Animating Images to Transfer CLIP for Video-Text Retrieval

Yu Liu*
DAMO Academy, Alibaba Group
ly103369@alibaba-inc.com

Huai Chen*
Shanghai Jiao Tong University
chenhuai@sjtu.edu.cn

Lianghua Huang
DAMO Academy, Alibaba Group
xuangen.hlh@alibaba-inc.com

Di Chen
DAMO Academy, Alibaba Group
guangpan.cd@alibaba-inc.com

Bin Wang
DAMO Academy, Alibaba Group
ganfu.wb@alibaba-inc.com

Pan Pan
DAMO Academy, Alibaba Group
panpan.pp@alibaba-inc.com

Lisheng Wang
Shanghai Jiao Tong University
lswang@sjtu.edu.cn

ABSTRACT

Recent works show the possibility of transferring the CLIP (Contrastive Language-Image Pretraining) model for video-text retrieval with promising performance. However, due to the domain gap between static images and videos, CLIP-based video-text retrieval models with interaction-based matching perform far worse than models with representation-based matching. In this paper, we propose a novel image animation strategy to transfer the image-text CLIP model to video-text retrieval effectively. By imitating the video shooting components, we convert widely used image-language corpus to synthesized video-text data for pretraining. To reduce the time complexity of interaction matching, we further propose a coarse to fine framework which consists of dual encoders for fast candidates searching and a cross-modality interaction module for fine-grained re-ranking. The coarse to fine framework with the synthesized video-text pretraining provides significant gains in retrieval accuracy while preserving efficiency. Comprehensive experiments conducted on MSR-VTT, MSVD, and VATEX datasets demonstrate the effectiveness of our approach.

CCS CONCEPTS

• **Information systems** → **Retrieval models and ranking**.

KEYWORDS

Video-Text Retrieval, CLIP, Image Animation, Coarse to Fine Retrieval

ACM Reference Format:
Yu Liu, Huai Chen, Lianghua Huang, Di Chen, Bin Wang, Pan Pan, and Lisheng Wang. 2022. Animating Images to Transfer CLIP for Video-Text Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3477495.3531776

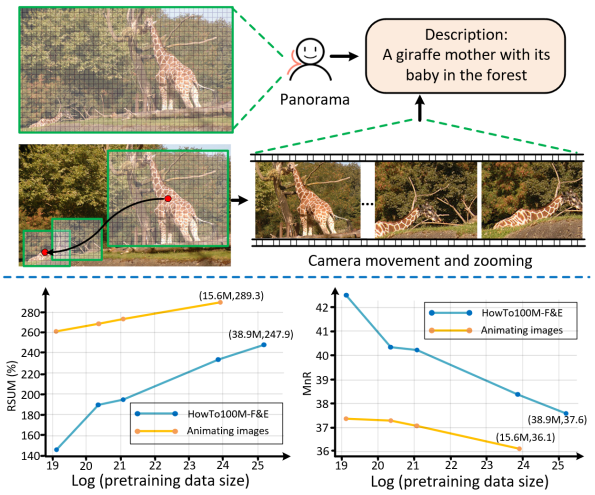

Figure 1: Top: the illustration of our image animation method. Bottom: the zero-shot retrieval results of our model on MSR-VTT 1k-A when pretrained with HowTo100M-F&E and our synthesized video-text data. We plot the RSUM (higher is better) and the MnR (lower is better) metrics in the y-axes respectively, and the x-axes show the pretraining data size using log scale. Obviously, our method shows superiority with significantly higher RSUM and lower MnR, and exhibits promising scaling property.

1 INTRODUCTION

Video-text retrieval aims to return target videos based on text descriptions. This task has received considerable attentions with the increasing amount of web videos and the rapid developments of multimedia technologies [1, 14, 17, 20, 21, 24, 36, 38]. Various methods have been proposed including multi-level encoding [5], global-local matching [4, 35], multi-modality fusion [9, 18, 25], knowledge distillation [21], hierarchical contrastive matching [17], and video-text pretraining [1, 19, 38]. From the perspective of the model architecture, existing retrieving methods can be roughly categorized into two classes: 1) representation-based matching [5, 7, 9, 17, 18, 25, 27,

*Both authors contributed equally.

SIGIR '22, July 11–15, 2022, Madrid, Spain.

