X-Pool: Cross-Modal Language-Video Attention for Text-Video Retrieval

Satya Krishna Gorti[1*] Noël Vouitsis[1,2*] Junwei Ma[1*]
Keyvan Golestan[1] Maksims Volkovs[1] Animesh Garg[2,3,4] Guangwei Yu[1]

[1]Layer 6 AI [2]University of Toronto [3]Vector Institute [4]NVIDIA

Abstract

*In text-video retrieval, the objective is to learn a cross-modal similarity function between a text and a video that ranks relevant text-video pairs higher than irrelevant pairs. However, videos inherently express a much wider gamut of information than texts. Instead, texts often capture sub-regions of entire videos and are most semantically similar to certain frames within videos. Therefore, for a given text, a retrieval model should focus on the text's most semantically similar video sub-regions to make a more relevant comparison. Yet, most existing works aggregate entire videos without directly considering text. Common text-agnostic aggregations schemes include mean-pooling or self-attention over the frames, but these are likely to encode misleading visual information not described in the given text. To address this, we propose a cross-modal attention model called X-Pool that reasons between a text and the frames of a video. Our core mechanism is a scaled dot product attention for a text to attend to its most semantically similar frames. We then generate an aggregated video representation conditioned on the text's attention weights over the frames. We evaluate our method on three benchmark datasets of MSR-VTT, MSVD and LSMDC, achieving new state-of-the-art results by up to 12% in relative improvement in Recall@1. Our findings thereby highlight the importance of joint text-video reasoning to extract important visual cues according to text. Full code and demo can be found at: **layer6ai-labs.github.io/xpool/**.*

1. Introduction

The advent of video content platforms like TikTok, YouTube and Netflix have enabled the mass outreach of videos around the world. The ability to retrieve videos that are most semantically similar to a provided text-based search query allows us to quickly find relevant information and to make sense of massive amounts of video data.

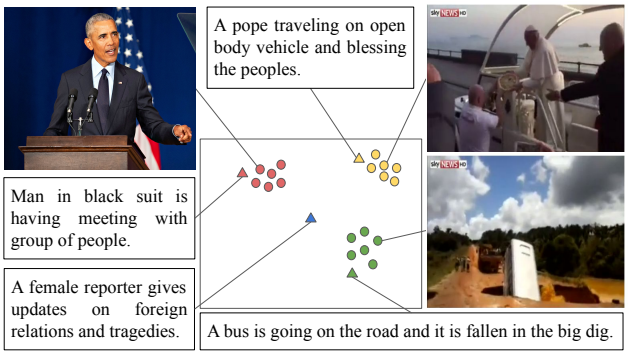


Figure 1. Illustration of the joint text and visual representations for a single video and its captions taken verbatim from the MSR-VTT dataset. Since the video is capturing more content than each individual text, aggregating the entire video regardless of the input text can be misleading.

The task of text-video retrieval is an approach to solve this problem wherein the objective is for a model to learn a similarity function between texts and videos. To compute the similarity between both modalities, a common technique is to first embed a text and a video into a joint latent space and then apply a distance metric such as the cosine similarity between the text and video embeddings [5,12,22].

However, there is an important discrepancy between both modalities that makes such a direct comparison challenging. Videos inherently express a much wider gamut of information than texts, so a text generally cannot fully capture the entire contents of a video. Instead, texts are most semantically similar to sub-regions of videos, represented as a subset of frames. Depending on the given text, the frames that are the most semantically similar would differ, so multiple equally valid texts can match a particular video. For example, in Figure 1, we show frames of a sample video from the MSR-VTT dataset [40]. The frames depict various scenes from international news and express different visual content. Moreover, we show multiple captions associated with this video, and observe that each caption best matches a different video frame but can seem irrelevant to others. In this example, we would expect the same video to be re-

*Authors contributed equally to this work.

trieved for any of these queries, even though the relevant content is limited to sub-regions of the video.

Based on this observation, we want a retrieval model to focus on the video sub-regions that are most relevant to the given text during retrieval. A model should therefore directly reason between texts and the frames of videos to extract the most relevant information as described in each text. However, most existing works do not apply direct cross-modal reasoning, and instead utilize the entire contents of a video such as through mean-pooling or self-attention [5, 12, 26, 30]. By encoding a video independently from a given text, a model is likely to encode superfluous or even distracting visual information that is not described in the text, which can reduce retrieval performance.

To address this gap, we design a cross-modal attention model that we call **X-Pool** to allow for joint reasoning between a text and a video's frames. Unlike previous works that pool the entire frames of a video, our model provides flexibility for a text to attend to its most semantically similar frames and then generates an aggregated video representation conditioned on those frames.

Our main contributions can be summarized as follows: (i) We show empirically through a proof of concept that text-conditioned video pooling allows a model to reason about the most relevant video frames to a given text, which outperforms baselines that use text-agnostic video pooling; (ii) We propose a cross-modal attention model that extends our proof of concept with parametric capacity for a text to attend to its most semantically similar video frames for aggregation which we call X-Pool. X-Pool obtains state-of-the-art results across the popular benchmark datasets of MSR-VTT [40], MSVD [8] and LSMDC [34]; (iii) We demonstrate the robustness of X-Pool to videos with increasing amounts of content diversity, such as videos with many scene transitions. We show how text-agnostic pooling methods are much more sensitive to such videos compared to our text–conditioned X-Pool model.

