Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval

Max Bain[1] Arsha Nagrani[1†] Gül Varol[1,2] Andrew Zisserman[1]
[1] Visual Geometry Group, University of Oxford
[2] LIGM, École des Ponts, Univ Gustave Eiffel, CNRS
{maxbain, arsha, gul, az}@robots.ox.ac.uk

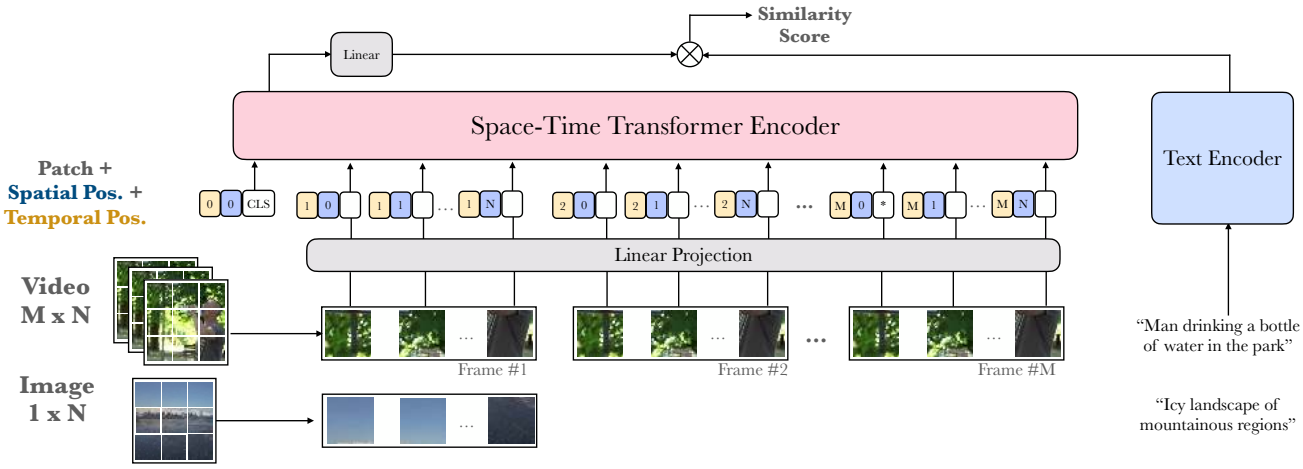


Figure 1: **Joint Image and Video Training:** Our dual encoding model consists of a visual encoder for images and video and a text encoder for captions. Unlike 2D or 3D CNNs, our space-time transformer encoder allows us to train flexibly on both images and videos with captions jointly, by treating an image as a single frame video.

Abstract

Our objective in this work is video-text retrieval – in particular a joint embedding that enables efficient text-to-video retrieval. The challenges in this area include the design of the visual architecture and the nature of the training data, in that the available large scale video-text training datasets, such as HowTo100M, are noisy and hence competitive performance is achieved only at scale through large amounts of compute.

We address both these challenges in this paper. We propose an end-to-end trainable model that is designed to take advantage of both large-scale `image` *and video captioning datasets. Our model is an adaptation and extension of the recent ViT and Timesformer architectures, and consists of attention in both space and time. The model is flexible and can be trained on both image and video text datasets, either independently or in conjunction. It is trained with a curriculum learning schedule that begins by treating images as 'frozen' snapshots of video, and then gradually learns to attend to increasing temporal context when trained on video datasets. We also provide a new video-text pretraining dataset WebVid-2M, comprised of over two million videos with weak captions scraped from the internet. Despite training on datasets that are an order of magnitude smaller, we show that this approach yields state-of-the-art results on standard downstream video-retrieval benchmarks including MSR-VTT, MSVD, DiDeMo and LSMDC.*

1. Introduction

Joint visual-text models have become increasingly popular as they enable a wide suite of downstream tasks, including text-to-visual retrieval [29, 31, 36, 59], visual captioning [24, 58, 66], and visual question and answering [4, 27]. Their rapid development is due to the usual improvements on three fronts: new neural network architectures (e.g. transformers [56] for both text and visual inputs); new

[†]Now at Google Research.

large-scale datasets; and new loss functions that are, for example, able to handle label noise [35]. However, their development mostly proceeds on two independent tracks: one for *images*, with its own architectures, training datasets and benchmarks [25, 29, 52]; and the other for *videos* with a similar separation of training datasets and benchmarks [3, 5, 24, 46, 65, 70]. The only common link between the two is that often video networks are initialized by pre-training image networks on image datasets [6, 8]. This separation of effort is suboptimal given the overlap in information that images and video convey over multiple tasks. For example, although classifying some human actions requires the temporal ordering of video frames, many actions can be classified from just their distribution over frames or even from a single frame [51].

