LaT: Latent Translation with Cycle-Consistency for Video-Text Retrieval

Abstract. Video-text retrieval is a class of cross-modal representation learning problems, where the goal is to select the video which corresponds to the text query between a given text query and a pool of candidate videos. The contrastive paradigm of vision-language pretraining has shown promising success with large-scale datasets and unified transformer architecture, and demonstrated the power of a joint latent space. Despite this, the intrinsic divergence between the visual domain and textual domain is still far from being eliminated, and projecting different modalities into a joint latent space might result in the distorting of the information inside the single modality. To overcome the above issue, we present a novel mechanism for learning the translation relationship from a source modality space $\mathcal{S}$ to a target modality space $\mathcal{T}$ without the need for a joint latent space, which bridges the gap between visual and textual domains. Furthermore, to keep cycle consistency between translations, we adopt a cycle loss involving both forward translations from $\mathcal{S}$ to the predicted target space $\mathcal{T}'$, and backward translations from $\mathcal{T}'$ back to $\mathcal{S}$. Extensive experiments conducted on MSR-VTT, MSVD, and DiDeMo datasets demonstrate the superiority and effectiveness of our LaT approach compared with vanilla state-of-the-art methods.

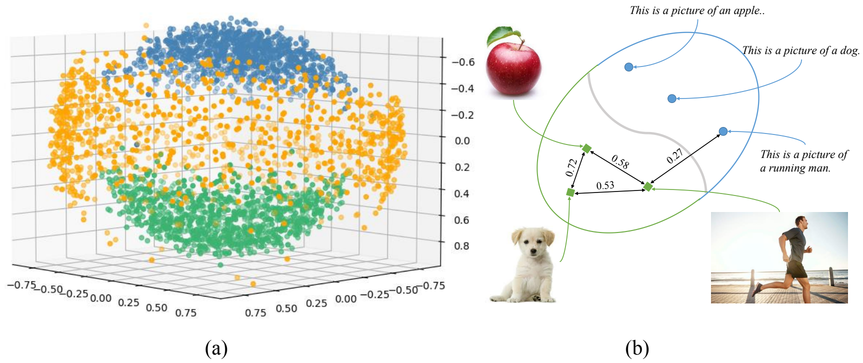
Keywords: Video-text retrieval, Latent translation, Cycle-consistency

1 Introduction

Video-text retrieval requires a bidirectional understanding of video and language, aiming to select the video which corresponds to the text query between a given text query and a pool of candidate videos, and vice versa. Cross-modal representation learning has witnessed an explosion of architectures to solve this task, be it single-stream [13] or dual-stream architectures [39], early fusion or late fusion. And the contrastive paradigm of vision-language pretraining [39] has shown promising success with large-scale datasets and unified transformer architecture, which simply and crudely maps embedding from different modalities to a common latent space via linear layers, and demonstrated the power of a joint latent space.

Despite this, the intrinsic gap between the visual domain and textual domain is still far from being eliminated, and projecting different modalities into a joint latent space might result in the distorting of intra-modal information (information inside the single modality). To show our conjecture more intuitively, we

[*] Interns at MMU, Kuaishou Technology.



(a) (b)

Fig. 1. Figure (a) is the visualization of embeddings from different latent space. The green points and blue points represent latent embedding from CLIP visual encoder and textual encoder, respectively. The orange points represent latent embedding from BERT. Figure (b) is an example for showing the intrinsic gap between visual domain and textual domain. We use **cosine similarity** (larger means more similar) as the measure. The similarity between picture *Running man* and picture *Apple* or picture *Dog* is 0.58, 0.53, which is more similar than picture *Running man* and text *This is a picture of a running man.* (0.27). The same is true for other pictures and texts, which means a huge divergence still exists between different modality domains.

choose 1,000 image-text pairs from Google Conceptual Captions 3M dataset [41], and obtain their embeddings. Fig. 1-(a) shows some visualization after dimensionality reduction with multidimensional scaling [7]. The green points and blue points represent latent embedding from CLIP [39] visual encoder and textual encoder, respectively. The orange points represent latent embedding from BERT. According to the huge divergence between the distribution of blue points and orange points, we can infer that the distribution obtained by the textual representation model changed during the feature alignment precess, which verify the conjecture of distortion of latent space. According to the huge divergence between the distribution of blue points and green points, we can infer that in the process of supervising the formation of the visual model through the natural language model, CLIP still does not merge the latent spaces of two modalities perfectly with a simple linear projection. Fig. 1-(b) shows more quantitative details about this intrinsic gap between the visual domain and textual domain.

One possible solution to the existing divergence is just to give up aligning them, to bridge them. Based on machine translation in different languages and image-to-image translation in CycleGAN [53], we explore to build the bridges to help transition and generation between different modalities. Our methods maximize the preservation of the uniqueness of each original modalities and minimize the cost of unifying different modalities.

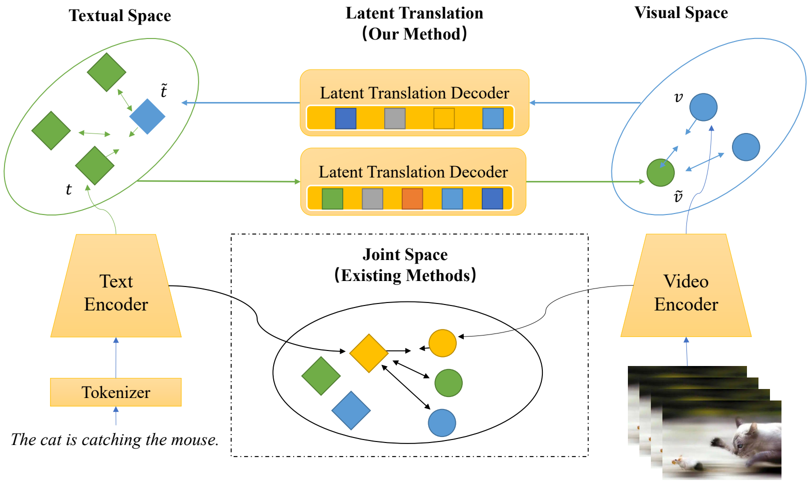


Fig. 2. Overview of our latent translation method. The encoder above can transform a video v to a meaningful embedding $\widetilde{t}$ in the textual space, and the encoder below can transform a sentence t to a meaningful embedding $\widetilde{v}$ in the visual space.

