FitCLIP: Refining Large-Scale Pretrained Image-Text Models for Zero-Shot Video Understanding Tasks

Santiago Castro*[1]
https://santi.uy/
Fabian Caba Heilbron[2]
https://fabiancaba.com/

[1] University of Michigan
[2] Adobe Research

Abstract

Large-scale pretrained image-text models have shown incredible zero-shot performance in a handful of tasks, including video ones such as action recognition and text-to-video retrieval. However, these models have not been adapted to video, mainly because they do not account for the time dimension but also because video frames are different from the typical images (*e.g.*, containing motion blur, less sharpness). In this paper, we present a fine-tuning strategy to refine these large-scale pretrained image-text models for zero-shot video understanding tasks. We show that by carefully adapting these models we obtain considerable improvements on two zero-shot Action Recognition tasks and three zero-shot Text-to-video Retrieval tasks. The code is available at https://github.com/bryant1410/fitclip

1 Introduction

Imagine is winter season and our quest is to develop an auto-tagging system that recognizes all the activities in our winter vacation footage. Luckily, there have been tremendous advances in the action recognition community [3, 7, 54]. For instance, we could leverage one of the existing models that recognize up to 700 human actions [31]. Sadly, it turns out that our family's favorite activity, sledding, is not on the list of categories that these models can recognize. In a traditional supervised setting, we would have to collect many sledding examples to train a new model. Such a process is labor-intensive, costly to create, and difficult to scale to recognize further new activities. Instead, zero-shot models [6, 32, 47] can alleviate such a burden by enabling recognition of unseen concepts.

Large pre-trained image-text models, such as CLIP [45] and ALIGN [29], have shown outstanding zero-shot capabilities on a handful of visual tasks, including video tasks such as Action Recognition and Text-to-Video Retrieval. Such models have overcome the limitations of traditional zero-shot learning algorithms by using abundant images (on the internet) with (free) natural language supervision. Despite their remarkable zero-shot performance in video tasks, there is room for improvement to close the image-to-video domain gap. For

*Work done as an intern at Adobe Research.

instance, recent studies have shown that fine-tuning CLIP yields significant improvements in target video tasks [36, 56]. Unfortunately, fine-tuning and improving performance in a target dataset comes with a cost: harshly penalizing the model's zero-shot capabilities [59].

There have been multiple efforts to train video-language models that can be employed for various downstream video understanding tasks. Even though these approaches use video data, their zero-shot capabilities remain poor compared to those exhibited by CLIP [45]. It would be unfair not to mention that video-language pretraining methods either train with clean yet two orders of magnitude smaller datasets [4], or with large datasets with unaligned natural language supervision [38]. The alternative is to scale up further the amount of unaligned natural language supervision abundant on internet videos. In comparison, ALIGN [29] (in the image space) has shown the ability to cope with noisy supervision by scaling up to the billion-samples scale. However, replicating such experiments with video data would only be possible for selected (if any) industrial players.

This work introduces FitCLIP, a fine-tuning strategy to adapt large-scale image-text pretrained models for zero-shot video understanding tasks. The goal of FitCLIP is to retain the knowledge of CLIP [45] while gently adapting and learning how video data looks. Our method leverages relatively small labeled and extensive pseudo-labeled video data to train a student network. To validate the effectiveness of FitCLIP, we designed and set zero-shot benchmarks for two popular video understanding tasks: action recognition and text-to-video retrieval. Our experiments empirically validate the effectiveness of distillation to better train and fine-tune multimodal video models and show that FitCLIP establishes a new state-of-the-art for zero-shot video recognition and retrieval. Our design incorporates weight-space ensembling in a strategic manner, which has not been explored before, as far as we know.

Contributions. Our key idea is to develop a method to refine large-scale pretrained image-language models to zero-shot video use-cases. Our work brings two contributions:

(1) We introduce FitCLIP, a refinement strategy, and model for zero-shot video understanding. The model leverages abundant knowledge in large-scale image models and a distillation strategy to learn *new* video knowledge. We describe FitCLIP in (Section 3).

(2) We evaluate FitCLIP and competitive baselines in a newly designed zero-shot benchmark (Section 4). Our experiments include results for two sets of video understanding tasks, action recognition, and text-to-video retrieval, where we show the value of FitCLIP (Section 5).

