Learning Semantic Alignment with Global Modality Reconstruction for Video-Language Pre-training towards Retrieval

Mingchao Li[1,2,*], **Xiaoming Shi**[3], **Haitao Leng**[2,†], **Wei Zhou**[2], **Hai-Tao Zheng**[4,5,†], **Kuncai Zhang**[2]

[1] Department of Computer Science and Technology, Tsinghua University
[2] Alibaba Group
[3] Shanghai Artificial Intelligence Laboratory
[4] Shenzhen International Graduate School, Tsinghua University
[5] Peng Cheng Laboratory
mingchaoli.lmc@gmail.com, {heiqie.lht, fayi.zw, kuncai.zkc}@alibaba-inc.com
shixiaoming@pjlab.org.cn, zheng.haitao@sz.tsinghua.edu.cn

Abstract

Video-language pre-training for text-based video retrieval tasks is vitally important. Previous pre-training methods suffer from semantic misalignments. The reason is that these methods ignore sequence alignments but focus on critical token alignment. To alleviate the problem, we propose a video-language pre-training framework, termed video-language pre-training **F**or l**E**arning s**E**mantic a**L**ignments (**FEEL**), to learn semantic alignments at the sequence level. Specifically, the global modality reconstruction and the cross-modal self-contrasting method are utilized to learn the alignments at the sequence level better. Extensive experimental results demonstrate the effectiveness of **FEEL** on text-based video retrieval and text-based video corpus moment retrieval.

Introduction

Text-based video retrieval tasks, including text-based video retrieval (Luo et al. 2020; Zhu and Yang 2020; Li et al. 2020) and text-based video corpus moment retrieval (Li et al. 2020) have shown significant potential and alluring technological value. Thanks to the ability of cross-modality alignments, video-language pre-training shows effectiveness on these retrieval tasks (Wang et al. 2022b; Huang et al. 2022). Cross-modal alignment is the key challenge in learning video-language pre-training.

Previous pre-training methods learn cross-modality alignments in two manners. Matching tasks are widely used for cross-modality alignments, such as using the special token [CLS] for binary classification with a cross-modal Transformer (Sun et al. 2019b; Zhu and Yang 2020; Luo et al. 2020) and contrastive learning between video and language with two unimodal Transformers (Miech et al. 2019, 2020; Ging et al. 2020; Patrick et al. 2020; Xu et al. 2021b; Wang et al. 2022b). Besides, reconstruction tasks (Li et al. 2020; Xu et al. 2021a; Luo et al. 2020) also prove their effectiveness in learning cross-modality alignments.

a **soldier** is **lying down**

(a)

yeah it's been a while since I've **battled chickens** while I was fixing **mowers** but we're gonna do that today give you guys some entertainment today have a Toro **lawnmower**

(b)

Figure 1: (a) shows a case of alignment on important tokens but not on the video sequence. (b) shows a case of a not wholly aligned video-language pair. The aligned objects are marked in red, and the misaligned objects are marked in blue. The narration words "battled chickens" are irrelevant to the visual content, while the important visual objects "stool" and "signboards" are not described by the words.

Despite the promising performances, previous methods suffer from semantic misalignments (the information expressed by video and language is inconsistent at the semantic level) in downstream tasks. The reason is that previous methods focus on major information but ignore secondary information. Secondary information includes verbs (Hendricks and Nematzadeh 2021) and secondary objective concepts (*e.g.*, position, size) (Salin et al. 2022). Secondary information unalignment leads to semantic unalignment in downstream tasks. For example, in Figure 1 (a), previous models align "a soldier is laying down" with the wrong video due to

[*] Part of work done during an internship at Alibaba Group.
[†] Corresponding authors.

the sole alignments of critical tokens, "solider" and "lying down," but lost the sequence semantic.

To alleviate the semantic unalignment issue, this paper proposes a novel video-language pre-training method, termed video-language pre-training **F**or l**E**arning s**E**mantic a**L**ignments (**FEEL**) to learn better video-language semantic alignments. Specifically, the global modality reconstruction module in FEEL is utilized to convert the representation of a whole video (text) into text (video). Global modality reconstruction reconstructs the entire text and video sequences from another modality, not solely major tokens, thus enhancing the sequence alignment.