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531776

35] and 2) interaction-based matching [4, 13, 19, 32]. For approaches in the first category, the visual and textual inputs are separately projected by dual encoders into the same embedding space where similarities can be computed by cosine distance efficiently. On the other hand, interaction-based approaches leverage a single-stream architecture to fuse visual and textual representations at the early stage, which produces accurate relevance estimation, but is slow and impractical for large-scale cross-modal retrieval.

Another key challenge of video-text retrieval is the lacking of large-scale video datasets with annotated texts [23], promoting deep research in utilizing pretrained models [1, 14, 19, 20, 22, 38, 42]. In the image-text domain, the CLIP model [29] learns aligned visual and textual representations by utilizing contrastive objectives on 400 million noisy image-text pairs. Recent works [8, 20, 28] have shown that it is possible to transfer image-text pretrained CLIP model for video-text retrieval. However, we find that CLIP-based video retrieval models with interaction-based matching (*e.g.* CLIP4Clip-tightTransf [20]) perform far worse than models with representation-based matching (*e.g.* CLIP2Video [8], CLIP4Clip-meanP [20]), which goes against our previous knowledge of the performance advantage of interaction-based methods in other retrieval tasks [13, 15, 19, 41]. We argue that the root cause comes from the transferring learning process. Due to the domain gap between static images and videos, it is inevitable for CLIP-based video-text retrieval methods to introduce additional modules, such as a video frames aggregator and a cross-modality model, to handle temporal correlations and video-text interactions. However, as pointed out by [20], it is hard to learn the cross-modality interaction without enough video-text dataset.

A straightforward way to solve this problem is to utilize a large-scale video-text dataset (*e.g.* HowTo100M [24], WebVid-2M [1]) for new introduced cross-modality modules pretraining. However, existing video-text datasets are either noisy or relatively small compared with commonly used image-text datasets [2, 29]. To this end, we propose a novel image animation approach to convert image-language corpus to synthesized video-text data for new modules pretraining. Specifically, by mimicking the shooting components (*e.g.* camera movement, camera zooming, and shoot scene changing), the generated video clips combined with the corresponding original texts easily compose large-scale video-text pairs. Besides single-view video by animating a single image, we also utilize multiple images to construct multi-view video clips which simulates the scene changes of natural videos. Benefiting from the image-text data source, the synthesized video has highly aligned caption. By pretraining the cross-modality modules of CLIP4Clip-tightTransf with the synthesized video-text data, we successfully boost the original mediocre performance by a large margin, surpassing the performance of state-of-the-art (SOTA) methods.

To reduce the inference cost of the pairwise interaction-based matching, we further propose a coarse to fine retrieval framework which combines representation-based dual encoders for fast candidates searching and a interaction-based cross-attention module for fine-grained re-ranking. The coarse-grained stage includes a video frame encoder and a text encoder which are instantiated from a pretrained CLIP [29] model. We adopt Multiple Instance Learning (MIL) to cope with the temporal misalignment between the video frames and sentences. The fine-grained stage is equipped with a new cross-attention module using low-level features for fine-grained cross-modality interaction modeling. The coarse to fine model with the synthesized video-text data pretraining provides significant gains in retrieval accuracy while preserving efficiency.

The contributions of this work are threefold: 1) To the best of our knowledge, we present the first approach that transfers image-language domain knowledge for video-text retrieval by explicitly converting static images to dynamic *videos*; 2) We propose a simple image animation method to synthesize video-text data for pretraining, which shows performance superiority and promising scaling property; 3) We present a coarse to fine retrieval framework which preserves both efficiency and accuracy. Our method achieves absolute RSUM gains of 2.4% $\sim$ 10.5% for text-to-video retrieval, and 5.4% $\sim$ 9.1% for video-to-text retrieval over previous SOTA works on MSR-VTT [39], MSVD [3], and VATEX [34] datasets.

2 METHOD

This work aims to solve the accuracy and efficiency issues of the CLIP-based, interaction-based video-text retrieval model by two novel mechanisms: 1) an image animation strategy as illustrated in Figure 1 for the pretraining of newly-introduced modules to boost the transferring performance, and 2) a coarse to fine framework as shown in Figure 2 to improve accuracy and preserve efficiency.

2.1 Animating Images

The purpose of animating images is to generate synthesized video-text pairs using the image-text dataset. We generate single-view videos by simulating a videographer's behaviors (*e.g.* camera movement, camera zooming in and zooming out) and synthesize multi-view videos by mimicking the scene changes of natural videos.

