# All in One: Exploring Unified Video-Language Pre-training

Jinpeng Wang[1], Yixiao Ge[2], Rui Yan[1], Yuying Ge[4], Kevin Qinghong Lin[1], Satoshi Tsutsui[1],
Xudong Lin[5], Guanyu Cai[1,6], Jianping Wu[7], Ying Shan[2], Xiaohu Qie[3], Mike Zheng Shou[1*]
[1]Show Lab, National University of Singapore [2]ARC Lab,[3]Tencent PCG
[4]The University of Hong Kong [5]Columbia University [6]Tongji University [7]Tsinghua University

## Abstract

*Mainstream Video-Language Pre-training (VLP) models [10, 26, 64] consist of three parts, a video encoder, a text encoder, and a video-text fusion Transformer. They pursue better performance via utilizing heavier unimodal encoders or multimodal fusion Transformers, resulting in increased parameters with lower efficiency in downstream tasks. In this work, we for the first time introduce an end-to-end VLP model, namely all-in-one Transformer, that embeds raw video and textual signals into joint representations using a unified backbone architecture. We argue that the unique temporal information of video data turns out to be a key barrier hindering the design of a modality-agnostic Transformer. To overcome the challenge, we introduce a novel and effective token rolling operation to encode temporal representations from video clips in a non-parametric manner. The careful design enables the representation learning of both video-text multimodal inputs and unimodal inputs using a unified model. Our pre-trained all-in-one Transformer is transferred to various downstream video-text tasks after fine-tuning, including text-video retrieval, video-question answering, multiple choice and video captioning. State-of-the-art performances with the minimal model FLOPs on ten datasets demonstrate the superiority of our method compared to the competitive counterparts. The code and pretrained models are available at https://github.com/showlab/all-in-one.*

## 1. Introduction

Science advances rather steadily for most of the time, but sometimes has a disruptive episode, where "an older paradigm is replaced in whole or in part by an incompatible new one." [24] In this regard, Video-Language Pre-training (VLP) models have recently experienced steady progress, where joint representations are generally produced with a multimodal fusion network after extracting the visual and language features through unimodal encoders [10,26,50,64]. We are here to break it and replace them with "an incompatible new one" that has NO unimodal encoders.

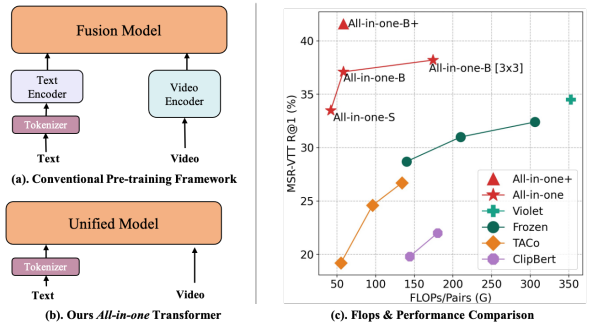


Figure 1. **Compare to mainstream video-language pre-training methods.** (a). Conventional methods [3, 10, 26, 64] use deep features from separate encoders before fusion. The fusion layer can be light [3] or heavy [10,26,64]. (b). Ours *All-in-one* Transformer learns video and text joint representations end-to-end from their raw inputs. We also support fast retrieval by feeding unimodal inputs during inference. (c). Comparison of FLOPs and retrieval performance on MSRVTT [53]. Our *All-in-one* brings excellent results with modest computational cost.

The pre-train and then fine-tune scheme, has become a standard paradigm to learn transferable video-language representations for a wide range of downstream video-text tasks [6, 15, 52, 52, 53, 61]. Mainstream methods attempt to boost the pre-training in two ways: *i*. adopting more expensive video/text encoders to obtain more powerful unimodal features [3, 10] *ii*. designing heavier fusion networks to enhance the association between modalities [61, 64].

Instead of following these trends, we fundamentally rethink design decisions and develop *the simplest and most lightweight* architecture that learns video-language representations from their raw inputs in an end-to-end manner. Our model does not need any unimodal encoders (*e.g.*, object detector in [64] or ResNet visual encoder in [26]) or complex fusion layers, but embeds visual and text signals in a unified manner, termed as *All-in-one* Transformer in our paper. Our design is inspired by recent studies [1, 21, 37]

*Corresponding Author.

that perform multimodal pre-training under the presumption that Transformer can process visual data in the same way as it processes text. However, our work is not the straightforward application of them. It is not trivial how to embed videos for our unified Transformer due to the unique challenge of modeling temporal information without adding much computational cost.

Existing works model temporal information by designing temporal attention layers [3] or using temporal-aware visual encoders (*e.g.*, 3D convnets in [64] or Video Swin [36] in [10]). We cannot simply use them in our unified *All-in-one* Transformer because they are modality-dependent and computationally too expensive. To address this issue, we design a novel, effective, and efficient method to model temporal information. Our model only needs three frames per video clip, which is much lower than other models (*e.g.*, 16 [26] or 32 [1]) but can achieve the comparable performance to them. Nevertheless, we are still not satisfied with the computational cost in the self-attention layer. To further reduce the computational cost, we propose the *temporal token rolling operation*, which is a cyclic attention between small proportions of the visual tokens in each frame (Fig. 3-right). This is much more efficient than a naive self-attention approach on flattened tokens (Fig. 3-bottom). Furthermore, our modality-agnostic design enables us to use our pre-trained model as a powerful unimodal feature extractor by feeding only video or text inputs. This can significantly reduce the computational cost for retrieval task because we can simply compute the cosine similarity of texts and videos soon after the pretraining, eliminating the need for training additional fusion module of projecting the disjoint text and visual features into a common space (Fig. 1-b). Taken together, our *All-in-one* architecture achieves much less FLOPs and better text-to-video performance than previous work (Fig. 1-c), despite the fact that we use the same pre-training objectives [10, 26].