Despite the promising performances, the global modality reconstruction suffers from semantic over-alignment. Semantic over-alignment means aligning semantics in two modalities too strictly. The reason is that texts and videos is pre-training data are not perfectly aligned. For example, in Figure 1 (b), "chickens" solely exists in the text, while "stool" and "signboards" solely exist in the video. In the example, the global modality reconstruction forces alignments of inconsistent information, thus leading to the semantic over-alignment issue. To alleviate the issue, Miech et al. (2020) consider a set of multiple text candidates. Besides, DECEMBERT(Tang, Lei, and Bansal 2021) completes missing texts by providing extra dense captions. These methods can only deal with the text loss issue while the video loss issue is ignored. To alleviate the issue, the self-contrasting method in FEEL is proposed to constrain the global modality reconstruction from over-alignment. Specifically, the representation in the original modality is converted to the other modality, and then converted back to the original modality. The distance between the original representation and the twice-converted representation is expected to be smaller than the one between the original representation and the once-converted representation. The reason is that the information in one modality is more consistent than in two modalities. To this end, the triplet loss is utilized by regarding the original representation as an anchor. Thus, the self-contrasting method alleviates the over-alignment issue in the global modality reconstruction.

FEEL is evaluated on two video retrieval tasks: text-based video retrieval (Xu et al. 2016) and text-based video corpus moment retrieval (Lei et al. 2020). The experimental results demonstrate the effectiveness of FEEL on 5 benchmarks compared to state-of-the-art methods. Our contributions can be concluded as follows:

- We propose a novel pre-training method, termed FEEL, to learn a better video-language pre-training model for retrieval-based tasks.
- We propose two novel video-language pre-training techniques to solve semantic misalignments in downstream tasks.
- We conduct extensive experiments on text-based video retrieval and text-based video corpus moment retrieval. The experimental results show the effectiveness of FEEL on 5 benchmarks.

Related Work

Thanks to the success of BERT (Devlin et al. 2019), video-language pre-training works have shown promising ability in cross-modality alignments. Previous pre-training methods learn cross-modality alignments in two manners. First, matching tasks are widely used for cross-modality alignments. VideoBERT (Sun et al. 2019b), ActBert (Zhu and Yang 2020), DECEMBERT (Tang, Lei, and Bansal 2021), and VIOLET (Fu et al. 2021) pre-train matching tasks using the special token [CLS] for binary classification (Ruan and Jin 2022) with a cross-modal encoder (Vaswani et al. 2017). Some methods (Zellers et al. 2021; Ge et al. 2022; Miech et al. 2019, 2020; Ging et al. 2020; Wang et al. 2022b; Yang, Bisk, and Gao 2021; Yan et al. 2021; Luo et al. 2021; Patrick et al. 2020; Cai et al. 2022; Li et al. 2020; Xu et al. 2021b; Cai et al. 2022; Cao et al. 2022) pre-train matching tasks with two-stream encoders by forcing the paired samples closer while pushing different ones away (Ruan and Jin 2022). The others (Luo et al. 2020; Li et al. 2022) combine cross-modal Transformer matching tasks and two-stream encoders matching tasks for more vital learning ability. Second, reconstruction tasks are also widely used (Li et al. 2020; Xu et al. 2021a; Sun et al. 2019b; Tang, Lei, and Bansal 2021; Fu et al. 2021), and some of them show solid abilities for cross-modality alignments. Some methods (Li et al. 2020; Xu et al. 2021a; Fu et al. 2021) reconstruct masked tokens by surrounding tokens with the same modality and information from another modality. Moreover, UniVL (Luo et al. 2020) and Victor (Lei et al. 2021a) introduce the video captioning task to strengthen the ability of text generation. Though the reconstruction process is sequence-oriented, the learning method is based on token-level vocabulary prediction. VLM (Xu et al. 2021a) argues that the information of the same modality accounts for too much and affects the cross-modality alignments in previous reconstruction tasks. Based on this, VLM (Xu et al. 2021a) proposes MMM to encourage the reconstruction tasks only using information from another modality. Despite the promising performances, previous reconstruction tasks focus on alignments of tokens(words/frames/objects) but fail in sequence alignments. However, the methods above focus on major information but ignore secondary information. This paper proposes global modality reconstruction, extending the token reconstructions to global reconstructions for better sequence alignments. Besides, we introduce self-contrasting to avoid semantic over-alignment.