In this paper we take a step towards unifying these two tracks, by proposing a dual encoder architecture which utilises the flexibility of a transformer visual encoder to train from images-with-captions, from video clips-with-captions, or from both (Fig. 1). We do this by treating images as a special case of videos that are 'frozen in time'. Using a transformer-based architecture allows us to train with variable-length sequences, treating an image as if it was a single frame video, unlike in standard 3D CNNs [8, 18, 64] where to train on images jointly with videos one must incur the cost of actually generating a static video. Furthermore, unlike many recent methods [16, 31, 36] for video-text dual encoding, we do not use a set of 'expert networks' that are pre-trained on external image datasets and then fixed, but instead train the model end-to-end.

This end-to-end training is facilitated by scraping the web for a new large-scale video-text captioning dataset of over two million video alt-text pairs (WebVid-2M). We also take advantage of large-scale image captioning datasets such as Conceptual Captions [52].

We make the following contributions: (i) we propose a new end-to-end model for video retrieval that does *not* rely on 'expert' features, but instead, inspired by [6] employs a transformer architecture with a modified divided space-time attention applied directly to pixels; (ii) because our architecture can gracefully handle inputs of different lengths, it is versatile and can be flexibly trained on both video and image datasets (by treating images as a single-frame video). We build on this flexibility by designing a curriculum learning schedule that begins with images and then gradually learns to attend to increasing temporal context when trained on video datasets through temporal embedding interpolation. We show that this increases efficiency, allowing us to train models with far less GPU time; (iii) we introduce a new dataset called WebVid-2M, consisting of 2.5M video-text pairs scraped from the web; and finally (iv) we achieve state-of-the-art performance by only using the video modality on MSR-VTT [65], MSVD [9], DiDeMo [3] and LSMDC [46] – outperforming works that use pre-extracted experts from multiple modalities, as well as those that are pretrained on the noisy HowTo100M, which is 20x larger than our dataset in the number of video-text pairs.

2. Related Works

Pretraining for video-text retrieval. Given that most video-text retrieval datasets tend to be small-scale, the dominant paradigm for video retrieval has been to use a combination of pre-extracted features from 'expert' models, including models trained for various diverse tasks and on multiple modalities such as face, scene and object recognition, action classification and sound classification. MoEE [36], CE [31], MMT [16] and concurrent work HiT [30] all follow this paradigm, with the overall similarity for a video-text pair obtained as a weighted sum of each expert's similarity with the text.

However, since the release of the HowTo100M dataset [37], a large-scale instructional video dataset, there has been a flurry of works leveraging large-scale pretraining to improve video-text representations for tasks such as video question-answering [50], text-video retrieval [41] and video captioning [71]. Although semantically rich and diverse, text supervision from instructional videos is extremely noisy, and hence incurs a large computational cost, as scale is required for competitive results. A few approaches have been proposed to combat the noise – *e.g.* using loss functions such as MIL-NCE [35] or using the raw audio [1, 48] directly to increase robustness. Given the large size of existing image-captioning datasets, some have naturally tried to overcome the lack of video-caption training data with joint image-text pretraining (such as in MoEE [36] and ClipBERT [26]). MoEE [36] trains on images jointly by feeding in zeros to all expert streams that require videos, such as the motion and audio features, while ClipBERT [26] restricts their feature extractors to 2D CNNs. Instead we propose an elegant transformer-based encoder that works well with either images or videos and can be trained effectively on both.

Similar to our work, although only suitable for images is CLIP [42], which learns an effective joint image-text representation from millions of text-image pairs scraped from the internet using contrastive loss.

End-to-end video representation learning. A large number of architectural developments have been driven by action recognition on datasets such as Kinetics [21] where manual labelling has been relatively easier than obtaining textual descriptions for datasets. For a long time this space was dominated by spatio-temporal CNNs such as I3D [8], 3D ResNets [18], S3D [64] or 'R(2+1)D' CNNs [55]. Here, images are used simply to initialise video models, through inflation [8]. Multigrid scheduling has been proposed for

efficient training [63].

Transformers for vision. A number of works use self-attention for images, either in combination with convolutions [7, 19, 56, 61] or even replacing them entirely.

Works that use only self-attention blocks tend to apply them at an individual pixel level [12, 39, 43], often requiring tricks to ensure computational tractability, including restricting the scope of self-attention to a local neighbourhood [43], adding global self-attention on heavily downsized versions, or sparse key-value sampling [11]. To increase efficiency, ViT [14] decompose images into a sequence of patches and then feeds linear embeddings of these patches as inputs to a transformer, effectively adding a single convolutional layer to the image at the start. This idea has been extended in DeiT [54]. For video, previous works also employ self-attention blocks together with CNN layers, for action recognition [17] and video classification [10].

In contrast, our architecture consists entirely of self-attention units and is heavily inspired by ViT [14] and particularly the Timesformer [6], which uses divided space and time attention. Unlike these works, we use expandable temporal embeddings to allow flexible training of variable-length videos and images both jointly and separately. We are unaware of any previous works that use self-attention to train on both images and videos in the same model.

3. Method

In this section, we describe our transformer-based spatio-temporal model architecture (Section 3.1), and our training strategy (Section 3.2). Details are given in the Appendix.