**Contributions.** (1) We introduce the simplest, most lightweight, and most efficient video-language model, namely *All-in-one* Transformer, which is the first to capture video-language representations from the raw visual and textual signals end-to-end in a unified backbone architecture. (2) We elucidate and tackle the difficulties of applying a unified and shared backbone for multimodal video and text data, that is, how to properly process the unique temporal information of videos. A novel temporal token rolling operation is proposed to capture the temporal representations of sparsely sampled frames without any extra parameters or increasing time complexity. (3) We propose a success practical to overcome the slow retrieval of one-stream model and explore how to cotrain the image and video data together in better ways. (4) Comprehensive experiments on five downstream video-text tasks of eleven datasets fully demonstrate the superiority of our pre-trained *All-in-one* Transformer on both effectiveness and efficiency compared to recent mainstream methods [3, 10, 26].

## 2. Related Work

**Video-Language Pre-training.** Pre-training on large-scale video-text pairs and fine-tuning on specific downstream tasks gradually become the standard paradigm in the video-language domain. Pre-trained models show strong transfer ability in a series of popular downstream video-language tasks including Text-to-Video Retrieval [6,53], Video Question Answering [15, 52], and Visual Storytelling [61]. Previous approaches [43, 58, 64] leverage offline video and text features extracted from off-the-shelf visual and language backbones. Some recent methods [3, 26, 33, 55] have attempted to train models in an end-to-end fashion but still rely on well-trained visual encoders for feature extraction. In addition, these works mainly pre-train models on the image-text datasets, like Google Conceptual Captions [42] and Visual genome [23], and fine-tune the pre-trained models for downstream video-language tasks. In this work, we try to challenge this paradigm and explore an effective strategies for pre-training on pure large-scale video-text benchmarks with only one network, and adapt our approach to various video-language downstream tasks.

**Temporal Modeling in Video Understanding.** Temporal modeling is a fundamental yet challenging topic in video representation learning. Several classic ideas including sparse sampling [49], 3D-type operations [5, 36] are proposed for temporal modeling in both convolution and Transformer architectures. 3D-type temporal modeling like Timesformer [4] is extremely time-consuming because of the increasing number of sampled frames, which can be disastrous for large-scale pre-training techniques. Sparse sampling along the temporal dimension, a type of data augmentation proposed in TSN [49], has been widely adopted to train video backbones. Based on this, more related works [31, 50] try to shift channels among different frames for temporal modeling in action recognition. Inspired by these works, we try to roll video tokens for better alignment between modalities. This work focuses on parameter-free temporal modeling based on sparsely sampled frames without heavy 3D-type operation.

**Unified Architecture Design for Multimodal Data.** Recently the unified model, which is capable of processing either unimodal or multimodal inputs with a shared encoder, has attracted a lot of attention. VATT [1] and Merlot Reserve [60] trains a shared transformer with unimodal inputs to process Video, Audio, and Text via multimodal contrastive learning. Omnivore [13] converts the image, video, and single-view 3D modalities into embeddings that are fed into a Transformer model and trains the model with multi-task learning, which focuses on image/video/scene classification. In image-text pre-training, the early work Unimo

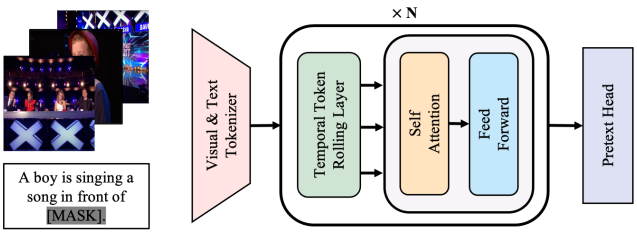


Figure 2. **Model overview.** Our model is simple, efficient, and based on the commonly-used ViT [8], where our additional parameters are only in the light text tokenizer and the task heads. Since the ViT cannot model the temporal information, we also introduce a parameter-free temporal token rolling layer before each self-attention block.

[29] solves both understanding and generation tasks with cross-modal contrastive learning. UFO [47] also uses contrastive learning and employs a momentum teacher to guide the pre-training of an image-text shared encoder, which incurs large computational costs. Based on cross-modal contrastive learning, our work can also process unimodal inputs and perform retrieval tasks in a dual-stream manner, which is very efficient. To the best of our knowledge, *All-in-one* Transformer is the first unified network for VLP.

## 3. Method

We propose *All-in-one* Transformer, a generic framework that enables end-to-end learning on video and language data, by learning joint representations directly from raw video pixels and raw text tokens, instead of the deeper feature from two separate deep embedders. *All-in-one* has a succinct architecture as a Video-Language Pre-training (VLP) model with a parameter-free temporal modeling layer. In model design, we make the pipeline as simple as possible so that the model can be used almost out of the box.

### 3.1. Unified Video-language Transformer

Fig.2 gives an overview of *All-in-one* framework, it mainly contains three parts: Video and Text Tokenizer, $N$ Transformer blocks and a pretext head. For each video, *All-in-one* uses a sparse sampling strategy with $S$ segments (one frame per segment) at each training step, rather than full-length videos. The sampled video clip and text are inputted into the same Transformer, as described below.

**Video & Text Tokenizer.** The video tokenizer slices an input video into patches, maps the patches into tokens with a linear projection, and adds learnable spatio-temporal position embeddings. The text tokenizer similarly maps the words into tokens with a word embedding layer. Following common practices [21, 26], we also add modality type embeddings to distinguish the token of video or texts.