More specifically, we called these bridges as a form of more general latent translation, which is natural and applicable in cross-modal learning. The query embeddings Q_G and Q_F are some learnable parameters to guide the translation processes $G : \mathcal{S} \rightarrow \mathcal{T}$ from source modality $\mathcal{S}$ to target modality $\mathcal{T}$, and $F : \mathcal{T} \rightarrow \mathcal{S}$ from $\mathcal{T}$ to $\mathcal{S}$. The scale of Q_G and Q_F are decided by the dimension of T and S, alternately. Under the guidance of these Q_G and Q_F, latent embeddings from domain $\mathcal{T}$ is easy to be translated to domain $\mathcal{V}$, and vice versa. So for each paired embedding $t \in \mathcal{T}$ and $v \in \mathcal{V}$, we add constraints $G(t) \approx v$ and $F(v) \approx t$ between the embedding it is translated to and the target embedding it should be. Besides, after the cycle translation, the final target domain should be consistent with its original domain, whatever the intermediate domain is. To keep this cycle consistency, we add the other more strict constraints $v = G(F(v))$ and $t = F(G(t))$ between the source embedding and the final target domain which is translated twice.

In this paper, we take a step towards building bridges between visual and textual modalities, by proposing a dual decoder architecture which utilizes the learnt queries [8] to translate from visual domain to textual domain, or from textual domain to visual domain (Fig. 2). As a result, one embedding from textual space can be translated to an embedding in visual space, which is then translated into the original embedding in textual space, and we call this latent translation process cycle-consistent.

In summary, our main contributions are in two aspects:

- We dive deep into the intrinsic modality gap between visual and textual modalities with qualitative and quantitative anaysis.

- We propose a novel latent translation framework, which is designed for eliminating the distortion from projecting different modalities to a common latent space. Besides, some constraints are adapted for our framework, i.e., cycle-consistent loss and intra-modal contrastive loss, which contribute to the cross-modal translation.
- Extensive experiments conducted on MSR-VTT, MSVD, and DiDeMo datasets demonstrate the superiority and effectiveness of our LaT approach compared with vanilla state-of-the-art methods for video-text retrieval.

2 Related Work

Our work builds on prior works in several domains: dual stream architecture for cross-modal retrieval, video representation based on image encoder, and cycle-consistent learning methods.

2.1 Cross-modal Retrieval

Existing methods for cross-modal learning can be roughly categorized as single stream [26–28, 13, 34, 42, 30, 21] and dual stream architectures [39, 23, 31]. Single stream architecture directly fuses visual and textual representations with a multi-modal transformer or independently extract visual and textual features then introduces a cross-modal attention to achieve the fusion of multi-modal information. Dual stream architecture applies two encoders to extract different modal representations and projecting these features into a common latent space with a similarity-based ranking loss, has become a recent trend for cross-modal retrieval. One illustrious work with dual stream architecture is CLIP [39].

Dual stream models have a merit of efficient inference for downstream tasks such as image-text retrieval, since they can decouple and offline store pre-computed image/text features from encoders [20]. Assuming that we have M videos and N texts, when it comes to large-scale cross-modal retrieval tasks, single stream method usually requires $\Theta(MN)$ time complexity of intra-modal information exchange while dual stream method requires only $\Theta(M+N)$ time complexity. As a result, it is impractical to apply cross-modal retrieval tasks real-time with single stream architectures.

2.2 Video Encoder Backbone

The early research on video representation learning utilize 2D convolutions on spatial features [24] with a temporal sampling method [47] to capture temporal information. Moreover, 3D convolutions are further proposed [43, 22] to jointly learn spatio-temporal information. The convergence of 3D kernels is further improved by initializing them from inflated 2D kernels [9] (known as I3D models) and decoupling spatial and temporal convolutions in each block [44] (known as R(2+1)D models). The success of the vision transformer on images also inspire the progress of the video encoder. [4, 6] propose pure transformer architectures to model spatial-temporal feature.

Another benefit of video transformers is that the main attention function that works on a group of local patches doesn't rely on inductive bias and makes it easy to be extended on a joint image and video input. The effective transformer proposed by Bain [5], which is based on TimeSformer[6], can gracefully handle inputs of different length of videos and images (by treating images as a single-frame video).

2.3 Cycle-Consistent Learning

Inspired by machine translation in different languages, Zhu *et al.* [53] exploited unpaired image-to-image translations between two different visual domains with a cycle-consistency loss and an adversarial loss [19]. Several works have also transferred the cycle-consistency loss to image-text retrieval tasks [32, 48, 15]. CycleMacth [32] designs some fully-connected layers which can cascade dual and reconstructed mappings together to maintain inter-modal correlations and intra-modal consistency. DGH [48] propose a noval deep generative approach to cross-modal retrieval to learn hash functions. The concept Cornia *et al.* [15] proposed is pretty similar to us. However, their translation frameworks are based on convolution neural networks and there is no satisfactory ablation and further attempt for more architectures.

Here we propose a brand new latent translation framework with cycle-consistency, which introduces the attention [45] architecture and learnable queries [8], for cross-modal retrieval.

