CMMT: Cross-Modal Meta-Transformer for Video-Text Retrieval

Yizhao Gao
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
gaoyizhao@ruc.edu.cn

Zhiwu Lu
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
luzhiwu@ruc.edu.cn

ABSTRACT

Video-text retrieval has drawn great attention due to the prosperity of online video contents. Most existing methods extract the video embeddings by densely sampling abundant (generally dozens of) video clips, which acquires tremendous computational cost. To reduce the resource consumption, recent works propose to sparsely sample fewer clips from each raw video with a narrow time span. However, they still struggle to learn a reliable video representation with such locally sampled video clips, especially when testing on cross-dataset setting. In this work, to overcome this problem, we sparsely and globally (with wide time span) sample a handful of video clips from each raw video, which can be regarded as different samples of a pseudo video class (i.e., each raw video denotes a pseudo video class). From such viewpoint, we propose a novel Cross-Modal Meta-Transformer (CMMT) model that can be trained in a meta-learning paradigm. Concretely, in each training step, we conduct a cross-modal fine-grained classification task where the text queries are classified with pseudo video class prototypes (each has aggregated all sampled video clips per pseudo video class). Since each classification task is defined with different/new videos (by simulating the evaluation setting), this task-based meta-learning process enables our model to generalize well on new tasks and thus learn generalizable video/text representations. To further enhance the generalizability of our model, we induce a token-aware adaptive Transformer module to dynamically update our model (prototypes) for each individual text query. Extensive experiments on three benchmarks show that our model achieves new state-of-the-art results in cross-dataset video-text retrieval, demonstrating that it has more generalizability in video-text retrieval. Importantly, we find that our new meta-learning paradigm indeed brings improvements under both cross-dataset and in-dataset retrieval settings.

CCS CONCEPTS

• **Information systems** → **Video search**.

KEYWORDS

Video-text retrieval, meta-learning, representation learning

ICMR '23, June 12–15, 2023, Thessaloniki, Greece

ACM ISBN 979-8-4007-0178-8/23/06…$15.00
https://doi.org/10.1145/3591106.3592238

ACM Reference Format:
Yizhao Gao and Zhiwu Lu. 2023. CMMT: Cross-Modal Meta-Transformer for Video-Text Retrieval. In *International Conference on Multimedia Retrieval (ICMR '23), June 12–15, 2023, Thessaloniki, Greece.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3591106.3592238

1 INTRODUCTION

With the prosperity of online video platforms (e.g., YouTube and TikTok), video-text data is massively and rapidly generated. To make better use of the video-text data and satisfy the demands of users, video-text modelling has become increasingly popular. One of the fundamental tasks for video-text modelling is video-text retrieval [5, 15, 19, 44, 46], which requires the models to align the embeddings between two separate modalities. However, it is difficult to learn reliable/generalizable video representations for video-text retrieval, because each raw video generally has a long series of image frames. Traditional methods [18, 23, 29, 50, 53] typically utilize dense sampling strategies to learn video representations which require costly computation resource. It is thus of great challenge to find a new paradigm to learn an effective video-text retrieval model with limited resource consumption.

Recent representative work ClipBERT [21] has attempted to overcome this challenge by proposing a sparse sampling strategy. Specifically, it chooses to sparsely and locally sample a few video clips (each has a narrow time span) from each raw video, which are then aligned with the query text to obtain clip-level predictions. However, those predictions may be inaccurate for they only consider the local information of raw video. In our opinion, simultaneously utilizing several global video clips can help the model learn more reliable/generalizable video representations. Therefore, in this paper, we propose to sparsely and globally sample a handful of video clips from each raw video, where the frames of each video clip are sampled throughout the entire video. Specifically, we first divide the whole video into several video segments (with equal length). Then for each video clip, we randomly sample one frame per segment. In this way, three video clips are sampled from each video (see the left part of Figure 1), which are inputted into a video encoder (e.g., ViT-Base [8]) to obtain the video clip embeddings. Since these video clips are sampled from the same raw video, they can be naturally regarded as different samples of a pseudo video class (i.e., each raw video denotes a pseudo video class).

Motivated by the concept of pseudo video class, we thus propose a novel Cross-Modal Meta-Transformer (CMMT) model that can be trained in a meta-learning paradigm. Meta-learning [10, 39] has seen tremendous success in many vision/language classification tasks since it helps the model obtain more generalization ability by training across different episodes (tasks). For the first time, we seamlessly extend it to video-text retrieval by defining each training task as a cross-modal fine-grained classification task, where

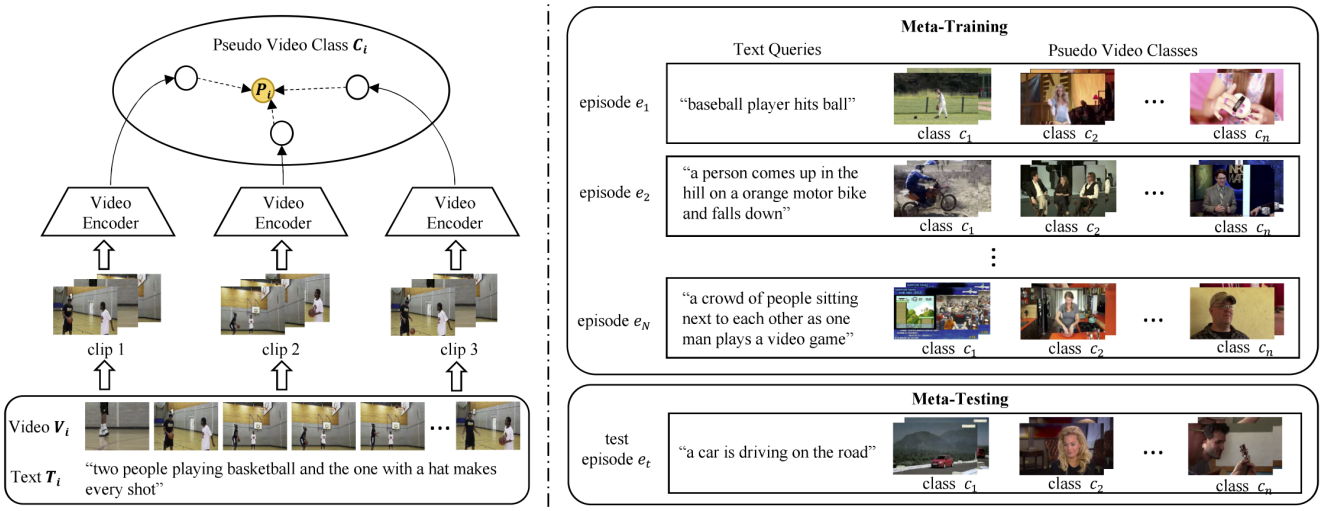

Figure 1: Schematic illustration of our meta-learning paradigm for video-text retrieval. In the left part, we show how to construct a pseudo video class. Concretely, for each raw video, we first randomly and globally sample three video clips and then regard them as samples of a pseudo video class (i.e., each raw video denotes a pseudo video class). In the right part, we show the meta-learning process of our paradigm. For each meta-training step, we randomly sample an episode (cross-modal classification task) to update our model, which is used to simulate the setting of the meta-testing task (test episode).