3.1. Model Architecture

Input. The visual encoder takes as input an image or video clip $X \in \mathrm{R}^{M\times 3\times H\times W}$ consisting of M frames of resolution $H \times W$, where $M = 1$ for images. The text encoder takes as input a tokenised sequence of words.

Spatio-temporal patches. Following the protocol in ViT and Timesformer [6], the input video clip is divided into $M\times N$ non-overlapping spatio-temporal patches of size $P\times P$, where $N = HW/P^2$.

Transformer input. The patches $\boldsymbol{x} \in \mathrm{R}^{M\times N\times 3\times P\times P}$ are fed through a 2D convolutional layer and the output is flattened, forming a sequence of embeddings $\boldsymbol{z} \in \mathrm{R}^{MN\times D}$ for input to the transformer, where D depends of the number of kernels in the convolutional layer.

Learned temporal and spatial positional embeddings, $\boldsymbol{E}^s \in \mathrm{R}^{N\times D}$, $\boldsymbol{E}^t \in \mathrm{R}^{M\times D}$ are added to each input token:

$$\boldsymbol{z}^{(0)}_{p,m} = \boldsymbol{z}_{p,m} + \boldsymbol{E}^s_p + \boldsymbol{E}^t_m, \tag{1}$$

such that all patches within a given frame m (but different spatial locations) are given the same temporal positional embedding E^t_m, and all patches in the same spatial location (but different frames) are given the same spatial positional embedding E^s_p. Thus enabling the model to ascertain the temporal and spatial position of patches.

In addition, a learned [CLS] token [13] is concatenated to the beginning of the sequence, which is used to produce the final visual embedding output embedding of the transformer.

Space-time self-attention blocks. The video sequence is fed into a stack of space-time transformer blocks. We make a minor modification to the Divided Space-Time attention introduced by [6], by replacing the residual connection between the block input and the temporal attention output with a residual connection between the block input and the spatial attention output, see the appendix for details. Each block sequentially performs temporal self-attention and then spatial self-attention on the output of previous block. The video clip embedding is obtained from the [CLS] token of the final block.

Text encoding. The text encoder architecture is a multi-layer bidirectional transformer encoder, which has shown great success in natural language processing tasks [13]. For the final text encoding, we use the [CLS] token output of the final layer.

Projection to common text-video space. Both text and video encodings are projected to a common dimension via single linear layers. We compute the simliarity between text and video by performing the dot product between the two projected embeddings.

Efficiency. Our model has independent dual encoder pathways (such as in MIL-NCE [35] and MMV networks [1]), requiring only the dot product between the video and text embeddings. This ensures retrieval inference is of trivial cost since it is indexable, i.e. it allows application of fast approximate nearest neighbour search, and is scalable to very large scale retrieval at inference time. Given t text queries and v videos in a target gallery, our retrieval complexity is $O(t+v)$. In contrast, ClipBERT [26] which inputs both text and video as input to a single encoder, has retrieval complexity $O(tv)$ since every text-video combination must be inputted to the model. Other expert-based retrieval methods such as MoEE [36], CE [31] and MMT [16] also contain a dual encoder pathway, however they still require query-conditioned weights to compute the similarity scores for each expert, while our model does not.

3.2. Training Strategy

Loss. We employ [68] in a retrieval setting, where matching text-video pairs in the batch are treated as positives, and all other pairwise combinations in the batch are treated as negatives. We minimise the sum of two losses, video-to-

text and text-to-video:

$$L_{v2t} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\exp(x_i^\top y_i/\sigma)}{\sum_{j=1}^{B}\exp(x_i^\top y_j/\sigma)} \quad (2)$$

$$L_{t2v} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\exp(y_i^\top x_i/\sigma)}{\sum_{j=1}^{B}\exp(y_i^\top x_j/\sigma)} \quad (3)$$

where x_i and y_j are the normalized embeddings of i-th video and the j-th text respectively in a batch of size B and σ is the temperature.

Joint image-video training. In this work, we train jointly on both image-text pairs as well as video-text pairs, taking advantage of both for larger-scale pretraining. Our joint training strategy involves alternating batches between the image and video datasets. Since the attention mechanism scales with the square of input frames $O(M^2)$, the alternate batch training allows the image batches ($M = 1$) to be far greater in size.

Weight initialisation and pretraining. Following [6], we initialise the spatial attention weights in the space-time transformer model with ViT [14] weights trained on ImageNet-21k, and initialise the temporal attention weights to zero. The residual connections mean that under these initialisation settings, the model is at first equivalent to ViT over each input frame – thereby allowing the model to learn to attend to time gradually as training progresses. Since transformer architectures have demonstrated most of their success from large-scale pretraining, we utilise two large-scale text-image/video datasets with a joint training strategy, resulting in large improvements in performance.

Temporal curriculum learning. The space-time transformer architecture allows a variable length input sequence and therefore a variable number of input video frames. If the model has only trained on videos up to length m however, then the temporal positional embedding $\boldsymbol{E}^t$ will only be learned up to $\boldsymbol{E}^t_{:m}$. Therefore, applying the model to input video of sequences up to length M will result the addition of $\boldsymbol{E}^t_{m:M}$, which would not yet be learned.