Methodology

Model Architecture

Model architecture of **FEEL** is illustrated in Figure 2, which takes the frames of a video clip and the textual tokens of subtitle sentences as inputs. They are fed into a Video Encoder and a Language Encoder to extract initial representations. **FEEL** computes cross-modal embeddings with a cross-modal transformer. Modal Converting Block is utilized to convert features in one modality into the other modality.

Language Encoder Following HERO (Li et al. 2020), WordPieces (Wu et al. 2016) is utilized to tokenize words.

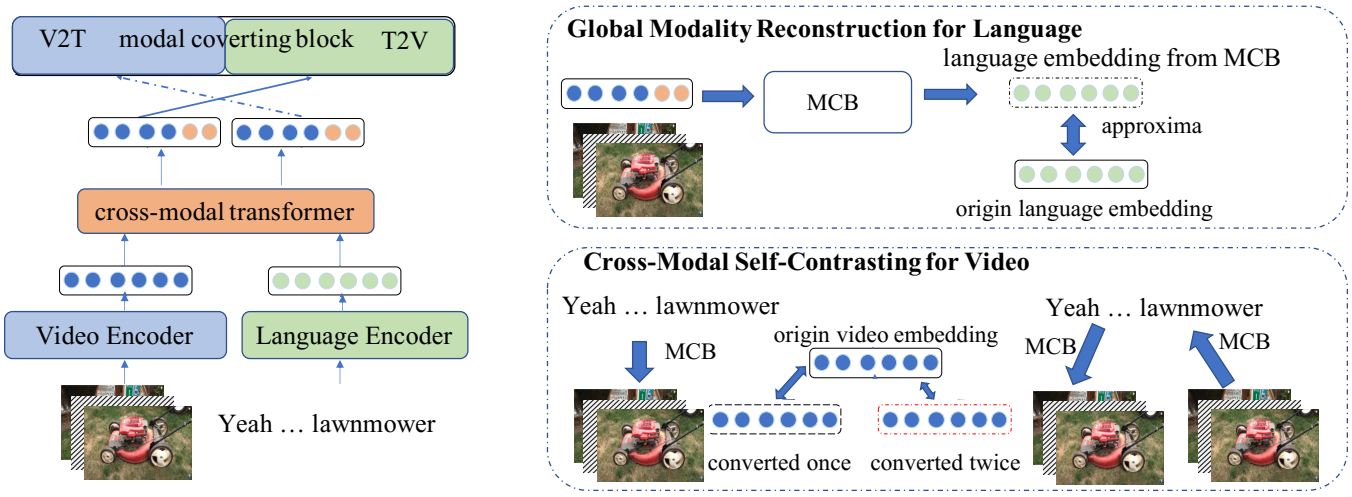

Figure 2: The architecture of **FEEL**. *T2V* converts text features into the video features. *V2T* converts video features into text features. Global modality reconstruction aims to minimize the distance between converted features and original features in one modality. Cross-modal self-contrasting is utilized to avoid language-video over-alignment.

Then, the token sequence is obtained, $\mathbf{t} = [t_1; \ldots; t_n]$, where $\mathbf{t}$ is the obtained token sequence, t_i is the i-th token, and n is the length of the token sequence. Tokens are encoded with pre-trained RoBERTa (Liu et al. 2019). Then, token embeddings are fed to a layer normalization layer. The language embedding for $\mathbf{t}$ is denoted as $\mathbf{e}_t \in \mathbb{R}^{n \times d}$, where d is the embedding dimension.

Video Encoder Video frames are extracted from videos at a fixed rate, $\mathbf{v} = [v_1; \ldots; v_j; \ldots; v_m]$, where $\mathbf{v}$ is the obtained frame sequence, v_j is the j-th frame, and m is the length of the obtained frame sequence. Then, the off-the-shelf visual feature extractors are utilized to obtain features of these frames. Following HERO (Li et al. 2020), Resnet (He et al. 2016) and SlowFast (Feichtenhofer et al. 2019) is utilized to extract 2D and 3D visual features for video frames. A fully connected layer converts the obtained video features to the video embeddings. Then, the embeddings are fed to a layer normalization layer. The video embedding for $\mathbf{v}$ is denoted as $\mathbf{e}_v \in \mathbb{R}^{m \times d}$, where d is the embedding dimension.