2. Related Work

Joint Language-Image Understanding. Joint language-image models are a form of multimodal learning [6] that aim to understand and relate the text and image modalities. Methods in text-image understanding such as [9, 13, 17–19, 25, 33, 37] are pre-trained to jointly reason about language and image semantics which make them suitable for downstream cross-modal tasks like visual question answering (VQA) [3], image captioning [41] and text-image retrieval [14]. Most recently, methods such as CLIP [33], ALIGN [13], DeCLIP [20] and ALBEF [17] employ unimodal encoders to learn a joint latent space that matches relevant text-image pairs via a contrastive loss. Our goal is to bootstrap from a pre-trained joint text-image model and extend it towards a joint text-video model for the task of text-video retrieval.

Text-Video Retrieval. The prototypical approach to text-video retrieval has been through a pre-trained language expert and often a combination of video experts pre-trained for various tasks and modalities, after which the language and vision streams are consolidated through late fusion. MoEE [29], CE [22], MMT [12] MDMMT [11], and TeachText [10] are all such works. The motivation for using pre-trained experts stems from the small-scale nature of the datasets used in text-video retrieval.

Some works have also benefited from pre-training their own models on either large-scale text-video datasets [5, 30, 46] or through text-image pre-training [15, 29]. Among them, ActBERT [46] and ClipBERT [15] are both single stream models that jointly embed text-video pairs through BERT-like architectures for early cross-modal fusion. However, these works do not allow for direct reasoning about the most semantically similar video sub-regions to a given text.

Recently, the works of CLIP4Clip [26] and Straight-CLIP [32] use the joint language-vision model of CLIP [33] pre-trained on a large-scale text-image dataset as a backbone. Even the trivial use of CLIP in a zero-shot manner outperforms most of the above recent works [32], highlighting how the rich joint text-image understanding of CLIP can be expanded towards videos. CLIP4Clip [26] proposes several video aggregation schemes including mean-pooling, self-attention and a multimodal transformer, yet none allow for direct matching of a text with its most relevant video sub-regions which motivates our cross-modal attention model. Cross-modal attention has been explored in previous related work such as [9, 15, 17–19, 25, 27, 37, 38, 42, 45, 46]. We design a cross-modal attention mechanism for the task of text-video retrieval that shows significant improvement over previous methods.

3. Problem Statement

In text-video retrieval, the objective is for a model to learn a scalar similarity function $s(t, v)$ between a text t and a video v. We want to assign higher similarity to relevant text-video pairs and assign lower similarity to irrelevant pairs. We define two retrieval tasks, text-to-video retrieval denoted as $t2v$ and video-to-text retrieval denoted as $v2t$. In $t2v$, we are given a query text t and a video index set $\mathcal{V}$. The goal is to rank all videos $v \in \mathcal{V}$ according to their similarities with the query text. Analogously, in $v2t$, we are given a query video v and a text index set $\mathcal{T}$. The goal is to rank all texts $t \in \mathcal{T}$ according to their similarities with the query video. In both of these tasks, we are under the assumption that only the index set is known ahead of time.

The inputs to our problem are a video v and a text t. We define a video $v \in \mathbb{R}^{F \times 3 \times H \times W}$ as a sequence of F sampled image frames in time. That is, $v = [v^1, v^2, \cdots, v^F]^T$ where v^f is the f^{th} image frame of resolution $H \times W$. We define

a text t as a sequence of tokenized words.

4. Methodology

In this section, we incrementally introduce the insights and methodologies that motivate our final model X-Pool. We first describe in Section 4.1 how the use of a pre-trained joint text-image model is an essential component of our model to match texts and images which we extend to match texts and videos. We then explain the drawbacks of aggregating a video into a text-agnostic embedding in Section 4.2, and present an alternative framework that aggregates frames conditioned on a given text in Section 4.3. We then introduce our X-Pool model in Section 4.4, a cross-modal attention model that enables joint reasoning between a text and the frames of a video. Our model learns to aggregate videos using the most semantically similar frames to a given text.

4.1. Expanding Joint Text-Image Models

Bootstrapping From Joint Text-Image Models. Jointly pre-trained text-image models have demonstrated the ability to match semantically similar texts and images [9, 13, 17, 20, 25, 33]. We can leverage the existing text-image reasoning of such models to bootstrap a joint text-video model. This allows us to learn language-video interactions with substantially less video data and offers a more compute efficient solution during training, while benefiting from the rich cross-modal understanding of pre-trained joint text-image models. In general, the idea of bootstrapping video models from image models stems from the importance of first understanding images in order to understand videos, as shown in [7].

CLIP as a Backbone. We bootstrap from CLIP [33] due to its strong downstream performance, its simplicity, and to more objectively compare with recent works that also leverage CLIP as a backbone [26, 32], although other pre-trained text-image models may be suitable backbone candidates. To bootstrap from CLIP for text-video retrieval, we first embed a text and individual video frames into its joint latent space and then pool the frame embeddings to obtain a video embedding [32]. Since the existing information extracted from a pre-trained CLIP model contains rich text-image semantics, we use CLIP as a backbone to learn a new joint latent space to match texts and videos instead of just images.