**Cross-modality Fusion.** The *All-in-one* fuses the text and video tokens using the $N$ transformer blocks as follows. It concatenates the text and vision tokens of each frame as $z^0 = [\hat{t}; \hat{v}]$. It then feeds the $z^0$ into the $N$ Transformer blocks, where each block consists of a temporal Token Rolling layer, a multi-head self-attention layer, and a multilayer perceptron, whose weights are initialized from pre-trained ViT [8] or DeiT [45]. Formally, for $d = 1...N$,

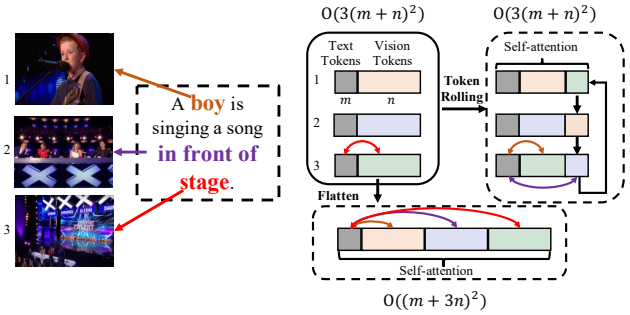


Figure 3. **The token rolling vs. flatten.** By simply rolling tokens, the text tokens can also see tokens from different frame in the same time. But the computation complex for Self-attention of Token Rolling is only one third of Flatten.

$$z_r^{d-1} = \mathrm{TTR}(\mathrm{z}^{\mathrm{d-1}}),$$
$$z^d = \mathrm{MLP}(\mathrm{MSA}(\mathrm{z}_{\mathrm{r}}^{\mathrm{d-1}})), \quad (1)$$

where MSA means multiheaded self-attention, MLP is multilayer perceptron and TTR is short for Temporal Token Rolling Module, which aims to learn temporal information.

### 3.2. Temporal Token Rolling

**Motivation.** In VLP, the common way to model the temporal information is to add additional time attention layers [3] in the vision encoder or to use the feature from a deep off-the-shelf video encoder [10, 64] (e.g, VideoSwin [36]). However, these techniques are particularly designed for video and thus cannot be applied to process text signals, as well as bringing a large amount of additional parameters. For example, simply adding one temporal attention layer to each block of the Transformer will increase the model's parameters from 86M to 121.7M (an increase of 42%) [4]. Thus, we cannot use these techniques in our unified framework in an affordable way, so we turn to finding new ways to learn temporal information with modest parameters.

**Approach.** A straightforward approach, denoted as "Flatten", is to concatenate video and text tokens together and *flatten* into one tensor, which will be fed into the self-attention blocks. Given a text token of length $m$ and a video token of length $S \times n$, we show the *flatten* version in Fig.3. However, as the self-attention layer has quadratic complexity, the computation cost will be $\mathcal{O}((m + Sn)^2)$, about $S^2$ times more than 1-frame *All-in-one*[1].To reduce the computational cost, *we exchange information for different time*

[1]The length of text tokens $m$ is much smaller than video tokens $n$

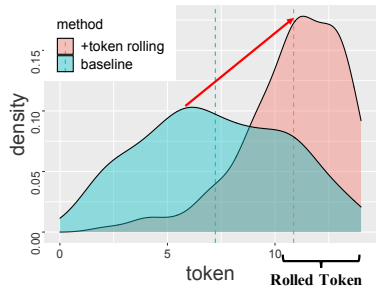


Figure 4. **Text to video attention weight distribution over tokens**: The Temporal Token Rolling layer causes text tokens to focus more on rolled tokens, contrasting with prior centric attention.

*segments using only a small part of tokens.* The proposed Token Rolling module is described in the right of Fig.3. Tokens at varying timestamps are represented by different colors in each row. A portion of tokens is rolled by 1 along the temporal dimension, while others remain unchanged. Self-attention is calculated for each $m+n$ token, treating them all identically. In this way, we reduce the computational complexity to $\mathcal{O}(S(m+n)^2)$, around $\frac{1}{S}$ of the Flatten.

**Discussion.** Our method takes advantage of Token Rolling, gradually modeling long-time dependencies between texts and videos in deeper layers, which helps to learn better video-text alignment. We visualize the density of the cross-modality attention weight between text and video tokens in Fig. 4. For each text token, we compute the similarity by dot product to reveal its corresponding highly-weighted video tokens. The baseline in the figure is *All-in-one* without rolling layers. Interestingly, we observe that text tokens in the baseline are severely biased to the centric visual tokens, having much more attention than others. This implies that objects appear mainly in the center of images. Our Temporal Token Rolling makes these rolled tokens contribute more to the attention value, which demonstrates that these tokens carry richer information.

## 3.3. Training Objectives

### 3.3.1 Pre-training

We pre-train *All-in-one* with two commonly used objectives to train VLP models: video-text matching (VTM) [26] and masked language modeling (MLM) [7].

**Video-text Matching.** Given a paired video-text input, we randomly replace the paired video with a different video with the probability of 0.5 and ask the model to distinguish them. For the *cls token* of the last block, a single linear layer VTM head projects tokens to logits over binary classes. We compute the negative log-likelihood loss as our VTM loss.

**Masked Language Modeling.** MLM [7] aims to predict the ground truth labels of masked text tokens from the remaining text and video tokens. Following common practices [7,21], we randomly mask text tokens with a probability of 0.15 and model it as a classification task.

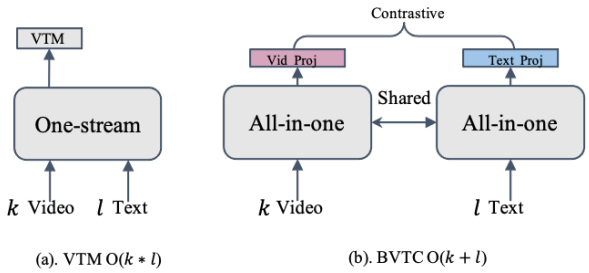


Figure 5. **The retrieval pipeline for VTM score and our BVTC.** Our BVTC significantly reduces the computation cost from $O(k*l)$ to $O(k+l)$.