3 Method

In this section, we define our tasks (Section 3.1), present architecture of our Latent Translation (LaT) (Section 3.2), then detail our latent translation decoder (Section 3.3) and the supervised methods we designed (Section 3.4), respectively.

3.1 Problem Definition

Let a vector $\bar{v}$ denote a raw video and a vector $\bar{t}$ denote a raw text. We use a vision encoder E_v and a language encoder E_l to get the latent feature $v \in \mathcal{V}$ and $t \in \mathcal{T}$, where $\mathcal{V}$ and $\mathcal{T}$ are visual and textual latent space respectively, formulated as

$$v = E_v(\bar{v}) \tag{1}$$

$$t = E_t(\bar{t}) \tag{2}$$

Given paired training samples $\{v_i\}_{i=1}^N$ and $\{t_i\}_{i=1}^N$ where $v_i \in \mathcal{V}$ and $t_i \in \mathcal{T}$, the video-text retrieval task can be defined as selecting the video $\bar{v}_i$ which corresponds to the text query $\overline{t_{\text{query}}}$ between a given text query and a pool of candidate videos (Text-To-Video Retrieval, T2V), or vice versa (Video-To-Text Retrieval, V2T). Previous methods apply two projection functions f_v and f_t

to project visual and textual space into a joint latent space and measure the distance within that, formulated as

$$\mathrm{T2V}(\overline{t_{\mathrm{query}}}) = \mathrm{argmin}_i\,(f_v(v_i), f_t(t_{query})) \tag{3}$$

$$\mathrm{V2T}(\overline{v_{\mathrm{query}}}) = \mathrm{argmin}_j\,(f_v(v_{query})), f_t(t_j)) \tag{4}$$

where $t_{query} = D_t(\overline{t_{\mathrm{query}}})$ and $v_{query} = D_v(\overline{v_{\mathrm{query}}})$ are embeddings of the query samples.

As shown in Fig. 2, our goal is to learn the translation function between visual space $\mathcal{V}$ and textual space $\mathcal{T}$ under the guidance of learnt query embeddings Q_G and Q_F.There are two translation relationships $G : t \longrightarrow \widetilde{v}$ and $F : v \longrightarrow \widetilde{t}$ in our mechanism. With our latent translation function, cross-modal retrieval can be attributed to inter-modal retrieval formulated as:

$$\mathrm{T2V}(\overline{t_{\mathrm{query}}}) = \mathrm{argmin}_i\,(v_i, G(t_{query})) \tag{5}$$

$$\mathrm{V2T}(\overline{v_{\mathrm{query}}}) = \mathrm{argmin}_j\,(F(v_{query}), t_j) \tag{6}$$

In this case, we can translate textual embedding from textual space $\mathcal{T}$ to visual space $\mathcal{V}$ and retrieve the most similar visual embedding in visual space $\mathcal{V}$, and vice versa.

3.2 Model Architecture

The previous works in video-language model tends to utilize exist image encoder and text encoder, i.e., Clip-Based framework: CLIP4Clip [35], CLIP2Video [17], CAMoE [14]. ViT-Bert-Based framework: HiT [31], Frozen [5], and so on [46], [50]. Due to the giant pretraining 400M dataset CLIP utilized, which is difficult for us to raise a comparable dataset for our decoder's pretrain, we choose the ViT-Bert-based framework Frozen [5] as our baseline*. The visual encoder is an optimized TimeSformer [6] initialized with ViT [16] weights trained on ImageNet-21k for spatial attention weights, and zero for temporal attention weights. The language encoder is DistilBERT base-uncased [40].

Another reason for adopting this structure is that the two encoders of CLIP are originally aligned, and the excellent zero-shot performance also shows that it can be well adapted to the video-text dataset without training. What we have to do is try to reduce the modality gap in the process of alignment. Because of the lack of data matching the clip, CLIP is trained with 400M data to achieve strong alignment, we do not have enough data to train a strong translation network.

*Another reason for adopting this structure is that the two encoders of CLIP are originally aligned, and the excellent zero-shot performance also shows that it can be well adapted to the video-text dataset without training. CLIP4CLIP also shows that with CLIP weights, good video representations can be learned without temporal features (using mean pooling). As a result, what we aim to do is trying to reduce the modality gap in the process of alignment. So we do not intend to use CLIP weights nor do we intend to compare with CLIP weight based methods.

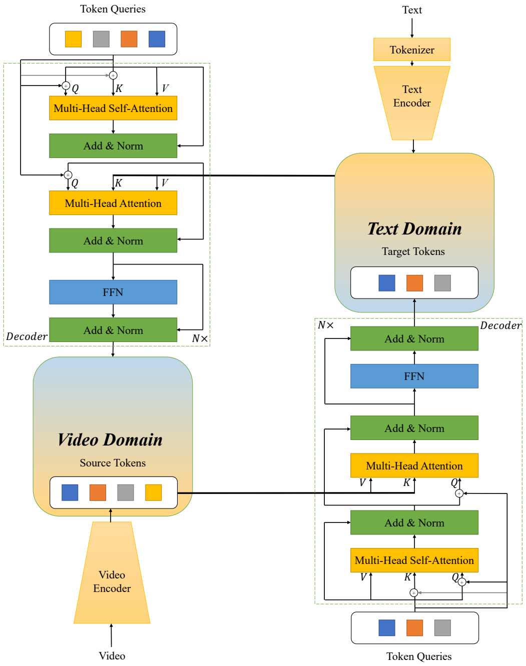


Fig. 3. The detailed latent translation decoder. **Token queries** are some learnable parameters. The figure shows how source tokens in video domain translated to target tokens in text domain.

CLIP4CLIP also shows that with CLIP weights, good video representations can be learned without temporal features (using mean pooling).