More precisely, given a text t a video frame v^f as input, CLIP outputs a text embedding $\mathbf{c}_t \in \mathbb{R}^D$ and a frame embedding $\mathbf{c}_v^f \in \mathbb{R}^D$ in a joint latent space:

$$\mathbf{c}_t = \psi(t) \tag{1}$$

$$\mathbf{c}_v^f = \phi(v^f) \tag{2}$$

where ψ is CLIP's text encoder and ϕ is CLIP's image encoder. By computing equation (2) for each frame in a video v, we obtain a sequence of frame embeddings $C_v = [\mathbf{c}_v^1, \mathbf{c}_v^2, \cdots, \mathbf{c}_v^F]^T \in \mathbb{R}^{F \times D}$.

Computing Text and Video Embeddings. As mentioned, we want to embed our given text and video into a joint space to compute similarity. That is, we want to compute a text embedding $\mathbf{z}_t \in \mathbb{R}^D$ and a video embedding $\mathbf{z}_v \in \mathbb{R}^D$. The text embedding is directly taken as the output from CLIP. On the other hand, we compute the video embedding by aggregating the frame embeddings in C_v using a temporal aggregation function ρ:

$$\mathbf{z}_t = \mathbf{c}_t \tag{3}$$

$$\mathbf{z}_v = \rho(C_v) \tag{4}$$

4.2. Gap: Text-Agnostic Pooling

In most existing works, the aggregation function ρ does not directly consider the input text and is purely a function of the frames of the videos such as through mean-pooling, self-attention or an LSTM [1, 5, 12, 26, 28, 30, 32].

While defining the temporal aggregation function as agnostic to text forms a simple baseline, there are important drawbacks with this approach. Videos are inherently much more expressive than texts, so the information captured in text generally cannot fully capture that of an entire video. Instead, texts are most semantically similar to certain sub-regions of videos which we define as subsets of frames, as shown in Figure 1. As such, common text-agnostic aggregation schemes that pool entire videos like mean-pooling and self-attention might encode spurious information that is not described in the input text.

We note that this effect is exacerbated when we consider videos that exhibit significant diversity in their visual content [21] which we refer to as content diversity. To elaborate, it is natural to find videos with scene transitions such as when the actor moves from an indoor setting to an outdoor setting, abrupt scene cuts like in movies, occlusions of key subjects or noise in the form of distractors for example. Since this is an intrinsic property of many videos "in the wild" [16, 35], we want a retrieval model to be robust to such content diversity by focusing its attention to the most relevant video sub-regions described in a given text. Intuitively, any text-agnostic pooling method will fail under this setting since it aggregates information from all scenes of the video, disregarding the input text for retrieval, as we empirically show in Section 5.3.

4.3. Key Insight: Text-Conditioned Pooling

We note that it is therefore important to match texts not with the entire contents of a video, but with those video frames that are most semantically similar to a given text. Depending on the given text, the frames that are most semantically similar would differ, so there could be multiple equally valid texts that match a particular video. As such,