the query text embeddings (encoded by BERT-Base [7]) are classified according to pseudo video class labels. Concretely, for each pseudo video class, we first obtain all video clip embeddings and then average them as a video prototype. Each video prototype is a reliable/generalizable video representation since it has aggregated all video clip embeddings of the same pseudo video class. Further, we follow the widely-used meta-learning method ProtoNet [39] (designed for image classification), which predicts the class labels of the query samples with a non-parameter nearest-neighbor classifier (i.e., a set of prototypes). Similarly, for video-text retrieval, we predict the class labels of each query text with all video prototypes. As shown in the right part of Figure 1, for each meta-training step, we randomly sample N pseudo video classes and their corresponding texts to form an N-way episode (task), which mimics the test episode of the meta-testing phase. After training in such episode-based way, our proposed model can generalize well on new tasks, including those from the same current dataset (in-dataset) and those from different unseen datasets (cross-dataset).

Moreover, we induce a token-aware adaptive Transformer (TAAT) module to further enhance the generalizability of our model. This TAAT module is motivated by the fact that each video is often corresponding to multiple text descriptions with very different tokens. For instance, one video we have met during training is described by both 'a woman is talking' and 'the audience is clapping'. Therefore, we propose to adaptively adjust the classifier (i.e., the set of video prototypes) for each query text according to the text tokens. Specifically, for the attention layer of our TAAT, 'queries' are set to the video prototypes, 'keys' and 'values' are text token embeddings of each query text. In this way, the video prototypes become the combination of original prototypes and the specific text tokens. Overall, our model enhanced by TAAT achieves better performance in video-text retrieval, which has been validated in Table 9.

Our main contributions are three-fold: (1) We propose a novel Cross-Modal Meta-Transformer model termed CMMT for video-text retrieval based on the concept of pseudo video class. We train our model in a meta-learning paradigm with different training episodes, where each episode represents a randomly constructed cross-modal classification task. The cross-modal classification is thus performed with a non-parameter classifier based on video class prototypes, each of which aggregates all sparsely and globally sampled video clips per pseudo video class. (2) We induce a token-aware adaptive Transformer module to adaptively update the cross-modal classifier (video class prototypes) according to the text tokens of each individual query text. This adaption leads to the improved generalizability of our model. (3) We conduct extensive experiments on three benchmarks with two settings (cross-dataset and in-dataset) to demonstrate that our model achieves new state-of-the-art and has more generalizability in video-text retrieval.

2 RELATED WORK

2.1 Video-Text Retrieval

Video-text retrieval has been a popular but challenging task. A typical way [1, 11, 14, 25, 28, 34, 37, 48] to video representation learning (for video-text retrieval) is utilizing expert models to pre-extract the video features. These expert models are pre-trained on various tasks and multiple modalities, including face/scene/object recognition and action/sound classification. However, they suffer from the lack of cross-modal interaction since models use pre-extracted single modal features. Recent works [3, 12, 18, 21, 23, 27, 29, 45, 50, 53] have started to address this problem by training models directly from raw videos/texts (without using pre-extracted features). Among

them, [18, 23, 29, 50, 53] encode videos with the dense sampling strategy, which requires costly computation. Differently, ClipBERT [21] first proposes to sparsely sample video clips with short time span to obtain clip-level predictions. Frozen [3] uniformly samples one video clip and proposes a space-time Transformer to model the video frames. In this paper, we propose to sparsely and globally sample several video clips from each raw video to compose a pseudo video class. More importantly, inspired by the concept of pseudo video class, we are able to devise a novel cross-modal meta-Transformer termed CMMT for video-text retrieval.

2.2 Meta-Learning

Meta-learning has made remarkable progress for vision/language classification tasks. Recent meta-learning approaches can be categorized into four groups: (1) Metric-based methods [39, 40, 49] learn shared embedding space with distance metrics, including Cosine and Euclidean distances. (2) Optimization-based methods [24, 31, 36] aim to meta-learn new optimization algorithms, instead of using the classic gradient decent, to quickly adapt to unseen tasks. (3) Generation-based methods [13, 22, 52] meta-learn generators on base tasks and then apply the generators to meta-testing. (4) Model-based methods [10, 32] aim to learn a good model initialization on seen tasks in order to quickly fine-tune them on new tasks. In this work, our CMMT for video-text retrieval is trained with a metric-based meta-learning approach. That is, we classify the query texts with pseudo video class prototypes by computing their Euclidean metric distances, inspired by the popular metric-based meta-learning method ProtoNet[39]. Moreover, to adapt our model to different query texts, we propose to update the video prototypes by a token-aware adaptive Transformer. Extensive results show that our CMMT is effective for video-text retrieval.

2.3 Cross-Modal Transformer

Transformer is first introduced in [2] for machine translation. It has now achieved great success in both natural language processing [7, 41] and computer vision [8, 47]. A number of recent works deploy Transformers for video-text retrieval [3, 11, 25, 53]. ActBERT [53] learns joint video-text embeddings by leveraging both global and local clues from video-text pairs. MMT [11] proposes to learn a multi-modal Transformer which utilizes many pre-extracted features from multiple modalities. HiT [25] matches both feature-level and semantic-level features by a hierarchical Transformer. Frozen [3] proposes a new space-time Transformer to capture the correlation among video frames. Although a cross-modal Transformer model is also deployed in this work, we are the first to train it in a meta-learning paradigm, to the best of our knowledge. Particularly, on top of the visual and text Transformer backbones, we design a token-aware adaptive Transformer module to update the video prototypes for each individual query text according to the text tokens. The induced meta-learning process is shown to improve the generalizability of our model.

3 METHODOLOGY

In this section, we first outline the framework overview of our proposed model by describing the video-text retrieval task and our video/text encoders. We then introduce our proposed CMMT and its full loss for model training.

3.1 Framework Overview

3.1.1 Video-Text Retrieval. Given a dataset of $\mathcal{N}$ video-text pairs $\mathcal{D} = \{V_i, T_i\}_{i=1}^{\mathcal{N}}$, where V_i denotes a video with S_i frames and T_i denotes its paired text. The video-text retrieval task is to retrieve the closest video (or text) given a query text (or video). Therefore, its main goal is to learn a video encoder f_v and a text encoder f_t that can project each input video and its paired text into a joint embedding space where they are well aligned.