Cross-modal Transformer Inspired by UniVL (Luo et al. 2020) and HERO (Li et al. 2020), a multi-layer cross-modal Transformer (Vaswani et al. 2017) is utilized for the video-language fusion. The cross-modal Transformer takes video embedding and language embedding as input to generate the video feature and language feature,

$$\mathbf{o}_v, \mathbf{o}_t = \mathrm{T}_{\mathrm{cross}}\left(\mathbf{e}_v, \mathbf{e}_t\right), \tag{1}$$

where $\mathbf{o}_v \in \mathbb{R}^{m \times d}$ is the video feature, $\mathbf{o}_t \in \mathbb{R}^{n \times d}$ is the language feature, and $\mathrm{T}_{\mathrm{cross}}$ denotes the cross-modal Transformer.

Modality Converting Block **M**odality **C**onverting **B**lock (MCB) consists of two sub-modules, named $\mathrm{MCB}_{\mathrm{T2V}}$ and $\mathrm{MCB}_{\mathrm{V2T}}$. $\mathrm{MCB}_{\mathrm{T2V}}$ and $\mathrm{MCB}_{\mathrm{V2T}}$ are two Transformers. $\mathrm{MCB}_{\mathrm{T2V}}$ converts language features into video embedding space, and $\mathrm{MCB}_{\mathrm{V2T}}$ converts video features into language embedding space,

$$\mathbf{e}'_t = \mathrm{MCB}_{\mathrm{V2T}}\left(\mathbf{o}_v\right), \tag{2}$$

$$\mathbf{e}'_v = \mathrm{MCB}_{\mathrm{T2V}}\left(\mathbf{o}_t\right), \tag{3}$$

where $\mathbf{e}'_t \in \mathbb{R}^{m \times d}$ is the corresponding embedding of the video in the language embedding space, and $\mathbf{e}'_v \in \mathbb{R}^{n \times d}$ is the corresponding embedding of the language in the video embedding space.

Pre-training Tasks

Masked Language Modeling **M**asked **L**anguage **M**odeling (MLM) (Devlin et al. 2019; Liu et al. 2019; Li et al. 2020; Xu et al. 2021a; Luo et al. 2020; Xue et al. 2022; Li et al. 2022; Wang et al. 2022a) aims to complete tokens based on their surrounding words. Following BERT, 15% of tokens are randomly masked. To be specific, if the i-th token is chosen, the i-th token is replaced with the `[MASK]` token with a probability of 80%, replaced with a random token with a probability of 10%, and not replaced with any other tokens with a probability of 10%. Besides, following HERO (Li et al. 2020), video frames $\mathbf{v}$ are also utilized to reconstruct masked tokens for better use of cross-modal information by minimizing the negative log-likelihood,

$$\mathcal{L}_{MLM}(\theta) = -E_{t_{mask} \sim \mathbf{t}} \log P_\theta\left(t_{mask} \mid t_{\neg mask}, \mathbf{v}\right), \tag{4}$$

where t_{mask} is the masked token, $t_{\neg mask}$ are the surrounding tokens around t_{mask} and θ denotes trainable parameters.

Masked Frame Modeling **M**asked **F**rame **M**odeling (MFM) (Li et al. 2020; Luo et al. 2020; Xu et al. 2021a; Fu et al. 2021) aims to complete video frames based on their surrounding frames. Rather than directly frame feature regression, contrastive learning is utilized to enhance the video understanding. **FEEL** regards the origin features of masked frames and the recovered features as positive samples. Features of unmasked frames are regarded as negative samples.