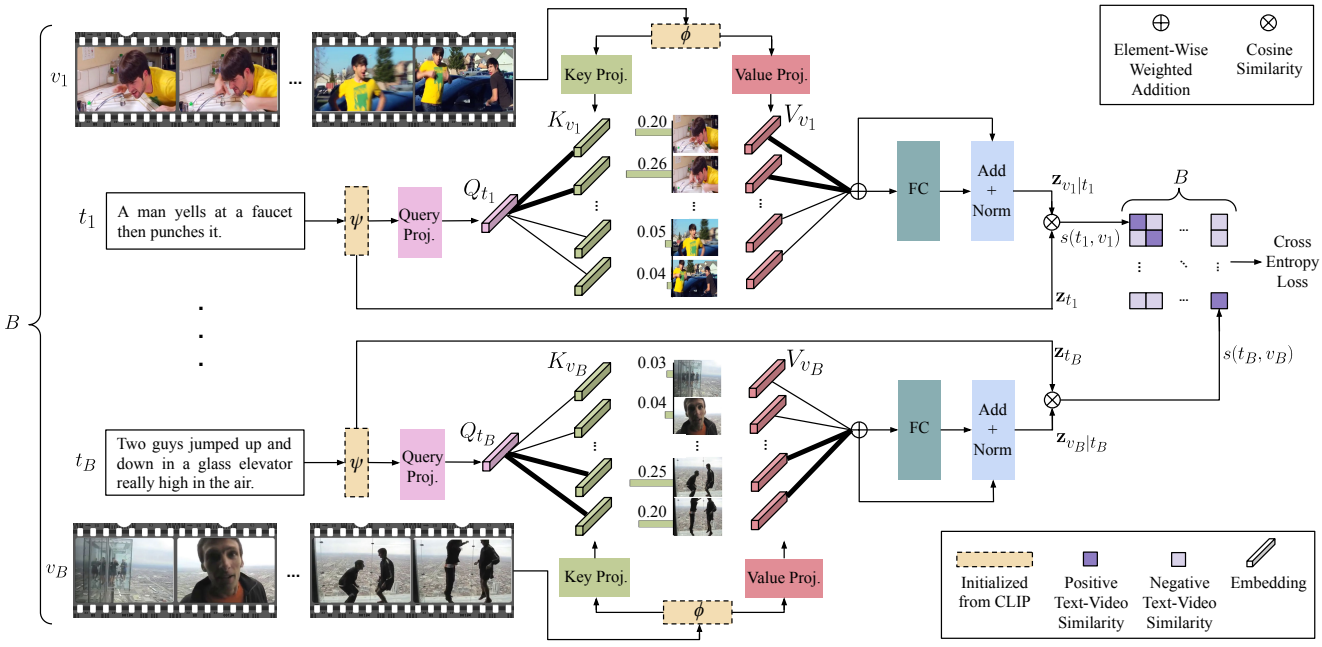


Figure 2. Diagram of X-Pool. For the given text t_1, we embed it with the text encoder ψ and then apply a query projection to obtain Q_{t_1}. We similarly embed the frames of the given video v_1 with the image encoder ϕ and then apply a key projection to obtain K_{v_1}. We compute the dot product attention between them as illustrated by the horizontal bar plot in the middle of the figure. Our attention mechanism allows X-Pool to focus on the most relevant frames given an input text. We aggregate a separate set of value-projected frame embeddings that we weight by the previously computed dot product attention scores to obtain an aggregated video embedding that we then pass through a fully connected layer (FC) with a residual connection to obtain $\mathbf{z}_{v_1|t_1}$. We compute the similarity score $s(t_1, v_1)$ as the cosine similarity between $\mathbf{z}_{v_1|t_1}$ and $\mathbf{z}_{t_1} = \psi(t_1)$. Finally, we compute a cross entropy loss after obtaining $s(t_i, v_j)$ as just described for each pair (t_i, v_j) within a batch of size B.

our temporal aggregation function should directly reason between a given text and the frames of a video.

To that end, we formulate a new temporal aggregation function π that allows us to aggregate the video frames that are most semantically similar to a given text t. By conditioning π on t, we can extract from a video v the most relevant information as described in t while suppressing noisy and misleading visual cues. We denote the resulting aggregated video embedding as $\mathbf{z}_{v|t}$ and define our similarity function $s(t, v)$ as:

$$\mathbf{z}_{v|t} = \pi(C_v \mid t) \tag{5}$$

$$s(t, v) = \frac{\mathbf{z}_t \cdot \mathbf{z}_{v|t}}{\|\mathbf{z}_t\| \|\mathbf{z}_{v|t}\|} \tag{6}$$

To demonstrate the efficacy of our idea, we first propose a top-k aggregation function $\pi_{\text{top-}k}(C_v \mid t)$ as:

$$\pi_{\text{top-}k}(C_v \mid t) = \frac{1}{k} \sum_{f \in \mathcal{K}} \mathbf{c}_v^f \tag{7}$$

where the set $\mathcal{K}$ is defined as:

$$\mathcal{K} = \underset{\substack{\mathcal{K} \subseteq \{1,\ldots,F\} \\ |\mathcal{K}|=k}}{\arg\max} \sum_{f \in \mathcal{K}} \frac{\mathbf{c}_t \cdot \mathbf{c}_v^f}{\|\mathbf{c}_t\| \|\mathbf{c}_v^f\|} \tag{8}$$

and the selected frames are those with the highest cosine similarity. Here, we directly select only the frames with the highest cosine similarity to a given text as a proxy for semantic similarity. Only the top-k most semantically similar frames to a given text are pooled while lower similarity frames are completely ignored.

We observe that even by just applying top-k pooling, there is already a significant improvement over baselines where the temporal aggregation function is text-agnostic. Detailed experiments can be found in Section 5.3.

4.4. Our Model: X-Pool

Towards Parametric Text-Conditioned Pooling. However, there are still drawbacks with the top-k method. Firstly, the tuning of the k hyperparameter can be task and instance specific as we show in Section 5.3. Secondly, deciding which frames to aggregate from can require more complex reasoning than simple cosine similarity. Lastly, completely suppressing frames with lower

similarity may be too restrictive. As such, we propose a parametric approach to address these additional considerations while incorporating our insights from applying text-conditioned pooling.

Cross-Modal Language-Video Attention. Our idea is to design a learned frame aggregation function with parametric capacity for cross-modal reasoning about a text's most semantically similar frames in a video, which we call X-Pool. The core mechanism is our adaptation of a scaled dot product attention [39] between a text and the frames of a video. Conditioned on these frames, we generate a video embedding that learns to capture the most semantically similar video sub-regions as described in a given text. Since the frames with highest semantic similarity can differ depending on the text, our scaled dot product attention mechanism can learn to highlight relevant frames to a given text while suppressing frames not described in said text. Our model's capacity to selectively pick frames based on relevance to a given text is motivated by the same text-conditioning insights as outlined in the previously described top-k approach. However, unlike the top-k approach, our proposed model learns the optimal amount of information to extract for a text-video pair, thereby removing the need to manually specify a k value. Furthermore, our cross-attention module handles both high and low relevancy frames rather than adopting a hard selection of relevant frames as in the top-k approach.

To elaborate, in our cross-modal attention module, we first project a text embedding $\mathbf{c}_t \in \mathbb{R}^D$ into a single query $Q_t \in \mathbb{R}^{1\times D_p}$ and a video's frame embeddings $C_v \in \mathbb{R}^{F\times D}$ into key $K_v \in \mathbb{R}^{F\times D_p}$ and value $V_v \in \mathbb{R}^{F\times D_p}$ matrices, where D is the size of our model's latent dimension and D_p is the size of the projection dimension. The projections are defined as:

$$Q_t = \text{LN}(\mathbf{c}_t^T)W_Q \tag{9}$$

$$K_v = \text{LN}(C_v)W_K \tag{10}$$

$$V_v = \text{LN}(C_v)W_V \tag{11}$$

where LN is a Layer Normalization layer [4] and W_Q, W_K and W_V are projection matrices in $\mathbb{R}^{D\times D_p}$. In order to learn flexible conditioning between the given text and the frames, we then adapt scaled dot product attention from the query-projected text embedding to the key-projected frame embeddings. The dot product attention gives relevancy weights from a text to each frame which we leverage to aggregate the value-projected frame embeddings:

$$\text{Attention}(Q_t, K_v, V_v) = \text{softmax}\left(\frac{Q_t K_v^T}{\sqrt{D_p}}\right)V_v \tag{12}$$

As such, the Q_t, K_v and V_v matrices can be interpreted akin to those in the original scaled dot product attention proposed in [39] except with cross-modal interactions. That is, the query-projected text embedding is used to seek from the key-projected frame embeddings to attend to frames with highest relevance. The value-projected embeddings represent the video's context from which we want to aggregate only certain sub-regions depending on the text.