### 3.3.2 BVTC for fast downstream retrieval

Vision-text Contrastive loss has shown great success for efficient retrieval task of dual-stream models,which encode image and text independently [16,40]. However, this objective cannot transfer to one-stream models easily because the input of these models are the joint of vision and text. For these models, the common way to do retrieval is to measure vision-text pairwise matching scores, which is very slow and cannot be applied to large-scale retrieval [29, 30]. To overcome the disadvantage of the low retrieval efficiency of one-stream models, we introduce a new paradigm to utilize a contrastive loss for retrieval.

**Backbone-shared Video-text Contrastive Loss.** As shown in Fig. 5, we input video and text independently to the shared encoder to obtain high-level features for video-text pairs. We then feed these features into a modality-specific projection head to project them into the shared embedding space. Following common practice [3,46], we use a symmetric (both text-to-video and video-to-text) contrastive loss based on these features. When doing retrieval tasks, we only need to extract unimodal features once, which significantly reduces the computational cost. More discussion is reported in the supplementary material.

## 3.4. Image Video Co-Training

Since image datasets often provide more comprehensive and fine-grained annotations than video datasets, we use both image and video datasets. Inspired by recent studies in action recognition that demonstrates a unified Transformer model can be extended to both image and video classification tasks [37, 63], we propose to leverage both image and video data to train *All-in-one* jointly.

The naive solution is to change the training pipeline with a minimal modification by considering an image as a single-frame video [13, 37]. However, we experimentally find that this simple solution damages the learning of temporal information and leads to unstable training (refer to supplementary material for more discussion). In this work, we propose a balanced sampling co-training strategy. Specifically, we

sample a half of image-text samples and a half of video-text samples for each batch. Then, we pass the images to the self-attention block directly. For video-text pairs, the input is first fed into the Temporal Token Rolling Layer and then each self-attention block. In contrast to previous work [63], we use a shared pretext head for both image and video. The weighted loss for co-training over both image and videos samples is computed as:

$$\mathcal{L}_{ct} = \sum_i w^i_{video} \mathcal{L}(y^i_{video}) + \sum_j w^j_{image} \mathcal{L}(y^j_{image}), \quad (2)$$

where $w_{image}$ means the weight for the image and $w_{video}$ for the video. We also analyze the other variations of co-training strategies in the supplementary material and show the superiority of our balanced sampling co-training.

| Model | Embed Dim | #Heads | #Params | Throughput |
|---|---|---|---|---|
| *All-in-one*-Ti | 192 | 3 | 12M | 745 |
| *All-in-one*-S | 384 | 6 | 33M | 285 |
| *All-in-one*-B | 768 | 12 | 110M | 89 |

Table 1. **Variants of our *All-in-one* architecture**. Embed Dim is short for Embedding Dimension. The throughput is measured for videos at a resolution of 224×224. We use *All-in-one*-B as default without specific explanation.

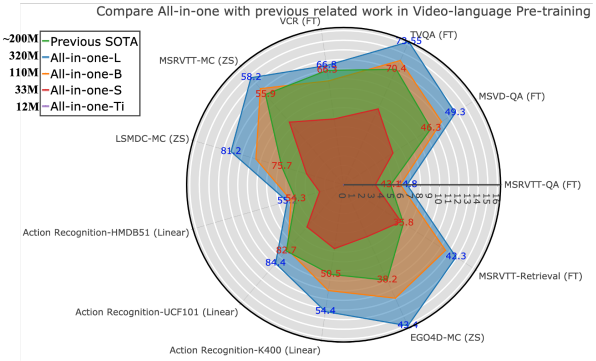


Figure 6. **The parameters & performance over eleven downstream datasets with the varying of model size.** The origin point represent *All-in-one*-Ti.

## 3.5. Setup

To explore model scalability, we use large-scale Webvid-2.5M [3], HowTo100M [39] and YT-Temporal 180M [61] for Pre-training. For Image and Video Co-training, we adopt additional image dataset CC3M [42]. We evaluate *All-in-one* on four popular downstream video-language tasks: text-to-video retrieval, video question answering, multiple-choice, video captioning across 9 different datasets. We also transfer our model to video action recognition. We also provide extensive ablation studies to analyze the key factors that contribute to *All-in-one*'s success, with insights and qualitative results.

### 3.5.1 Model Variants.

When considering the generality of *All-in-one*, we consider using three configurations based on ViT [8] and DeiT [45], as summarized in Tab. 1. To simplify, we use the brief notation to indicate the model size: for instance, *All-in-one*-B/16 means the "Base" variant with 16×16 input patch size. We varying the model size from tiny to base. Following ViLT [21], we use the *bert-base-uncased* tokenizer [7] to tokenize text inputs. For input video, we random sample 3 frames and resize each frame to 224 × 224.

### 3.5.2 Pre-training & Fine-tuning.

Considering YT-Temporal 180M [61] partially overlaps with HowTo100M [39], we pre-train on WebVid2.5M + Howto100M by default. If the model is trained with additional YT-Temporal 180M, we named it as *All-in-one* *. For *All-in-one*+, we pre-train on WebVid2.5M, HowTo100M, and CC3M as default. When we train *All-in-one*+ on more datasets, we also list the datasets for reference. We refer readers to supplementary for more pre-training details.