As shown in Fig. 2, we remove the projection layers after two encoders which may result in information distortion. And add two decoders framework following DETR [8]'s setting to achieve cross-modal translation. The only difference between these two decoders is the number of queries stored in them, which is adapted to the target modality. Besides, due to the limit of GPU memories, we adapt the memory bank [12] to offset the decrease in batch size caused by the increase of the network parameters.

3.3 Decoder

The detailed description of the decoder is given in Fig. 3. The decoder follows the standard architecture of the DETR [8] transformer, transforming M embeddings of size d using multi-head self-attention and decoder part of encoder-decoder attention mechanisms. Since the decoder is permutation-invariant, the M input embeddings must be different to produce different results. These input embeddings are learnt priors that we refer to as *token queries*, and we add them to the input of each attention layer. The perceiver resampler from Flamingo [1] and Clipcap [37] also applied this framework.

In briefly, visual or textual features are extracted from the encoder, as the source tokens, and then passed into the multi-head attention as V and K. For the decoder itself, there are some learnable queries (the number of these queries is always determined by the target domain), which can be initially randomly, passing a multi-head self-attention before passed into the multi-head attention as Q. We repeat the multi-head self-attention and multi-head attention for N times, then get the target tokens which belong to the target domain.

Intuitively, the queries are like some eyes from different angles, translating the source tokens to target tokens. And the self-attention plays a role that they see each other to make sure they translate different tokens. We can add some constraints to make the first query token translate the global information (means global token from visual or textual embeddings if they have, always comes from the first or last token) and the other tokens to translate the detailed information. That is how our latent translation decoder works.

3.4 Supervised Methods

Let $v \in \mathbb{R}^{N \times L_1 \times D}$ and $t \in \mathbb{R}^{N \times L_2 \times D}$ be the paired embeddings extracted from training samples in visual modality and textual modality, where L_1 and L_2 denotes the token numbers, N denotes the batch size and D denotes the dimension of each feature. $v \xrightarrow{F} t$ denote the translation network from visual modality to language modality, and $t \xrightarrow{G} v$ is the opposite. v_{i,l_1} and t_{i,l_2} denotes each embedding of each video-text pair, where $i = 1...N$, $l_1 = 1...L_1$, $l_2 = 1...L_2$ ($l_1 = 1$ or $l_2 = 1$ denotes the [CLS] token).

$$v = \begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,L_1} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,L_1} \\ \vdots & \vdots & \ddots & \vdots \\ v_{N,1} & v_{N,2} & \cdots & v_{N,L_1} \end{pmatrix}$$

$$t = \begin{pmatrix} t_{1,1} & t_{1,2} & \cdots & t_{1,L_2} \\ t_{2,1} & t_{2,2} & \cdots & t_{2,L_2} \\ \vdots & \vdots & \ddots & \vdots \\ t_{N,1} & t_{N,2} & \cdots & t_{N,L_2} \end{pmatrix}$$

First we will introduce the components of our loss funtion, an then we will describe how we deal with global and local information. Finally, we will show our overall training objective.

Loss. Our loss includes inter-modal loss and intra-modal loss.

Inter-modal loss aims to learn the right translation relationship. Follow the settings of Frozen [5], we apply infoNCE loss as our cross-modal loss.

$$\mathcal{L}_{v2t} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(sim(v_{i,l_1}, t_{i,l_2})/\tau)}{\sum_{j=1}^{N}\exp(sim(v_{i,l_1}, t_{j,l_2})/\tau)} \tag{7}$$

$$\mathcal{L}_{t2v} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(sim(v_{i,l_1}, t_{i,l_2})/\tau)}{\sum_{j=1}^{N}\exp(sim(v_{i,l_1}, t_{j,l_2})/\tau)} \tag{8}$$

where τ is the temperature and sim is a similarity function(i.e., dot product). Specifically, when applying the general global infoNCE loss, $l_1 = l_2 = 1$. And when applying the local infoNCE loss, there will be a MeanPooling method described in the next subsection.

As a result, the final inter-modal loss is

$$\mathcal{L}_{inter} = \frac{1}{2}\big(\mathcal{L}_{v2t} + \mathcal{L}_{t2v}\big) \tag{9}$$

Intra-modal loss aims to make sure that after a cycle translation, the final target objects keep the same as the source target objects. To achieve this cycle-consistent propose, we apply a more strict constraint: mean square error (MSE) loss.

$$\mathcal{L}_{intra} = \frac{1}{2}\big(||G(F(v_{i,l_1})) - v_{i,l_1}||_2^2 + ||F(G(t_{i,l_2})) - t_{i,l_2}||_2^2\big) \tag{10}$$

where F denotes the translation network from visual modality to textual modality, and G denotes the translation network form textual modality to visual modality.

Global and Detailed Information For paired videos and texts, passing them into the visual and textual encoders will obtain paired videos and texts embeddings, where there is more than one channel for each modality ($L > 1$). For instance, The output of visual transformer for the input of an image is usually being setted to $197 \times D$, and the first one of 197 is [CLS] token.

Previous cross-modal learning methods, i.e., CLIP [39], Frozen [5], HiT [31] tends to ignore the fine-grained information and interact via only the global features ([CLS] token) of the entire video or sentence. We named this choice which only employ the [CLS] token as global level interaction.

$$\mathcal{L}_{global} = \lambda_{inter}\mathcal{L}_{inter} + \lambda_{intra}\mathcal{L}_{intra} \tag{11}$$

where λ_{inter} and λ_{intra} are two hyper-parameters to balance two objectives. We set both λ_{inter} and λ_{intra} to 1 in our experiments.