To embed a video into a joint space with a text, we project the aggregated video representation from the attention module back into $\mathbb{R}^D$ by applying a weight $W_O \in \mathbb{R}^{D_p\times D}$ to obtain:

$$\mathbf{r}_{v|t} = \text{LN}(\text{Attention}(Q_t, K_v, V_v)W_O) \tag{13}$$

where the resulting output $\mathbf{r}_{t|v}$ is an aggregated video embedding conditioned on the text t. We can thereby learn this embedding such that a text can attend to its most semantically similar frames through parametric reasoning in the dot product attention. Our final text-conditioned pooling is defined as:

$$\pi_{\text{X-Pool}}(C_v \mid t) = \text{LN}(\text{FC}(\mathbf{r}_{v|t})) + \mathbf{r}_{v|t})^T \tag{14}$$

where FC is a fully connected network which together with the residual connection provides additional capacity for more complex reasoning in our aggregation function.

Figure 2 shows a diagram of our model. We show how X-Pool performs text-conditioned video aggregation over frames by allowing a text to learn to attend to its most semantically similar frames for pooling. In the top example, the input text t_1 is most relevant to the first few frames displayed of video v_1 of a man yelling at and punching a sink, whereas the final displayed frames of a man near a car do not capture what is described in the text and instead act as misleading visual distractors. We show how our model can reason about semantic similarity by assigning higher attention weights to the text's most relevant frames for aggregation. We emphasize that any text-agnostic pooling method such as mean-pooling would have aggregated the contents from this entire video. The resulting aggregation would thereby capture noisy distractors not described in the input text which could hamper the similarity score for retrieval. In the bottom example, we show a similar behaviour wherein X-Pool can attend to the most relevant frames of two guys jumping in an elevator as described in the text, whereas text-agnostic methods would capture non-relevant content from this video.

Loss. We train models using a dataset $\mathcal{D}$ consisting of N text and video pairs $\{(t_i, v_i)\}_{i=1}^N$. In each pair, the text t_i is a matching text description of the corresponding video v_i. We employ the cross entropy loss from [44] by considering matching text-video pairs as positives and by considering all other pairwise text-video combinations in the batch as negatives. Specifically, we jointly minimize the symmetric text-to-video and video-to-text losses:

$$\mathcal{L}_{t2v} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{s(t_i,v_i)\cdot\lambda}}{\sum_{j=1}^{B}e^{s(t_i,v_j)\cdot\lambda}} \tag{15}$$

$$\mathcal{L}_{v2t} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{s(t_i,v_i)\cdot\lambda}}{\sum_{j=1}^{B}e^{s(t_j,v_i)\cdot\lambda}} \tag{16}$$

$$\mathcal{L} = \mathcal{L}_{t2v} + \mathcal{L}_{v2t} \tag{17}$$

where $s(t_i, v_j)$ is the cosine similarity between the text t_i and the video v_j, B is the batch size and λ is a learnable scaling parameter. By bootstrapping from a pre-trained CLIP model and through our cross-modal attention mechanism, training with this loss enables our model to learn to match a text with its most semantically similar sub-regions of the ground-truth video.

5. Experiments

We perform experiments on the commonly used benchmark text-video retrieval datasets of MSR-VTT [40], MSVD [8] and LSMDC [34] and evaluate our performance following existing literature [5, 12, 22, 29, 43] by reporting Recall@1 (R@1), Recall@5 (R@5), Recall@10 (R@10), Median Rank (MdR), and Mean Rank (MnR).

5.1. Datasets

MSR-VTT is comprised of 10,000 videos, each paired with about 20 human-labeled captions. We note that the multiple captions for each video in MSR-VTT often describe different video sub-regions, which supports our motivation for matching a given text with its most relevant frames in a video. The lengths of videos in this dataset range from 10 to 32 seconds, and we use two training splits which we call *7k-Train* and *9k-Train* to effectively compare with previous works. *7k-Train* is a subset of roughly 7k videos as defined in [30], while *9k-Train* consists of approximately 9k videos following the split in [12]. Unless otherwise stated, we use the *9k-Train* split for training. To evaluate our models, we use the *1K-A* test set from [43] consisting of 1,000 selected caption-video pairs.

MSVD contains about 120k captions that each describe one of 1,970 videos ranging in length from 1 to 62 seconds. Again, videos are paired with multiple captions and each may describe different sub-regions of the same video. In MSVD, the training, validation and test splits are comprised of 1,200, 100 and 670 videos respectively. Our final results are evaluated on the test split that has a varying number of captions per video. To that end, we follow recent methods for evaluation by treating all the provided caption-video pairs as separate instances for evaluation [26, 32].