Then, the **n**oise **c**ontrastive **e**stimate (NCE) loss (Sun et al. 2019a) is utilized,

$$\mathcal{L}_{MFM}(\theta) = -E_{v_{mask}\sim\mathbf{v}} \log \mathrm{NCE}\left(v_{mask} \mid v_{\neg mask}, \mathbf{t}\right), \tag{5}$$

$$\mathrm{NCE}\left(v_{mask} \mid v_{\neg mask}, \mathbf{t}\right) = \frac{\exp\left(\mathbf{e}_{v_{mask}}, \mathbf{o}_{v_{mask}}^{\top}\right)}{\exp\left(\mathbf{e}_{v_{mask}}, \mathbf{o}_{v_{mask}}^{\top}\right) + \sum_{v_j\in\mathcal{N}} \exp\left(\mathbf{e}_{v_{mask}}, \mathbf{o}_{v_j}^{\top}\right)}, \tag{6}$$

where $v_{\neg mask}$ is the surrounding frames around v_{mask}. Other frames in the same batch are regarded as negative cases, denoted as $\mathcal{N}$.

Global Modality Reconstruction To explicitly model the alignments between video and language at the sequence level, a global modality reconstruction is introduced based on masked modality modeling. As for the video reconstruction, the $\mathrm{MCB}_{\mathrm{T2V}}$ is utilized to construct the video embedding $\mathbf{e}'_v$ with the language feature $\mathbf{o}_t$. For global video-language alignments at the sequence level, $\mathbf{e}'_v$ is expected to be close to the original video embedding $\mathbf{e}_v$. The **M**ean **S**quared **E**rror (MSE) loss is utilized to calculate the distance between two video embeddings $\mathbf{e}'_v$ and $\mathbf{e}_v$,

$$\mathcal{L}_{G_{t2v}} = \mathrm{MSE}\left(\mathrm{m}\left(\mathbf{e}_v\right), \mathrm{m}\left(\mathbf{e}'_v\right)\right), \tag{7}$$

where m is the mean-pooling operation, and $\mathcal{L}_{G_{t2v}}$ is the MSE loss between $\mathbf{e}'_v$ and $\mathbf{e}_v$.

Similarly, $\mathcal{L}_{G_{v2t}}$ is calculated,

$$\mathcal{L}_{G_{v2t}} = \mathrm{MSE}\left(\mathrm{m}\left(\mathbf{e}_t\right), \mathrm{m}\left(\mathbf{e}'_t\right)\right), \tag{8}$$

where $\mathcal{L}_{G_{v2t}}$ is the MSE loss between $\mathbf{e}'_t$ and $\mathbf{e}_t$.

For global modality reconstruction, the sum of the above two losses is utilized as the global modality reconstruction loss,

$$\mathcal{L}_G = \mathcal{L}_{G_{t2v}} + \mathcal{L}_{G_{v2t}}, \tag{9}$$

Cross-Modal Self-Contrasting The cross-modal self-contrasting is introduced to avoid the over-alignment. For language features, language features are converted twice by MCB. Specifically, language feature $\mathbf{o}_t$ is firstly converted to video embedding $\mathbf{e}'_v$, and then $\mathbf{e}'_v$ is converted to language embedding $\mathbf{e}''_t$,

$$\mathbf{o}_v^{sc\_t}, \mathbf{o}_t^{sc\_t} = \mathrm{T}_{\mathrm{cross}}\left(\mathbf{e}'_v, \mathbf{e}_t\right), \tag{10}$$

$$\mathbf{e}''_t = \mathrm{MCB}_{\mathrm{V2T}}\left(\mathbf{o}_v^{sc\_t}\right), \tag{11}$$

where $\mathbf{o}_v^{sc\_t}$ and $\mathbf{o}_t^{sc\_t}$ are the features obtained by encoding $\mathbf{e}'_v$ and $\mathbf{e}_t$ with the cross-modal Transformer, $\mathbf{e}''_t$ donates the language embedding converted twice by MCB.

To avoid over-alignments, the distance between $\mathbf{e}''_t$ and original language embedding $\mathbf{e}_t$ are expected to be less than the distance between $\mathbf{e}'_t$ and $\mathbf{e}_t$. The reason is that the source information of the former is more consistent. A triplet loss is utilized,

$$\mathcal{L}_{SC_{t2v}} = \delta\left(\mathrm{m}\left(\mathbf{e}''_t\right), \mathrm{m}\left(\mathbf{e}_t\right), \mathrm{m}\left(\mathbf{e}'_t\right)\right), \tag{12}$$

where δ is the triplet loss.