3.1.2 Video Encoder. We follow recent works [3, 21] to learn video representations based on sparsely-sampled video clips, but with different frame sampling strategy (see more details in following sections). For each raw video V_i with S_i frames, we randomly sample K video clips $V_{i,j}^v$ with $s < S_i$ frames per video clip, where $j = 1, 2, \cdots, K$. We first extract the visual embeddings $\widehat{F}_{i,j}^v \in \mathbb{R}^{s \times d_v}$ of all sampled frames through a pre-trained image encoder f_{img} (e.g., ViT-Base model [8]), with d_v being the output dimension:

$$\widehat{F}_{i,j}^v[r] = f_{img}(V_{i,j}^v[r]), r = 1, \cdots, s, \tag{1}$$

where $V_{i,j}^v[r]$ denotes the r-th frame of the video clip $V_{i,j}^v$ and $\widehat{F}_{i,j}^v[r]$ is the r-th row of the extracted visual embeddings $\widehat{F}_{i,j}^v$. Before we align the video and text embeddings, we need to make sure that the dimensions between video and text embeddings are equal. Thus, we adopt a linear projection layer:

$$\widetilde{F}_{i,j}^v[r] = \text{Linear}(\widehat{F}_{i,j}^v[r]), r = 1, \cdots, s, \tag{2}$$

where $\widetilde{F}_{i,j}^v[r] \in \mathbb{R}^d$ denotes the r-th projected visual embeddings of $\widetilde{F}_{i,j}^v$ with output dimension d. Linear$(\cdot)$ denotes a linear projection layer. Temporal correlation across the sampled frames is captured by a Transformer [42] module f_{att}, and the final video clip embedding $F_{i,j}^v$ is obtained by averaging the output embeddings:

$$F_{i,j}^v = \text{Avg}(f_{att}(\widetilde{F}_{i,j}^v[1], \widetilde{F}_{i,j}^v[2], \cdots, \widetilde{F}_{i,j}^v[s])), \tag{3}$$

where Avg$(\cdot)$ denotes the average pooling function. The video clip embedding $F_{i,j}^v$ is a d-dimensional vector, which has the same dimension as the text embedding. Overall, Eqs. (1)–(3) denote the process of encoding a video clip by our entire video encoder f_v.

3.1.3 Text Encoder. For each raw text T_i, we first tokenize it into a token list $[t_1^i, t_2^i, \cdots, t_{l_i}^i]$, where l_i denotes the length of the text T_i. We then project the token list into a sequence of text token embeddings through a pre-trained language model f_{lang} (e.g., BERT-Base model [7]):

$$\widehat{F}_i^t = f_{lang}(t_1^i, t_2^i, \cdots, t_{l_i}^i), \tag{4}$$

where $\widehat{F}_i^t \in \mathbb{R}^{l_i \times d_t}$, with d_t being the output dimension. Then we project all text token embeddings into the d-dimensional space (d is also the dimension of the video clip embedding $F_{i,j}^c$):

$$\widetilde{F}_i^t[r] = \text{Linear}(\widehat{F}_i^t[r]), r = 1, \cdots, l_i, \tag{5}$$

where Linear$(\cdot)$ is a linear projection layer with the output dimension d. We finally obtain the text embedding F_i^t by averaging all

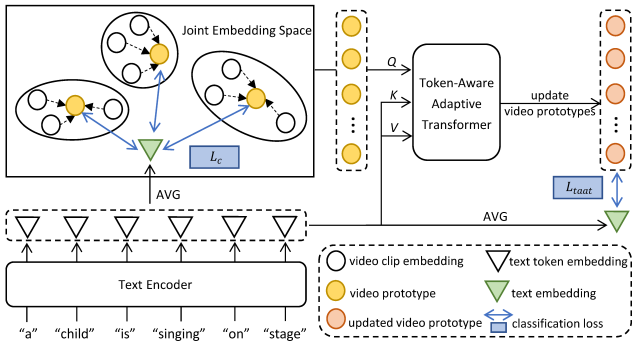

Figure 2: Schematic illustration of our CMMT model. Each video prototype (colored in yellow) is computed for a pseudo video class. The token-aware adaptive Transformer is an attention module to update the video prototypes for each individual query text. Based on both the original video prototypes and the updated video prototypes, we devise two losses (i.e., L_{cls} and L_{taat}) for video-text retrieval.

text token embeddings:

$$F_i^t = \text{Avg}(\widetilde{F}_i^t[1], \widetilde{F}_i^t[2], \cdots, \widetilde{F}_i^t[l_i]), \tag{6}$$

where $\widetilde{F}_i^t[r] \in \mathbb{R}^d$ denotes the r-th text token embedding of $\widetilde{F}_i^t$. Overall, Eqs. (4)–(6) denote the process of encoding a raw text by our entire text encoder f_t.

3.2 Cross-Modal Meta-Transformer

3.2.1 Pseudo Video Class. To effectively and efficiently learn video representations, we sample video clips in a sparse and global way. Specifically, for each raw video V_i with S_i frames, we uniformly divide it into s equal fragments and randomly sample one frame from each fragment to compose a video clip. Since video clips sampled from the same video have similar semantic contents, they can be regarded as different samples of the same class. In other words, they are similar/positive samples w.r.t. each other and compose a pseudo video class together. With the concept of pseudo video class, we thus propose a new meta-learning paradigm for video-text retrieval. Concretely, we are given a training set $\mathcal{D}_s$ and a test set $\mathcal{D}_u$, where $\mathcal{D}_s \cap \mathcal{D}_t = \emptyset$. Our model is trained on a set of seen pseudo video classes C_s from D_s and evaluated on a set of unseen pseudo video classes C_u from D_u. Following the standard meta-learning process, our CMMT aims to bridge the gap between C_s and C_u by simulating the test setting while training.

For each training step, we thus define a N-way K-shot Q-query fine-grained classification episode (task) e, where N is the number of classes (raw videos), K is the number of support samples (video clips) per class, and Q denotes the number of query samples (texts) per class. Formally, we randomly sample N classes to form a support set $\mathcal{S}_e = \{(V_{i,j}^v, y_{i,j}) | y_{i,j} \in C_e, i = 1, \cdots, N, j = 1, \cdots, K\}$ and a query set $\mathcal{Q}_e = \{(T_i, y_i) | y_i \in C_e, i = 1, \cdots, N \times Q\}$, where C_e is a subset of C_s ($|C_e| = N$). Note that all the query texts corresponding to a video should be assigned with the same pseudo class label.