2 Related Work

Zero-shot Video Understanding. Multiple zero-shot methods have been proposed to tackle popular tasks such as action recognition [6, 8], text-to-video retrieval [60], and localization-related tasks [28, 66]. Most of the zero-shot action recognition literature either follows an attribute-based approach or leverages word embedding to transfer knowledge [6, 8, 18, 19, 28, 34, 37]. Differently, in the text-to-video retrieval task, zero-shot methods leverage large-scale natural language supervision to pre-train video-language models. After pretraining, these models can then be employed and tested in text-to-video retrieval tasks. Similar to [4, 60], our work leverages natural language supervision from video titles to unlock zero-shot capabilities. However, we focus on adapting already well-trained image-text models to videos rather than learning a video-language model from scratch.

One of our goals is to establish a benchmark for zero-shot action recognition and text-to-video retrieval. Previous efforts have devoted insightful analyses to creating *true* zero-shot evaluation for action recognition [24]. These efforts are valuable for the traditional zero-shot setting where methods use a close vocabulary of (seen) actions, but they do not fit when zero-shot models learn with natural language supervision. Instead, we follow standard (full) tests on popular action recognition datasets and well-established text-to-video retrieval datasets.

Visual-Language Pretraining. Pretraining visual models with natural language became a popular learning strategy in the image domain [10, 30, 41, 45, 50]. The idea of matching images with text dates back to the late 90s when Mori *et al*. trained models to predict nouns and adjectives from image-text pairs [41]. Others modernized this idea using large-scale datasets to train CNNs [30]. However, only recently, Radford *et al*. took this idea to the next level [45]. They trained CLIP, a dual image-text encoder, with more than 400M images and text descriptions using a contrastive objective [42]. Our work builds upon CLIP and adapts it to video use-cases while preserving its zero-shot capabilities.

Video-language pretraining also gained traction in the video space. Despite the progress, it has been hard for video-language methods to compete in zero-shot settings with image-language pre-trained models. We argue this is due to the limited availability of videos with clean (and aligned) natural language supervision. For instance, Frozen in Time [4] trains a transformer-based architecture on the WebVid dataset, which contains 2.5M humanly curated video-title pairs. The dataset is at least two orders of magnitude smaller than the dataset to pre-train CLIP [45]. The importance of large and diverse data emerges when we compare Frozen in Time with CLIP in zero-shot video tasks. Others [38, 39] have trained with the relatively larger HowTo100M dataset, which contains 100M unaligned video-text pairs. Still, the zero-shot capabilities of these models remain subpar to what CLIP can provide. Our approach, FitCLIP, leverages the WebVid dataset [4] as a rich source to adapt CLIP for zero-shot video understanding tasks.

Refining Large-scale Image Models. DistInit [22] explored distilling image models for video. More recently, CLIP's strong visual representation inspired multiple researchers to explore its usage for video tasks [9, 15, 36, 44, 45, 56]. CLIP4Clip, for instance, proposed a straightforward strategy to fine-tune CLIP for the text-to-video retrieval task [36]. Surprisingly, their simple method sets a new state-of-the-art in various datasets. Similarly, ActionCLIP introduced a novel paradigm to action recognition harnessing CLIP's general visual knowledge [56]. While existing approaches effectively boost performance on target datasets, and tasks, they have not show to preserve the original CLIP zero-shot capabilities (also based on early experiments we ran).

3 Method: FitCLIP

Our goal is to train a model that *expands and complements* large image-language models [29, 45] for zero-shot video (see Figure 1). To do so, we introduce FitCLIP, a refinement strategy that leverages small labeled and large pseudo-labeled data together with existing knowledge acquired from large image-text pairs. FitCLIP includes two steps. The first step trains a model, in a Teacher-Student fashion, leveraging both: labeled video-text pairs and pseudo-labels generated by a teacher model. The second step fuses the existing knowledge of the teacher, a large-scale pre-trained image-language model, with the student trained on video data. We call the resulting model the same as our refinement strategy, FitCLIP.

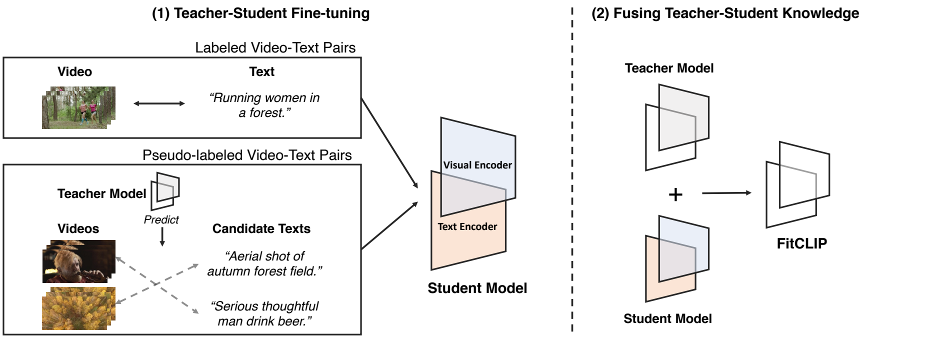

Figure 1: **FitCLIP refinement strategy and model.** We propose a refinement strategy to adapt large-scale image-text pretrained models. Our strategy first trains a model in a Teacher-Student fashion. To do so, we leverage labeled and pseudo-labeled (with a teacher) video-text pairs. This process, Step (1), yields a Student model that captures video-relevant knowledge while being compatible with the teacher. In step (2), similar to [59], we combine the Teacher and Student weights to create our final model, FitCLIP.