Single-view video-text synthesis. Given an image-text pair (x_i, t_i), the image animation generator $G(\cdot)$ will first randomly generate r focuses $\{p_1, p_2, ..., p_r\}$ in image x_i to identify the positions of key frames. Then in each focus, a square bounding box with random side length will be set to determine the focused area. To simulate the smooth camera movement, d moving frames, whose center position and side length are linearly interpolated between two continual focuses from image x_i, are generated. And it is worth noting that all the moving frames and key frames are resized to the same shape, which simulates the zooming effect. As illustrated in the top row of Figure 1, the synthesizing process can be formulated as:

$$v'_i = G(x_i) = \{v'_{i,1}, v'^{1}_{i,1,2}, , v'^{2}_{i,1,2}, ..., v'^{d}_{i,1,2}, v'_{i,2}, ..., v'_{i,r}\}, \quad (1)$$

where $v'_{i,r}$ means the r-th key frame identified by the focus p_r, and $v'^{d}_{i,r-1,r}$ denotes the d-th moving frame between $v'_{i,r-1}$ and $v'_{i,r}$. The synthesized v'_i combined with t_i forms the video-text pair (v'_i, t_i).

Multi-view video-text synthesis. Natural videos usually contain multiple scenes and can be explained by different descriptions. Furthermore, adding noises in the type of different views is helpful for the cross-attention module to learn robust multimodal interaction. Therefore, we further propose a multi-view video generation method where multiple image-text pairs are provided. For simplicity, we only describe the generator for combining two image-text pairs and it can be easily extended to create samples based on more than two image-text pairs. Besides (x_i, t_i), another image-text pair

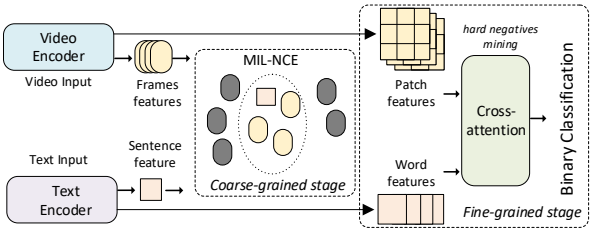

Figure 2: A brief overview of our coarse to fine framework.

(x_j, t_j) is provided. Following the description of single-view situation, the synthesized videos v'_i and v'_j can be firstly generated. Then the multi-view video based on these two images can be created by simply combining corresponding synthesized video clips, *i.e.*:

$$\begin{aligned} v'_{i,j} &= G(x_i, x_j) = [G(x_i), G(x_j)] \\ &= \{v'_{i,1}, v'^{1}_{i,1,2}, ..., v'_{i,r}, v'_{j,1}, v'^{1}_{j,1,2}, ..., v'_{j,r}\} \end{aligned} \tag{2}$$

where $[\cdot]$ means concatenating video clips. And the final generated video-text pairs can be defined as $\{(v', t)|v' \in \{v'_{i,j}, v'_{j,i}\}, t \in \{t_i, t_j\}\}$. Similarly, when providing more than two image-text pairs, the generated multi-view video is the combination of corresponding single-view synthesized videos, while one of the original image captions is chosen randomly as the text.

In practice, we take the CLIP4Clip-tightTransf [20] as our baseline model. Utilizing image animation to generate synthesized video-text dataset for cross-attention module pretraining, we successfully boost the original mediocre CLIP4Clip-tightTransf by a large margin, surpassing the SOTA retrieval performance.

2.2 Coarse to Fine Retrieval

CLIP4Clip-tightTransf is slow at test due to the pairwise cross-modality matching, limiting its application in practice. To this end, we further propose a coarse to fine framework to combine dual encoders' ability of fast candidates searching and cross-attention's ability of accurate multimodal interaction modeling.