3.2.2 Meta-Learning with Video Prototypes. The schematic illustration of our CMMT is presented in Figure 2. We instantiate our model based on ProtoNet [39], a widely-used meta-learning method for image classification. We seamlessly transfer it to video-text retrieval with pseudo video classes. Concretely, for each pseudo video class $c \in C_e$, we randomly sample K video clips and separately encode them to video clip embeddings through the video encoder f_v. We then define the pseudo video prototype $\mathcal{P}_c$ as the final video embedding which aggregates all video clip embeddings to represent the pseudo video class c for the classification task:

$$\mathcal{P}_c = \frac{1}{K} \sum_{(V_{i,j}^v, y_{i,j}) \in \mathcal{S}_e} f_v(V_{i,j}^v) \cdot I(y_{i,j} = c), \tag{7}$$

where $y_{i,j}$ denotes the class label of $V_{i,j}^v$, $I(\cdot)$ denotes an indicator function, and the pseudo video class $c \in C_e$. Note that $f_v(\cdot)$ is the video clip encoding process defined by Eqs. (1)–(3).

For each query text T_i, we encode its tokens through the entire text encoder f_t to gain the text token embeddings and average them to obtain the text embedding. Then we compute the metric distance (e.g., Euclidean distance) between each query and the video prototypes to construct a cross-entropy loss for meta-learning. Formally, the cross-modal classification loss is defined as:

$$L_{cls} = \frac{1}{NQ} \sum_{(T_i, y_i) \in \mathcal{Q}_e} -\log \frac{\exp(-d(f_t(T_i), \mathcal{P}_{y_i}))}{\sum_{c \in C_e} \exp(-d(f_t(T_i), \mathcal{P}_c))}, \tag{8}$$

where $d(\cdot)$ is a distance function (Euclidean distance is used in this work). $f_t(\cdot)$ is the text encoding process defined in Eqs. (4)–(6).

3.2.3 Token-Aware Adaptive Transformer. In the video-text datasets, each video is typically associated with many texts (e.g., 20 text descriptions per video). These texts often describe the video from different angles/viewpoint, resulting in that different texts have very different text tokens. However, the classic prototype-based loss L_{cls} applies the same set of video prototypes to every (query) text, which may bring harm on the performance of our model. To tackle this problem, we devise a token-aware adaptive Transformer (TAAT) module to adjust the video prototypes for each individual query text based on its text token embeddings.

Concretely, for each query text T_i, we first obtain all the text token embeddings $\widetilde{F}_i^t$ by Eqs. (4)–(5). We then adopt a Transformer which takes the triplet $(\mathcal{P}_{all}, \widetilde{F}_i^t, \widetilde{F}_i^t)$ as input (i.e., as the queries, keys, and values, respectively), where $\mathcal{P}_{all}$ denotes all video prototypes in an episode. As a result, $\mathcal{P}_{all}$ is updated by:

$$\widehat{\mathcal{P}}_{all} = \mathcal{P}_{all} + \text{softmax}(\frac{\mathcal{P}_{all} W_Q (\widetilde{F}_i^t W_K)^T}{\sqrt{d}}) \widetilde{F}_i^t W_V, \tag{9}$$

where W_Q, W_K, and W_V denote the parameters of the fully-connected layers in the Transformer. The text token embeddings are used to define both keys and values for the Transformer module, but the video prototypes are used to define the queries, since we expect it to be updated by the token embeddings of each query text. With the obtained token-aware class prototypes, we can now define the adaptive classification loss by:

$$L_{taat} = \frac{1}{NQ} \sum_{(T_i, y_i) \in \mathcal{Q}_e} -\log \frac{\exp(-d(f_t(T_i), \widehat{\mathcal{P}}_{y_i}))}{\sum_{c \in C_e} \exp(-d(f_t(T_i), \widehat{\mathcal{P}}_c))}. \tag{10}$$

Algorithm 1 CMMT for video-text retrieval

Input: Video encoder f_v (with parameters θ_v)
Text encoder f_t (with parameters θ_t)
A dataset $\mathcal{D}$ of video-text pairs
The hyper-parameters λ
Output: The learned f_v^* and f_t^*

for all iteration = 1, 2, $\cdots$, MaxIteration **do**
  Sample an N-way K-shot Q-query episode e from $\mathcal{D}_s$;
  Obtain video clip embeddings $F_{i,j}^v$ with Eq. (3);
  Obtain text embeddings F_i^t with Eq. (6);
  Obtain pseudo video class prototypes $\mathcal{P}_c$ with Eq. (7);
  Obtain updated video prototypes $\widetilde{\mathcal{P}}_c$ with Eq. (9);
  Compute L_{cls} with Eq. (8);
  Compute L_{taat} with Eq. (10);
  Compute L_{total} with Eq. (11);
  Compute the gradients $\nabla_{f_v} L_{total}$ and $\nabla_{f_t} L_{total}$;
  Update f_v and f_t using Adam;
end for
return the found best f_v^* and f_t^*.

3.3 Full Loss for Model Training

The token-aware adaptive Transformer module updates the video prototypes by computing the attention map between the original video prototypes and text token embeddings, which requires video and text embeddings are reliably aligned in the joint embedding space. Therefore, we need to train our CMMT with both L_{cls} and L_{taat} (experimental evidence is presented in Table 9):

$$L_{total} = L_{cls} + \lambda * L_{taat}, \tag{11}$$

where λ is the weight hyper-parameter. The full algorithm for training our CMMT is presented in Algorithm 1.

4 EXPERIMENTS

4.1 Experimental Setup

In this section, we will introduce the datasets used in our work, the metrics to evaluate our model and the pre-training details. More implementation details can be found in our supplementary material.

4.1.1 Datasets. We evaluate our SST-VLM model on three benchmarks (including cross-dataset and in-dataset settings). (1) **MSR-VTT** [46] contains 10k videos with 200k paired texts. Following recent works [3, 11, 26], we use the 1k-A split with 9k training videos and 1k test videos. (2) **DiDeMo** [15] has 10k Flickr videos annotated with 40k texts. We follow [3, 21] to evaluate our model on paragraph-to-video retrieval, where all texts for each video are concatenated into one query paragraph. (3) **MSVD** [5] consists of 1,970 videos from YouTube with 80k English texts, where each video has about 40 corresponding texts. Following [3, 26, 34], we use the standard split: 1,200 videos for training, 100 videos for validation, and 670 ones for testing. (4) **ActivityNet** [19] has 20k videos collected from YouTube annotated with 100K texts. We follow [3, 14, 21], using 10K training videos and 4.9K test videos (the val1 split). All texts are concatenated into one query paragraph.