3.1 Teacher-Student Fine-tuning

Our goal is to train a model using video-text pairs while leveraging knowledge from image-language representations. One alternative is to reuse image-language encoder weights and fine-tune them in a target dataset [36, 56]. Such an approach is effective in boosting performance for in-distribution datasets but tends to fail at preserving the zero-shot capabilities of the original model's weights due to catastrophic forgetting [16]. Instead, we focus on gently refining the original image-language model's weights by incorporating a two-fold strategy. We use a small sample of labeled data to avoid model drift [46] (because of using a much smaller batch size and less diverse dataset), and we regularize the learning process by adding pseudo-labels generated with the original image-language model. Note that our strategy shares intuitions with the Knowledge Distillation literature [25], where a Teacher-Student analogy is used to describe the process of training a Student with priors derived from a strong Teacher model. Figure 1 (step 1) illustrates the process to train our Student model.

Data Subsets. Our fine-tuning strategy relies upon two subsets of data: a small labeled dataset of video-text pairs and an unlabeled (unaligned) set of video-text candidate pairs. The labeled subset contains a collection of videos matched with one text describing their visual content. These video-text pairs are of high quality and made by a human. The unlabeled subset also contains a list of videos and a list of text descriptions. However, the match between a video and the best describing text does not exist in this subset.

Teacher Model. The goal of the teacher is to provide *soft* pseudo-labels on unlabeled sets of video-text candidate pairs. We adopt CLIP [45] as a teacher. CLIP includes an image encoder and a text encoder, which were trained to predict the correct pairing of image-text pairs using a contrastive objective [42]. In practice, we use CLIP to compute the similarity between a subset of videos (within a large unlabeled set) and a set of candidate texts. Given that CLIP only takes individual images as input, we pass N frames from the video through its visual encoder and mean-pool the outputs into a single visual feature. We then use these similarity scores as target soft pseudo-labels.

Student Model. We aim to train a student model that learns from video-text pairs and distills knowledge from large pretrained image-language models. As the student, we choose the same dual architecture proposed by CLIP [45]. To train the model, we leverage two types of supervision: samples from the manually labeled video-text pairs dataset and soft pseudo-labels from the unlabeled set. Like the Teacher model, the student's visual stream takes N frames from each video and mean-pool the resulting representations into a single feature.

Student's Training Objective. We train the student model with two losses: a loss to learn from labeled samples, and a loss to distill the teacher knowledge via pseudo-labels. Given a video-text pair denoted (v,t) our student's dual encoder extracts a video representation z_v and a text representation z_t. For labeled samples, we use the InfoNCE [42] loss to learn a video-text correspondence. We follow [4, 60] and minimize the text-to-video and video-to-text contrastive losses:

$$\mathcal{L}_{v2t} = \sum_{(v,t)\in B_l} \log \frac{e^{z_v \cdot z_t^+ / \sigma}}{\sum_{z\in\{z_t^+, z_t^-\}} e^{z_v \cdot z / \sigma}} \tag{1}$$

$$\mathcal{L}_{t2v} = \sum_{(v,t)\in B_l} \log \frac{e^{z_t \cdot z_v^+ / \sigma}}{\sum_{z\in\{z_v^+, z_v^-\}} e^{z_t \cdot z / \sigma}} \tag{2}$$

where σ is the temperature hyper-parameter, B_l is a batch of video-text pairs, z_t^+ the positive text for the candidate video z_v, z_v^+ the positive video for candidate text z_t, and $\{z_v^-, z_t^-\}$ the negatives sets to contrast the candidate video and text representations. Then $(\mathcal{L}_{v2t} + \mathcal{L}_{t2v})$ is the final labeled (contrastive) loss.