Two temporal expansion methods are investigated: *interpolation* and *zero-padding*. Zeros can be filled in, $\mathbf{0} \rightarrow \boldsymbol{E}^t_{m:M}$, allowing the model to learn the additional temporal positions from scratch during training. Alternatively, interpolation could be used to upsample the temporal embeddings in the temporal dimension, $\boldsymbol{E}^t_{:m} \rightarrow \boldsymbol{E}^t_{:M}$. We investigate two methods of interpolation: nearest neighbour and bilinear. The effects of these different initialisations can be found in the Supplementary Materials.

We employ this expansion strategy in order to perform curriculum learning in the number of input frames. Initially training on fewer frames has drastic savings in computation, whilst having comparable or even better performance (see Section 4.5).

Frame sampling. Given a video containing L frames, we subdivide it into M equal segments where M is the desired number of frames for the video encoder. During training, we sample a single frame uniformly from each segment (in a similar manner to TSN [60] and GST [32]). At test time, we sample the i^{th} frame in every segment, to get a video embedding v_i. The values for i are determine using a stride S, resulting in an array of video embeddings $\boldsymbol{v} = [v_0, v_S, v_{2S}, v_M]$. The mean of these video embeddings is used as the final embedding for the video.

4. Experiments

We first describe the pretraining datasets including our WebVid-2M video-text dataset (Section 4.1), followed by the downstream datasets used for the evaluations in our experiments (Section 4.2). We then describe implementation details of our model (Section 4.3). Next, we ablate various training components on the MSR-VTT dataset, in particular the effects of pretraining and our space-time attention modification (Section 4.4), and our proposed curriculum strategy (Section 4.5). Then, we compare to the state of the art on four benchmarks: MSR-VTT, MSVD, DiDeMo and LSMDC (Section 4.6).

4.1. Pretraining Datasets

We jointly pretrain our model on image and video data.

Video pretraining: The WebVid-2M Dataset. We scrape the web for a new dataset of videos with textual description annotations, called WebVid-2M. Our dataset consists of 2.5M video-text pairs, which is an order of magnitude larger than existing video captioning datasets (see Table 1).

The data was scraped from the web following a similar procedure to Google Conceptual Captions [52] (CC3M). We note that more than 10% of CC3M images are in fact thumbnails from videos, which motivates us to use such video sources to scrape a total of 2.5M text-video pairs. The use of data collected for this study is authorised via the Intellectual Property Office's Exceptions to Copyright for Non-Commercial Research and Private Study[1]. We are currently performing further analysis of the dataset on its diversity and fairness.

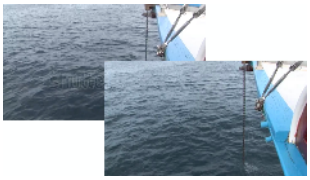
Figure 2 provides sample video-caption pairs. There are a variety of different styles used in caption creation, as can be seen from Figure 2 (left to right) where the first video has a longer, poetic description compared to the succinct description for the second video. The third video caption has a less defined sentence structure, with keywords appended to the end, while the fourth video mentions a specific place (maldives). Time-specific information is important for the second and third example, where details such as "talking on

[1] www.gov.uk/guidance/exceptions-to-copyright/

"Lonely beautiful woman sitting on the tent looking outside. wind on the hair and camping on the beach near the colors of water and shore. freedom and alternative tiny house for traveler lady drinking"

"Billiards, concentrated young woman playing in club"

"Female cop talking on walkie-talkie, responding emergency call, crime prevention"



"Get anchor for departure safari dive boat scuba diving maldives"

Figure 2: **Example video-caption pairs from the WebVid2M dataset:** Note the different captioning styles: from left to right, captions can be (i) long, slightly poetic, with disjoint sentences and phrases, (ii) succint and to the point, (iii) have a less defined sentence structure with keywords appended to the end, (iv) mention specific places ('maldives'). We show two randomly sampled frames for each video.

walkie-talkie" or "playing billiards" would be missed when looking at certain frames independently.

Table 1: **Dataset Statistics:** We train on a new dataset mined from the web called WebVid2M. Our dataset is an order of magnitude larger than existing video-text datasets in the number of videos and captions. HowTo100M (highlighted in blue) is a video dataset with noisy, weakly linked text supervision from ASR.

dataset	domain	#clips	avg dur. (secs)	#sent	time (hrs)
MPII Cook [47]	cooking	44	600	6K	8
TACos [44]	cooking	7K	360	18K	15.9
DideMo [3]	flickr	27K	28	41K	87
MSR-VTT [65]	youtube	10K	15	200K	40
Charades [53]	home	10K	30	16K	82
LSMDC15 [46]	movies	118K	4.8	118K	158
YouCook II [70]	cooking	14K	316	14K	176
ActivityNet [24]	youtube	100K	180	100K	849
CMD [5]	movies	34K	132	34K	1.3K
WebVid-2M	open	**2.5M**	18	**2.5M**	**13K**
HT100M [37]	instruction	136M	4	136M	134.5K

We note that our video dataset is 10x smaller than HowTo100M in video duration and over 20x smaller in the number of paired clip-captions (Table 1). Our dataset consists of manually generated captions, that are for the most part well formed sentences. In contrast, HowTo100M is generated from continuous narration with incomplete sentences that lack punctuation. The clip-text pairs are obtained from subtitles and may not be temporally aligned with the video they refer to, or indeed may not refer to the video at all [37]. Our captions, on the other hand, are aligned with the video and describe visual content.