The overall architecture is constructed with a text encoder $e_t(\cdot)$, a video frame encoder $e_v(\cdot)$ and a cross-attention module $h(\cdot)$. We extract video frame sequence $v_i = \{v_i^1, v_i^2, ..., v_i^{|v_i|}\} \in \mathcal{V}$ as input, and $e_v(\cdot)$ will output frame representations $\mathbf{F}_i = \{\mathbf{f}_i^1, \mathbf{f}_i^2, ..., \mathbf{f}_i^{|v_i|} | \mathbf{f}_i^j \in \mathbb{R}^C\}$ and temporal-spatial patch features $\mathbf{Z}_i = \{\mathbf{z}_i^1, \mathbf{z}_i^2, ..., \mathbf{z}_i^{|v_i|} | \mathbf{z}_i^j \in \mathbb{R}^{H\times W\times C}\}$. As for text containing tokens $t_i = \{t_i^1, t_i^2, ..., t_i^{|t_i|}\} \in \mathcal{T}$, we use $e_t(\cdot)$ to extract global sentence representation $\mathbf{g}_i \in \mathbb{R}^C$ and local word features $\mathbf{W}_i = \{\mathbf{w}_i^1, \mathbf{w}_i^2, ..., \mathbf{w}_i^{|t_i|} | \mathbf{w}_i^j \in \mathbb{R}^C\}$.

Coarse-grained Stage: After extracting the representations $\mathbf{F}_i$ and $\mathbf{g}_i$, the coarse-grained stage tries to calculate the video-text similarity efficiently without introducing new parameters, *i.e.* by a *parameter-free* manner. The natural method is to aggregate the features of all frames by mean pooling to generate the video representation as done in [20]. However, videos are composed of diverse views and not all frames are correctly aligned with the corresponding sentences. Therefore, we introduce MIL-NCE loss [22] which includes Multiple Instance Learning (MIL) and Noise Contrastive Estimation (NCE) to copy with the temporal misalignment between the video frames and sentences. Specifically, given a batch of B video-text pairs, the text-to-video *MIL-NCE* loss is applied:

$$\begin{aligned} \mathcal{L}^c_{t2v} &= -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\sum_{k=1}^{|v_i|} e^{\cos(\mathbf{f}_i^k, \mathbf{g}_i)/\tau}}{\sum_{j=1}^{B}\sum_{k=1}^{|v_j|} e^{\cos(\mathbf{f}_j^k, \mathbf{g}_i)/\tau}}, \\ \mathcal{L}^c_{v2t} &= -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\sum_{k=1}^{|v_i|} e^{\cos(\mathbf{f}_i^k, \mathbf{g}_i)/\tau}}{\sum_{j=1}^{B}\sum_{k=1}^{|v_i|} e^{\cos(\mathbf{f}_i^k, \mathbf{g}_j)/\tau}}, \\ \mathcal{L}^c &= \mathcal{L}^c_{t2v} + \mathcal{L}^c_{v2t} \end{aligned} \tag{3}$$

where $\cos(\mathbf{f}_i^k, \mathbf{g}_j) = ((\mathbf{f}_i^k)^{\mathrm{T}}\mathbf{g}_j)/(\|\mathbf{f}_i^k\|_2 \cdot \|\mathbf{g}_j\|_2)$ denotes the cosine similarity between $\mathbf{f}_i^k$ and $\mathbf{g}_j$, τ is a temperature adjusting the scale of cosine similarities, and L^c is the overall coarse-grained loss.

Fine-grained stage: The fine-grained stage predicts whether a pair of video and text candidate is matched in a more detailed granularity. Specifically, we concatenate the temporal-spatial patch features $\mathbf{Z_i}$ and word features $\mathbf{W_i}$ together, and then feed them into a new multimodal cross-attention Transformer Encoder [33] followed by a fully-connect (FC) layer to make the final prediction. The predicted similarity $s_f(\cdot)$ is defined as follows:

$$s_f(v_i, t_j) = h(\mathbf{Z}_i, \mathbf{W}_j) = g_u(g_t([\mathbf{Z}_i, \mathbf{W}_j] + \mathbf{P} + \mathbf{T})) \tag{4}$$

where $[\mathbf{Z}_i, \mathbf{W}_j]$ means concatenated features of video and text, $\mathbf{P}$ is the positional embedding, $\mathbf{T}$ is the data type embedding, g_t is the cross-attention blocks and g_u is a FC layer to get the final similarity score. The binary cross entropy loss $\mathcal{L}^f$ is applied to classify the positive and negative video-text pairs. In practice, we adopt a hard negatives sampling strategy according to the dual encoders' similarity score distribution for the classification task.