To distill knowledge from soft pseudo-labels generated by the teacher, we use the teacher's predictions as pseudo-labels [25] and minimize the cross-entry of the student's scores relative to those from the teacher:

$$\mathcal{L}_{distill,v2t} = \sum_{(v,t)\in B_l} \frac{e^{x_v \cdot x_t / \sigma}}{\sum_{x\in T} e^{x_v \cdot x / \sigma}} \log \frac{e^{z_v \cdot z_t / \sigma}}{\sum_{z\in T} e^{z_v \cdot z / \sigma}} \tag{3}$$

$$\mathcal{L}_{distill,t2v} = \sum_{(v,t)\in B_l} \frac{e^{x_v \cdot x_t / \sigma}}{\sum_{x\in V} e^{x \cdot x_t / \sigma}} \log \frac{e^{z_v \cdot z_t / \sigma}}{\sum_{z\in V} e^{z \cdot z_t / \sigma}} \tag{4}$$

where x_v and x_t are the teacher's video and text representations, and V and T are the sets of videos and texts in the batch.

Our final objective combines the contrastive and distillation losses as in Equation (5). We scale the distillation loss, with λ, to prevent over-fitting to noisy pseudo-labels.

$$\mathcal{L} = \lambda(\mathcal{L}_{distill,v2t} + \mathcal{L}_{distill,t2v}) + (1-\lambda)(\mathcal{L}_{v2t} + \mathcal{L}_{t2v}) \tag{5}$$

3.2 Fusing Teacher-Student Knowledge

We aimed to train a competent student compared to the teacher. However, it is hard to compete with the 400M image-text pairs that were used to train CLIP [45]. Therefore, our goal is, instead, to fuse both: the general visual knowledge encapsulated by the teacher and the video-specific properties learned by the student. There are multiple ways to ensemble models [12]; however, given that our fine-tuning strategy gently adapts the teacher to video use cases, we can leverage elegant weight-space ensembling techniques [20, 59]. We follow the same approach in [59] to linearly combine the teacher and student weights (by α) and create our final model, FitCLIP.

3.3 FitCLIP's Implementation Details

We uniformly sample $N = 4$ frames from each video, similarly to TSN [55]. The Teacher and Student models both use a VIT-B/16 architecture initialized with OpenAI's publicly released weights [45]. We empirically set $\lambda = 10^{-4}$ to smooth the training process (note the labeled and pseudo-labeled loss magnitudes may be wildly different). We consistently use $\sigma = 0.05$ as temperature value. At training time, we randomly crop the frames to a size of 224×224, and perform random horizontal flips. We use the AdamW optimizer with a learning rate equal to 3×10^{-5}. We use the same tokenizer as in CLIP [45]. We conduct our experiments using 8x A100 (40GB) GPUs. We use 4.5K labeled videos, randomly sampled from the WebVid-2.5M dataset [4], to compute the losses in Equations (1) and (2). The entire WebVid-2.5M dataset is used to compute the distillation losses – Equations (3) and (4). We choose the (labeled) validation loss in the WebVid-2.5M dataset as a criterion to select the best student models. Finally, to fuse the teacher and the student weights, we use $\alpha = 0.4$. We encourage the reader to read the supplementary material for analyses to some of the hyperparameter values. We wrote our code on Python using PyTorch [43] and Lightning [14].

4 Zero-shot Video Understanding Benchmark

4.1 Baselines

CLIP [45]. This model has been pre-trained with the WIT dataset [49], which contains about 400M image-text pairs. We re-implement the zero-shot inference of this baseline model. To deal with video, we encode $N = 4$ uniformly sampled frames per video and average their features to obtain the final video representation. In all our experiments, we use the publicly released CLIP ViT-B/16 [13] model. Note that our CLIP adaptation is equivalent to ActionCLIP [56] (see *Supplementary Material*).
CLIP4Clip [36]. This method proposes changes on top of CLIP. In particular, they propose something the authors call *post-pretraining* that fine-tunes CLIP on the category "Food and Entertaining" (380k videos) from the HowTo100M [38] dataset. The authors have not provided this checkpoint, so we cannot evaluate it on our benchmarks. Still, we decide to include the results they report. Nevertheless, note the evaluation conditions are not the same to constitute a fair comparison (*e.g.*, the authors sample more than 4 frames per video clip).
Frozen in Time [4] (Frozen). This model was pre-trained leveraging video-text pairs from the WebVid dataset. There are multiple versions pre-trained versions for this model, including one that leverages the well-curated CC3M image-text pairs dataset. In our (main) experiments, we use the model that trains using the WebVid-2.5M, COCO, and CC3M dataset (note this is much less data than CLIP's pretraining dataset). Results for other versions of Frozen in Time can be found in the supplementary material.
VideoCLIP [60]. This baseline uses a Transformer [53] on top of a frozen HowTo100M-pre-trained S3D [65] video model from MIL-NCE [39] and a fine-tuned BERT [11] text model. This method trains on HowTo100M. A notable difference is that VideoCLIP samples 32 clips of size 32 frames (1024 frames) for each video, while we sample only 4 frames for each video.
VIOLET [17]. This method uses a video-language transformer trained end-to-end by masking discrete visual tokens. The authors use multiple training datasets including CC3M and WebVid.

