MuMUR : Multilingual Multimodal Universal Retrieval

Avinash Madasu (✉ avinash.madasu@intel.com)

Intel (United States)

Estelle Aflalo

Intel (United States)

Gabriela Ben Melech Stan

Intel (United States)

Shachar Rosenman

Intel (United States)

Shao-Yen Tseng

Intel (United States)

Gedas Bertasius

University of North Carolina at Chapel Hill

Vasudev Lal

Intel (United States)

MuMUR : <u>Mu</u>ltilingual <u>M</u>ultimodal <u>U</u>niversal <u>R</u>etrieval

Avinash Madasu[1*], Estelle Aflalo[1], Gabriela Ben Melech Stan[1], Shachar Rosenman[1], Shao-Yen Tseng[1], Gedas Bertasius[2], Vasudev Lal[1]

[1]Cognitive Computing Research, Intel Labs.
[2]Department of Computer Science, University of North Carolina at Chapel Hill.

*Corresponding author(s). E-mail(s): avinash.madasu@intel.com;
Contributing authors: estelle.aflalo@intel.com;
gabriela.ben.melech.stan@intel.com; shachar.rosenman@intel.com;
shao-yen.tseng@intel.com; gedas@cs.unc.edu; vasudev.lal@intel.com;

Abstract

Multi-modal retrieval has seen tremendous progress with the development of vision-language models. However, further improving these models require additional labelled data which is a huge manual effort. In this paper, we propose a framework MuMUR, that utilizes knowledge transfer from a multilingual model to boost the performance of multi-modal (image and video) retrieval. We first use state-of-the-art machine translation models to construct pseudo ground-truth multilingual visual-text pairs. We then use this data to learn a joint vision-text representation where English and non-English text queries are represented in a common embedding space based on pretrained multilingual models. We evaluate our proposed approach on a diverse set of retrieval datasets: five video retrieval datasets such as MSRVTT, MSVD, DiDeMo, Charades and MSRVTT multilingual, two image retrieval datasets such as Flickr30k and Multi30k . Experimental results demonstrate that our approach achieves state-of-the-art results on all video retrieval datasets outperforming previous models. Additionally, our framework MuMUR significantly beats other multilingual video retrieval dataset. We also observe that MuMUR exhibits strong performance on image retrieval. This demonstrates the universal ability of MuMUR to perform retrieval across all visual inputs (image and video) and text inputs (monolingual and multilingual).

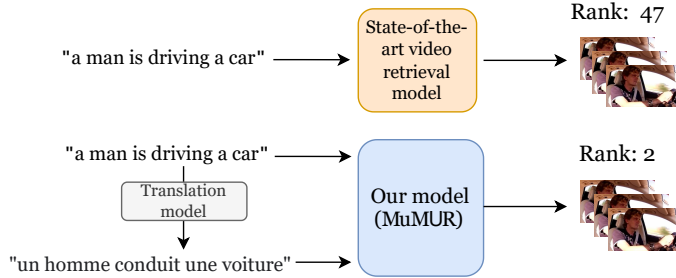**Keywords:** Video-retrieval, Image-retrieval, Multilingual, Multimodal

Fig. 1 Illustration of the improved video retrieval ranking for a test sample in MSRVTT dataset. For the current state-of-the-art video retrieval model, the rank of the ground truth video is 47. When multilingual data is used as knowledge transfer, the ranking of the ground truth video improved significantly from 47 to 2.

1 Introduction

The task of multimodal retrieval aims to retrieve images or videos that are semantically similar to a given text query and vice-versa. In the world of computer vision, image retrieval and video retrieval have been treated as separate tasks. For example, transformer models pretrained on large amounts of image-text pairs are then fine tuned for the task of image retrieval [10, 24, 31–33]. In contrast, video retrieval models are developed in two parallel directions. The first line of work [4, 19, 31] focused on video-text pretraining on large-scale datasets like Howto100M [44] and WebVid-2M [4]. While the second line of work [41] focused on using pretrained image features like CLIP [48] for video retrieval often surpassing the models pretrained on video datasets.

Most of the current multimodal retrieval datasets typically contain around 10k visual inputs and the corresponding captions of maximum lengths ranging from 30 to 60. In multimodal retrieval, the text encoder projects the input text caption and visual inputs (image or video) into a common embedding space. With longer captions, the text embeddings might lose required contextual information resulting in incorrect retrievals. One could address this by incorporating more structured knowledge (e.g., parts of speech, dependency graphs) in the text encoder [7]. However, a drop in performance is observed [7] with the addition of structural knowledge in a text-to-video retrieval setting. One could argue that the reason might be the smaller retrieval datasets and creating meaningful structural knowledge becomes a challenging task.

There are over 6000 languages in the world each having its own vocabulary of words, grammar and morphology. However, there exists an overlap of knowledge among these languages [3, 22, 59]. In Natural Language Processing, some works [12] explored this idea of using multilingual data to improve the performance on monolingual English datasets. Conneau et al. introduced a new pretraining objective: Translation Language Modeling (TLM) in which random words were masked in the concatenated sentences of English and multilingual data and the model predicts the masked words. There, the objective was to use multilingual context to predict masked English words if the English context was not sufficient and vice-versa.

Most of the current works in multimodal retrieval are primarily focused on monolingual English datasets. Recently, few works have proposed models that can perform retrieval in a multilingual setting [6, 23, 45, 66]. However, these works use large amounts of multilingual data for pretraining and are tuned for retrieval on each language. Moreover, these models are also limited to performing either image retrieval or video retrieval. In addition, these models show a drop in performance on English retrieval datasets even though there is a boost in performance on multilingual data.

In this work, we are interested in two objectives **(i) design a model capable of performing retrieval on both images and videos (ii) design a model capable of performing retrieval on multiple languages including English**. Multilingual data serves as a powerful knowledge augmentation for monolingual models [12]. Nevertheless, creating multilingual data requires huge human effort. To overcome this, we use state-of-the-art machine translation models [52] to convert English text captions into other languages. Specifically, we choose languages whose performance on XNLI benchmark [13] is comparable to that of English (i.e., French, German, Spanish). With this, we create high quality multilingual data without requiring human labelling. To the best of our knowledge, this is the first work that uses multilingual knowledge transfer to improve multimodal retrieval.

We propose a new framework MuMUR : <u>Mu</u>ltilingual <u>M</u>ultimodal <u>U</u>niversal <u>R</u>etrieval that is capable of performing both image and video retrieval on multilingual datasets. This framework is based on CLIP [48] to effectively utilize and adapt the multilingual knowledge transfer. Our model takes a visual input, English text caption and multilingual text caption as inputs and extracts joint visual-text representations. The multilingual text representations should act as a knowledge augmentation to the English text representations aiding in retrieval. For this purpose, we introduce a dual cross-modal (DCM) encoder block which learns the similarity between English text representations and visual representations. In addition, the DCM encoder block also associates the visual representations with the multilingual text representations. In the common embedding space, our model learns the important contextual information from multilingual representations which is otherwise missing from the English text representations effectively serving as knowledge transfer.