The fine-grained stage contains newly-introduced parameters, which contradicts the well pretrained dual encoders inherited from CLIP [29] model in the coarse-grained stage. We pretrain the cross-attention modules with the generated video-text pairs, and then finetune the whole network on the downstream video-text retrieval datasets. During evaluation, top-K candidates are retrieved from the coarse-grained stage, and the similarity scores are refined by weighted outputs of the two models. More precisely, given a text query t and the video database $\mathcal{V}$ with $|\mathcal{V}|$ videos, we first compute the coarse-grained similarity between t and any video $v_i \in \mathcal{V}$ by $s_c(v_i, t) = \sum_{k=1}^{|v_i|} e^{\cos(\mathbf{f}_i^k, \mathbf{g})/\tau}$, and then re-rank the top-K videos $\mathcal{V}_K$ as:

$$\arg\max_{v_i \in \mathcal{V}_K} s_c(v_i, t) + \beta s_f(v_i, t) \tag{5}$$

where β is a scalar hyper-parameter that weights the outputs of the coarse-grained and fine-grained models.

3 EXPERIMENTS

3.1 Experiments Settings

Datasets and metrics. As for the pretraining datasets, we construct synthesized video-text data using the collections of five image-text datasets including COCO [16], Visual Genome [12], SBU Captions [26], Conceptual 3M [31], and Conceptual 12M [2]. The total amount of synthesized video-text pairs is about 15.6M. We also conduct comparative experiments using WebVid-2M [1] and HowTo100M [24]. We follow the settings in [20] to use the 'Food and Entertaining' category which contains around 38.9M clip-text pairs (called

HowTo100M-F&E in this paper). We verify our method on three video-text retrieval datasets: MSR-VTT [39], MSVD [3], and VATEX [34]. We report the retrieval result on both the 1k-A split [9, 24, 40] and full split [7, 19] of MSR-VTT. We adopt the standard retrieval metrics: recall at top-K (*abbr.* R@K), RSUM[37], which is defined as the sum of recalls at K = 1, 5, 10, median rank (*abbr.* MdR), and mean rank (*abbr.* MnR). The higher recall metrics are better, while the lower rank metrics are better.

Networks. The video frame encoder is a vision transformer [6] initialized by ViT-B/32 [29] with 12 layers and the input frame size is 224×224. We use the output from the [*CLS*] as the frame feature and outputs of all the visual tokens as the patch features. We apply the text encoder in CLIP [29] as text feature extractor which is a Transformer [33] with 12 layers and 512 hidden channels. We use the embedding from the [*EOS*] token as the global sentence representation, and take every tokens' outputs as word features. The cross-attention module in the fine-grained stage is a 6-layers, 8-heads and 512-channels Transformer [33].

Image animation settings. We use the multi-view video-text generator by default with the scenes number ranging from 1 to 3. For each single view, we randomly set the focuses number ranging from 1 to 4, meaning no camera moving at least and three camera movements at most. Between two continual focuses, $6 \sim 10$ moving frames are sampled. All the cropped areas are resized to 224×224.

Pretraining and finetuning settings. We pretrain models for 15 epochs with a total batch size of 640. We use Adam optimizer with the weight decay of 0.05. We set the initial learning rate as 1e-3 and a cosine annealing rule is applied to decay the learning rate smoothly. As for the finetuning, we set the initial learning rate as 1e-5 for dual encoders and 3e-4 for the cross-attention module. The other settings, *e.g.* the caption token length and video frame length *etc.* follow [20]. We set the temperature in Equation 3 as $\tau = 0.01$, and the hyper-parameter K during evaluation as 15.

3.2 Ablation Study

Ablation: effectiveness of animating images for pretraining. In order to evaluate the effectiveness of image animation, we compare our method with the two conventional video synthesis approaches: 1) 'Static Video', *i.e.* repeating the same image over time, and 2) 'Random Video', where *video frames* are generated by random crops of the static image. We also include two popular real video-text datasets: HowTo100M-F&E [24] and WebVid-2M [1] for comparison. As for the image-text data source, we combine COCO [16] and Visual Genome [12] datasets, leading to 1.3M synthesized video-text pairs. As for the HowTo100M-F&E and WebVid-2M datasets, we randomly sample the same data volume of 1.3M video-text pairs to facilitate comparison. We also test our approach by exploring the whole 15.6M image-text pairs. Table 1 shows the zero-shot and finetuning retrieval performance of CLIP4Clip-tightTransf [20] model on MSR-VTT 1k-A split when pretraining the cross-attention module by different video-text datasets. We observe that: 1) Our animated video-text data has clear advantages over the conventional approaches, indicating the necessity of camera movement simulation; 2) With the same data volume, the performance of our method surpasses that using HowTo100M-F&E and lags behind that utilizing WebVid-2M. The reason may be that HowTo100M is noisy and our synthesized videos cannot cover all the motion patterns of real videos; 3) But when we sale up the data size, our method can compensate for the limitation and re-exceed the other approaches by a large margin. Considering the easy accessibility of much larger-scale image-text datasets [10, 30], our method has promising applications.