BridgeFormer [21] (BF). This model leverages a multimodal encoder on top of the unimodal encoders and a method that masks the main verb and nouns as a for of multiple-choice questions as a pre-text task. The authors find this method to be more sample efficient than vanilla NCE.

4.2 Zero-shot Tasks and Datasets

Action Recognition. Our goal is to classify a video with one of C possible action classes. To do so, we form pretext language queries with predefined prompts. An illustrative example is the following prompt: "a video of a person $\{c_i\}$", where c_i is the i-th class out of the C candidate action categories. Given the visual representation of the target video, we compute its similarity with the language feature of each candidate action class prompt. We predict the action class by selecting the visual-text pair with the highest similarity. We report the top-1 and top-5 accuracy. We evaluate zero-shot action recognition in two datasets:

- *Moments in Time (MiT) [40]* consists of 3-second YouTube clips that capture the dynamics of actions performed by varied subjects including animals and humans. The dataset includes 339 categories and $33,900$ validation videos.

- *UCF101 [48]* contains 101 action classes. Our zero-shot experiments in this dataset aim to classify all the 1794 available test videos from the split 1.

Text-to-video Retrieval. Given a text query, the goal of text-to-video retrieval is to find a video, from a collection, that visually matches the text description. Given that the concept of classes does not exist in this task, previous methods [4, 36] denote experiments as zero-shot when the visual-language models are not fine-tuned on the downstream datasets. To measure performance, we report recall at $k = \{1, 5, 10\}$ and the median ranking (MdR). We evaluate zero-shot text-to-video retrieval in three datasets:

- *MSR-VTT [61]* contains video clips with a duration of up to 30 seconds paired with captions. We adopt the 1K-A test split [64], which contains $1,000$ video-text pairs.

- *YouCook2 [67]* comprises challenging cooking videos depicting fine-grained human actions. We test on 3305 clip-text pairs [39].

- *DiDeMo [2]* contains mostly unedited video clips from Flickr. We follow [4, 33, 35] and cast a video-paragraph retrieval problem. We evaluate on 4021 test samples.

5 Experimental Results

In this section, we conduct zero-shot experiments in two popular video understanding tasks, and then a diagnostic analysis of FitCLIP. First, we study the performance of the zero-shot baselines described in Section 4.1 in the task of action recognition. The second analysis summarizes the baseline performance in diverse datasets for text-to-video retrieval. We run diagnostic experiments to validate the importance of fusing the teacher knowledge, as in [59], to a competent zero-shot model. Finally, we run performance analyses on FitCLIP that study per-class gains in the action recognition task, and the shift in ranking distributions for the text-to-video retrieval tasks.

5.1 Zero-shot Action Recognition Results

We compare the zero-shot performance of FitCLIP and different baselines using two popular action recognition datasets. We describe the results and provide our analysis.

Analysis on Moments in Time. Table 1a summarizes the zero-shot results in the moments in time dataset. To establish a reference point, we also report VATT [1], the state-of-the-art using full supervision. In this dataset, FitCLIP outperforms both baselines, CLIP and Frozen, by a significant margin. It is noteworthy that CLIP, without seeing video data at training time, still outperforms Frozen by 11% at top-5 accuracy. Despite CLIP's good performance, we observe that FitCLIP further improves performance by 4.3% (top-5) setting a new state-of-the-art in this dataset. While FitCLIP achieves outstanding zero-shot results, a *e.g.* 44.6% top-5 accuracy, there is still an ample gap with respect to approaches that leverage supervision from the target dataset.

Analysis on UCF101. Table 1b shows the results on the UCF101 zero-shot benchmark. FitCLIP outperforms CLIP at Top 5 accuracy and sligthly underperforms at Top 1. All the findings remain consistent: a not-so-large gap between the best zero-shot and supervised approaches and Frozen under-performing with respect to CLIP-based methods. We attribute FitCLIP and CLIP close performance (when looking at both top-1 and top-5) due to the characteristics of UCF101, which contains a lot of common actions, including many sport-related actions. These types of actions often appear in photographs, and chances are, they are well-represented in CLIP training set.