Table 1: Cross-dataset results for text-to-video retrieval on MSVD. Models are all trained on the MSR-VTT training set and then directly evaluated on the MSVD test set.

Method	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
VSE++ [9]	13.8	34.6	46.1	13.0
Dual [30]	12.7	32.0	43.8	15.0
HGR [26]	16.4	38.3	49.8	11.0
HCGC [16]	17.4	39.6	52.9	9.0
CMMT (ours)	**35.1**	**64.4**	**75.9**	**3.0**

4.1.2 Performance Evaluation. We evaluate the video-text retrieval performance with the widely-used evaluation metrics in information retrieval, including Recall at K (shortened as R@K, K=1, 5, 10) and Median Rank (shortened as MedR). R@K refers to the percentage of queries that are correctly retrieved in the top-K most related candidates, where higher score indicates better performance. MedR computes the median rank of correct answers in the retrieved ranking list, where lower score indicates better performance.

4.1.3 Model Pre-Training. Note that our CMMT is also applicable for image-text retrieval when we remove the temporal Transformer module from the video encoder. Concretely, for each training step, we sample N image-text pairs which can be regarded as a N-way 1-shot 1-query classification task. Therefore, similar to ClipBERT [21] and Frozen [3], our CMMT is pre-trained on a pure image-text dataset. In this paper, our pre-training dataset has 5.3M image-text pairs, which consists of CC3M [38], SBU [33], Flickr30k [35], VisGenome [20], and COCO [6]. Since our computation resource is very restricted, we do not pre-train our CMMT on the large-scale video-text dataset HowTo100M [29]. Although we only pre-train our model with a pure image-text dataset that has the smallest number of visual-text pairs, our CMMT still achieves new state-of-the-art on three benchmarks.

4.1.4 Implementation Details. Our CMMT adopts ViT-Base [8] as the pre-trained image encoder (the basis of video encoder) and BERT-Base [7] as the pre-trained language encoder (the basis of the text encoder). During the training stage, all frames are resized to 384×384, and augmented by random-crop, horizontal-flip, gray-scaling, and color-jitter. The last eight layers of our image and language encoders are set to be learnable and other layers are frozen during model training, due to limited computation resource. We train our model for 1,500 iterations, and the total training time is around 1.5 hours with 8 Tesla V100 GPUs. We empirically set the hyper-parameters as: $\lambda = 0.5$, and the initial learning rate $= 5e-5$ (reduced to $5e-6$ after 750 iterations). Unless otherwise specified, for each training step, we sample a 48-way 3-shot 10-query classification task.

4.2 Cross-Dataset Text-to-Video Retrieval

Table 1 shows the text-to-video retrieval results of our CMMT under the cross-dataset retrieval setting. Following recent works [9, 16, 17, 26, 30], we train our model on the MSR-VTT training set with the full split (which consists of 6.5k videos) and then directly evaluate it on the MSVD test set (without fine-tuning on

Table 2: Cross-dataset results for text-to-video retrieval on DiDeMo. We train our models (different variants) on the MSR-VTT 1k-A training set and then directly evaluate them on the DiDeMo test set. ‡ denotes that the model is fine-tuned on the DiDeMo training set.

Method	Frames	R@1 ↑	R@5 ↑	R@10 ↑
ClipBERT [21]‡	16	20.4	48.0	60.8
Frozen [3]‡	32	34.6	65.0	74.7
3-shot (w/ TAAT, ours)‡	3×3	36.9	66.3	75.7
1-shot (w/o TAAT, ours)	1×9	24.1	51.4	64.2
1-shot (w/ TAAT, ours)	1×9	25.5	52.2	63.2
1-shot (w/o TAAT, ours)	1×3	18.3	38.8	49.5
1-shot (w/ TAAT, ours)	1×3	19.9	40.2	50.6
2-shot (w/o TAAT, ours)	2×3	23.6	50.4	61.1
2-shot (w/ TAAT, ours)	2×3	25.4	52.1	62.4
3-shot (w/o TAAT, ours)	3×3	24.3	50.9	63.9
3-shot (w/ TAAT, ours)	3×3	**26.3**	**52.4**	**64.8**

the training set). We can see that our CMMT outperforms all other methods with a large margin, indicating that our meta-learning paradigm indeed *enhances the generalization ability* of our model for video-text retrieval.

Furthermore, we also provide the generalization results on the DiDeMo test set obtained by our CMMT in Table 2, where our CMMT is trained on the MSR-VTT training set. Since other works [3, 21] do not release their fine-tuned model on MSR-VTT, we only list their results for models *fine-tuned* on the DiDeMo training set. We notice that our zero-shot results on DiDeMo even beat the fine-tuning results of [21] (see the last row vs. the 1st row). In addition, we conduct extensive experiments with different variants of our CMMT (4th row to the last row) to further demonstrate the effectiveness of our new meta-learning paradigm (including the TAAT module) under the cross-dataset retrieval setting.

4.3 Visualization Results

To directly show that our CMMT model has learned to align the video and text embeddings and also has great generalization ability, we present the attention maps of the input frames (given text descriptions) and the video-text retrieval examples on the MSVD test set under both cross-dataset and in-dataset retrieval settings.

4.3.1 Attention Visualization. We adopt a recent Transformer visualization method [4] to visualize the Transformer-based video encoder of our CMMT model. It can highlight the relevant regions of input image frames by computing the gradients of training losses with the input text. This enables us to find out which part of the input image frame is more relevant to the input text according to our CMMT's understanding. Specifically, in Figure 3, we present a video-text pair sampled from the MSVD test set and the attention visualization for our CMMT under both cross-dataset (trained on MSR-VTT and then directly evaluated on MSVD) and fine-tuning/in-dataset (trained on MSR-VTT and then fine-tuned on MSVD) settings. In the 1st row, it is a 4-frame video clip with the text description 'A gymnast is doing back flips then falls'. The 2nd and 3rd rows present the attention maps of our CMMT, where they have correctly

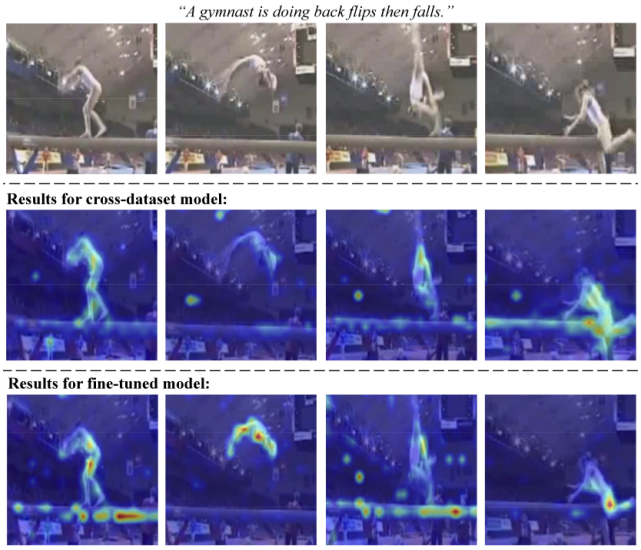

Figure 3: Attention visualization for our CMMT model. The heatmaps are shown for a video-text pair sampled from the MSVD test set. We present the visualization results for our CMMT model under both cross-dataset (the 2nd row) and fine-tuning (the 3rd row) settings.

captured the 'gymnast' of all frames. Note that our cross-dataset model can obtain comparable/similar results against our fine-tuned model. Therefore, these visualization results demonstrate that our CMMT is able to generalize well on new datasets and has learned to understand the semantic content in videos.