## 3.6. Downstream Tasks Settings

*All-in-one* is evaluated on five video-language tasks: text-to-video retrieval, video question answering, multiple-choice, captioning, and action recognition across 10 datasets. We also provide extensive ablation studies to analyze the key factors that contribute to *All-in-one*'s success, with insights and qualitative results. We refer readers to supplementary material for the datasets and evaluation setting for downstream tasks.

# 4. Main Results

We extensively evaluate the capabilities of *All-in-one* on a wide range of downstream tasks as a pretrained foundation model. We mainly consider core tasks of two categories that examine (1) video-text understanding capabilities, (2) video-text alignment, and (3) video captioning and action recognition capabilities. Fig. 6 summarizes the performance on key benchmarks of *All-in-one* compared to other foundation models. In addition, we also transfer our model to more downstream image-text tasks (refer to supplementary for more details).

## 4.1. Multimodal Understanding Tasks

### 4.1.1 Video-question Answering.

In this experiment, we compare three variations of our *All-in-one* to state-of-the-art methods from the literature. For multiple-choice VQA, we evaluate our *All-in-one* on two sub splits of TGIF-QA and report the result in Tab. 2. We find *All-in-one* especially good at this type of VQA. With only 1 frame input, our *All-in-one*-B outperforms previous

| Method | Nets | Params | Pre-training Data | Frames | Action | Transition | FrameQA |
|---|---|---|---|---|---|---|---|
| Heterogeneous [9] | *T+V+LSTM* | - | - | 35 | 73.9 | 77.8 | 53.8 |
| HCRN [25] | *T+V+LSTM* | - | - | 16 | 75.0 | 81.4 | 55.9 |
| QueST [17] | *T+V+LSTM* | - | - | 16 | 75.9 | 81.0 | 59.7 |
| ClipBERT [26] | *T+V+CE* | 137M | COCO + Visual Genome | $1 \times 1$ | 82.9 | 87.5 | 59.4 |
| VIOLET [10] | *T+V+CE* | 198M | CC3M + WebVid2.5M | 16 | 87.1 | 93.6 | - |
| *All-in-one*-Ti | *CE* | 12M | WebVid2.5M + HowTo100M | 3 | 80.6 | 83.5 | 53.9 |
| *All-in-one*-B | *CE* | 110M | WebVid2.5M + HowTo100M | 1 | 92.9 | 94.2 | 62.5 |
| *All-in-one*-B+ | CE | 110M | CC3M + WebVid | 3 | 94.4(7.3↑) | 94.5(0.9↑) | 66.4(7.0↑) |
| *All-in-one*-B+ | CE | 110M | CC3M + WebVid2.5M + HowTo100M | 3 | **96.3**(9.2↑) | **95.5**(1.9↑) | **67.3** (7.9↑) |
| *All-in-one*-B [384] | *CE* | 110M | WebVid2.5M + HowTo100M | 3 | 94.7 | 95.1 | 65.4 |
| *All-in-one*-B * | *CE* | 110M | CC3M + WebVid2.5M + YT-Temporal180M | 3 | 95.5 | 94.7 | 66.3 |

(a) Three sub-tasks on TGIF-QA test set (the first row are methods w/o. pre-training). "*T*" refers to text encoder, "*V*" is video encoder and "*CE*" is cross-modality encoder. 384 means the resolution is $384 \times 384$ for each frame while the default is $224 \times 224$.

| Method | Frames | Accuracy |
|---|---|---|
| Heterogeneous [9] | 35 | 33.0 |
| ClipBERT [26] | $4 \times 2$ | 37.4 |
| VIOLET [10] | 16 | 43.1 |
| GIT [48] | 6 | 43.2 |
| FrozenBiLM [57] | 10 | 47.0 |
| *All-in-one*-S | 3 | 39.5 |
| *All-in-one*-B | 3 | 42.9 (0.2↓) |
| *All-in-one*-B+ | 3 | **44.6** (1.5↑) |
| *All-in-one*-B * | 3 | 46.8 |

(b) MSRVTT-QA test set.

| Method | Frames | Accuracy |
|---|---|---|
| QueST [17] | 10 | 36.1 |
| HCRN [25] | 16 | 36.1 |
| SSML [2] | 16 | 35.1 |
| CoMVT [41] | 30 | 42.6 |
| Just-Ask † [56] | 32 | 46.3 |
| *All-in-one*-S | 3 | 41.7 |
| *All-in-one*-B | 3 | 46.5 (0.2↑) |
| *All-in-one*-B+ | 3 | **48.2** (1.9↑) |
| *All-in-one*-B * | 3 | 48.3 |

(c) MSVD-QA test set.

| Method | Frames | Accuracy |
|---|---|---|
| PAMN [19] | 32 | 66.3 |
| Multi-task [18] | 16 | 66.2 |
| STAGE [27] | 16 | 70.5 |
| CA-RN [12] | 32 | 68.9 |
| MSAN [20] | 40 | 70.4 |
| *All-in-one*-S | 3 | 63.5 |
| *All-in-one*-B | 3 | 69.8 |
| *All-in-one*-B+ | 3 | 71.5 |
| *All-in-one*-B * | 3 | 72.0 |

(d) TVQA val set.