Table 1: Zero-shot and finetuning performance of CLIP4Clip-tightTransf [20] on MSR-VTT 1k-A pretrained with different video-text datasets.

Pretraining datasets	# of V-T Pairs	T2V R@1 ↑	T2V MnR ↓	V2T R@1 ↑	V2T MnR ↓
Zero-shot					
HowTo100M-F&E	1.3M	15.7	82.4	16.2	94.1
WebVid-2M	1.3M	27.7	41.0	26.9	39.5
Static Video	1.3M	21.9	48.1	19.8	61.2
Random Video	1.3M	23.3	48.1	21.8	53.0
Animated Video	1.3M	24.3	47.1	23.7	50.7
	15.6M	**32.4**	**32.8**	**27.4**	**37.6**
Finetuning					
Baseline [20]	-	40.2	13.4	40.6	13.6
HowTo100M-F&E	1.3M	42.3	14.9	41.0	14.4
WebVid-2M	1.3M	43.6	14.5	43.2	12.3
Static Video	1.3M	42.0	18.0	42.7	17.5
Random Video	1.3M	43.1	16.6	41.1	18.9
Animated Video	1.3M	43.3	13.7	42.7	11.7
	15.6M	**45.6**	**13.0**	**45.8**	**9.4**

Table 2: Part 1: Comparison of retrieval performance on MSR-VTT 1k-A with different architectures; Part 2: Comparison of single-view *vs.* multi-view image animation.

Method	T2V R@1 ↑	T2V MnR ↓	V2T R@1 ↑	V2T MnR ↓	M-Qt
Part 1: Different architectures comparison.					
Coarse-grained (C-meanP [20])	43.1	16.2	43.1	12.4	46ms
Fine-grained (C-tightTransf [20])	40.2	13.4	40.6	13.6	37,899ms
Coarse-grained (Ours)	43.2	14.4	45.4	9.3	113ms
Fine-grained (Ours)	45.6	**13.0**	45.8	9.4	37,899ms
Coarse to fine (Ours)	**47.2**	13.9	**46.3**	**8.8**	6,737ms
Part 2: Single-view *vs.* multi-view image animation.					
Single-view	31.6	36.8	29.5	34.7	-
Multi-view	33.1	36.5	31.1	34.1	-

Ablation: effectiveness of coarse to fine framework. The first part of table 2 compares performances of different retrieval architectures on MSR-VTT 1k-A, where CLIP4Clip-meanP [20] and CLIP4Clip-tightTransf [20] are baselines. Our coarse-grained approach adopts the same model architecture as CLIP4Clip-meanP, but replaces the original NCE loss with MIL-NCE loss, which can deal with misaligned video frames with the corresponding caption. Our fine-grained model pretrains the cross-attention module of CLIP4Clip-tightTransf with synthesized video-text data, instead of reusing similar weights from CLIP encoders [8, 20]. In our final coarse to fine model, we replace the original global frame and caption representation with local patch and word features for the detailed cross-modality interaction, and we also explore online hard negative mining to compute the cross-entropy loss. We compute the query time of video retrieval on MSR-VTT 1k-A (*abbr.* M-Qt) using 1×V100 GPU. We find that, our coarse to fine framework gets remarkable 7.0% R@1 gains for text-to-video retrieval and accelerates the inference by up to 5.6× compared with CLIP4Clip-tightTransf.

Table 3: Comparison with SOTA approaches on MSR-VTT, MSVD and VATEX datasets.