Moreover, there is no noise from imperfect ASR transcription and grammatical errors as is the case for HowTo100M. Our dataset also has longer captions on average (12 vs 4 words for HowTo) which are more diverse (Measure of Textual Lexical Diversity, MTLD [34] = 203 vs 13.5).

Image pretraining: Google Conceptual Captions [52]. This dataset consists of about 3.3M image and description pairs. Unlike the curated style of COCO images, Conceptual Captions (CC3M) images and their raw descriptions are harvested from the web, and therefore represent a wider variety of styles. The raw descriptions are harvested from the Alt-text HTML attribute associated with web images.

4.2. Downstream Datasets

We now describe the downstream text-video datasets that our model is evaluated on.

MSR-VTT [65] contains 10K YouTube videos with 200K descriptions. Following other works [31], we train on 9K train+val videos and report results on the 1K-A test set.

MSVD [9] consists of 80K English descriptions for 1,970 videos from YouTube, with each video containing 40 sentences each. We use the standard split of 1200, 100, and 670 videos for training, validation, and testing [31, 41].

DiDeMo [3] contains 10K Flickr videos annotated with 40K sentences. Following [26, 31], we evaluate paragraph-to-video retrieval, where all sentence descriptions for a video are concatenated into a single query. Since this dataset comes with localisation annotations (ground truth proposals), we report results with ground truth proposals (where only the localised moments in the video are concatenated and used in the retrieval set as done by [26]) as well as without (as done by [31]).

LSMDC [45] consists of 118,081 video clips sourced from 202 movies. The validation set contains 7,408 clips and evaluation is done on a test set of 1,000 videos from movies disjoint from the train and val sets. This follows the protocol outlined in [46].

For downstream datasets with separate `val` and `test` splits, we train all models for 75 epochs and use the epoch with the lowest validation loss for reporting test results. For downstream datasets without a `val` set we report results at 50 epochs.

4.3. Implementation Details

All experiments are conducted with PyTorch [40]. Optimization is performed with Adam, using a learning rate of 1×10^{-5}, we use batch sizes of 16, 24, and 96 for 8, 4, and 1-frame inputs respectively. The temperature hyperparameter σ for the loss defined in Eq. 2 & 3 is set to 0.05. The default pretraining is WebVid-2M and CC3M.

For the visual encoder, all models have the following: $|\ell| = 12$ attention blocks, patch size $P = 16$, sequence dimension $D = 768$, 12 heads and takes 4-frames as downstream input.

The text encoder of all models, unless specified otherwise, is instantiated as DistilBERT base-uncased [49] pretrained on English Wikipedia and Toronto Book Corpus. The dimensionality of the common text-video space is set to 256. For visual augmentation, we randomly crop and horizontally flip during training, and center crop the maximal square crop at test time. All videos are resized to 224×224 as input. At test-time we compute clip-embeddings for the video with a stride of 2 seconds. For paragraph-retrieval settings, we employ text augmentation during training by randomly sampling and concatenating a variable number of corresponding captions per video.

Finetuning time. A large motivation for using pre-extracted expert models for video retrieval is to save computational cost. Finetuning our 4-frame model for 50 epochs on MSR-VTT takes 10 hours on 2 Quadro RTX 6000k GPUs (with 24GB RAM each), which is similar to other works using *pre-extracted expert features* [41]. This shows that our model is lightweight and can be finetuned end-to-end on the downstream video datasets quickly with sufficient pretraining (which is of one-time cost).

4.4. Ablation Study

In this section we study the effect of different pretraining strategies. In the Supplementary Materials, we provide architectural ablations on different temporal expansion methods, different visual backbones, different text backbones and the improvement when using our modified space-time attention block.

Effect of pretraining. We compare performance on MSR-VTT with our model (i) trained from scratch, (ii) initialised with ImageNet weights and then finetuned, as well as (iii) initalised with ImageNet, and then pretrained on a number of different visual-text datasets before finetuning. For the video data, 4 frames are sampled at both pretraining and finetuning. Results on the MSR-VTT 1KA test set are shown in Table 2. For HowTo100M, we pretrain on a random 17M subset due to computational constraints (the largest subset we could obtain at the time of writing) totalling 19K hours. To generate text-video pairs, we sample 5 contiguous speech-video pairs and concatenate them to form a longer video. This allows for robustness to the noisy alignment of speech and vision. We find that training on CC3M alone does reasonably well, outperforming the HowTo-17M subset. This demonstrates the benefit of our flexible encoder that can be cheaply trained on images and easily applied to videos. Training on WebVid2M also outperforms training on the HowTo17M subset, despite being much smaller, confirming that the HowTo100M dataset is noisy. The best performance is achieved by jointly training on both CC3M and WebVid2M, effectively exploiting image and video data.