Table 2. **Comparison with state-of-the-art methods on VQA.** The columns with gray color are **open-ended VQA** and the others are **multiple-choice VQA**. † means use additional large-scale VQA dataset HowToVQA60M [56] for pre-training. ⋆ means pre-training with additional YT-Temporal180M [61]. The parameter of FrozenBiLM [3] is 890M, eight times larger than *All-in-one*-B+.

| Method | Frames | MSRVTT | LSMDC |
|---|---|---|---|
| JSFusion [59] | 40 | 83.4 | 73.5 |
| ActBERT [64] | 32 | 85.7 | - |
| ClipBERT [26] | $8 \times 2$ | 88.2 | - |
| MERLOT [61] | 8 | - | 81.7 |
| VIOLET [10] | 16 | - | 82.9 |
| *All-in-one*-B | 3 | 91.4 | 83.1 |
| *All-in-one*-B+ | 3 | **91.9** (3.8↑) | **83.9** (1.0↑) |
| *All-in-one*-B+ (zero-shot) | 3 | 82.2 | 58.1 |

Table 3. **Comparison with state-of-the-art methods on multiple-choice task.**

| Method | Parameters | #Frames | Zero-shot | Fine-tune |
|---|---|---|---|---|
| Frozen [3] | 232M | 8 | 32.47 | 60.32 |
| VATT † [1] | 264M | 3 | 27.34 | 59.44 |
| *All-in-one*-B | 110M | 3 | **36.52** | **65.89** |

Table 4. **The multiple-choice result on first-view ego-4d.** † means our implementation.

VIOLET [10] about 5.8% on the Action subset. Interestingly, we find more frames do not benefit Action and Transition split but FrameQA. We also report the result of *All-in-one* on the three open-ended datasets. Even though Just-Ask [56] is specifically designed for VQA and pre-trained on a large-scale HowToVQA69M, our method still achieves a similar even better result than Just-Ask on MSVD-QA.

#### 4.1.2 Multiple-choice.

Tab. 3 shows that *All-in-one* improves the ClipBERT model by 3.2% on accuracy, on MSRVTT multiple-choice test task. We also report the zero-shot results for comparison and find that zero-shot accuracy already close to JSFusion [59] in MSRVTT multiple-choice with only three frames as input.

**Extending to Egocentric Video.** Ego4d [14] is a egocentric dataset that has a large domain gap with our third-view video from Youtube. We test multiple-choice (5 options) tasks on this dataset. We report both the zero-shot result and fine-tune result in Tab. 4. Comparing with other multiple-choice benchmarks such as LSMDC and MSR-VTT, this dataset is more challenge and hards to tell sample

| Method | Nets | PT Data | Params | Flops | Frames | 9K Train | | | 7K Train | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| ClipBERT [26] | *T+V+CE* | COCO + Visual Genome | 137M | 183.2G | 8 × 2 | - | - | - | 22.0 | 46.8 | 59.9 |
| TACo [58] | *T+V+CE* | HowTo100M | 212M | 140.5G | 48 | 28.4 | 57.8 | 71.2 | 24.8 | 52.1 | 64.0 |
| VIOLET [10] | *T+V+CE* | CC3M + WebVid2.5M | 198M | 351.4G | 16 | 34.5 | 63.0 | 73.4 | - | - | - |
| CLIP-ViP [54] | *T+V+CE* | WIT300M+HD-VILA-100M | - | 225.1G | 12 | 50.1 | 74.8 | 84.6 | - | - | - |
| Frozen [3] | *T+V* | CC3M + WebVid2.5M | 232M | 217.3G | 8 | 31.0 | 59.5 | 70.5 | - | - | - |
| OA-Trans [46] | *T+O+V* | CC3M + WebVid2.5M | 232M | 217.3G | 8 | 35.8 | 63.4 | 76.5 | 32.1 | 61.0 | 72.9 |
| MILES [11] | *T+V* | CC3M + WebVid2.5M | 295M | 771.0G | 4 | 37.7 | 63.6 | 73.8 | - | - | - |
| *All-in-one*-B | *CE* | HowTo100M | 110M | 58.7G | 3 | 29.5 | 63.3 | 71.9 | 26.5 | 59.4 | 69.8 |
| *All-in-one*-B+ | *CE* | CC3M + WebVid2.5M | 110M | 58.7G | 3 | 39.7 | 67.8 | 76.1 | 35.9 | 66.1 | 75.1 |
| *All-in-one*-B+ | *CE* | + HowTo100M | 110M | 58.7G | 3 | 41.8 | 68.5 | 76.7 | 37.3 | 66.4 | 75.6 |

(a) The retrieval performance on MSR-VTT 9K and 7K training split. For Nets, "*O*" is object extractor. Notice that COCO [34], CC3M [42]) and Visual Genome are all image-text datasets, which are not suitable for temporal modeling during pre-training.

| Method | Frames | R@1 | R@5 | R@10 | MdR |
|---|---|---|---|---|---|
| Dense [22] | 32 | 14.0 | 32.0 | - | 34.0 |
| FSE [62] | 16 | 18.2 | 44.8 | - | 7.0 |
| HSE [62] | 8 | 20.5 | 49.3 | - | - |
| ClipBERT [26] | 4 × 2 | 20.9 | 48.6 | 62.8 | 6.0 |
| *All-in-one*-B | 3 | 21.5 | 50.3 | 65.5 | 6.0 |
| *All-in-one*-B+ | 3 | **22.4** | **53.7** | **67.7** | **5.0** |

(b) ActivityNet Caption val1 set.

| Method | Frames | R1 | R5 | R10 | MdR |
|---|---|---|---|---|---|
| FSE [62] | 16 | 13.9 | 36.0 | - | 11.0 |
| CE [35] | 16 | 16.1 | 41.1 | - | 8.3 |
| ClipBERT [26] | 8 × 2 | 20.4 | 48.0 | 60.8 | 6.0 |
| Frozen [3] | 8 | 31.0 | 59.8 | 72.4 | 3.0 |
| *All-in-one*-B | 3 | 31.2 | 60.5 | 72.1 | 3.0 |
| *All-in-one*-B+ | 3 | **32.7** | **61.4** | **73.5** | **3.0** |

(c) DiDeMo test set.