		Text ⇒ Video						Video ⇒ Text					
Dataset	Method	R@1 ↑	R@5 ↑	R@10 ↑	RSUM ↑	MdR ↓	MnR ↓	R@1 ↑	R@5 ↑	R@10 ↑	RSUM ↑	MdR ↓	MnR ↓
MSR-VTT 1k-A	CE [18]	20.9	48.8	62.4	132.1	6.0	28.2	20.6	50.3	64.0	134.9	5.3	25.1
	MMT [9]	26.6	57.1	69.6	153.3	4.0	24.0	27.0	57.5	69.7	154.2	3.7	21.3
	SSB [27]	27.4	56.3	67.7	151.4	3.0	-	26.6	55.1	67.5	149.2	3.0	-
	FROZEN [1]	31.0	59.5	70.5	161.0	3.0	-	-	-	-	-	-	-
	CLIP4Clip-meanP [20]	43.1	70.4	80.8	194.3	2.0	16.2	43.1	70.5	81.2	194.8	2.0	12.4
	CLIP4Clip-seqTransf [20]	44.5	71.4	81.6	197.5	2.0	15.3	42.7	70.9	80.6	194.2	2.0	11.6
	CLIP2Video [8]	45.6	72.6	81.7	199.9	2.0	14.6	43.5	72.3	82.1	197.9	2.0	10.2
	Ours	**47.2**	**73.0**	**82.8**	**203.0**	**2.0**	**13.9**	**46.3**	**74.1**	**84.8**	**205.2**	**2.0**	**8.8**
MSR-VTT Full	CE [18]	10.0	29.0	41.2	80.2	16.0	86.8	15.6	40.9	55.2	111.7	8.3	38.1
	HT [24]	14.9	40.2	52.8	107.9	9.0	-	-	-	-	-	-	-
	UNiVL [19]	21.2	49.6	63.1	133.9	6.0	-	-	-	-	-	-	-
	CLIP2Video [8]	29.8	55.5	66.2	151.5	4.0	45.5	54.6	82.1	90.8	227.5	1.0	5.3
	Ours	**32.0**	**58.2**	**68.6**	**158.8**	**3.0**	**44.1**	**61.3**	**82.4**	**90.9**	**234.6**	**1.0**	**5.2**
MSVD	VSE [11]	12.3	30.1	42.3	84.7	14.0	-	34.7	59.9	70.0	164.6	3.0	-
	CE [18]	19.8	49.0	63.8	132.6	6.0	23.1	-	-	-	-	-	-
	SSB [27]	28.4	60.0	72.9	161.3	4.0	-	-	-	-	-	-	-
	FROZEN [1]	33.7	64.7	76.3	174.7	3.0	-	-	-	-	-	-	-
	CLIP4Clip-meanP [20]	46.2	76.1	84.6	206.9	2.0	10.0	56.6	79.7	84.3	220.6	1.0	7.6
	CLIP2Video [8]	47.0	76.8	85.9	209.7	2.0	9.6	58.7	85.6	91.6	235.9	1.0	**4.3**
	Ours	**47.5**	**78.0**	**86.6**	**212.1**	**2.0**	**9.3**	**70.2**	**88.1**	**92.7**	**251.0**	**1.0**	6.0
VATEX	Dual Enc. [5]	31.1	67.5	78.9	177.5	3.0	-	-	-	-	-	-	-
	HGR [4]	35.1	73.5	83.5	192.1	2.0	-	-	-	-	-	-	-
	SSB [27]	44.9	82.1	89.7	216.7	1.0	-	58.4	84.4	91.0	233.8	1.0	-
	C-seqTransf [20]	55.9	89.2	95.0	240.1	1.0	3.9	73.2	97.1	99.1	269.4	1.0	1.7
	CLIP2Video [8]	57.3	90.0	95.5	242.8	1.0	3.6	76	**97.7**	**99.9**	273.6	1.0	1.5
	Ours	**64.7**	**92.3**	**96.3**	**253.3**	**1.0**	**3.1**	**82.5**	97.3	99.2	**279.0**	**1.0**	**1.5**

Ablation: multi-view *vs.* single-view synthesized video-text pretraining. The second part of table 2 shows the zero-shot retrieval performance of our coarse to fine model on MSR-VTT 1k-A, when pretraining the cross-attention module using 1.3M single-view synthesized video-text pairs and multi-view video-text pairs respectively. Clearly, the multi-view video synthesis has obvious advantages where it simulates the scene changes of natural videos.

Ablation: scaling property of animating images. To verify the scaling ability of our image animation method, we start from COCO[16] dataset, and then accumulate the image-text pairs by gradually adding Visual Genome[12], SBU Captions[26], and Conceptual dataset[2, 31]. To facilitate comparison, we sample video-text pairs randomly from HowTo100M-F&E and scale up the data size according to the growing schedule of image animation. The bottom row of Figure 1 visualizes the comparison, where we show the sum of all the R@K, and the average MnR. We observe that: 1) both data types show good scaling property, 2) our animating image method has clear advantages, where the RSUM of pretrained model with 1.3M synthesized video-text pairs outperforms that using 38.9M HowTo100M-F&E video-text pairs by a large margin. And when scaling up our data to 15.6M, the absolute gain of RSUM compared with HowTo100M-F&E reaches at 41.4%, showing promising scaling property.