We validate our proposed model on a comprehensive set of image retrieval datasets: Flickr30k [47] and video retrieval datasets: MSRVTT-9k [61], MSRVTT-7k [61], MSVD [9], DiDeMo [2] and Charades [50]. We show that our approach achieves state-of-the-art results, outperforming previous models on most datasets. In addition to the evaluation on monolingual retrieval datasets, we also compare the performance of our model on multilingual datasets Multi30k [16] and MSRVTT multilingual.[23]. Experimental results demonstrate that MuMUR achieves state-of-the-art results on all English video retrieval datasets and significantly outperforms previous models on multilingual video retrieval in a zero-shot setting. Furthermore, our model MuMUR establishes new benchmark on multilingual image retrieval while achieving strong performance on English image retrieval datasets. These results demonstrate the universal capability of MuMUR to perform all types of multimodal and multilingual retrieval.

To summarize, our contributions are as follows: (i) We generate multilingual data using external state-of-the-art machine translation models. (ii) We propose a model

3

that is capable of knowledge transfer from multilingual data to improve the performance of multimodal retrieval. (iii) We evaluate the proposed framework on six English retrieval benchmarks and achieve state-of-the-art results in both text-to-visual and visual-to-text retrieval settings. (iv) Finally, we demonstrate that our model significantly outperforms previous approaches on multilingual retrieval datasets.

2 Related work

2.1 Multimodal retrieval

Pre-train and then fine-tune is the most popular paradigm involving image retrieval [10, 32–34]. These models are pre-trained on huge amounts of image text pairs such as Conceptual Captions [8], Visual Genome [28] and SBU [40] and tested on image retrieval datasets such as Flickr30k [47] and COCO [25].

The task of video retrieval has seen tremendous progress in the recent years. This is partly due to the availability of large-scale video datasets like HowTo100M [44] and WebVid-2M [4]. Besides the adaption of transformers to image tasks like image classification [14] spurred the development of models based on transformers. However, videos require computationally more memory and compute power and can be infeasible to compute self-attention matrices. With the introduction of more efficient architectures [5] large-scale pretraining on videos became a possibility. In this direction, several transformer based architectures [4, 19, 42, 43] were proposed and pretrained on large video datasets which achieved state-of-the-art results on downstream video retrieval datasets in both zero-shot and fine-tuning settings.

In a parallel direction, a few works [41] have adopted image level features pretrained on large scale image-text pairs to perform video retrieval. Surprisingly, these works have performed significantly better than the models that are pretrained from scratch on large scale video datasets. Compared to these models, our approach completely differs in the architecture and the training methodology.

2.2 Multilingual training

The recent success of multimodal image-text models on a variety of tasks, such as retrieval and question-answering, has been mostly limited to monolingual models trained on English text. This is mainly due to the availability and high-quality of English-based multimodal datasets. Recent work indicates that incorporating a second language or a multilingual encoder, thus creating a shared multilingual token embedding space, can improve monolingual pure-NLP downstream tasks [12]. This concept was rapidly embraced for training multimodal models. Previous works had used images as a bridge for translating between two languages, without using a language-to-language shared dataset for training [9, 49, 51].

Recent work has focused on multimodal tasks, such as image retrieval, aiming to add multilingual capabilities to multimodal models [6, 21]. The work often indicates that incorporating a second language during training of multimodal models, improves performance on single-language multimodal tasks such as image retrieval, compared to multimodal models that were trained on a single language [21, 26, 57]. MULE [26],
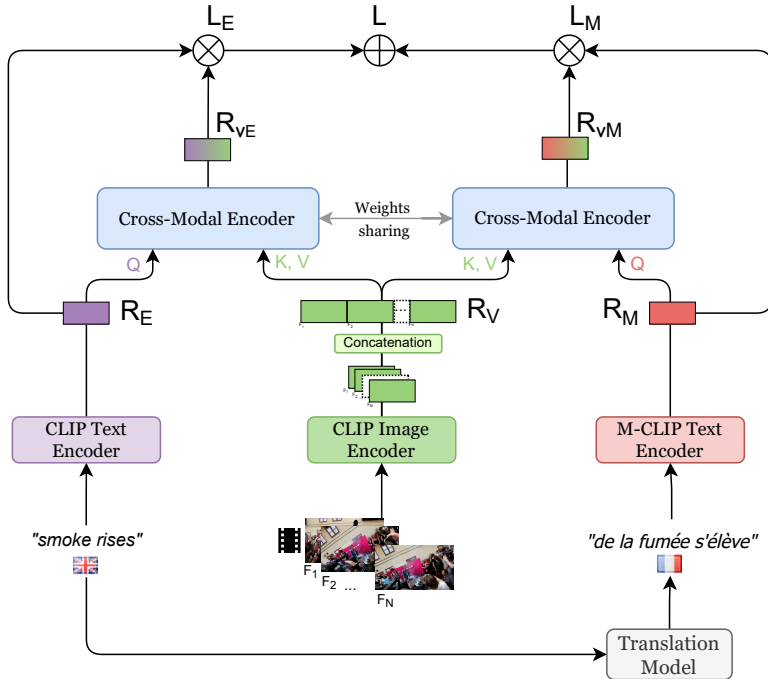
Fig. 2 Illustration of the proposed MuMUR model. The model takes as visual input (image or video), a corresponding English text query and a translated multilingual query. The multilingual text query is obtained using the off-the-shelf machine translation model. It is used only for the inference and is not part of the architecture. The video and English text features are extracted using CLIP model whereas multilingual text features are extracted using M-CLIP model. The features are then passed onto a cross-model encoder to learn the association in a common embedding space. Cross-entropy loss is then applied to measure the similarity between text features R_E and R_{vE}, R_M and R_{vM}. The final loss is the sum of both the losses.

which is a multilingual universal language encoder trained on image-multilingual text pairs, showed an improvement on image-sentence retrieval tasks of up to 20% compared to monolingual models. Nevertheless, all these previous works focus on designing models separately for image and video retrieval. Our objective is to use multilingual knowledge transfer to improve the performance on current image and video retrieval datasets. In addition, our model is capable of performing multilingual retrieval on more than 10 languages.

3 MuMUR : Multilingual Multimodal Universal Retrieval

In this section, we introduce our framework MuMUR : Multilingual Multimodal Universal Retrieval. We first describe the problem statement, then the multilingual data augmentation strategy and finally go over the proposed approach that enables knowledge transfer from multilingual data for video retrieval.

3.1 Problem statement

Given a set of visual data V, their corresponding English text captions E and related multilingual text captions M, our goal is to learn similarity functions $s_1(v_i, e_i)$ and $s_2(v_i, m_i)$ ($v_i \in V$, $e_i \in E$ and $m_i \in M$). In other words, we propose a framework MuMUR that enables end-to-end learning on a tuple of visual input, English text caption and multilingual text caption by bringing closer the joint representations of those three elements. Specifically, for each visual input V and the English text captions E, we generate the multilingual translations M using external state-of-the-art machine translation models [52]. Next, we present the proposed approach that facilitates the end-to-end learning using multilingual data.

3.2 Approach

Our model, illustrated in Figure 2, is comprised of three components: (i) visual encoder (ii) text encoder (iii) dual cross modal encoder. Next, we describe the framework in detail.

