 2015. 2

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 3

[6] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large scale holistic video understanding. In *European Conference on Computer Vision*, pages 593–610. Springer, 2020. 2, 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021. 2, 3, 6

[8] Deniz Engin, François Schnitzler, Ngoc QK Duong, and Yannis Avrithis. On the hidden treasure of dialog in video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2064–2073, 2021. 2

[9] Linxi Fan*, Shyamal Buch*, Guanzhi Wang, Ryan Cao, Yuke Zhu, Juan Carlos Niebles, and Li Fei-Fei. RubiksNet: Learnable 3D-Shift for Efficient Video Action Recognition. In *European Conference on Computer Vision*, pages 505–521. Springer, 2020. 2, 3, 5

[10] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *CVPR*, 2020. 2, 3

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 2, 3, 6

[12] Rohit Girdhar and Deva Ramanan. Cater: A diagnostic dataset for compositional actions and temporal reasoning. *arXiv preprint arXiv:1910.04744*, 2019. 2

[13] 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 *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 2, 3

[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2, 3

[16] 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)*, 2018. 2, 5, 7

[17] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7366–7375, 2018. 1, 3

[18] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 5

[19] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017. 2

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *ICML*, 2021. 2, 3

[21] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 3, 7

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[23] 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. 2, 3, 5, 7

[24] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 1

[25] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learningvia sparse sampling. In *CVPR*, 2021. 3, 5, 6, 7

[26] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 2

[27] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. In *EMNLP*, 2020. 2

[28] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), pages 2046–2065, Online, Nov. 2020. Association for Computational Linguistics. 3, 5, 6

[29] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, et al. Value: A multi-task benchmark for video-and-language understanding evaluation. *NeurIPS (Benchmarks and Datasets Track)*, 2021. 1, 2, 5, 6

[30] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 2, 3

[31] Xin Liu, Silvia L Pintea, Fatemeh Karimi Nejadasl, Olaf Booij, and Jan C van Gemert. No frame left behind: Full video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14892–14901, 2021. 2, 3

[32] 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 preprint arXiv:1907.11692*, 2019. 3

[33] Mandy Lu, Qingyu Zhao, Jiequan Zhang, Kilian M Pohl, Li Fei-Fei, Juan Carlos Niebles, and Ehsan Adeli. Metadata normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10917–10927, 2021. 2

[34] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 3

[35] 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 *Proceedings of the IEEE international conference on computer vision*, pages 2630–2640, 2019. 3

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6

[37] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F. Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *International Conference on Learning Representations*, 2021. 7

[38] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. An information divergence measure between neural text and human text. *NeurIPS*, 2021. 2

[39] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *CVPR*, 2017. 3

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2, 3, 5, 6, 7

[41] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018. 2, 6

[42] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *CVPR*, 2015. 2

[43] Konrad Schindler and Luc Van Gool. Action snippets: How many frames does human action recognition require? In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 1, 3

[44] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021. 3

[45] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014. 3

[46] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864*, 2020. 2, 3

[47] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[48] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473, 2019. 3

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3, 5, 6

[50] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 1, 2, 3

[51] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross Girshick. Long-Term Feature Banks for Detailed Video Understanding. In *CVPR*, 2019. 2, 3

[52] Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S Davis. Adaframe: Adaptive frame selection for fast video recognition. In *CVPR*, 2019. 3

[53] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: next phase of question-answering to explaining

temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, June 2021. 1, 2, 3, 5, 6, 7, 8

[54] 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 *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, Nov. 2021. Association for Computational Linguistics. 3, 5, 6, 7

[55] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 1, 2, 5, 6, 7

[56] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2678–2687, 2016. 3

[57] 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 preprint arXiv:1910.01442*, 2019. 2

[58] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134, 2019. 2

[59] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *NeurIPS*, 2021. 6

[60] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 3

[61] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 2

[62] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8746–8755, 2020. 6