Method	Top 1	Top 5
Supervised		
VATT [1]	41.1	67.7
Zero-shot		
Frozen	14.0	31.8
CLIP	19.9	40.3
FitCLIP	**21.8**	**44.6**

(a) **Moments in Time (MiT)**

Method	Top 1	Top 5
Supervised		
SMART [23]	98.6	–
Zero-shot		
Frozen	51.9	76.1
BF [21]	51.1	–
CLIP	**74.5**	94.3
FitCLIP	73.3	**95.3**

(b) **UCF101**

Table 1: **Zero-shot action recognition results.** (a) FitCLIP improves performance upon CLIP, and significantly outperforms Frozen. (b) FitCLIP shows slight improvements upon CLIP; Frozen lags behind in terms of zero-shot performance. Reported numbers in both tables are percentages and compute the top-1 and top-5 accuracy.

5.2 Zero-shot Text-to-video Retrieval

To compare FitCLIP and the baselines, here we report the experimental results and analysis for the text-to-video retrieval task.

Analysis on MSR-VTT. Table 2a summarizes results in the MSR-VTT dataset. We observe that Frozen performance is poor compared to that of CLIP and FitCLIP. Even though Frozen was trained on video data with similar properties to MSR-VTT, it is hard for this model to compete with the general knowledge encoded in CLIP-like models. We observe FitCLIP consistently improves performance upon CLIP across all the retrieval metrics. These results suggest that FitCLIP captures complementary video-language information that CLIP lacks. Concerning the gap to reach the performance of the best-supervised approach, CAMoE [9],

Method	R@1	R@5	R@10	MdR
Supervised				
CAMoE [9]	52.9	78.5	86.5	1
Zero-shot				
VideoCLIP [60]	10.4	22.2	30.0	–
Frozen	21.3	43.6	55.9	7
VIOLET [17]	25.9	49.5	59.7	–
BF [21]	26.0	46.4	56.4	7
CLIP via [36]	30.6	54.4	64.3	4
CLIP4Clip [36]	32.0	57.0	66.9	4
CLIP	30.4	55.1	64.1	4
FitCLIP	**33.8**	**59.8**	**69.4**	**3**

(a) **MSR-VTT**

Method	R@1	R@5	R@10	MdR
Supervised				
TACo [63]	29.6	59.7	72.7	4
Zero-shot				
VideoCLIP [60]	**22.7**	**50.4**	**63.1**	–
Frozen	3.2	10.1	16.2	135
CLIP	5.3	14.6	20.9	94
FitCLIP	5.8	15.5	22.1	75

(b) **YouCook2**

Method	R@1	R@5	R@10	MdR
Supervised				
CAMoE [9]	43.8	71.4	79.9	2
Zero-shot				
VideoCLIP [60]	16.6	46.9	–	–
Frozen	23.2	45.8	56.8	7
VIOLET [17]	23.5	49.8	59.8	–
BF [21]	25.6	50.6	61.1	5
CLIP	26.2	49.9	60.6	5
FitCLIP	**28.5**	**53.7**	**64.0**	**4**

(c) **DiDeMo**

Table 2: **Zero-shot text-to-video retrieval results.** In all datasets, FitCLIP improves upon CLIP by significant margins. (a) FitCLIP's shows the best zero-shot results though there is an important gap with the supervised state of the art. (b) In this dataset, YouCook2, FitCLIP exhibits the largest gap concerning fully supervised approaches and with VideoCLIP, that pretrained on HowTo100M. We attribute this result to the fine-grained nature of the dataset. (c) FitCLIP shows consistent boosts upon CLIP even for the DiDeMo (paragraph-retrieval) task, which includes long language queries. R@k denotes recall at the top-$k = \{1, 5, 10\}$ predictions, and MdR refers to the Median Ranking metric.

FitCLIP does not lag that behind. Even though there is a 16.1% gap at R@10, we see that FitCLIP closely approaches supervised performance at the MdR metric.

Analysis on YouCook2. We report zero-shot results for the YouCook2 dataset in Table 2b. From the get-go, we observe the difficulty of this dataset. Even the state-of-the-art, TACo [63], struggles to achieve more than 30% R@1. While we observe that FitCLIP's performance consistently outperforms other zero-shot baselines, we have observed a large overall gap between our method and those that are supervised or pretrained on HowTo100M [38] (VideoCLIP [60] in the table). We hypothesize this is due to the fine-grained nature of the language descriptions contained in YouCook2 and HowTo100M. Moreover, lots of videos in this dataset are captured from an egocentric view.

Analysis on DiDeMo. Table 2c summarizes the results in DiDeMo's paragraph retrieval task. First, we observe that the performance of Frozen, VIOLET [17], and BridgeFormer [21] approach the one achieved by CLIP in this dataset. Contrary to other datasets, DiDeMo contains unedited, human-centric footage that shares commonalities with the WebVid dataset used to train Frozen. Conversely, FitCLIP, which leverages both: the knowledge from CLIP and the WebVid dataset, achieves the best performance overall. For completeness, we report the CAMoE's supervised performance [9], which is 15.9% better than FitCLIP, the most competitive zero-shot alternative.