3.2.1 Visual Encoder

Given a visual input V, we consider uniformly sampled clips $C \in R^{N_v \times H \times W \times 3}$ where N_v is the number of frames (1 for images), H and W are the spatial dimensions of a RGB frame. We then use a pretrained CLIP-ViT image encoder [48] to extract the frame embeddings $F_v \in R^{N_v \times D_v}$ where D_v denotes the dimensions of the frame embeddings. The frame embeddings are concatenated to obtain the final representation for the video V.

3.2.2 Text Encoder

Let the inputs English text caption be E and multilingual text caption M of lengths p and q respectively. We use a pretrained CLIP-ViT text encoder to convert the English text caption into a sequence of embeddings $R_E = R^{E_p \times D_E}$ where D_E denote the embedding dimensions. We consider the representation of the token $[EOS]$ as the final representation of English text caption. To encode multilingual text caption M, we use a M-CLIP[1] model which is a multilingual clip model pretrained on multilingual text and image pairs. Specifically, the multilingual text caption is converted into a sequence of embeddings $R_M = R^{M_q \times D_M}$ where D_M denote the embedding dimensions. Similar to the CLIP model, we consider the $[EOS]$ representation as the final representation of M-CLIP model.

3.2.3 Dual cross-modal encoder (DCM)

Our goal is to closely associate the visual embeddings R_v, English text embeddings R_E and multilingual text embeddings R_M in a common embedding space. For this purpose, we propose a dual cross-modal encoder (DCM). To incorporate textual information into visual features and to learn visual features that are semantically most similar to

---

[1]https://github.com/FreddeFrallan/Multilingual-CLIP

text features, we use multi-head attention. The text features are used as the queries whereas the visual features are used as the keys and values.

$$r_{vE} = Attention(T_E, F_v, F_v) \tag{1}$$
$$r_{vM} = Attention(M_E, F_v, F_v) \tag{2}$$

where multi-head attention ($Attention$) is defined as:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}})V \tag{3}$$

Here Q, K and V are same as the original multi-head attention matrices in the transformer encoder. We then apply a fully connected layer on the attention outputs and finally layer normalization to obtain R_{vE} and R_{vM}.

$$R_{vE} = LN(FC(r_{vE}) + r_{vE}) \tag{4}$$

$$R_{vM} = LN(FC(r_{vM}) + r_{vM}) \tag{5}$$

where FC is the fully connected layer and LN is the layer normalization layer.

3.2.4 Loss

We use the standard image-text or video-text matching loss [58] to train the model. It is measured as the dot product similarity between matching text embeddings and visual embeddings in a batch. First, we compute the loss L_E between R_{vE} and R_E and then compute the loss L_M between R_{vM} and R_M. The final loss is the sum of losses L_E and L_M.

$$L = L_E + L_M. \tag{6}$$

where $L_E = L_E^{t2v} + L_E^{v2t}$ and $L_M = L_M^{t2v} + L_M^{v2t}$

$$\mathcal{L}_E^{\text{v2t}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(R_E^{(i)} \cdot R_{vE}^{(i)})}{\sum_{j=1}^{B} \exp(R_E^{(i)} \cdot R_{vE}^{(j)})}, \tag{7}$$

$$\mathcal{L}_E^{\text{t2v}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(R_{vE}^{(i)} \cdot R_E^{(i)})}{\sum_{j=1}^{B} \exp(R_{vE}^{(i)} \cdot R_E^{(j)})}. \tag{8}$$

$$\mathcal{L}_M^{\text{v2t}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(R_M^{(i)} \cdot R_{vM}^{(i)})}{\sum_{j=1}^{B} \exp(R_M^{(i)} \cdot R_{vM}^{(j)})}, \tag{9}$$

$$\mathcal{L}_M^{\text{t2v}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(R_{vM}^{(i)} \cdot R_M^{(i)})}{\sum_{j=1}^{B} \exp(R_{vM}^{(i)} \cdot R_M^{(j)})}. \tag{10}$$

3.2.5 Inference

During inference for English retrieval datasets, we freeze the multilingual text encoder and measure the retrieval performance only using R_E and R_{vE}. Similarly, for multilingual datasets, we freeze the English text encoder and calculate the retrieval score using R_M and R_{vM}.

4 Experiments

4.1 Datasets

Our goal is to design a universal multimodal multilingual retrieval model. Therefore, our experiments focus on evaluating on both image and video retrieval datasets comprising of monolingual and multilingual captions.

4.1.1 Video Retrieval

We perform experiments on six standard text-video retrieval datasets: MSRVTT-9k and MSRVTT-7k splits [61], MSVD [9], DiDeMo [2], Charades [50] and MSRVTT multilingual [23].

MSRVTT contains 10K videos with each video ranging from 10 to 32 seconds and 200K captions. We report the results both on MSRVTT-9k and MSRVTT-7k datasets following [41].

MSVD consists of 1970 videos and 80K descriptions. We use the standard training, validation and testing splits following [41]. In this dataset, each video has multiple captions and are treated as independent samples during testing.

DiDeMo is made up of 10K videos and 40K localized descriptions of the videos. We concatenate all the sentences for each video and evaluate the paragraph-to-video retrieval following [35, 41].

Charades contains of 9848 videos and each video is associated with a caption. We use the standard training and test splits following [35].

MSRVTT multilingual is a multilingual version of MSRVTT in which the English captions are translated into nine different languages. We use the standard splits following [23].

4.1.2 Image Retrieval

We evaluate the proposed approach MuMUR on the following image retrieval datasets:

Flickr30k contains 31000 images with each image containing single caption for training and validation data and 5 captions for testing data. We follow the standard splits of 29k/1k/1k [32].

Multi30K [16] is a multilingual version of Flickr30k [47] in which the English text captions are translated into German (de), French (fr) and Czech (cs) languages. We use the training, validation and testing splits of 29k/1k/1k following [45].

4.2 Metrics

For evaluating the performance of models, we use recall at rank K ($R@1$, $R@5$, $R@10$), median rank (MedR) and mean rank (MnR). Unless specified, the values reported are the mean of three runs with different seeds.

Type	Model	Text-to-Video Retrieval					Video-to-Text Retrieval				
		R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Others	JsFusion [64]	10.2	31.2	43.2	13.0	-	-	-	-	-	-
	HT [44]	14.9	40.2	52.8	9.0	-	-	-	-	-	-
	HERO [34]	20.5	46.8	60.9	-	-	-	-	-	-	-
	CE [37]	20.9	48.8	62.4	6.0	28.2	20.6	50.3	64.0	5.3	25.1
	ClipBERT [30]	22.0	46.8	59.9	-	-	-	-	-	-	-
	SupportSET [46]	27.4	56.3	67.7	3.0	-	-	-	-	-	-
	VideoCLIP [60]	30.9	55.4	66.8	4.0	-	-	-	-	-	-
	FrozenInTime [4]	31	59.5	70.5	3.0	-	-	-	-	-	-
	CLIP [48]	31.2	53.7	2.6	4.0	-	-	-	-	-	-
	HIT [36]	30.7	60.9	73.2	2.6	-	32.1	62.7	74.1	3.0	-
	AlignPrompt [31]	33.9	60.7	73.2	-	-	-	-	-	-	-
	All-in-one [54]	34.4	65.4	75.8	-	-	-	-	-	-	-
	MDMMT [15]	38.9	69.0	79.7	**2.0**	-	-	-	-	-	-
CLIP based	CLIP4Clip [41]	44.5	71.4	81.6	-	15.3	43.1	70.5	81.2	**2.0**	12.4
	VCM [7]	43.8	71.0	80.9	**2.0**	14.3	45.1	72.3	82.3	**2.0**	10.7
	MCQ [19]	44.9	71.9	80.3	**2.0**	15.3	-	-	-	-	-
	MILES [20]	44.3	71.1	80.8	**2.0**	14.7	-	-	-	-	-
	CAMoE [11]	44.6	**72.6**	81.8	**2.0**	**13.3**	45.1	72.4	83.1	**2.0**	10.0
	CLIP2Video [17]	45.6	**72.6**	81.7	**2.0**	14.6	43.5	72.3	82.1	**2.0**	10.2
	CLIP2TV [18]	46.1	72.5	**82.9**	**2.0**	15.2	43.9	70.9	82.2	**2.0**	12.0
Ours	**MuMUR**	**46.6**	**72.6**	82.2	**2.0**	13.9	**45.5**	**73.4**	**84.7**	**2.0**	**8.07**