LSMDC is a movie clip dataset containing 118,081 videos each paired with a single caption description. The lengths of videos range from 2 to 30 seconds. 101,079 videos are used for training while 7,408 and 1,000 videos are used for validation and testing respectively. We report all results on the test set.

5.2. Implementation Details

We use CLIP's ViT-B/32 image encoder as ϕ and CLIP's transformer base text encoder as ψ, and initialize all encoder parameters from CLIP's pre-trained weights. We set the query, key and value projection dimension size as $D_p = 512$ to match CLIP's output dimension and initialize our logit scaling paramter λ with that from a pre-trained CLIP model. We apply a linear layer with $D = 512$ output units and dropout [36] of 0.3 as our FC. Finally, we initialize all new projection weight matrices with identity and all new biases with zeros to bootstrap our entire model from the existing text-image semantic reasoning of a pre-trained CLIP. Our models are fine-tuned end-to-end on each dataset. To that end, we set our batch size to 32 for all experiments and set the learning rate for CLIP-initialized weights to 1e-6 and for all other parameters to 1e-5. We optimize our model for 5 epochs using the AdamW optimizer [24] with weight decay set to 0.2 and decay the learning rate using a cosine schedule [23] following CLIP [33]. For all experiments, we uniformly sample 12 frames from every video and resize each frame to 224x224 following previous works [5, 22, 26].

5.3. Results

To evaluate our method, we compare its performance with recent works from the literature. We tabulate the $t2v$ retrieval performance of our model trained on the MSR-VTT *9k-Train* and *7k-Train* splits in Table 1 and Table 2 respectively. Tables 3 and 4 similarly compare the performance of X-Pool on the MSVD and LSMDC datasets respectively. We note that on all datasets and across all metrics, our text-conditioned X-Pool model outperforms all other works that use text-agnostic pooling [5, 26, 32] includ-

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
CE [22]	20.9	48.8	62.4	6.0	28.2
MMT [12]	26.6	57.1	69.6	4.0	24.0
Straight-CLIP [32]	31.2	53.7	64.2	4.0	-
Support Set [31]	30.1	58.5	69.3	3.0	-
MDMMT [11]	38.9	69.0	79.7	**2.0**	16.5
Frozen [5]	31.0	59.5	70.5	3.0	-
TeachText-CE+ [10]	29.6	61.6	74.2	3.0	-
CLIP4Clip-meanP [26]	43.1	70.4	80.8	**2.0**	16.2
CLIP4Clip-seqTransf [26]	44.5	71.4	81.6	**2.0**	15.3
X-Pool (ours)	**46.9**	**72.8**	**82.2**	**2.0**	**14.3**

Table 1. $t2v$ results on the MSR-VTT-9K dataset.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
HowTo100M [30]	14.9	40.2	52.8	9.0	-
ActBERT [46]	8.6	23.4	33.1	36.0	-
NoiseE [2]	17.4	41.6	53.6	8.0	-
ClipBERT [15]	22.0	46.8	59.9	6.0	-
CLIP4Clip-meanP [26]	42.1	71.9	81.4	**2.0**	15.7
CLIP4Clip-seqTransf [26]	42.0	68.6	78.7	**2.0**	16.2
X-Pool (ours)	**43.9**	**72.5**	**82.3**	**2.0**	**14.6**

Table 2. $t2v$ results on the MSR-VTT-7K dataset.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
CE [22]	19.8	49.0	63.8	6.0	23.1
Support Set [31]	28.4	60.0	72.9	4.0	-
NoiseE [2]	20.3	49.0	63.3	6.0	-
Straight-CLIP [32]	37.0	64.1	73.8	3.0	-
Frozen [5]	33.7	64.7	76.3	3.0	-
TeachText-CE+ [10]	25.4	56.9	71.3	4.0	-
CLIP4Clip-meanP [26]	46.2	76.1	84.6	**2.0**	10.0
CLIP4Clip-seqTransf [26]	45.2	75.5	84.3	**2.0**	10.3
X-Pool (ours)	**47.2**	**77.4**	**86.0**	**2.0**	**9.3**

Table 3. $t2v$ results on the MSVD dataset.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
CE [22]	11.2	26.9	34.8	25.3	-
MMT [12]	12.9	29.9	40.1	19.3	75.0
NoiseE [2]	6.4	19.8	28.4	39.0	-
Straight-CLIP [32]	11.3	22.7	29.2	56.5	-
MDMMT [11]	18.8	38.5	47.9	12.3	58.0
Frozen [5]	15.0	30.8	39.8	20.0	-
TeachText-CE+ [10]	17.2	36.5	46.3	13.7	-
CLIP4Clip-meanP [26]	20.7	38.9	47.2	13.0	65.3
CLIP4Clip-seqTransf [26]	22.6	41.0	49.1	11.0	61.0
X-Pool (ours)	**25.2**	**43.7**	**53.5**	**8.0**	**53.2**

Table 4. $t2v$ results on the LSMDC dataset.

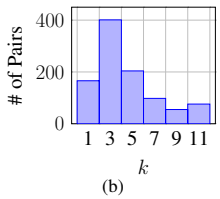
ing those using video experts in multiple video modalities [1,12,28]. Most notably, our model outperforms the hitherto state-of-the-art methods CLIP4Clip-meanP and CLIP4Clip-seqTransf [26] which are the most directly comparable to X-Pool since they also use CLIP as a backbone. Therefore, we can directly attribute the performance gains of our model to the fact that we use text-conditioned pooling compared to the text-agnostic pooling schemes of CLIP4Clip-meanP and CLIP4Clip-seqTransf.