Table 5. **Comparison with state-of-the-art methods on text-to-video retrieval.** We gray out dual-stream networks that only do retrieval tasks. Notice that OA-Trans [46] uses additional offline object features.

apart with static cues only, but our *All-in-one* still outperforms Frozen [3] clearly with half the parameters and less frames. With same pretrain data, our method also outpeforms VATT [1] clearly.

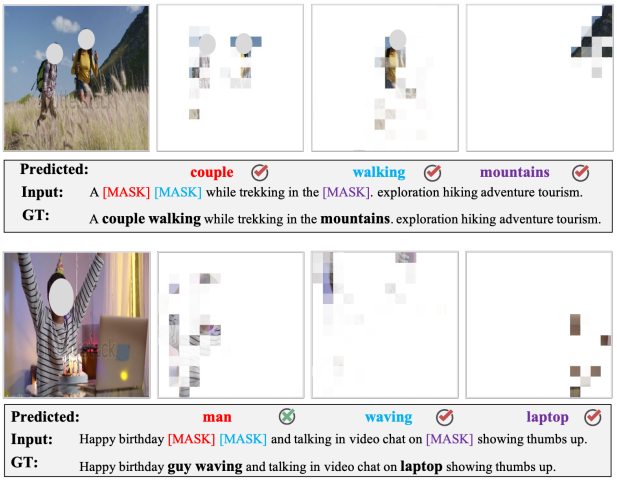


Figure 7. **Cloze evaluation**: Given a video and masked text pair, the model fills masked words and displays their corresponding high-attention patch. These samples are randomly selected from the Webvid [3] dataset's validation set.

## 4.2. Video-text Alignment Task

### 4.2.1 Text-to-video Retrieval.

In this experiment, we fine-tune *All-in-one* on MSRVTT, ActivityNet Caption, and DiDeMo datasets. Tab. 5 summarizes results on text-to-video retrieval. In Tab. 5(a), *All-in-one* achieves significant performance gain over existing methods on MSRVTT retrieval in both 9K and 7K train settings. Compare with these related works, we only use one Cross-modality Encoder and the parameter is half of the Frozen [3]. *All-in-one* even leads to 2.1% relative improvement on R@1 when compare with OA-Trans [46], which use additional offline object feature and only focus on retrieval. When adopt to LSMDC and DiDeMo dataset, our method also show competitive performance.

## 4.3. Video Captioning and Action Recognition

### 4.3.1 Video Captioning.

Follow SwinBERT [32], we add a light Language Modeling Head [7] on top of *All-in-one*. For fair comparison, we compare with related pre-training works on TVC and YouCook2 datasets in Tab. 7. We observe *All-in-one* outperforms ActBert in terms of CIDEr metric by a large margin on YouCook2. *All-in-one* even leads to better result on TVC which use additional text script as input. These results showcase the generative capability of *All-in-one* as an

| Method | Parameters | #Frames | K400 | | | HMDB51 | | | UCF101 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Top-1 | Top-5 | Top-10 | Top-1 | Top-5 | Top-10 | Top-1 | Top-5 | Top-10 |
| MIL-NCE [38] | 157M | 32 | - | - | - | 53.1 | 87.2 | 92.8 | 82.7 | - | - |
| Frozen [3] | 232M | 8 | 50.5 | 80.7 | 90.2 | 54.3 | 88.0 | 94.8 | 81.3 | 94.3 | 96.2 |
| VATT † [1] | 264M | 3 | 44.3 | 75.2 | 87.3 | 43.1 | 75.5 | 90.5 | 77.6 | 86.4 | 90.9 |
| *All-in-one*-B | 110M | 3 | 49.8 | 79.8 | 90.7 | 51.9 | 84.1 | 93.4 | 81.1 | 93.8 | 95.5 |
| *All-in-one*-B | 110M | 8 | 52.4 | 83.2 | **92.9** | 54.7 | 88.2 | 95.2 | 82.8 | 95.1 | 96.9 |
| *All-in-one*-B+ (Not Shared) | 110M | 8 | **53.2** | **83.5** | 92.7 | **55.2** | **89.1** | **95.8** | **84.1** | **95.7** | **97.8** |
| *All-in-one*-B+ (Shared) | 110M | 8 | 51.4 | 78.5 | 89.9 | 53.1 | 87.1 | 93.2 | 82.0 | 94.0 | 96.0 |

Table 6. **The linear probe results on action recognition benchmarks over kinetics 400, hmdb51 and UCF101 datasets.** Notice that not shared means two pre-text heads are not shared for image-text and video-text pairs, and the video-text head are used for fine-tuning.

| Method | TVC | | | YouCook2 | | |
|---|---|---|---|---|---|---|
| | B4 | M | C | B4 | M | C |
| VideoBERT [44] | - | - | - | 4.3 | 11.9 | 55.0 |
| ActBERT [64] | - | - | - | 5.4 | 13.3 | 65.0 |
| VALUE [28] | 11.6 | 17.6 | 50.5 | 12.4 | 18.8 | 130.3 |
| *All-in-one*-B | 11.3 | 19.2 | 54.3 | 10.7 | 13.5 | 109.4 |
| *All-in-one*-B+ | 12.5 | 20.4 | 56.3 | 11.2 | 13.9 | 114.5 |

Table 7. **Video captioning results on the test split of TVC and YouCook2.** We gray out VALUE which use both Video and subtitle sentences from the original TV show scripts modality for input while the others only use Video modality.

video-text foundation model.

#### 4.3.2 Action Recognition via Linear Probe.

To evaluate the transfer ability of our model on the single-modality task. We transfer the learned representation to downstream linear probe results on K400, UCF101 and HMDB51 datasets. Specifically, we *frozen* the overall unified model and only learn linear layers based on the *cls* token of the last layer. By pre-training model on these two datasets, we compare the base model with Time Average and the previous pre-training method Frozen.