Table 1 Text-to-video and video-to-text retrieval results on MSR-VTT dataset 9k split. Recall at rank 1 (R@1)↑, rank 5 (R@5)↑, rank 10 (R@10)↑, Median Rank (MdR)↓ and Mean Rank (MnR)↓ are reported. Results of other methods taken from mentioned references. Our model surpasses previous state-of-the-art performance. In video-to-text retrieval, our model achieved 1.6 points boost in performance.

4.3 Implementation Details

We use translations of French for the multilingual inputs to train the MuMUR model. The visual encoder and the English text encoder are initialized with CLIP-ViT-B-32. The multilingual text encoder is initialized with M-CLIP-ViT-B-32. The dimension size of the video, English caption and multilingual caption representations is 512. The dual cross-model encoder is initialized randomly and trained from scratch. The dimension size of the key, query and value projection layers is 512. The fully connected layer in the transformer has a size of 512 and a dropout of 0.4 is applied on this layer. We use 16 frames for MSRVTT-9k, MSRVTT-7k and MSVD datasets, 42 frames for DiDeMo and Charades datasets. The maximum sequence length is set to 32 for MSRVTT-9k and MSRVTT-7k, 64 for DiDeMo and 30 for charades dataset. The model is trained using AdamW [39] a learning rate of 1e-4 and a cosine decay of 1e-6. The MSRVTT-9k and MSRVTT-7k datasets are trained with a batch size of 32 and for 15 epochs. The MSVD dataset is trained with a batch size of 32 and for 5 epochs.

The DiDeMo and charades datasets are trained with a batch size of 16 for 12 and 15 epochs respectively.

Model	R@1 (↑)	R@5 (↑)	R@10(↑)	MdR (↓)
HowTo100M [44]	10.2	31.2	43.2	13.0
ActBERT [67]	8.6	23.4	33.1	36.0
NoiseE [1]	17.4	41.6	53.6	8.0
ClipBERT [30]	22.0	46.8	59.9	6.0
CLIP4clip- [41]	42.1	71.9	81.4	**2.0**
Singularity [29]	42.7	69.5	78.1	**2.0**
MuMUR	**44.8**	**72.0**	**82.5**	**2.0**

Table 2 Text-to-video retrieval results on MSR-VTT - 7k split. Recall at rank-1 (R@1), rank-5 (R@5), rank-10 (R@10), Median Rank (MdR) are reported. Results of other methods taken from mentioned references.

Model	R@1 (↑)	R@5 (↑)	R@10(↑)	MdR (↓)
VSE [17]	12.3	30.1	42.3	14.0
CE [37]	19.8	49.0	63.8	6.0
SSML [1]	20.3	49.0	63.3	6.0
SUPPORT-SET [46]	28.4	60.0	72.9	4.0
FROZEN [4]	33.7	64.7	76.3	3.0
CLIP [48]	37.0	64.1	73.8	3.0
CLIP4Clip [41]	46.2	76.1	84.6	2.0
CLIP2Video [17]	47.0	76.8	85.9	2.0
MuMUR	**47.2**	**77.3**	**86.2**	**2.0**

Table 3 Text-to-video retrieval results on MSVD dataset (multi-caption evaluation). Recall at rank-1 (R@1), rank-5 (R@5), rank-10 (R@10), Median Rank (MdR) are reported. Results of other methods taken from mentioned references.

5 Results and Discussion

5.1 Evaluation on English video retrieval datasets

In Table 1 we report the results of our proposed approach on MSVRTT-9k dataset. It can be observed that the difference between CLIP based models and other models is very significant ($> 5\%$). Therefore, it explains the incentive to build our model using CLIP features. On MSRVTT-9k split, our model significantly outperforms CLIP4Clip model on all the metrics in both text-to-video and video-to-text retrieval settings. VCM employs a knowledge graph between video and text modalities making its performance superior to other models in a video-to-text retrieval task. Our model surpasses VCM significantly in all the metrics elucidating that the multilingual representations serve as a powerful knowledge transfer. Moreover, our approach outperforms MCQ and MILES which are pretrained on WebVid-2M data, initialized with CLIP features,

Model	R@1 (↑)	R@5 (↑)	R@10 (↑)	MdR (↓)
S2VT [53]	11.9	33.6	-	13.0
FSE [65]	13.9	36	-	11.0
CE [37]	16.1	41.1	-	8.3
ClipBERT [30]	20.4	48.0	60.8	6.0
FrozenInTime [4]	31.0	59.8	72.4	3.0
OA-Trans [55]	34.8	64.4	75.1	3.0
CLIP4clip [41]	43.4	70.2	80.6	**2.0**
CLIP2TV [18]	43.9	70.5	79.8	**2.0**
TS2-Net [38]	41.8	71.6	82.0	**2.0**
ECLIPSE [35]	44.2	70.0	80.2	**2.0**
MuMUR	**44.4**	**74.3**	**83.1**	**2.0**

Table 4 Text-to-video retrieval result on DiDeMo dataset. Recall at rank-1 (R@1), rank-5 (R@5), rank-10 (R@10), Median Rank (MdR) are reported. Results of other methods taken from mentioned references.

Model	R@1 (↑)	R@5 (↑)	R@10 (↑)	MdR	MnR
ClipBERT [30]	6.7	17.3	25.2	32.0	149.7
FrozenInTime [4]	11.9	28.3	35.1	17.0	103.8
CLIP4clip [41]	13.9	30.4	37.1	14.0	98.0
ECLIPSE [35]	15.7	32.9	42.4	16.0	84.9
MuMUR	**16.6**	**37.5**	**50.0**	**10.0**	**52.7**

Table 5 Text-to-video retrieval result on charades dataset. Recall at rank-1 (R@1), rank-5 (R@5), rank-10 (R@10), Median Rank (MdR) are reported. Results reported are taken from [35].

employing additional semantic information like parts-of-speech. This validates that our model doesn't require any pretraining on videos and structural knowledge injection. The multilingual text representations in our model effectively serves this purpose.