The results on these three datasets empirically demonstrate the value of FitCLIP to push the limits of zero-shot text-to-video retrieval. FitCLIP establishes a new state-of-the-art in zero-shot text-to-video retrieval across three different datasets. Despite such a milestone, there is still room for improvement, especially in fine-grained datasets such as YouCook2. We hope this benchmark promotes more work on zero-shot text-to-video retrieval.

5.3 Diagnostic Analysis

Impact of Fusing the Teacher-Student Knowledge (Table 3). One of the key properties of FitCLIP is the ability to incorporate the student learning from video data, and knowledge of the CLIP teacher. Here we report the performance of both our Student and Teacher

	Action Recognition		Text-to-video Retrieval		
	UCF101	MiT	MSR-VTT	YouCook2	DiDeMo
Teacher (CLIP)	74.5	19.9	55.1	14.6	49.9
Student	64.7	17.7	52.6	9.7	42.4
FitCLIP	73.3	21.8	59.8	15.5	53.7
△	↓ 1.2	↑ 1.9	↑ 4.7	↑ 0.9	↑ 3.8
Err. rate red.	↓ 4.7	↑ 2.4	↑ 10.5	↑ 1.1	↑ 7.6

Table 3: **Impact of fusing teacher-student knowledge.** △ denotes the absolute difference in performance between FitCLIP and the Teacher model. We report the top-1 accuracy for the zero-shot action recognition datasets, and the top-5 recall for the zero-shot text-to-video retrieval ones. We observe that even though the Student model is weaker than the Teacher, it still provides complementary information to FitCLIP, yielding consistent improvements (△) across datasets. Full results with all the metrics in the supplementary material.

(CLIP) and contrast that with the final zero-shot performance obtained with FitCLIP. Table 3 summarizes the results. We observe that although the Student's performance remains inferior to that of the Teacher, it is close enough in various datasets, *e.g.* MiT, MSR-VTT, and DiDeMo. △ denotes the difference in performance between FitCLIP and the teacher and indirectly measures the contribution of the student learning. We observe that improvements are consistent across all tasks and datasets. These results suggest that the Student effectively passes complementary information to the teacher after weight assembling.

Additional Ablations. Due to space limitations, we include additional analysis in the *Supplementary Material*. We compare the properties of FitCLIP vs. CLIP, do a deep-dive on the impact of fusing the Teacher-Student knowledge, ablate weight-ensembling parameters, and report comparisons with additional methods trained on HowTo100M.

6 Conclusions

This paper presents a fine-tuning strategy to adapt large-scale image-text pre-trained models for zero-shot video understanding tasks, dubbed FitCLIP. FitCLIP performs well on zero-shot settings for three Text-to-Video Retrieval and two Action Recognition tasks that we evaluated. We show the importance of doing the weight-space ensembling step of our method to keep or improve the teacher's robust performance across different datasets, even when the student was trained on different data. We highlight our method introduces no extra inference costs while improving CLIP results overall.

Acknowledgements

We thank Christine Feak for revising a draft of this document. Santiago thanks Adobe Research for providing financial support to continue working on this project after the internship finished.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34, 2021.

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, October 2021.

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in Time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

[5] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., 2009.

[6] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4613–4623, 2020.

[7] 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 (CVPR)*, July 2017.

[8] Shizhe Chen and Dong Huang. Elaborative rehearsal for zero-shot action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13638–13647, 2021.

[9] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021.

[10] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11162–11173, June 2021.

[11] 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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[12] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[13] 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. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[14] William Falcon and The PyTorch Lightning team. PyTorch Lightning, 3 2019. URL https://github.com/Lightning-AI/lightning.

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

[16] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.

[17] 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 preprint arXiv:2111.12681*, 2021.

[18] Chuang Gan, Yi Yang, Linchao Zhu, Deli Zhao, and Yueting Zhuang. Recognizing an action using its name: A knowledge-based approach. *International Journal of Computer Vision*, 120(1):61–77, 2016.

[19] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8303–8311, Jul. 2019. doi: 10.1609/aaai.v33i01.33018303. URL https://ojs.aaai.org/index.php/AAAI/article/view/4843.

[20] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8803–8812, 2018.

[21] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16167–16176, June 2022.

[22] Rohit Girdhar, Du Tran, Lorenzo Torresani, and Deva Ramanan. DistInit: Learning video representations without a single labeled video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[23] Shreyank N Gowda, Marcus Rohrbach, and Laura Sevilla-Lara. Smart frame selection for action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1451–1459, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16235.