The linear probe results are given in Tab. 6. We observe the number of frames has a large impact on this task. When adopting the same 8 frames, our *All-in-one*-B clearly outperforms Frozen [3] especially on the large-scale K400 dataset. We also outperform MIL-NCE [38] clearly on UCF101 and HMDB51 datasets. Interestingly, we find the model have more stable results on this temporal-related task if we train two pre-text heads independently for image-text and video-text inputs . But there have no large difference for both shared or not shared for other tasks.

## 5. Visualization

To better understand the pre-trained *All-in-one*, we analyze its internal representations. Specifically, given paired ground truth text and raw video, we mask some keywords (both *verb* and *nouns*) and ask the model to predict these masked words and further find out which video patch has strong correlations with the masked words. We use optimal transports [51] to calculate the correlation between video and text. We only show the attention weight that is larger than the given threshold and give some examples of cross-modal alignment in Fig. 7. We find the model can predict correct *nouns* and *verbs* in most cases. Sometimes, it predicts the wrong word but with a similar meaning to the correct word. e.g. "guy" and "man". Benefiting from temporal modeling, we also find that the model attends to the motion regions for *verbs* like "waving" and "walking".

## 6. Conclusions & Future Work

In this paper, we present the first unified end-to-end Video-Language Pre-training (VLP) architecture with raw video and text as input, *All-in-one*. *All-in-one* is able to compete with contemporaries who are equipped with additional robust off-the-shelf video visual embedding networks and shows potential for the future by learning just one cross-modality fusion network. Instead of solely concentrating on heavier single-modality embedders or larger fusion models, we expect that the VLP community would place more emphasis on lightweight end-to-end modal interactions within Transformer modules. Although these preliminary findings are promising, this novel approach to unified video-language interaction also poses additional difficulties, particularly with regard to fine-grained word region matching. Additionally, the temporal modeling has yet to be completely investigated, and we hope the usage of *All-in-one* for other single-modality tasks in future research.

## Acknowledgement

This project is supported by the National Research Foundation, Singapore under its NRFF Award NRF-NRFF13-2021-0008. The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore.

## 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] Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. In *AAAI*, 2021.

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021.

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding. *arXiv preprint arXiv:2102.05095*, page 4, 2021.

[5] 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.

[6] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] 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. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019.

[10] 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.

[11] Yuying Ge, Yixiao Ge, Xihui Liu, Jinpeng Wang, Jianping Wu, Ying Shan, Xiaohu Qie, and Ping Luo. Miles: visual bert pre-training with injected language semantics for video-text retrieval. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 691–708. Springer, 2022.

[12] Shijie Geng, Ji Zhang, Zuohui Fu, Peng Gao, Hang Zhang, and Gerard de Melo. Character matters: Video story understanding with character-aware relations. *arXiv preprint arXiv:2005.08646*, 2020.

[13] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. *arXiv preprint arXiv:2201.08377*, 2022.

[14] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. *arXiv preprint arXiv:2110.07058*, 2021.

[15] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017.

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[17] Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11101–11108, 2020.

[18] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. Gaining extra supervision via multi-task learning for multi-modal video question answering. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[19] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. Progressive attention memory network for movie story question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8337–8346, 2019.

[20] Junyeong Kim, Minuk Ma, Trung Pham, Kyungsu Kim, and Chang D Yoo. Modality shifting attention network for multi-modal video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10115, 2020.

[21] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[22] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, pages 32–73, 2017.

[24] Thomas S. Kuhn. The structure of scientific revolutions. University of Chicago Press, 1962.

[25] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF con-*

*ference on computer vision and pattern recognition*, pages 9972–9981, 2020.

[26] 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*, pages 7331–7341, 2021.

[27] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*, 2019.

[28] 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. *arXiv preprint arXiv:2106.04632*, 2021.

[29] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

[30] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[31] 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.

[32] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17958, 2022.

[33] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. 2022.

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[35] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019.

[36] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021.

[37] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[38] 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.

[39] 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*, pages 2630–2640, 2019.

[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. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[41] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16877–16887, 2021.

[42] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[43] 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/CVF International Conference on Computer Vision*, pages 7464–7473, 2019.

[44] 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/CVF International Conference on Computer Vision*, pages 7464–7473, 2019.

[45] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[46] Alex Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Object-aware video-language pre-training for retrieval. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[47] Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. Ufo: A unified transformer for vision-language representation learning. *arXiv preprint arXiv:2111.10023*, 2021.

[48] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

[49] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions*

*on pattern analysis and machine intelligence*, pages 2740–2755, 2018.

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

[51] Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. A fast proximal point method for computing exact wasserstein distance. In *Uncertainty in artificial intelligence*, pages 433–453. PMLR, 2020.

[52] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.

[53] 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.

[54] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022.

[55] Rui Yan, Mike Zheng Shou, Yixiao Ge, Alex Jinpeng Wang, Xudong Lin, Guanyu Cai, and Jinhui Tang. Video-text pre-training with learned regions. *arXiv preprint arXiv:2112.01194*, 2021.

[56] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021.

[57] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. In *Advances in Neural Information Processing Systems*, 2022.

[58] 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*, pages 11562–11572, 2021.

[59] 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)*, pages 471–487, 2018.

[60] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022.

[61] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34, 2021.

[62] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 374–390, 2018.

[63] Bowen Zhang, Jiahui Yu, Christopher Fifty, Wei Han, Andrew M Dai, Ruoming Pang, and Fei Sha. Co-training transformer with videos and images improves action recognition. *arXiv preprint arXiv:2112.07175*, 2021.

[64] 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.