In Tables 2, 3, 4 and 5 we report the results on MSRVTT-7k, MSVD, DiDeMo and Charades datasets respectively. Our model outperforms all the previous approaches across all the metrics on all the datasets. For the MSRVTT-7k split, our model achieves a significant boost of 2.1%, 2.5% and 4.4% in R@1, R@5 and R@10 respectively compared to the previous baselines. For the MSVD dataset, we notice an improvement of 0.2%, 0.5% and 0.3% in R@1, R@5 and R@10 respectively. MSVD is a relatively smaller dataset with test size of 670 videos and hence, the improvements are relatively marginal.

For the DiDeMo dataset, our model showed a marginal boost of 0.2% in R@1 but a significant boost of 4.3% and 2.9% in R@5 and R@10 respectively compared to the previous approaches. For the Charades dataset, our model outperformed previous approaches by 0.9% in R@1 and by a significant margin of 4.6%, 7.6% and 6.0% in R@5, R@10 and MedianR respectively. ECLIPSE uses audio as additional information for video retrieval. We showed that multilingual text acts as a better knowledge transfer input.

11

	Image-to-text Retrieval			Text-to-Image Retrieval		
Model	R@1 (↑)	R@5 (↑)	R@10 (↑)	R@1 (↑)	R@5 (↑)	R@10 (↑)
ViLT [27]	83.5	96.7	98.6	64.4	88.7	93.8
UNITER [10]	87.3	98.0	99.2	75.6	94.1	96.8
Frozen [4]	-	-	-	61.0	87.5	92.7
ALBEF [33]	94.3	99.4	99.8	82.8	96.7	98.4
BLIP [32]	**96.6**	**99.8**	100.0	87.2	97.5	98.8
ALIGN [31]	95.3	99.8	100.0	84.9	97.4	98.6
SINGULARITY [29]	93.3	99.4	99.8	81.4	95.8	97.9
BridgeTower [63]	94.7	99.61	100.0	85.8	97.6	98.9
ManageTower [62]	95.6	-	-	86.5	-	-
MuMUR	90.5	99.7	99.8	**92.0**	**99.5**	**99.9**

Table 6 Text-to-image and image-to-text retrieval results on Flickr30k [47] dataset split. Recall at rank 1 (R@1)↑, rank 5 (R@5)↑, rank 10 (R@10)↑ are reported. Results of other methods taken from mentioned references. MuMUR significantly outperforms previous models in text-to-image retrieval while achieving comparable results in image-to-text retrieval.

Model	de	cs	zh	ru	sw	es
m-BERT (zero-shot)	11.1	8.2	6.9	7.9	1.4	12
m-BERT MMP (zero-shot)	15	11.2	8.4	11	3.4	15.1
XLM-R (zero-shot)	16.3	16	14.9	15.4	7.7	17.3
XLM-MMP (zero-shot)	19.4	19.3	18.2	19.1	8.4	20.4
m-BERT (fine-tune)	18.2	16.9	16.2	16.5	13	18.5
XLM- R + MMP (fine-tune)	21.1	20.7	20	20.5	14.4	21.9
MuMUR - fr (zero-shot)	**27.4**	**28.2**	**24.1**	**26.6**	**22.5**	**26.5**

Table 7 Text-to-video retrieval (R@1 metric) results on MSR-VTT - multilingual [23]. Results of other methods taken from [23]. Our model is trained on Charades dataset and using only french language and evaluated in a zero-shot setting on MSRVTT multilingual dataset. In zero-shot evaluation on other languages, our model significantly outperforms previous models trained in both zero-shot and fine-tuning setting.

5.2 Evaluation on English image retrieval datasets

Next, we measure the performance of MuMUR on English image retrieval dataset Flickr30k. The results are reported in the table 6 in both image-to-text and text-to-image settings. As shown in the table, MuMUR achieves comparable performance to previous approaches on image-to-text retrieval. Moreover, we observe that MuMUR significantly outperforms these models by 5.5% in image-to-text retrieval. Note that, these models are pretrained on large amount of image-text pairs and fine-tuned on Flickr30k dataset. In contrast, our model uses a small amount of multilingual data and achieves remarkable results in both the settings. This validates that multilingual data acts as superior knowledge transfer even for image retrieval.

5.3 Evaluation on multilingual video retrieval datasets

In addition to the monolingual datasets, we also evaluate the proposed approach on multilingual video retrieval datasets. Specifically, we use the model trained only

Model	de	fr	cs
EmbN [56]	60.3	54.8	46.3
PAR. EmbN [21]	62.6	60.6	54.1
S-LIWE [57]	72.1	63.4	59.4
MULE [26]	64.1	62.3	57.7
SMALR [6]	69.8	65.9	64.8
M3P [45]	82.1	67.3	65
MuMUR	**85.2**	**70.5**	**69.3**

Table 8 Text-to-image retrieval (R@1 metric) results on Multi30K [16] dataset (de - German, fr - French and cs - Czech). Results of other methods are taken from [45]. Our model is fine-tuned for each language and evaluated on the corresponding test data.

using French captions and test on 6 languages such as German (de), Czech (cs), Chinese (zh), Swahili (sw), Russian (ru) and Spanish (es) in a zero-shot setting. Table 7 shows the results on MSRVTT-multilingual dataset. Our model achieved a significant boost of 8.2% (average) in R@1 in a zero-shot setting. It is worth noting that our model in a zero-short evaluation outperformed the previous approaches fine-tuned on these languages by a huge margin of 6.1% (average). MMP [23] is pretrained on the large scale multilingual dataset HowTo100M on 9 languages. However, our model trained on just 1 language outperformed MMP. This shows that our dual cross-modal (DCM) encoder block can effectively learn the association among video, English and multilingual representations even when large video pretraining is not involved.

5.4 Evaluation on multilingual image retrieval datasets

We also evaluate the proposed model MuMUR on the multilingual image retrieval dataset Multi30K. Table 8 show the results on 3 languages German (de), French (fr) and Czech (cs) fine-tuned for these languages. As shown in the table, MuMUR significantly outperforms previous models by 3.1% for German (de), 3.2% for French (fr) and 4.3% for Czech (cs) languages.

5.5 Ablation studies

5.5.1 Effect of multilingual knowledge transfer

We investigate the effect of multilingual knowledge transfer on the video-retrieval performance. Precisely, we train a model without the multilingual text encoder keeping the rest of the architecture intact. As shown in Figure 3, using multilingual data as knowledge transfer significantly improved the performance on DiDeMo and Charades datasets. The improvement is 2.4% for DiDeMo and 3.62% for Charades datasets.