Besides, we also concern about the fine-grained information stored in other tokens. There are some latest work trying to propose some new cross-modal interaction mechanism to capture the fine-grained representations, FILIP[51] applies an max-mean method for similarity calculation, while [50] and [46] applies some auxiliary networks or introduces existing object detection networks to assist similarity calculation.

Here we just simply average all tokens except the [CLS] token, then apply token-level loss as

$$\mathcal{L}_{token} = \lambda_{inter}\mathcal{L}_{inter\_token} + \lambda_{intra}\mathcal{L}_{intra\_token} \tag{12}$$

Overall Training Objective. Thus, the overall objective function is the summary of global level loss and token level loss:

$$\mathcal{L} = \lambda_{global}\mathcal{L}_{global} + \lambda_{token}\mathcal{L}_{token} \tag{13}$$

where λ_{global} and λ_{token} are two hyper-parameters to balance two objectives. We set both λ_{global} and λ_{token} to 1 in our experiments.

4 Experiments

In this section, we first introduce the prertaining and finetuning datasets (Section 4.1) and evaluation metrics (Section 4.2) we used in our experiments with some implementation details (Section 4.3). Then, we compare our performance with the state-of-the-art methods to show the effectiveness and generality of our model (Section 4.4). Finally, we perform a number of ablation studies to understand the effects of proposed components in our model (Section 4.5).

4.1 Datasets

HowTo100M [36] is a large-scale dataset of narrated videos with an emphasis on instructional videos where content creators teach complex tasks with an explicit intention of explaining the visual content on screen. It contains 136M video clips with captions sourced from 1.2M Youtube videos, and most existing works pretrain with it. However, HowTo100M is heavily noisy and only contains instructional videos, while WebVid2M [5] shows its efficiency with one tenth scale of the data. Meanwhile, although there is more than 10% data from Google Conceptual Captions 3M (CC3M) [41] lost, applying the latest Google Conceptual Captions 12M (CC12M) [10] as our pretraining dataset is unfair for comparing with existing methods.

As a result, we conduct the experiments on 2 pretraining datasets (CC3M [41] and WebVid2M [5]) and 3 downstream datasets (MSR-VTT [49], MSVD [11], DiDeMo [3]). The followings are the descriptions of these video-text (image-text) datasets.

- **Conceptual Captions 3M (CC3M) [41]** contains approximately 3.3 millions images annotated with captions. Conceptual Captions images and their raw descriptions are harvested from the web, and therefore represent a wider variety of styles.

- **WebVid2M [5]** is a large scale text-video dataset, containing 2.5 millions video-text pairs scraped from the web. The videos are diverse and rich in their content.
- **MSR-VTT [49]** contains 10,000 videos, where each video is annotated with 20 captions in English. We follow the training protocol defined in [31][5][35] to train on 9k videos, and to evaluate on text-to-video and video-to-text retrieval tasks on the 1k-A testing split with 1,000 video and text candidates defined by [52].
- **MSVD [11]** contains 1,970 videos, and each video has approximately 40 captions in English and range form 1 to 62 seconds. The train, validation, and test splits contain 1,200, 100, and 670 videos, respectively.
- **DiDeMo [3]** contains 10k Flickr videos annotated with 40k sentences. We evaluate video-paragraph retrieval following [25], [5], [35], where all sentence descriptions for a video are concatenated into a single query.

4.2 Evaluation Metrics

We use standard retrieval metrics: recall at rank K (R@K, K=1, 5, 10, higher is better), median rank (MedR, lower is better) to evaluate the performance of our model.

Table 1. The experimental results on MSR-VTT.

Methods	**Video-to-Text Retrieval**				**Text-to-Video Retrieval**			
	R@1	**R@5**	**R@10**	**MedR**	**R@1**	**R@5**	**R@10**	**MedR**
HowTo100M[36]	12.2	33.5	47.5	13	12.6	36.2	48.1	13
ActBert[54]	–	–	–	–	16.3	42.8	56.9	10
Noise-Estimation[2]	–	–	–	–	17.4	41.6	53.6	8
CE[33]	20.6	50.3	64.0	5.3	20.9	48.8	62.4	6
ClipBERT[25]	–	–	–	–	22.0	46.8	59.9	6
MMT[18]	27.0	57.5	69.7	3.7	26.6	57.1	69.6	4
Support-Set[38]	28.5	58.6	71.6	3	30.1	58.5	69.3	3
HiT[31]	32.1	62.7	74.1	3	30.7	60.9	73.2	2.6
Frozen[5]	–	–	–	–	31.0	59.5	70.5	3
Ours(LaT)	**35.4**	**61.3**	**72.4**	**3**	**35.3**	**61.3**	**72.9**	**3**
Zero-shot								
Support-Set[38]	8.7	23.0	31.1	31	12.7	27.5	36.2	24
Frozen[5]	–	–	–	–	18.7	39.5	51.6	10
Ours(LaT)	**17.2**	**36.2**	**47.9**	**12**	**23.4**	**44.1**	**53.3**	**8**

4.3 Implementation Details

Following the code Bain *et al.* [5] released, our experiments are conducted with PyTorch, optimized with Adam. We also apply the author's training strategy for

Table 2. The experimental results on MSVD.

Methods	**Video-to-Text Retrieval**				**Text-to-Video Retrieval**			
	R@1	**R@5**	**R@10**	**MedR**	**R@1**	**R@5**	**R@10**	**MedR**
CE[33]	–	–	–	–	19.8	49.0	63.8	6
Support-Set[38]	34.7	59.9	70.0	3	28.4	60.0	72.9	4
Frozen[5]	–	–	–	–	33.7	64.7	76.3	3
Ours(LaT)	**39.7**	**75.6**	**85.4**	**2**	**40.0**	**74.6**	**84.2**	**2**
Zero-shot								
Frozen[5][Our Imp.]	32.4	65.5	76.9	3	35.7	63.9	77.8	3
Ours(LaT)	**34.4**	**69.0**	**79.2**	**3**	**36.9**	**68.6**	**81.0**	**2**

pretraining, which means first training images and 1 frame videos with a learning rate of 3e-5, to capture the image content, then training 4 frames videos with a learning rate of 1e-5, to capture the video content. Experiments show that this schedule will not only decrease the pretraining time but also improve the pretraining effect slightly.