More precisely, on the MSR-VTT dataset, we observe a relative improvement of 5% in Recall@1 compared to CLIP4Clip-seqTransf. For the MSVD dataset, we outperform CLIP4Clip-meanP by over 2% in relative improvement in Recall@1. In the case of the LSMDC dataset, the retrieval problem is more challenging since the movie scene text descriptions are much more ambiguous, which can be observed by the overall lower retrieval scores of all previous methods. Yet, our method notably outperforms CLIP4Clip-seqTransf by 12% in relative improvement in Recall@1. Our results thereby highlight the importance of our model's text-conditioned aggregation that can learn to match a text with its most relevant frames while suppressing distracting visual cues from other video sub-regions.

Top-k Experiments. To better understand the merits and intuition for our X-Pool model, we first revisit our top-k temporal aggregation function defined in equation (7) that we introduce as a proof of concept for our proposed idea of text-conditioned video pooling. To validate this idea, we compare top-k pooling with a mean-pooling baseline as in [26, 32] across two settings: first we apply a pre-trained CLIP model in a zero-shot manner similar to [32]

Aggr.	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
Zero-Shot CLIP					
Mean	31.5	52.8	63.6	5.0	42.9
Top-k	**33.6**	**54.0**	**64.3**	**4.0**	**42.5**
Fine-Tuned CLIP					
Mean	42.1	69.8	80.7	**2.0**	15.7
Top-k	**44.6**	**70.9**	**82.4**	**2.0**	**14.9**

(a)



(b)

Figure 3. Top-k analysis on MSR-VTT. (a) $t2v$ retrieval performance comparing mean-pooling with top-k text-conditioned pooling. (b) Histogram showing the k value where each ground truth text-video pair in the MSR-VTT test set achieves the highest cosine similarity when using top-k pooling.

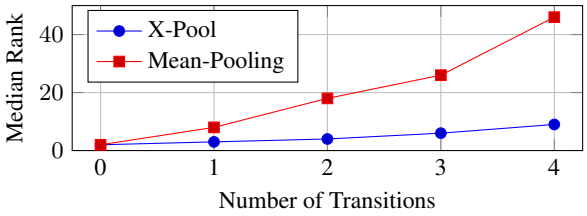


Figure 4. Robustness to content diversity. We show the $t2v$ Median Rank results on MSR-VTT for different amounts of content diversity measured by the number of scene transitions. Our X-Pool approach remains robust whereas mean-pooling significantly deteriorates as we increase the content diversity.

to compare mean-pooling and top-k aggregation, and second we fine-tune a pre-trained CLIP model on the MSR-VTT dataset and then measure retrieval performance for mean-pooling and top-k pooling. In both settings, we set k=3 which empirically yields the best overall performance. We compare the $t2v$ results in Table 3a and observe that even by using cosine similarity in top-k pooling as a proxy for semantic similarity between a text and frames, we can outperform mean-pooling across all listed metrics by up to 6% of relative improvement in Recall@1 through our text-conditioned pooling scheme.

Yet, the top-k aggregation function still presents some drawbacks as mentioned in Section 4.2, most notably relating to the tuning of the k hyperparameter. To analyze this shortcoming, we run an experiment wherein for a zero-shot pre-trained CLIP, we find the optimal k of each individual text-video pair in the MSR-VTT test set and report the results in a histogram in Figure 3b. Here, we define optimal as the the k value that yields the highest similarity score between a ground-truth text-video pair as defined in equation (6). We observe that the optimal choice of k varies widely between text-video pairs, which makes k difficult to select in general. Our proposed X-Pool model therefore addresses the drawbacks of top-k pooling while being motivated by our derived insights of text-conditioned pooling.

Robustness to Content Diversity in Videos. We now analyze the robustness of our model to content diversity as

we described in Section 4.2. As explained, many videos inherently exhibit diverse visual content such as scene transitions or changes in object appearance for example. While current datasets such as MSR-VTT, LSMDC and MSVD already display these traits to an extent, they are curated by choosing only small video clip segments extracted from larger videos. Therefore, in order to more effectively test the robustness of text-video retrieval methods to content diversity, one way is to introduce additional diversity in visual content with more scene transitions. That is, we augment a video's visual content by randomly injecting another video from the dataset to simulate an abrupt scene transition. By performing retrieval on such augmented videos and their original text captions, we can better evaluate a retrieval model's ability to handle diverse videos in the wild.

To that end, we construct augmented versions of the MSR-VTT test set by adding scene transitions from each video to other videos in the test set. The number of transitions is defined as the number of random videos that are added to the original video at a random location. We compare the $t2v$ retrieval performance of our X-Pool model to the baseline of mean-pooling, and plot the results in Figure 4. Here, we measure performance using the metric of Median Rank. We can clearly observe that as the number of video transitions increases and we add video content diversity, there is a sharp performance decline in mean-pooling as the Median Rank increases from 2 to 46, whereas our X-Pool model is significantly more robust to content diversity as Median rank only increases from 2 to 9. The performance gap is because any text-agnostic pooling method like mean-pooling aggregates content from all scenes of a video regardless of their relevance to an input text. Therefore, the more diverse a video is in terms of scene transitions, the more possibly noisy distractors are being aggregated. Conversely, X-Pool can extract only the most relevant visual cues as described in a text through text-conditioned pooling.