Similarly, for video features, the triplet loss is calculated,

$$\mathbf{o}_v^{sc\_v}, \mathbf{o}_t^{sc\_v} = \mathrm{T}_{\mathrm{cross}}\left(\mathbf{e}_v, \mathbf{e}'_t\right), \tag{13}$$

$$\mathbf{e}''_v = \mathrm{MCB}_{\mathrm{T2V}}\left(\mathbf{o}_t^{sc\_v}\right), \tag{14}$$

$$\mathcal{L}_{SC_{v2t}} = \delta\left(\mathrm{M}\left(\mathbf{e}''_v\right), \mathrm{M}\left(\mathbf{e}_v\right), \mathrm{M}\left(\mathbf{e}'_v\right)\right), \tag{15}$$

where $\mathbf{o}_v^{sc\_v}$, $\mathbf{o}_t^{sc\_v}$ are the features obtained by encoding $\mathbf{e}_v$ and $\mathbf{e}'_t$ with the cross-modal Transformer, $\mathbf{e}''_v$ donates the video embedding converted twice by MCB.

For cross-modal self-contrasting, the total loss is calculated as the sum of the above two losses,

$$\mathcal{L}_{SC} = \mathcal{L}_{SC_{t2v}} + \mathcal{L}_{SC_{v2t}}. \tag{16}$$

Experiments

Pre-training Datasets

The method is pre-trained on Howto100M (Miech et al. 2019), which is collected from YouTube and widely used for video-language pre-training (Li et al. 2020; Xu et al. 2021a,b; Yang, Bisk, and Gao 2021; Gabeur et al. 2020). The videos are cut into 60-second clips and exclude videos in non-English languages and appearing in the downstream tasks to avoid contamination in evaluation. Finally, a subset of 7.56M video clips with ASR (Automatic Speech Recognition) transcripts are obtained.

Application to Text-based Video Retrieval

Fine-tuning Datasets Text-based video retrieval tasks are typical for industry and academia. For a given query, text-based video retrieval needs the model to find the most relevant one from lots of videos. We conduct experiments on two benchmarks, MSR-VTT (Xu et al. 2016) and TVR (Lei et al. 2020).

MSR-VTT (Xu et al. 2016) is a popular dataset for text-based video retrieval.

"Training-7K" follows the data splits from (Yu, Kim, and Kim 2018; Miech et al. 2019, 2020; Tang, Lei, and Bansal 2021; Luo et al. 2020; Li et al. 2020), and "Training-9K" follows the data splits from (Gabeur et al. 2020; Xu et al. 2021a,b; Patrick et al. 2020). The two splits have the same test set but different training sets.

TVR (Lei et al. 2020) is drawn from 6 long-running TV shows, consisting of 109K queries on 21.8K videos accompanied with timestamped subtitles. The average length of videos of TVR is 76.2 seconds. For each video, there are 5 corresponding queries. The dataset is split into four parts: 80% training, 10% validation, 5% testing-public, 5% testing-private.

Comparison with SOTA We use TVR and MSR-VTT to evaluate the performances on text-based video retrieval tasks. The results are shown in Tables 1(a), 2 and 3, respectively. The results show **FEEL** has good performances on the benchmarks. For the MSR-VTT benchmark, **FEEL** outperforms the SOTA method TACO (Yang, Bisk, and Gao 2021) by +0.7% in R@1, +1.4% in R@5, +2.1% in R@10 with the "training-7k" split. Besides, **FEEL** also shows promising results with the "training-9k" split. Following HERO (Li

Mehod	R@1	R@5	R@10
MEE(2018)	7.56	20.78	29.88
XML(2020)	16.54	38.11	50.41
ReLoNet(2021)	16.96	39.28	51.34
ReLoCLNet(2021)	22.13	45.85	57.25
Cascaded MPN(2022)	28.54	-	61.73
*CONQUER(2021)	29.29	-	-
*HERO (Li et al. 2020)	30.11	-	62.69
*FEEL	**31.15**	**55.01**	**64.66**

(a)