4.3.2 Retrieval Examples. Figure 4 presents a text-to-video retrieval example sampled from the MSVD test set obtained by our CMMT model under both cross-dataset and fine-tuning/in-dataset settings. We list top-3 retrieved videos (with the same query text "A person is writing with a pencil") under both settings. It can be clearly seen that: (1) Given the query text, the ground-truth video is correctly retrieved at the 1st place by both the cross-dataset and fine-tuning models. (2) The 2nd and 3rd retrieved videos have similar semantic contents that are (partially) related to the query text under both settings. Concretely, in the 2nd and 3rd retrieved videos, we can see key contents of the query text including "a person" and "a pencil". These observations show that our CMMT has learned to align the video and text embeddings for video-text retrieval and indeed has great generalization ability (even without fine-tuning).

4.4 Further Evaluation

4.4.1 In-Dataset Text-to-Video Retrieval. Since most of recent works [3, 21, 28, 48] focus on in-dataset text-to-video retrieval (i.e., fine-tuning setting), we follow their settings to evaluate our model. Table 3 summarizes the comparative results for in-dataset text-to-video retrieval on the DiDeMo benchmark dataset. We compare our CMMT model with recent representative/latest methods. Although our CMMT model is pre-trained on the smallest dataset which has

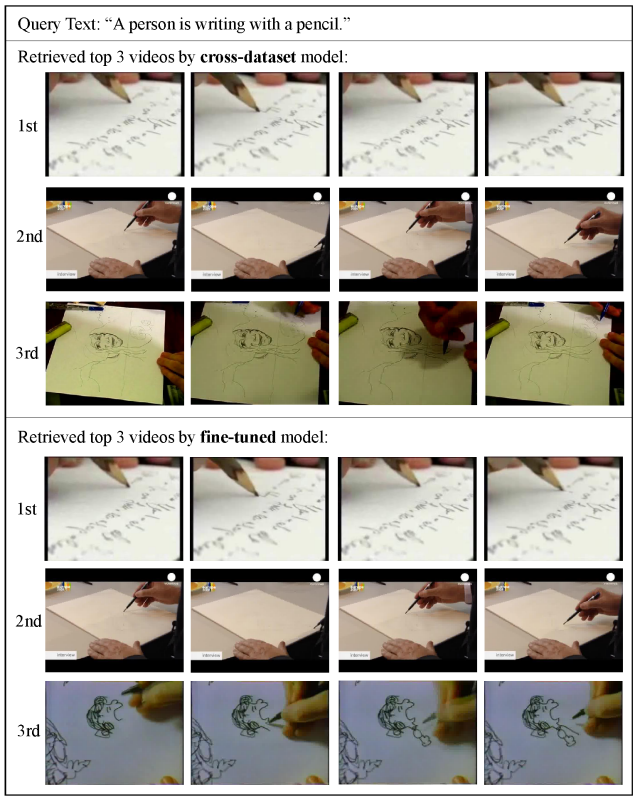

Figure 4: Text-to-video retrieval examples for our CMMT model. All examples are sampled from the MSVD test set. We present the retrieval examples under both cross-dataset (upper part) and fine-tuning (lower part) settings.

Table 3: Comparison to the state-of-the-art results for text-to-video retrieval on the DiDeMo test set. * denotes that localization annotations are used.

Method	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
S2VT [43]	11.9	33.6	-	13.0
FSE [51]	13.9	36.0	-	11.0
CE [26]	16.1	41.1	-	8.3
ClipBERT [21]*	20.4	48.0	60.8	6.0
Frozen [3]*	34.6	65.0	74.7	3.0
BridgeFormer [12]	37.0	62.2	73.9	3.0
CMMT (ours)	**37.3**	**66.3**	**75.7**	**2.0**

only 5.3M image-text pairs in total, it still achieves the best performance. This suggests that our CMMT has great potential even with limited pre-training data. Concretely, we observe that: our CMMT outperforms the second-best results by 0.3% on R@1, 1.3% on R@5, 1.0% on R@10, and 1.0 on MedR (2.0 vs. 3.0). As compared with ClipBERT [21] (the current best method pre-trained on image-text datasets), our model leads to even larger margins.

Table 4 presents the comparative results for text-to-video retrieval on MSR-VTT. On both 7k and 1k-A splits, we compare our

Table 4: Comparison to the state-of-the-art results for text-to-video retrieval on the MSR-VTT test set. The upper and bottom blocks show the results on 7k split and 1k-A split, respectively. Notations: * denotes that extra modalities (e.g., motion and audio) are used; # PT Pairs: the number of vision-text pairs in the pre-training cross-modal datasets.

Model	# PT Pairs	R@1↑	R@5↑	R@10↑	MedR↓
JSFusion [50]	-	10.2	31.2	43.2	13.0
HT MIL-NCE [29]	>100M	14.9	40.2	52.8	9.0
ActBERT [53]	>100M	16.3	42.8	56.9	10.0
HERO [23]	>100M	16.8	43.4	57.7	-
VidTranslate [18]	>100M	14.7	-	52.8	-
NoiseEstimation* [1]	>100M	17.4	41.6	53.6	8.0
UniVL* [28]	>100M	21.2	49.6	63.1	6.0
ClipBERT [21]	5.6M	22.0	46.8	59.9	6.0
TACo* [48]	>100M	24.8	52.1	64.5	5.0
CMMT (ours)	5.3M	**31.4**	**58.6**	**71.4**	**4.0**
1k-A split:					
CE [26]	-	20.9	48.8	62.4	6.0
AVLnet* [37]	>100M	27.1	55.6	66.6	4.0
MMT* [11]	>100M	26.6	57.1	69.6	4.0
CMGSD [14]	>100M	26.1	56.8	69.7	4.0
HiT [25]*	>100M	30.7	60.9	73.2	**2.6**
TACo* [48]	>100M	28.4	57.8	71.2	4.0
Support Set* [34]	>100M	30.1	58.5	69.3	3.0
Frozen [3]	5.5M	31.0	59.5	70.5	3.0
CMMT (ours)	5.3M	**34.1**	**61.7**	**74.6**	3.0

Table 5: Comparison to the state-of-the-art results for text-to-video retrieval on ActivityNet.