Table 2: **Pretraining sources:** The effect of different pretraining sources. We use 4 frames per video in both pretraining and finetuning. Pretraining is performed for 1 full epoch only. Results are presented on the 1K-A MSR-VTT test set for text-video retrieval. **R@k:** Recall@K. **MedR:** Median Rank

Pre-training	**#pairs**	**R@1**	**R@10**	**MedR**
-	-	5.6	22.3	55
ImageNet		15.2	54.4	9.0
HowTo-17M subset	17.1M	24.1	63.9	5.0
CC3M	3.0M	24.5	62.7	5.0
WebVid2M	2.5M	26.0	64.9	5.0
CC3M + WebVid2M	5.5M	**27.3**	**68.1**	**4.0**

4.5. Curriculum strategy

Next, we evaluate the ability of our curriculum schedule to gradually learn the temporal dimension of videos by increasing the input number of frames. Table 3 summarises the results. Here, we show performance when pretraining on WebVid2M and finetuning on MSR-VTT. We explore two types of expansion in time: at pretraining and at finetuning stages. First, we observe that a single frame is not sufficient to capture the video content (18.8 R@1). Performing the temporal expansion at pretraining stage is better than doing so at finetuning (26.0 vs 24.9 R@1 with 4 frames). Finally, we obtain similar performance (slightly better at R@5) at half the computational cost in GPU hours by employing a curriculum strategy at pretraining (26.6 R@1). For 8 frames, the curriculum is even more useful, as we start training on 1 frame and then move to 4 before finally moving to 8 frames. Here, we obtain similar or better performance than training on 8 frames from the start, with almost a third of the computational cost. This is to be expected, as

Table 3: **Effect of #frames and curriculum learning:** The effect of a different number of input frames at pretraining and finetuning. ⇒ indicates a within-dataset curriculum learning strategy. Results are presented on the 1K-A MSR-VTT test set for text-video retrieval. Pretraining here is done on WebVid2M only, with a total budget of one epoch through the entire dataset. **PTT:** total pretraining time in hours.

PT #frames	FT #frames	R@1	R@10	MedR	PTT (hrs)
1	1	18.8	56.6	7.0	16.2
1	4	24.9	67.1	5.0	16.2
4	4	26.0	64.9	5.0	45.6
1⇒4	4	26.6	65.5	5.0	22.1
8	8	25.4	67.3	4.0	98.0
1⇒4⇒8	8	27.4	67.3	4.0	36.0

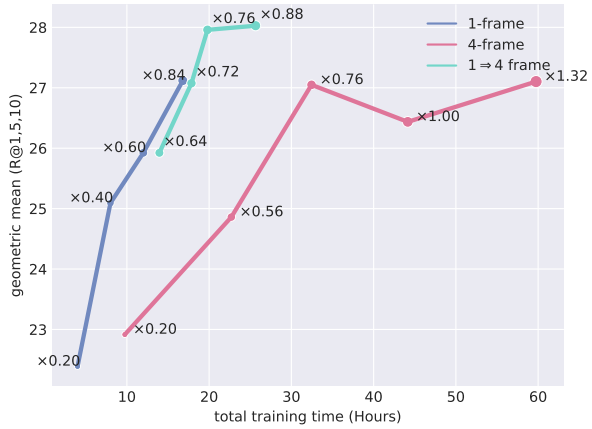


Figure 3: Plot showing the zero-shot performance (geometric mean of R@1,5,10) of various models on the MSR-VTT test set against their total training time in hours. ⇒ denotes a curriculum learning strategy. × denotes the multiple of dataset epochs completed.

fewer frames significantly reduces forward pass times and enables larger batch sizes. Note that for a fair comparison, we allow the same number of training iterations for each row in the table. We further analyse our proposed temporal curriculum strategy and its effects on training time and accuracy. Figure 3 shows the zero-shot results on MSR-VTT for various checkpoints with and without curriculum. It shows that our curriculum method yields a significant training speedup with a gain in accuracy. Shorter frame models are able to pass through more of the dataset in a shorter amount of time, which can lead to significant performance benefits in a constrained setting.

Expansion of temporal embeddings. We experiment with both zero padding and interpolation, and find that our model is robust to the type of temporal expansion strategy. More detailed results are provided in the Supplementary Materials.

4.6. Comparison to the State of the Art

Results on MSR-VTT can be seen in Table 4. We outperform all previous works, including many that pretrain on HowTo100M which is an order of magnitude larger than our pretraining dataset both in the number of hours (135K vs 13K) and in the number of caption-clip pairs (136M vs 5.5M). We also note that we outperform works that extract expert features (CE uses 9 experts, MMT uses 7) including object, motion, face, scene, sound and speech embeddings. We even outperform/perform on par with Support Set [41], which uses expert features from a 34-layer, R(2+1)-D model pretrained on IG65M, concatenated with ImageNet ResNet152 features, after which they add a transformer network and train end-to-end on HowTo100M.