5.5.2 Using only multilingual text encoder

Next, we ablate the choice of using an English text encoder. We validated previously that multilingual data improves the performance of video retrieval. This raises the
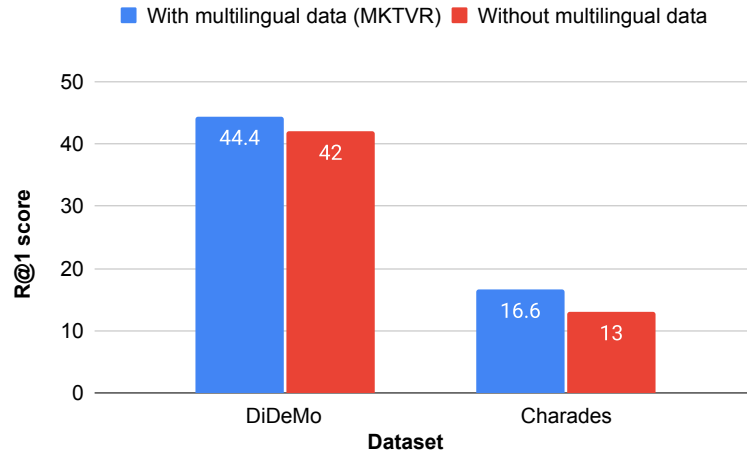
With multilingual data (MKTVR) ■ **Without multilingual data** ■

Fig. 3 Comparison of models with and without using multilingual data as input. The first model takes as input only video and English text captions whereas the second model takes video, English text and multilingual text captions as input. As shown in the figure, using multilingual text data as knowledge transfer significantly improved the performance.
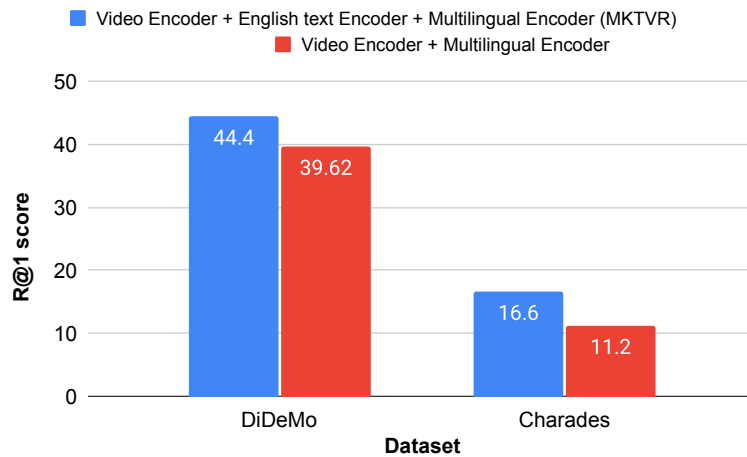

■ **Video Encoder + English text Encoder + Multilingual Encoder (MKTVR)**
■ **Video Encoder + Multilingual Encoder**

Fig. 4 Comparison of models consisting of only multilingual text encoder and multilingual text encoder + English text encoder. Using a separate English text encoder for encoding English text captions outperforms the model using multilingual text encoder to encode English text captions

question: *Why a separate English text encoder is required if multilingual text encoder can be used for both English text and multilingual text representations?* In Figure 4, we show the results of two different model variants. The first model uses a separate English text encoder to encode English text captions whereas in the second model, both the English text and multilingual text are encoded using the same multilingual text encoder. Results show that encoding English text captions using a separate
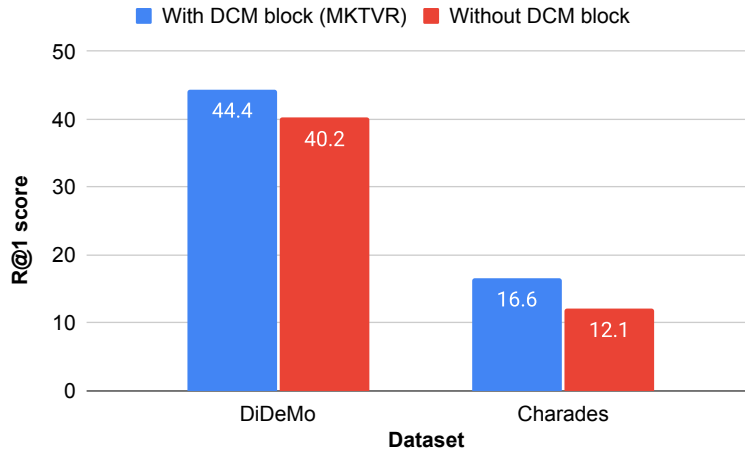
Fig. 5 Comparison of models with and without DCM block in the architecture. Using DCM block in the architecture showed superior performance to models without the DCM block.
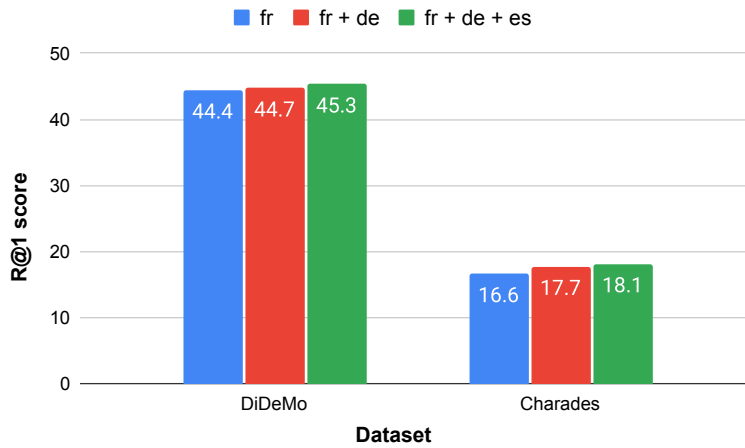

Fig. 6 Comparison of MuMUR trained with different multilingual caption data. It is evident from the figure that training with more languages improved the performance.

English text encoder surpasses the model using multilingual text encoder to encode both English text and multilingual text. Multilingual pretraining employs a part of English data whereas the English text encoder is pretrained on a comparatively larger English data. Hence, leveraging a separate English text encoder showed much superior performance to using multilingual text encoder for English text.

15
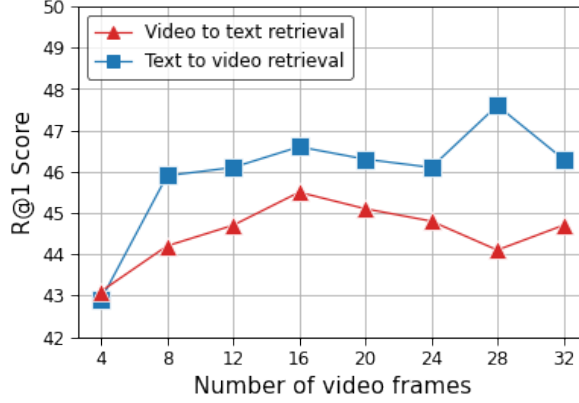
Fig. 7 Figure shows the effect of the number of video frames used to train MuMUR model on MSRVTT-9k dataset. Results demonstrate that R@1 score is the highest when 16 frames are used in text-to-video and video-to-text settings.
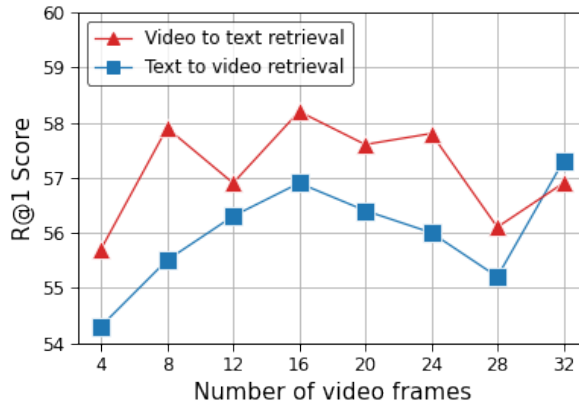

Fig. 8 Figure demonstrates the performance of MuMUR for varying number of video frames on MSVD dataset (single caption evaluation). It is evident that the R@1 score is maximum at 16 frames when evaluated in both text-to-video and video-to-text settings.