3.3 Comparison with State of the Art

Table 3 compares our work with previous video-text retrieval methods on MSVD, MSR-VTT, and VATEX datasets. Our full coarse to fine framework with synthesized video-text pretraining achieves absolute RSUM gains of 2.4% ∼ 10.5% for text-to-video retrieval, and 5.4% ∼ 9.1% for video-to-text retrieval over previous SOTA results. Among all the methods, CLIP4Clip [20], CLIP2Video [8] and our method can be categorized into the same class which transfers image-text pretrained CLIP [29] model for video-text retrieval. These CLIP-based methods surpass the other non-CLIP-based approaches significantly, indicating the benefits of large scale image-text pretraining for video-text retrieval. Due to the domain gap between static images and videos, these CLIP-based methods usually introduce additional modules, such as a video frames aggregator and a cross-attention model, to handle temporal correlations and video-text interactions. Existing approaches either train those new uninitialized weights from scratch [20] or reuse similar parameters from CLIP encoders [8, 20], which may yield suboptimal results. Our method converts widely used image-text corpus to synthesized video-text data by image animation for new introduced modules pretraining. Our image animation strategy has three main advantages: 1) providing a simple alternative to collecting annotated videos; 2) offering an explicit method to transfer image-language domain knowledge to video-related tasks; 3) benefiting from the data source, the synthesized video has highly aligned captions and is distributed in the open domain (*vs.* HowTo100M [24] in the instructional domain). All the consistent improvements across different benchmarks verify the effectiveness of our framework and the image animation strategy.

4 CONCLUSION

This paper solves the accuracy and efficiency issues of the CLIP-based, interaction-based video-text retrieval model with a novel image animation method for transferring learning and a coarse to fine framework to improve accuracy and efficiency. In the future, we would like to scale up the animated images and explore the limits of large-scale noisy image-text corpus [10] for video-text retrieval. And on the other hand, we would like to apply our synthesized video-text data for general video-language representation learning and transfer the learned knowledge to more video-text related tasks.

REFERENCES

[1] Max Bain, Arsha Nagrani, Gul Varol, and Andrew Zisserman. 2021. 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 (ICCV)*. 1728–1738.

[2] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. 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*. 3558–3568.

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

[4] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10638–10647.

[5] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. 2019. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9346–9355.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

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

[8] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. 2021. CLIP2Video: Mastering Video-Text Retrieval via Image CLIP. *arXiv preprint arXiv:2106.11097* (2021).

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

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

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

[12] 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. *International journal of computer vision* 123, 1 (2017), 32–73.

[13] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.

[14] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7331–7341.

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

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

[17] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. 2021. HiT: Hierarchical Transformer With Momentum Contrast for Video-Text Retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 11915–11925.

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

[19] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353* (2020).

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

[21] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2021. Thinking Fast and Slow: Efficient Text-to-Visual Retrieval with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9826–9836.

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

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

[24] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2630–2640.

[25] 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 *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. 19–27.

[26] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems* 24 (2011), 1143–1151.

[27] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F Henriques, and Andrea Vedaldi. 2020. Support-set bottlenecks for video-text representation learning. In *International Conference on Learning Representations*.

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

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

[30] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* (2021).

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

[32] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7464–7473.

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

[34] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4581–4591.

[35] Xiaohan Wang, Linchao Zhu, and Yi Yang. 2021. T2vlad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5079–5088.

[36] Michael Wray, Hazel Doughty, and Dima Damen. 2021. On Semantic Similarity in Video Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3650–3660.

[37] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. 2019. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6609–6618.

[38] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, and Florian Metze Luke Zettlemoyer Christoph Feichtenhofer. 2021. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. *arXiv preprint arXiv:2109.14084* (2021).

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

[40] Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 471–487.

[41] Yufeng Zhang, Jinghao Zhang, Zeyu Cui, Shu Wu, and Liang Wang. 2021. A graph-based relevance matching model for ad-hoc retrieval. *arXiv preprint arXiv:2101.11873* (2021).

[42] Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8746–8755.