We also report zero-shot results (Table 4) with no finetuning on MSR-VTT, outperforming both MIL-NCE and Support Set that trains on HowTo100M. This shows that our model is more generalisable, and can be used out of the box, and also perhaps that the domain of WebVid-2M is closer to that of MSR-VTT than HowTo100M. We will release the weights of our models publicly.

For both the zero-shot and finetuned setting we show that the addition of the COCO Captions image dataset further boosts our state-of-the-art MSR-VTT performance, indicating that the model is not yet saturated and additional pretraining dataset will lead to even better downstream performance.

For MSVD [9], we outperform all previous methods (Table 5). In particular, we outperform Support Set [41] even though they train on an order of magnitude more data.

Results on DiDeMo can be found in Table 6. Note that on this dataset, our zero-shot performance is equivalent to CLIPBERT's results with finetuning, and after we finetune our model on the DiDeMo training set we get an additional 14.2% boost in R@1.

In the Supplementary Materials, we demonstrate further state-of-the-art results on LSMDC text-to-video retrieval.

5. Conclusion

To conclude, we introduce a dual encoder model for end-to-end training of text-video retrieval, designed to take advantage of both large-scale image and video captioning datasets. Our model achieves state-of-the-art performance on a number of downstream benchmarks, however we note that the performance of our model is not saturated yet, and performance could be further improved by training on

Table 4: Comparison to state-of-the-art results on MSR-VTT for text-to-video retrieval, 1k-A split. †**E2E:** Works trained on pixels directly, without using pre-extracted expert features trained for other tasks. **Vis Enc. Init.:** Datasets used for pretraining visual encoders for tasks *other than visual-text retrieval*, eg object classification. **Visual-Text PT:** Visual-text pretraining data. Rows highlighted in blue use additional modalities such as sound and speech from the MSR-VTT test videos. † Object, Motion, Face, Scene, Speech, OCR and Sound classification features.

Method	E2E†	Vis Enc. Init.	Visual-Text PT	#pairs PT	R@1	R@5	R@10	MedR
JSFusion [67]	✓	-	-	-	10.2	31.2	43.2	13.0
HT MIL-NCE [37]	✓	-	HowTo100M	136M	14.9	40.2	52.8	9.0
ActBERT [72]	✓	VisGenome	HowTo100M	136M	16.3	42.8	56.9	10.0
HERO [28]	✓	ImageNet, Kinetics	HowTo100M	136M	16.8	43.4	57.7	-
VidTranslate [23]	✓	IG65M	HowTo100M	136M	14.7	-	52.8	
NoiseEst. [2]	✗	ImageNet, Kinetics	HowTo100M	136M	17.4	41.6	53.6	8.0
CE [31]	✗	Numerous experts†	-		20.9	48.8	62.4	6.0
UniVL [33]	✗	-	HowTo100M	136M	21.2	49.6	63.1	6.0
ClipBERT [26]	✓	-	COCO, VisGenome	5.6M	22.0	46.8	59.9	6.0
AVLnet [48]	✗	ImageNet, Kinetics	HowTo100M	136M	27.1	55.6	66.6	4.0
MMT [16]	✗	Numerous experts†	HowTo100M	136M	26.6	57.1	69.6	4.0
T2VLAD [62]	✗	Numerous experts†	-		29.5	59.0	70.1	4.0
Support Set [41]	✗	IG65M, ImageNet	-	-	27.4	56.3	67.7	3.0
Support Set [41]	✗	IG65M, ImageNet	HowTo100M	136M	30.1	58.5	69.3	**3.0**
Ours	✓	ImageNet	CC3M	3M	25.5	54.5	66.1	4.0
Ours	✓	ImageNet	CC3M,WV-2M	5.5M	**31.0**	**59.5**	**70.5**	**3.0**
Ours	✓	ImageNet	CC3M,WV-2M,COCO	6.1M	**32.5**	**61.5**	**71.2**	**3.0**
Zero-shot								
HT MIL-NCE [37]	✓	-	HowTo100M	136M	7.5	21.2	29.6	38.0
SupportSet [41]		IG65M, ImageNet	HowTo100M	136M	8.7	23.0	31.1	31.0
Ours	✓	ImageNet	CC3M,WV-2M	5.5M	**23.2**	**44.6**	**56.6**	**7.0**
Ours	✓	ImageNet	CC3M,WV-2M,COCO	6.1M	**24.7**	**46.9**	**57.2**	**7.0**

Table 5: Text-to-video retrieval results on the MSVD [9] test set.