5.5.3 Effectiveness of dual cross encoder block

Next, we ablate the effectiveness of dual cross encoder block. We train a model without the DCM block and directly compute the loss between video representations and English text representations and video representations and multilingual text representations. From Figure 5, we can see that the model using DCM block achieves better performance than the model without the encoder block. This justifies our motivation to use DCM block in our model.
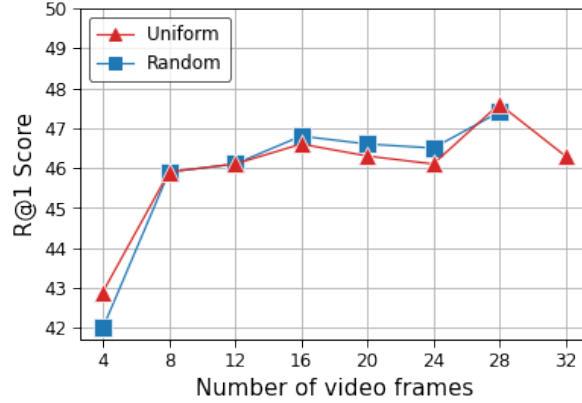
Fig. 9 We ablate the sampling strategy used for selecting video frames on MSRVTT-9k dataset. We observe that uniform and random sampling techniques achieve similar performance for all the video frames.
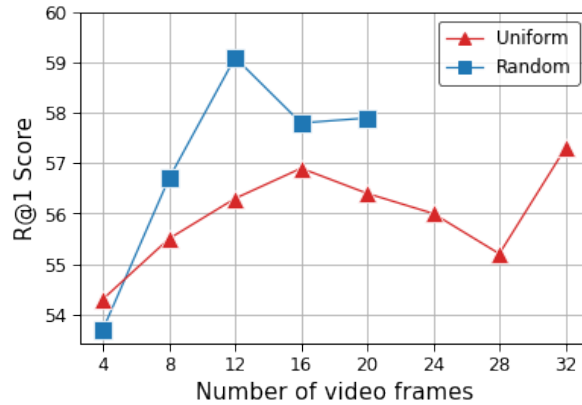

Fig. 10 Figure illustrates the comparative performance of random and uniform video sampling techniques on MSVD dataset. It is evident that random significantly outperforms uniform sampling but fails for larger number of video frames.

5.5.4 Training with more languages

Next, we ablate training our model with more than one language. Concretely we train our model with German (de) and Spanish (es) captions. These languages are chosen because their performance on XNLI dataset [13] is comparable to English. The results are shown in Figure 6 and it is seen that training with more languages improved the performance on video retrieval. These results validate that multilingual data act as an effective knowledge transfer mechanism for improving video retrieval.

17

5.5.5 Effect of video frames

We investigate the impact of the frequency of video frames on the retrieval performance. We train the MuMUR model with increasing number of video frames in intervals of 4. Figures 7 and 8 demonstrate the results of this ablation study on MSRVTT-9k and MSVD datasets respectively. As shown in the figure, the R@1 score is the highest when the model is trained with 16 video frames in both text-to-video and video-to-text retrieval settings. Therefore, these results motivate the use of 16 frames for training MuMUR on video retrieval datasets.

5.5.6 Effect of video frame sampling

Following, we study the impact of choosing video frames at random vs with an uniform manner. Figures 9 and 10 illustrate the results for varying video frames on MSRVTT-9k and MSVD datasets respectively. It is clear from the figures that uniform and random sampling strategies achieve nearly the same performance. However in case of MSVD, we observe that random sampling performed much better compared to uniform sampling. Nevertheless we see that random sampling fails for video frames greater than 20.

6 Conclusion

In this paper we introduced MuMUR, a multilingual knowledge transfer framework to improve the performance of multimodal retrieval. We constructed multilingual captions using off-the-shelf state-of-the-art machine translation models. We then proposed a CLIP-based model that enables multilingual knowledge transfer using a dual cross-modal encoder block. Experiment results on six standard multimodal retrieval datasets showed that our framework achieved state-of-the-art results on all the datasets. Finally, our model also showed superior performance to previous approaches on multilingual retrieval datasets in a zero-shot and fine-tune settings. In the future, we will focus on more efficient ways of multilingual knowledge transfer for multimodal retrieval.

7 Acknowledgment

8 Declarations

8.1 Ethical Approval

This declaration is not applicable.

8.2 Competing interests

The authors declare that they do not have competing interests.

8.3 Authors' contributions

A.M designed the model, performed major experiments and wrote the first draft. E.A helped in experiments and manuscript writing. G.B.M.S helped in experiments and manuscript writing. S.R helped in experiments and manuscript writing. S-Y.T helped manuscript writing and supervised the project. G.B supervised the project and reviewed the manuscript. V.L supervised the project and reviewed the manuscript.

8.4 Funding

This declaration is not applicable.

8.5 Availability of data and materials

The datasets used in this work are publicly accessible and code will be made public after paper acceptance.

References

[1] Amrani, E., Ben-Ari, R., Rotman, D., Bronstein, A.: Noise estimation using density estimation for self-supervised multimodal learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 6644–6652 (2021)

[2] Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Proceedings of the IEEE international conference on computer vision. pp. 5803–5812 (2017)

[3] Artetxe, M., Ruder, S., Yogatama, D.: On the cross-lingual transferability of monolingual representations

[4] Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1728–1738 (2021)

[5] Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. vol. 2, p. 4 (2021)

[6] Burns, A., Kim, D., Wijaya, D., Saenko, K., Plummer, B.A.: Learning to scale multilingual representations for vision-language tasks. In: European Conference on Computer Vision. pp. 197–213. Springer (2020)

[7] Cao, S., Wang, B., Zhang, W., Ma, L.: Visual consensus modeling for video-text retrieval (2022)

[8] Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3558–3568 (2021)

[9] Chen, D., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. pp. 190–200 (2011)

[10] Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations (2019)

[11] Cheng, X., Lin, H., Wu, X., Yang, F., Shen, D.: Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. arXiv preprint arXiv:2109.04290 (2021)

[12] Conneau, A., Lample, G.: Cross-lingual language model pretraining. Advances in neural information processing systems **32** (2019)

[13] Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., Stoyanov, V.: Xnli: Evaluating cross-lingual sentence representations. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2475–2485 (2018)

[14] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)

[15] Dzabraev, M., Kalashnikov, M., Komkov, S., Petiushko, A.: Mdmmt: Multidomain multimodal transformer for video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3354–3363 (2021)

[16] Elliott, D., Frank, S., Sima'an, K., Specia, L.: Multi30k: Multilingual english-german image descriptions. In: Proceedings of the 5th Workshop on Vision and Language. pp. 70–74 (2016)

[17] Fang, H., Xiong, P., Xu, L., Chen, Y.: Clip2video: Mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097 (2021)