The whole pretraining takes 1 days on 8 Tesla V100 GPUs. Unless otherwise specified, all results of downstream datasets reported in this paper adopt the best pretraining model. Both pretraining and fine-tuning, only 4 video frames are sampled.

4.4 Compare to state of the art

Table 3. The experimental results on DiDeMo.

Methods	**Video-to-Text Retrieval**				**Text-to-Video Retrieval**			
	R@1	**R@5**	**R@10**	**MedR**	**R@1**	**R@5**	**R@10**	**MedR**
ClipBERT[25]	–	–	–	–	20.4	44.5	56.7	7
Frozen[5]	–	–	–	–	31.0	59.8	72.4	3.0
Ours(LaT)	**32.7**	**61.1**	**72.7**	**3**	**32.6**	**61.3**	**71.6**	**3**
Zero-shot								
Frozen[5]	–	–	–	–	21.1	46.0	56.2	7
Ours(LaT)	**22.5**	**45.2**	**56.8**	**7**	**22.6**	**45.9**	**58.9**	**7**

We compare the proposed LaT with several vanilla state-of-the-art (SOTA) methods on MSR-VTT [49], MSVD [11] and DiDeMo [3] datasets and report the results in Table 1,2 and 3 respectively. The upper part of the tables show the fine-tuning results of methods. We can see that LaT outperforms all comparison methods by a clear margin on all three datasets, especially in terms of R@1. Apart from the fine-tuning results, we provide the results of zero-shot version of LaT. Compared with baseline methods Frozen [5] and Support-Set [38], the proposed LaT also shows better performance.

4.5 Ablation Study

In this part, we study the effect of different details in our latent translation framework to further demonstrate the effectiveness and robustness of LaT. All the experiments are pretrained on WebVid2M with the same training epochs and batch size without special instruction. We report our text-to-video zero-shot result on MSR-VTT 1k-A testing split.

Translation Methods We experiment with three translation methods compared with no translation framework by pretraining on WebVid2M with 200 training epochs, and report our text-to-video zero-shot result on MSR-VTT 1k-A testing split. **None** means without using latent translation network. **Linear** means using a 3-layer linear translation architecture. **Transformer** means using a 3-layer transformer architecture. **Decoder** means using a 3-layer query-guided transformer (shown in Figure 3) architecture. The results are given in Table 4, which show the decoder framework achieves better on R@1 compared to other translation methods, and applying a simple transformer-based translation network works better than with nothing.

Table 4. The ablation study on methods of latent translation.

Methods	R@1	R@5	R@10	MedR
None	19.4	38.6	48.4	11.5
Linear	19.2	38.9	46.9	12
Transformer	19.4	**41.2**	50.4	10
Decoder	**20.3**	41.1	**51.4**	**9.5**

Depth We experiment with different depth of the translation framework by pretraining on WebVid2M with 200 training epochs. The results are given in Table 5, which show the recall rate increases as the number of layers deepens. However, the deepening of the number of layers will lead to the increase of model parameters and GFLOPs, and a decease in batch size. For instance, when the depth of layers is 4, the batch size of each GPU is reduced from 26 to 22 (with the depth is 3).

Number of Queries We experiment with different number of queries of the translation framework by pretraining on WebVid2M with 100 training epochs. The results are given in Table 6, which show that when the number of guiding queries closer to the number of tokens in target domain, the better recall rate shows. In this experiment, the number of tokens in textual domain is close to 30 (The number of words from each text).

Token Usage We experiment with different methods for utilizing detailed tokens by pretraining on WebVid2M with 100 training epochs. **Global** means using

Table 5. The ablation study on depth of latent translation network. Depth = 4 utilizes a smaller batch size.

Depth	**R@1**	**R@5**	**R@10**	**MedR**
1	20.6	40.2	50.0	10.5
2	**20.7**	40.7	50.2	10
3	20.3	**41.1**	**51.4**	**9.5**
4*	**20.7**	39.9	48.2	11

Table 6. The ablation study on number of queries. $\boldsymbol{N_q}$ represents **Number of Queries**.

$\boldsymbol{N_q}$	**R@1**	**R@5**	**R@10**	**MedR**
15	19.6	38.8	47.5	12
30	**20.1**	**40.0**	**49.8**	**11**
60	19.3	37.7	47.9	12

[CLS] token only. **Detailed** means to average all tokens except [CLS] token as the detailed token. The results are given in Table 7, which show that **Global** + **Detailed** works slightly better.

Table 7. The ablation study on token usage.

Global	**Detailed**	**R@1**	**R@5**	**R@10**	**MedR**
✓		20.1	**40.0**	**49.8**	**11**
	✓	20.3	39.3	49.3	**11**
✓	✓	**20.5**	39.8	49.7	**11**

Parameters Quantities We experiment with different parameter quantites by pretraining on WebVid2M with 100 training epochs. The results are given in Table 8, which show that our method does not rely on improvements in the number of parameters, but on better supervision.

5 Limitations and Discussions

Quantitative examples: Figure 4 shows more quantitative details about how our LaT works. The triangle represents a translation result from a text embedding through decoder g. Similar to figure-1-(b), we use use **cosine similarity** (larger means more similar) as the measure. And it is evidently the picture *Running man* is more similar to the translated text *This is a picture of a running man* than picture *Apple* and picture *Dog*. To some extent, this means different modalities have been fused better than before.