Method	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
FSE [51]	18.2	44.8	89.1	7.0
CE [26]	18.2	47.7	91.4	6.0
HSE [51]	20.5	49.3	–	–
MMT [11]	22.7	54.2	93.2	5.0
Support Set [34]	26.8	58.1	93.5	**3.0**
CMGSD [14]	24.2	56.3	94.0	4.0
CMMT (ours)	**29.4**	**58.9**	**94.9**	**3.0**

CMMT with a wide variety of representative/latest methods. Results show that our CMMT has great potential even with limited data. Concretely, we can observe that: (1) For the 7k split, our CMMT outperforms TACo (the current best method pre-trained on HowTo100M) by 6.6% on R@1, 6.5% on R@5, 6.9% on R@10, and 1.0 on MedR (4.0 vs. 5.0). (2) For the 1k-A split, our CMMT has the best performance on R@1, R@5, and R@10, and achieves competitive MedR w.r.t. HiT [25] (which is pre-trained on HowTo100M and also adopts pre-extracted expert features).

Table 5 shows the comparative results for text-to-video retrieval on ActivityNet [19]. This dataset has longer videos (average length is 180 seconds) than other datasets. Since video clips are globally sampled, our CMMT can capture more complete content from each

Table 6: Comparison to the state-of-the-art results for text-to-video retrieval on the MSVD test set.

Method	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
VSE++ [9]	15.4	39.6	53.0	9.0
Multi. Cues [30]	20.3	47.8	61.1	6.0
CE [26]	19.8	49.0	63.8	6.0
Support Set [34]	28.4	60.0	72.9	4.0
Frozen [3]	33.7	64.7	76.3	3.0
CMMT (ours)	**36.9**	**67.9**	**78.4**	**2.0**

Table 7: Comparison to the state-of-the-arts for video-to-text retrieval on the MSR-VTT 1k-A test set.

Method	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
CE [26]	20.9	48.8	62.4	6.0
AVLnet [37]	28.5	54.6	65.2	4.0
MMT [30]	28.0	57.5	69.7	3.7
Support Set [34]	28.5	58.6	71.6	**3.0**
CMGSD [14]	27.2	58.0	69.5	3.9
CMMT (ours)	**31.2**	**59.3**	**72.4**	**3.0**

Table 8: Comparison to the state-of-the-art results for video-to-text retrieval on ActivityNet.

Method	R@1 ↑	R@5 ↑	R@50 ↑	MedR ↓
FSE [51]	16.7	43.1	88.4	7.0
CE [26]	17.7	46.6	90.9	6.0
HSE [51]	18.7	48.1	–	–
MMT [11]	22.9	54.8	93.1	4.3
Support Set [34]	25.5	57.3	93.5	**3.0**
CMGSD [14]	24.6	56.8	93.8	4.0
CMMT (ours)	**27.4**	**58.1**	**94.2**	**3.0**

video. The results show that our CMMT outperforms the second best method by 2.6% on R@1, 0.8% on R@5, and 1.4% on R@10. Meanwhile, our CMMT also achieves the best MedR=3.0 (equal to that of [34]). To further verify the effectiveness of our CMMT, we conduct extra experiments on MSVD [5] in Tables 6. We find that our CMMT still achieves new state-of-the-art.

4.4.2 Video-to-Text Retrieval. To further demonstrate the effectiveness of our CMMT, We evaluate it on two benchmarks for the video-to-text retrieval task: MSR-VTT and ActivityNet. Table 7 presents the comparative results for video-to-text retrieval on MSR-VTT. We observe that our CMMT outperforms the recent state-of-the-art (i.e., Support Set [34] pre-trained on Howto100M [29]) by 2.7% on R@1, 0.7% on R@5, and 0.8% on R@10. It also leads to the best MedR=3.0. Moreover, Table 8 shows the comparative results on ActivityNet. It can be seen that our CMMT still achieves state-of-the-art. Overall, the superior performance of our model for video-to-text retrieval (including text-to-video retrieval) indicates that our model has been well-trained to align the video and text embeddings.

Table 9: Ablation study for different training losses used in our CMMT. Text-to-video retrieval results are reported on the MSR-VTT 1k-A test set. All experiments are conducted under the 48-way setting. K-shot: K clips per video/class; Frames: the number of frames per video/class.

K-shot	+L_{cls}	+L_{taat}	Frames	R@1 ↑	R@5 ↑	R@10 ↑
1-shot	✓	×	1×9	32.8	60.2	72.6
1-shot	×	✓	1×9	30.7	58.9	69.7
1-shot	✓	✓	1×9	33.3	61.0	73.7
1-shot	✓	×	1×3	30.0	59.4	71.5
1-shot	✓	✓	1×3	30.4	60.1	72.0
2-shot	✓	×	2×3	32.4	60.7	72.3
2-shot	✓	✓	2×3	32.7	61.1	72.5
3-shot	✓	×	3×3	32.6	61.1	73.0
3-shot	✓	✓	3×3	**34.1**	**61.7**	**74.6**

4.4.3 Ablation Study Results. We have two cross-modal losses during our CMMT training process: L_{cls} and L_{taat}. In Table 9, we conduct extensive experiments (in-dataset) to analyze their contributions. It can be observed that: (1) Since L_{taat} is a cross-modal classification loss based on updating the video prototypes learned by L_{cls}, training CMMT with only L_{taat} (without L_{cls}) may not obtain promising performance in video-text retrieval. This is empirically verified by conducting ablative experiments with only L_{taat} (the 2nd row) and only L_{cls} (the 1st row). Therefore, in all other experiments, training the model with (at least) the loss L_{cls} becomes our default setting. (2) Training with both L_{cls} and L_{taat} can achieve better performance than the default setting, which validates the effectiveness of our TAAT module. (3) Sampling more samples (shots) from each pseudo video class can consistently improve the performance of our CMMT. That is, with the number of shots increases, our CMMT is achieving better results (4th row to 9th row). Interestingly, even without using more frames, sampling more shots can help our CMMT learn more generalizable representations for video-text retrieval (comparing the 3rd row and 9-th row, both with 9 frames sampled).