[18] Gao, Z., Liu, J., Chen, S., Chang, D., Zhang, H., Yuan, J.: Clip2tv: An empirical study on transformer-based methods for video-text retrieval. arXiv preprint arXiv:2111.05610 (2021)

[19] Ge, Y., Ge, Y., Liu, X., Li, D., Shan, Y., Qie, X., Luo, P.: Bridging video-text retrieval with multiple choice questions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16167–16176 (2022)

[20] Ge, Y., Ge, Y., Liu, X., Wang, A.J., Wu, J., Shan, Y., Qie, X., Luo, P.: Miles: Visual bert pre-training with injected language semantics for video-text retrieval. arXiv preprint arXiv:2204.12408 (2022)

[21] Gella, S., Sennrich, R., Keller, F., Lapata, M.: Image pivoting for learning multilingual multimodal representations. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2839–2845 (2017)

[22] Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., Johnson, M.: Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In: International Conference on Machine Learning. pp. 4411–4421. PMLR (2020)

[23] Huang, P.Y., Patrick, M., Hu, J., Neubig, G., Metze, F., Hauptmann, A.G.: Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2443–2459 (2021)

[24] Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)

[25] Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3128–3137 (2015)

[26] Kim, D., Saito, K., Saenko, K., Sclaroff, S., Plummer, B.: Mule: Multimodal universal language embedding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11254–11261 (2020)

[27] Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. pp. 5583–5594. PMLR (2021)

[28] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017)

[29] Lei, J., Berg, T.L., Bansal, M.: Revealing single frame bias for video-and-language learning. arXiv preprint arXiv:2206.03428 (2022)

[30] Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., Liu, J.: 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. pp. 7331–7341 (2021)

[31] Li, D., Li, J., Li, H., Niebles, J.C., Hoi, S.C.: Align and prompt: Video-and-language pre-training with entity prompts. In: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition. pp. 4953–4963 (2022)

[32] Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)

[33] Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)

[34] Li, L., Chen, Y.C., Cheng, Y., Gan, Z., Yu, L., Liu, J.: Hero: Hierarchical encoder for video+ language omni-representation pre-training. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2046–2065 (2020)

[35] Lin, Y.B., Lei, J., Bansal, M., Bertasius, G.: Eclipse: Efficient long-range video retrieval using sight and sound. arXiv preprint arXiv:2204.02874 (2022)

[36] Liu, S., Fan, H., Qian, S., Chen, Y., Ding, W., Wang, Z.: Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11915–11925 (2021)

[37] Liu, Y., Albanie, S., Nagrani, A., Zisserman, A.: Use what you have: Video retrieval using representations from collaborative experts. arXiv preprint arXiv:1907.13487 (2019)

[38] Liu, Y., Xiong, P., Xu, L., Cao, S., Jin, Q.: Ts2-net: Token shift and selection transformer for text-video retrieval. arXiv preprint arXiv:2207.07852 (2022)

[39] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)

[40] Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S.: 12-in-1: Multi-task vision and language representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10437–10446 (2020)

[41] Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860 (2021)

[42] Madasu, A., Aflalo, E., Ben Melech Stan, G., Tseng, S.Y., Bertasius, G., Lal, V.: Improving video retrieval using multilingual knowledge transfer. In: European Conference on Information Retrieval. pp. 669–684. Springer (2023)

[43] Madasu, A., Oliva, J., Bertasius, G.: Learning to retrieve videos by asking questions. arXiv preprint arXiv:2205.05739 (2022)

[44] Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2630–2640 (2019)

[45] Ni, M., Huang, H., Su, L., Cui, E., Bharti, T., Wang, L., Zhang, D., Duan, N.: M3p: Learning universal representations via multitask multilingual multimodal pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3977–3986 (2021)

[46] Patrick, M., Huang, P.Y., Asano, Y., Metze, F., Hauptmann, A.G., Henriques, J.F., Vedaldi, A.: Support-set bottlenecks for video-text representation learning. In: International Conference on Learning Representations (2020)

[47] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015)

[48] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)

[49] Sigurdsson, G.A., Alayrac, J.B., Nematzadeh, A., Smaira, L., Malinowski, M., Carreira, J., Blunsom, P., Zisserman, A.: Visual grounding in video for unsupervised word translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10850–10859 (2020)

[50] Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European Conference on Computer Vision. pp. 510–526. Springer (2016)

[51] Surís, D., Epstein, D., Vondrick, C.: Globetrotter: Connecting languages by connecting images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16474–16484 (2022)

[52] Tang, Y., Tran, C., Li, X., Chen, P.J., Goyal, N., Chaudhary, V., Gu, J., Fan, A.: Multilingual translation with extensible multilingual pretraining and finetuning. arXiv preprint arXiv:2008.00401 (2020)

[53] Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence-video to text. In: Proceedings of the IEEE international conference on computer vision. pp. 4534–4542 (2015)

[54] Wang, A.J., Ge, Y., Yan, R., Ge, Y., Lin, X., Cai, G., Wu, J., Shan, Y., Qie, X., Shou, M.Z.: All in one: Exploring unified video-language pre-training. arXiv

preprint arXiv:2203.07303 (2022)

[55] Wang, J., Ge, Y., Cai, G., Yan, R., Lin, X., Shan, Y., Qie, X., Shou, M.Z.: Object-aware video-language pre-training for retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3313–3322 (2022)

[56] Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 394–407 (2018)

[57] Wehrmann, J., Souza, D.M., Lopes, M.A., Barros, R.C.: Language-agnostic visual-semantic embeddings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5804–5813 (2019)

[58] Wu, H.Y., Zhai, A.: Classification is a strong baseline for deep metric learning. In: Sidorov, K., Hicks, Y. (eds.) Proceedings of the British Machine Vision Conference (BMVC). pp. 224.1–224.12. BMVA Press (September 2019). https://doi.org/10.5244/C.33.224, https://dx.doi.org/10.5244/C.33.224

[59] WU SJ, D.M.: The surprising cross-lingual effectiveness of bert. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China. pp. 833–844 (2019)

[60] Xu, H., Ghosh, G., Huang, P.Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: Videoclip: Contrastive pre-training for zero-shot video-text understanding. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 6787–6800 (2021)

[61] Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5288–5296 (2016)

[62] Xu, X., Li, B., Wu, C., Tseng, S.Y., Bhiwandiwalla, A., Rosenman, S., Lal, V., Che, W., Duan, N.: ManagerTower: Aggregating the insights of uni-modal experts for vision-language representation learning. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 14507–14525. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.811, https://aclanthology.org/2023.acl-long.811

[63] Xu, X., Wu, C., Rosenman, S., Lal, V., Che, W., Duan, N.: Bridgetower: Building bridges between encoders in vision-language representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 10637–10647 (2023)

[64] Yu, Y., Kim, J., Kim, G.: A joint sequence fusion model for video question answering and retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 471–487 (2018)

[65] Zhang, B., Hu, H., Sha, F.: Cross-modal and hierarchical modeling of video and text. In: Proceedings of the european conference on computer vision (ECCV). pp. 374–390 (2018)

[66] Zhou, M., Zhou, L., Wang, S., Cheng, Y., Li, L., Yu, Z., Liu, J.: Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4155–4165 (2021)

[67] Zhu, L., Yang, Y.: Actbert: Learning global-local video-text representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8746–8755 (2020)