Table 8. The ablation study on parameters quantites.

Encoder Size	Decoder Size	R@1	R@5	R@10	MedR
6-layer-Bert	None	19.4	38.6	48.4	11.5
12-layer-Bert	None				
6-layer-Bert	2 3-layer-Decoder	**20.3**	**41.1**	**51.4**	**9.5**

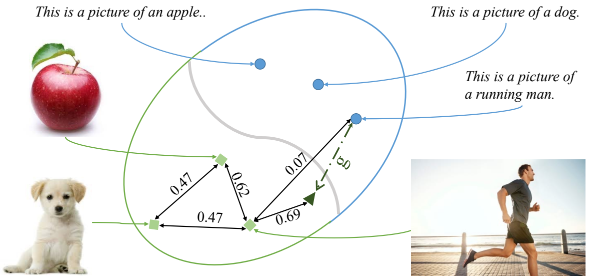

This is a picture of an apple..
This is a picture of a dog.
This is a picture of a running man.
0.47
0.62
0.47
0.07
0.69
g

Fig. 4. An example for showing the effectiveness of our LaT.

Fixed embedding space: Although our method can achieve compelling results in many cases, the results are still far from good enough. During the training schedule, the encoder itself has been changed, so as latent space. The ideal solution for avoiding the change is to fix the parameters from both encoders, to directly train the latent translation network. We have explored it, with little success. Nonetheless, this will be future work for us to explore a more suitable framework to achieve the latent translation, which is of great importance in large-scale cross-modal retrieval with existing encoders. Some existing works [20] have tried fixing the visual encoder and won some state-of-the-art performance on multiple downstream tasks. We believe, that with existing textual encoders and visual encoders, a learnable bridge will lead to more efficient training and communication between different modalities.

In addition, we found that freezing the backbones of visual encoder and textual encoder but retaining the learnability of tokenizer and conv operations (which used to convert raw data to embedding input of backbone) will have a relatively comparable performance.

Influence: When doing search and recall in the industry, the confidence of the recall results may be low. In this situation, it is hard to judge whether it is the modality gap of the model itself or the lack of high-quality recall data. Besides, reducing the modality gap can also bring more benefits to the modality-fused tasks, i.e., VQA, visual dialog.

6 Conclusions

In this paper, we dive deep into the intrinsic gap between visual and textual domains with qualitative and quantitative analysis. And conjecture projecting them into a joint embedding space may lead to a distortion of intra-modal information. To tackle the above challenges, we propose a latent translation mechanism (LaT) for cross-modal translation to solve the difficulty from aligning different modal spaces. Besides, we employ some learnable guiding parameters Q for better translation. Extensive experiments conducted on MSR-VTT, MSVD, and DiDeMo datasets demonstrate the superiority and effectiveness of our LaT approach compared with vanilla state-of-the-art methods. Finally, we discuss the possible impacts and limitations of our approach.

7 Acknowledgement

We would like to thank Dr. Yi Zhang and anonymous friends for their kindly helpful discussions and feedback on writing.

8 Appendix A: Quantities Details about CLIP and LaT

Table 9. Cosine Similarity in CLIP.

		T			**V**		
		Apple	**Dog**	**Man**	**Apple**	**Dog**	**Man**
T	**Apple**	1.	0.84	0.77	0.31	0.21	0.17
	Dog	0.84	1.	0.84	0.23	0.29	0.17
	Man	0.77	0.84	1.	0.21	0.24	0.27
V	**Apple**	0.31	0.23	0.21	1.	0.72	0.58
	Dog	0.21	0.29	0.24	0.72	1.	0.53
	Man	0.17	0.17	0.27	0.58	0.53	1.

Table 9 shows **cosine similarity** (larger means more similar) between different embeddings in different latent space from CLIP. **T** denotes textual space and **V** denotes visual space. **Apple**, **Dog** and **Man** denote embeddings with different semantic meanings from images or texts. The numbers denote the cosine similarity between different embeddings. And the cosine similarities for the same semantic embeddings between different latent spaces are significantly smaller than those for different semantic embeddings in the same latent space, which means there is still a huge gap between visual space and text space in CLIP. Liang *et al.* [29] also observed this problem at the same time.

An example that can be intuitively understood is: the intersection of two parallel planes with a line perpendicular to these two planes results in two points, which can be regarded as paired points from different modalities (planes). In this way we get many pairs of points, each with the shortest distance. But this way doesn't fuse the two modalities (planes) together.

Table 10 shows **cosine similarity** (larger means more similar) between different embedding in different latent spaces from LaT. **T** denotes textual space, **V** denotes visual space, **GT** denotes visual space translated from textual space

Table 10. Cosine Similarity in LaT.

		T			**V**			**GT**			**FV**		
		Apple	**Dog**	**Man**	**Apple**	**Dog**	**Man**	**Apple**	**Dog**	**Man**	**Apple**	**Dog**	**Man**
T	**Apple**	1.	0.67	0.53	0.10	0.02	0.01	0.04	-0.00	-0.02	0.59	0.34	0.38
	Dog	0.67	1.	0.59	0.08	0.13	0.01	0.04	0.07	0.00	0.40	0.59	0.42
	Man	0.53	0.59	1.	0.07	0.01	0.07	0.05	0.02	0.03	0.40	0.30	0.70
V	**Apple**	0.10	0.08	0.07	1.	0.47	0.62	0.57	0.36	0.38	0.15	0.05	0.09
	Dog	0.02	0.13	0.01	0.47	1.	0.47	0.29	0.60	0.27	0.03	0.17	0.02
	Man	0.01	0.01	0.07	0.62	0.47	1.	0.33	0.37	0.69	0.06	0.02	0.11
GT	**Apple**	0.04	0.04	0.05	0.57	0.29	0.33	1.	0.59	0.46	0.03	0.06	0.02
	Dog	-0.00	0.07	0.02	0.36	0.60	0.37	0.59	1.	0.55	-0.01	0.06	0.02
	Man	-0.02	0.00	0.03	0.38	0.27	0.69	0.46	0.55	1.	-0.05	0.01	0.03
FV	**Apple**	0.59	0.40	0.40	0.57	0.29	0.33	0.03	-0.01	-0.05	1.	0.47	0.62
	Dog	0.34	0.59	0.30	0.36	0.60	0.37	0.06	0.06	0.01	0.47	1.	0.47
	Man	0.38	0.42	0.70	0.38	0.27	0.69	0.02	0.02	0.03	0.62	0.47	1.