5 CONCLUSIONS

This paper presents a novel cross-modal meta-Transformer (CMMT) model for video-text retrieval. Firstly, for each raw video, we propose to sparsely and globally sample video clips which are regarded as different samples of a pseudo video class (i.e., each raw video denotes a pseudo video class). Further, we train our CMMT among cross-modal classification tasks with video prototypes, each of which aggregates all video clips of a pseudo video class. To improve the generalizability of our model, we induce a token-aware adaptive Transformer (TAAT) module. Extensive experiments on several benchmarks show that our model achieves new state-of-the-art under the cross-dataset setting. These generalization results also indicate the high generalizability of our CMMT.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (61976220). Zhiwu Lu is the corresponding author.

REFERENCES

[1] Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. 2021. Noise Estimation Using Density Estimation for Self-Supervised Multimodal Learning. In *AAAI*. 6644–6652.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*. http://arxiv.org/abs/1409.0473

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. *arXiv preprint arXiv:2104.00650* (2021). https://arxiv.org/abs/2104.00650

[4] Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer Interpretability Beyond Attention Visualization. In *CVPR*. 782–791.

[5] David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *ACL*. 190–200.

[6] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325* (2015). http://arxiv.org/abs/1504.00325

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*. 4171–4186.

[8] 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. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*. https://openreview.net/forum?id=YicbFdNTTy

[9] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *BMVC*. 12.

[10] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *ICML*. 1126–1135.

[11] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *ECCV*. 214–229.

[12] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. 2022. Bridging Video-Text Retrieval With Multiple Choice Questions. In *CVPR*. 16167–16176.

[13] Bharath Hariharan and Ross B. Girshick. 2017. Low-Shot Visual Recognition by Shrinking and Hallucinating Features. In *ICCV*. 3037–3046.

[14] Feng He, Qi Wang, Zhifan Feng, Wenbin Jiang, Yajuan Lü, Yong Zhu, and Xiao Tan. 2021. Improving Video Retrieval by Adaptive Margin. In *SIGIR*. ACM, 1359–1368.

[15] Anne Lisa Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *ICCV*. 5804–5813.

[16] Weike Jin, Zhou Zhao, Pengcheng Zhang, Jieming Zhu, Xiuqiang He, and Yueting Zhuang. 2021. Hierarchical Cross-Modal Graph Consistency Learning for Video-Text Retrieval. In *SIGIR*. ACM, 1114–1124.

[17] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014). http://arxiv.org/abs/1411.2539

[18] Bruno Korbar, Fabio Petroni, Rohit Girdhar, and Lorenzo Torresani. 2020. Video Understanding as Machine Translation. *arXiv preprint arXiv:2006.07203* (2020). https://arxiv.org/abs/2006.07203

[19] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *ICCV*. 706–715.

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

[21] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is More: ClipBERT for Video-and-Language Learning via Sparse Sampling. *CVPR* (2021), 7331–7341.

[22] Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. 2020. Adversarial Feature Hallucination Networks for Few-Shot Learning. In *CVPR*. 13467–13476.

[23] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. HERO: Hierarchical encoder for video+ language omni-representation pre-training. *EMNLP* (2020), 2046–2065.

[24] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. 2017. Meta-SGD: Learning to Learn Quickly for Few Shot Learning. *arXiv preprint arXiv:1707.09835* (2017). http://arxiv.org/abs/1707.09835

[25] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. 2021. HiT: Hierarchical Transformer with Momentum Contrast for Video-Text Retrieval. *arXiv preprint arXiv:2103.15049* (2021). https://arxiv.org/abs/2103.15049

[26] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019. Use What You Have: Video retrieval using representations from collaborative experts. In *BMVC*. 279.

[27] Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. 2022. COTS: Collaborative Two-Stream Vision-Language Pre-Training Model for Cross-Modal Retrieval. In *CVPR*. 15692–15701.

[28] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. 2020. UniVL: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353* (2020). https://arxiv.org/abs/2002.06353

[29] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*. 2630–2640.

[30] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. 2018. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *ICMR*. 19–27.

[31] Tsendsuren Munkhdalai and Hong Yu. 2017. Meta Networks. In *ICML*. 2554–2563.

[32] Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999* (2018). http://arxiv.org/abs/1803.02999

[33] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2Text: Describing images using 1 million captioned photographs. In *NeurIPS*. 1143–1151.

[34] Mandela Patrick, Po-Yao Huang, Yuki Markus Asano, Florian Metze, Alexander G. Hauptmann, João F. Henriques, and Andrea Vedaldi. 2021. Support-set bottlenecks for video-text representation learning. In *ICLR*. https://openreview.net/forum?id=EqoXe2zmhrh

[35] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *ICCV*. 2641–2649.

[36] Sachin Ravi and Hugo Larochelle. 2017. Optimization as a Model for Few-Shot Learning. In *ICLR*. https://openreview.net/forum?id=rJY0-Kcll

[37] Andrew Rouditchenko, Angie Boggust, David Harwath, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, et al. 2020. AVLnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199* (2020). https://arxiv.org/abs/2006.09199

[38] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*. 2556–2565.

[39] Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical Networks for Few-shot Learning. In *NeurIPS*. 4080–4090.

[40] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. 2018. Learning to Compare: Relation Network for Few-Shot Learning. In *CVPR*. 1199–1208.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS*. 5998–6008.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS*. 5998–6008.

[43] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond J. Mooney, and Kate Saenko. 2015. Translating Videos to Natural Language Using Deep Recurrent Neural Networks. In *NAACL-HLT*. 1494–1504.

[44] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In *ICCV*. 4580–4590.

[45] Xiaohan Wang, Linchao Zhu, and Yi Yang. 2021. T2VLAD: Global-Local Sequence Alignment for Text-Video Retrieval. In *CVPR*. 5079–5088.

[46] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*. 5288–5296.

[47] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*. 2048–2057.

[48] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. 2021. TACo: Token-aware Cascade Contrastive Learning for Video-Text Alignment. *arXiv preprint arXiv:2108.09980* (2021). https://arxiv.org/abs/2108.09980

[49] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. 2020. Few-Shot Learning via Embedding Adaptation with Set-to-Set Functions. In *CVPR*. 8805–8814.

[50] Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018. A joint sequence fusion model for video question answering and retrieval. In *ECCV*. 487–503.

[51] Bowen Zhang, Hexiang Hu, and Fei Sha. 2018. Cross-modal and hierarchical modeling of video and text. In *ECCV*. 385–401.

[52] Hongguang Zhang, Jing Zhang, and Piotr Koniusz. 2019. Few-shot learning via saliency-guided hallucination of samples. In *CVPR*. 2770–2779.

[53] Linchao Zhu and Yi Yang. 2020. ActBERT: Learning global-local video-text representations. In *CVPR*. 8743–8752.