Qualitative Results. In Figure 4, we show qualitative examples of our X-Pool model. For each example, we show four sampled frames from a video along with a bar plot representing the associated attention weights of X-Pool from the given text to each frame. In the top example, we can see that our model outputs a higher attention weight for the middle frames when the input text describes a brain animation and lower attention weights everywhere else. On the other hand, when the input text instead describes a fictional character looking at a machine, the attention weight correspondingly activates for the last frame where the text is most relevant. The second example in the middle shows a singing competition. Here, the text of "a judge hearing the voice of competitors" describes an event that requires reasoning over all of the frames. Indeed, we observe that X-Pool attends to the entire video, indicating the flexibility of our approach.

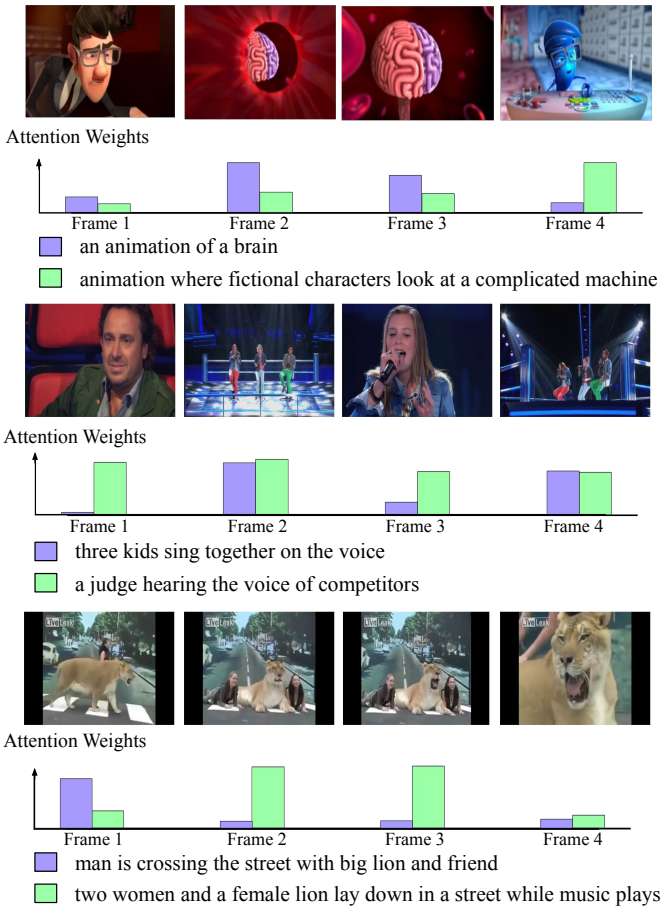


Figure 5. Qualitative results of X-Pool from the MSR-VTT dataset. For each displayed frame above, the bar plot shows its attention weights in our model given a particular text.

6. Conclusion

In this work, we highlight the drawbacks of text-agnostic video pooling and present an alternative framework for text-conditioned pooling for text-video retrieval. We then extend our idea and derived insights to design a parametric model for cross-modal attention between a text and video frames called X-Pool. We show how X-Pool can learn to attend to the most relevant frames to a given a text, which also makes our model substantially more robust to video content diversity such as in the form of scene transitions, a property that is common in videos in the wild. As part of future work, we plan on applying text-conditioned video pooling to other cross-modal tasks like video question answering.

7. Acknowledgments

Animesh Garg is supported by CIFAR AI Chair, NSERC Discovery Award, University of Toronto XSeed award, and gifts from LG.

References

[1] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 2(6):7, 2020. 3, 7

[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*, 8, 2020. 6, 7

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. *arXiv preprint arXiv:2104.00650*, 2021. 1, 2, 3, 6, 7

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 2

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

[8] 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, 6

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2, 3

[10] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11583–11593, 2021. 2, 6, 7

[11] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. Mdmmt: Multidomain multimodal transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3363, 2021. 2, 6, 7

[12] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 214–229. Springer, 2020. 1, 2, 3, 6, 7

[13] 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. 2, 3

[14] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2

[15] 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. 2, 6

[16] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2351–2359, 2019. 3

[17] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 3

[18] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[19] 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. 2

[20] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 2, 3

[21] Jingen Liu, Jiebo Luo, and Mubarak Shah. Recognizing realistic actions from videos "in the wild". In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1996–2003. IEEE, 2009. 3

[22] 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. 1, 2, 6, 7

[23] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[25] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 2, 3

[26] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 2, 3, 6, 7

[27] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2021. 2

[28] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 3, 7

[29] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018. 2, 6

[30] 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. 2, 3, 6

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

[32] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using clip. In *Mexican Conference on Pattern Recognition*, pages 3–12. Springer, 2021. 2, 3, 6, 7

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

[34] 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, 6

[35] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3

[36] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 6

[37] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[38] Reuben Tan, Bryan Plummer, Kate Saenko, Hailin Jin, and Bryan Russell. Look at what i'm doing: Self-supervised spatial grounding of narrations in instructional videos. *Advances in Neural Information Processing Systems*, 34, 2021. 2

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

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

[41] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 2

[42] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515, 2015. 2

[43] 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. 6

[44] Andrew Zhai and Hao-Yu Wu. Classification is a strong baseline for deep metric learning. *arXiv preprint arXiv:1811.12649*, 2018. 5

[45] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4486–4496, 2021. 2

[46] 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. 2, 6