by decoder G and **FV** denotes textual space translated from visual space by decoder F. **Apple**, **Dog** and **Man** denote different embeddings with different semantic meanings from images or texts. The numbers denote the cosine similarity between different emebeddings. We can observe that, the cosine similarities for the same semantic embeddings between **GT** and **V** (or **FV** and **T**) are the same or slightly larger than those for different semantic embeddings in the same latent space, which means that LaT fuses different modalities better than the previous ways.

9 Appendix B: Visualization on LaT

Figure 5 shows the visualization of embeddings from different latent space. The seagreen points, steelblue points, skyblue points and mediumseagreen points represent latent embeddings from **T**, **V**, **GT** and **FV**, respectively. We can clearly observe that LaT fuses **T** and **FV** (**V** and **GT**) better than CLIP.

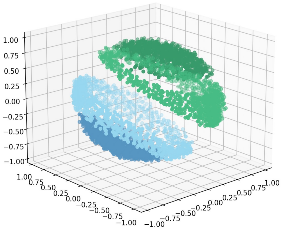


Fig. 5. Visualization of embeddings from different latent space of LaT.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. arXiv preprint arXiv:2204.14198 (2022)
2. Amrani, E., Ben-Ari, R., Rotman, D., Bronstein, A.: Noise estimation using density estimation for self-supervised multimodal learning. arXiv preprint arXiv:2003.03186 **8** (2020)
3. 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)
4. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6836–6846 (2021)
5. 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)
6. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding. arXiv preprint arXiv:2102.05095 **2**(3), 4 (2021)
7. Borg, I., Groenen, P.J.: Modern multidimensional scaling: Theory and applications. Springer Science & Business Media (2005)
8. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
9. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
10. 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)
11. 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)
12. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9640–9649 (2021)
13. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020)
14. 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)
15. Cornia, M., Baraldi, L., Tavakoli, H.R., Cucchiara, R.: Towards cycle-consistent models for text and image retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
16. 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. arXiv preprint arXiv:2010.11929 (2020)

17. Fang, H., Xiong, P., Xu, L., Chen, Y.: Clip2video: Mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097 (2021)
18. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: European Conference on Computer Vision. pp. 214–229. Springer (2020)
19. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
20. Gu, J., Meng, X., Lu, G., Hou, L., Niu, M., Xu, H., Liang, X., Zhang, W., Jiang, X., Xu, C.: Wukong: 100 million large-scale chinese cross-modal pre-training dataset and a foundation framework. arXiv preprint arXiv:2202.06767 (2022)
21. Hu, R., Singh, A.: Unit: Multimodal multitask learning with a unified transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1439–1449 (2021)
22. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence **35**(1), 221–231 (2012)
23. 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)
24. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014)
25. 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)
26. Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11336–11344 (2020)
27. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
28. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: European Conference on Computer Vision. pp. 121–137. Springer (2020)
29. Liang, W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. arXiv preprint arXiv:2203.02053 (2022)
30. Lin, X., Bertasius, G., Wang, J., Chang, S.F., Parikh, D., Torresani, L.: Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7005–7015 (2021)
31. 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)
32. Liu, Y., Guo, Y., Liu, L., Bakker, E.M., Lew, M.S.: Cyclematch: A cycle-consistent embedding network for image-text matching. Pattern Recognition **93**, 365–379 (2019)

33. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
34. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019)
35. 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)
36. 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)
37. Mokady, R., Hertz, A., Bermano, A.H.: Clipcap: Clip prefix for image captioning. arXiv preprint arXiv:2111.09734 (2021)
38. Patrick, M., Huang, P.Y., Asano, Y., Metze, F., Hauptmann, A., Henriques, J., Vedaldi, A.: Support-set bottlenecks for video-text representation learning. arXiv preprint arXiv:2010.02824 (2020)
39. 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)
40. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
41. Sharma, P., Ding, N., Goodman, S., Soricut, R.: 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). pp. 2556–2565 (2018)
42. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7464–7473 (2019)
43. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
44. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems (2017)
46. Wang, A.J., Ge, Y., Cai, G., Yan, R., Lin, X., Shan, Y., Qie, X., Shou, M.Z.: Object-aware video-language pre-training for retrieval. arXiv preprint arXiv:2112.00656 (2021)
47. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Gool, L.V.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016)
48. Wu, L., Wang, Y., Shao, L.: Cycle-consistent deep generative hashing for cross-modal retrieval. IEEE Transactions on Image Processing **28**(4), 1602–1612 (2018)
49. 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)

50. Yan, R., Shou, M.Z., Ge, Y., Wang, A.J., Lin, X., Cai, G., Tang, J.: Video-text pre-training with learned regions. arXiv preprint arXiv:2112.01194 (2021)
51. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. arXiv preprint arXiv:2111.07783 (2021)
52. 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)
53. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
54. 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)