Method	R@1	R@5	R@10	MedR
VSE [22]	12.3	30.1	42.3	14.0
VSE++ [15]	15.4	39.6	53.0	9.0
Multi. Cues [38]	20.3	47.8	61.1	6.0
CE [31]	19.8	49.0	63.8	6.0
Support Set [41]	23.0	52.8	65.8	5.0
Support Set [41] (HowTo PT)	28.4	60.0	72.9	4.0
Ours	**33.7**	**64.7**	**76.3**	**3.0**

Table 6: Text-to-video retrieval results on the DiDeMo test set. We show results with and without ground truth proposals (GT prop.) as well as with finetuning and without (zero-shot).

Method	GT prop.	R@1	R@5	R@10	MedR
S2VT [57]		11.9	33.6	-	13.0
FSE [69]		13.9	36.0	-	11.0
CE [31]		16.1	41.1	-	8.3
ClipBERT [26]	✓	20.4	44.5	56.7	7.0
Ours		**31.0**	**59.8**	**72.4**	**3.0**
Ours	✓	**34.6**	**65.0**	**74.7**	**3.0**
Zero-shot					
Ours		21.1	46.0	56.2	7.0
Ours	✓	20.2	46.4	58.5	7.0

the full HowTo100M dataset, larger weakly paired image datasets such as Google3BN [20], as well as multi-dataset combinations thereof.

Acknowledgements. The authors would like to thank Samuel Albanie for his useful feedback. We are grateful for funding from a Royal Society Research Professorship, EPSRC Programme Grant VisualAI EP/T028572/1, and a Google PhD Fellowship.

References

[1] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-

supervised multimodal versatile networks. In *NeurIPS*, 2020. 2, 3

[2] Elad Amrani, Rami Ben Ari, Daniel Rotman, and Alex Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. *arXiv preprint arXiv:2003.03186*, 2020. 8

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 2, 5

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

[5] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *ACCV*, 2020. 2, 5

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv:2102.05095*, 2021. 2, 3, 4

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3

[8] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. In *CVPR*, 2017. 2

[9] 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. 2, 5, 7, 8

[10] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. A2-nets: Double attention networks. *arXiv preprint arXiv:1810.11579*, 2018. 3

[11] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019. 3

[12] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*, 2019. 3

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

[14] 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 *ICLR*, 2021. 3, 4

[15] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 8

[16] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020. 2, 3, 8

[17] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *CVPR*, 2017. 3

[18] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In *CVPR*, 2018. 2

[19] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *CVPR*, 2018. 3

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021. 8

[21] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 2

[22] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 8

[23] Bruno Korbar, Fabio Petroni, Rohit Girdhar, and Lorenzo Torresani. Video understanding as machine translation. *arXiv preprint arXiv:2006.07203*, 2020. 8

[24] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 1, 2, 5

[25] 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*, 123(1):32–73, 2017. 2

[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. *arXiv preprint arXiv:2102.06183*, 2021. 2, 3, 5, 8

[27] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 1

[28] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *EMNLP*, 2020. 8

[29] 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 *ECCV*, 2014. 1, 2

[30] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. Hit: Hierarchical transformer with momentum contrast for video-text retrieval, 2021. 2

[31] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *Proc. BMVC*, 2019. 1, 2, 3, 5, 8

[32] Chenxu Luo and Alan Yuille. Grouped spatial-temporal aggretation for efficient action recognition. In *ICCV*, 2019. 4

[33] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. UniVL: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint*

arXiv:2002.06353, 2020. 8

[34] Philip M McCarthy and Scott Jarvis. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392, 2010. 5

[35] 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 *CVPR*, 2020. 2, 3

[36] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv*, 2018. 1, 2, 3

[37] 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 *ICCV*, 2019. 2, 5, 8

[38] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27, 2018. 8

[39] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, 2018. 3

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

[41] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, João Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020. 2, 5, 6, 7, 8

[42] 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, 2021. 2

[43] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019. 3

[44] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 5

[45] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *CVPR*, 2015. 5

[46] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123(1):94–120, 2017. 2, 5

[47] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012. 5

[48] Andrew Rouditchenko, Angie Boggust, David Harwath, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, et al. AVLnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*, 2020. 2, 8

[49] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. 6

[50] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. *CVPR*, 2021. 2

[51] Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. In *WACV*, 2021. 2

[52] 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. 2, 4, 5

[53] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 5

[54] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 3

[55] Du Tran, Heng Wang, L. Torresani, Jamie Ray, Y. LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 2

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1, 3

[57] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014. 8

[58] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2016. 1

[59] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016. 1

[60] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2740–2755, 2019. 4

[61] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 3

[62] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: Global-local sequence alignment for text-video retrieval, 2021. 8

[63] Chao-Yuan Wu, Ross B. Girshick, Kaiming He, Christoph

Feichtenhofer, and Philipp Krahenbuhl. A multigrid method for efficiently training video models. In *CVPR*, 2020. 3

[64] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018. 2

[65] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 2, 5

[66] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016. 1

[67] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, 2018. 8

[68] Andrew Zhai and Hao-Yu Wu. Classification is a strong baseline for deep metric learning. In *BMVC*, 2019. 3

[69] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *ECCV*, 2018. 8

[70] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 2, 5

[71] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, 2018. 2

[72] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, 2020. 8