[24] Shreyank N. Gowda, Laura Sevilla-Lara, Kiyoon Kim, Frank Keller, and Marcus Rohrbach. A new split for evaluating true zero-shot action recognition. In Christian Bauckhage, Juergen Gall, and Alexander Schwing, editors, *Pattern Recognition*, pages 191–205, Cham, 2021. Springer International Publishing. ISBN 978-3-030-92659-5.

[25] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *Deep Learning and Representation Learning Workshop at the Twenty-eighth Conference on Neural Information Processing Systems*, 2014. URL https://arxiv.org/abs/1503.02531.

[26] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020.

[27] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

[28] Mihir Jain, Jan C Van Gemert, Thomas Mensink, and Cees GM Snoek. Objects2action: Classifying and localizing actions without any video example. In *Proceedings of the IEEE international conference on computer vision*, pages 4588–4596, 2015.

[29] 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. In *International Conference on Machine Learning*, 2021.

[30] Armand Joulin, Laurens Van Der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pages 67–84. Springer, 2016.

[31] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. MoViNets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16020–16030, June 2021.

[32] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, page 646–651, 2008. ISBN 9781577353683.

[33] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7331–7341, June 2021.

[34] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing human actions by attributes. In *CVPR 2011*, pages 3337–3344. IEEE, 2011.

[35] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *30th British Machine Vision Conference (BMVC 2019)*. British Machine Vision Association, April 2020.

[36] 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.

[37] Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed, Fahad Shahbaz Khan, and Ling Shao. Out-of-distribution detection for generalized zero-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9985–9993, 2019.

[38] 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/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[39] 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 (CVPR)*, June 2020.

[40] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Moments in time dataset: One million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):502–508, 2020. doi: 10.1109/TPAMI.2019.2901464.

[41] Yasuhide Mori, Hironobu Takahashi, and Ryuichi Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *First international workshop on multimedia intelligent storage and retrieval management*, pages 1–9. Citeseer, 1999.

[42] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[43] 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 H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learni pdf.

[44] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using clip. In Edgar Roman-Rangel, Ángel Fernando Kuri-Morales, José Francisco Martínez-Trinidad, Jesús Ariel Carrasco-Ochoa, and José Arturo Olvera-López, editors, *Pattern Recognition*, pages 3–12, Cham, 2021. Springer International Publishing. ISBN 978-3-030-77004-4.

[45] 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 Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings*

of Machine Learning Research, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[46] Amelie Royer and Christoph H Lampert. Classifier adaptation at prediction time. In *CVPR*, pages 1401–1409, 2015.

[47] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/2d6cc4b2d139a53512fb8cbb3086ae2e-Paper.pdf.

[48] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CRCV-TR-12-01*, November 2012.

[49] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. *WIT: Wikipedia-Based Image Text Dataset for Multimodal Multilingual Machine Learning*, page 2443–2449. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450380379. URL https://doi.org/10.1145/3404835.3463257.

[50] Nitish Srivastava, Ruslan Salakhutdinov, et al. Multimodal learning with deep boltzmann machines. In *NIPS*, volume 1, page 2. Citeseer, 2012.

[51] O. Tange. Gnu parallel - the command-line power tool. *;login: The USENIX Magazine*, 36(1):42–47, Feb 2011. URL http://www.gnu.org/s/parallel.

[52] The pandas development team. pandas-dev/pandas: Pandas, 2021. URL https://github.com/pandas-dev/pandas.

[53] 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 I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[54] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.

[55] 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.

[56] Mengmeng Wang, Jiazheng Xing, and Yong Liu. ActionCLIP: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[57] Michael Lawrence Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 60(6), April 2021. doi: 10.21105/joss.03021. URL https://joss.theoj.org/papers/10.21105/joss.03021.

[58] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

[59] Mitchell Wortsman, Gabriel Ilharco, Mike Li, Jong Wook Kim, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*, 2021.

[60] 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*, pages 6787–6800, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.544.

[61] 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 (CVPR)*, June 2016.

[62] Omry Yadan. Hydra - a framework for elegantly configuring complex applications. Github, 2019. URL https://github.com/facebookresearch/hydra.

[63] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. TACo: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11562–11572, October 2021.

[64] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[65] Da Zhang, Xiyang Dai, Xin Wang, and Yuan-Fang Wang. S3D: Single shot multi-span detector via fully 3d convolutional network. In *Proceedings of the British Machine Vision Conference*, 2018.

[66] Lingling Zhang, Xiaojun Chang, Jun Liu, Minnan Luo, Sen Wang, Zongyuan Ge, and Alexander Hauptmann. Zstad: Zero-shot temporal activity detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 879–888, 2020.

[67] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17344.