Method	Validation			Testing		
	R@1	R@10	R@100	R@1	R@10	R@100
XML(2020)	2.62	9.05	22.47	3.32	13.41	30.52
FLAT(2020)	4.61	11.29	16.24	-	-	-
SAN(2022)	3.64	15.32	34.73	-	-	-
ReLoCLNet(2021)	4.15	14.06	32.42	-	-	-
HAMMER(2020)	5.13	11.38	16.71	-	-	-
Cascaded MPN(2022)	5.27	16.12	**35.11**	-	-	-
*HERO(2020)	5.13	16.26	24.55	6.21	19.34	36.66
*CONQUER(2021)	7.76	22.49	35.17	9.24	28.67	41.98
*FEEL	**6.02**	**19.00**	32.92	**6.91**	**20.70**	**42.20**

(b)

Table 1: Results on TVR benchmark: (a) results of text to video retrieval, (b) results of text-based corpus video moment retrieval. * represents using a pre-trained model. CONQUER is greyed out because of unfair comparison.

Method	R@1	R@5	R@10
JSFusion(2018)	10.2	31.2	43.2
HT(2019)	14.9	40.2	52.8
NoiseE(2021)	17.4	41.6	53.6
DECEMBERT(2021)	17.5	44.3	58.6
HERO(2020)	20.5	47.6	60.9
EAO(2022)	21.0	49.3	60.1
UniVL(2020)	21.2	49.6	63.1
TACO(2021)	24.8	52.1	64.0
FEEL	**25.5**	**53.5**	**66.1**

Table 2: Results of text-video retrieval on MSR-VTT testing set with the training-7k split.

Method	R@1	R@5	R@10
MMT(2020)	26.6	57.1	69.6
UniVL(2020)	27.2	55.7	68.7
VLM(2021a)	28.1	55.5	67.4
VideoClip(2021b)	**30.9**	55.4	66.8
SupportSet(2020)	30.1	58.5	69.3
TACO(2021)	28.4	57.8	71.2
FEEL	30.4	**58.9**	**71.7**

Table 3: Results of text-video retrieval on MSR-VTT testing set with the training-9k split.

et al. 2020) and MMT (Gabeur et al. 2020), ASR information is used. **FEEL** achieves good results because global modality reconstruction can learn the alignments at the sequence level. Besides, cross-modal self-contrasting can handle the misaligned information during pre-training, and forced alignments have been avoided. Note that methods with different pre-training data will not be compared with our method. Some methods (Xue et al. 2022; Fu et al. 2021; Li et al. 2022; Zellers et al. 2021; Lei et al. 2021b) use different dataset for pre-training. For example, ClipBERT (Lei et al. 2021b) use image-text pairs for pre-training, and Rouditchenko et al. (2020) and Gabeur et al. (2022) use additional audio modality.

Application to Text-based Video Corpus Moment Retrieval

Fine-tuning Datasets Compared with text-based video retrieval, text-based video corpus moment retrieval needs to locate the most relevant moment from large video corpus and is more difficult. Three benchmarks are included for text-based video moment retrieval tasks: DiDeMo (Anne Hendricks et al. 2017), How2R (Li et al. 2020), TVR (Lei et al. 2020).

DiDemo (Anne Hendricks et al. 2017) consists of 41.2K unique moments of 10.6K videos from YFCC100M Flickr videos (Thomee et al. 2016). The average length for query and moment is 8 tokens and 6.5 seconds, respectively. The dataset is split into three parts: 80% training, 10% validation, and 10% testing.

How2R (Li et al. 2020) is a new and challenging text-based video moment retrieval benchmark collected from 9K instructional videos in HowTo100M. On average, each clip is accompanied by 2-3 queries.

TVR (Lei et al. 2020) is a closed-world dataset. The moment length is between 0.29 seconds and 123.02 seconds, with an average of 9.1 seconds. The dataset is split into four parts: 80% training, 10% validation,5% testing-public, 5% testing-private.

Comparison with SOTA We use TVR (Lei et al. 2020)[1], How2R (Li et al. 2020), and DiDemo (Anne Hendricks et al. 2017) to evaluate text-based video corpus moment retrieval. The results are shown in Tables 1(b) and 4, respectively. Compared with text-based video retrieval, text-based video corpus moment retrieval is more challenging and requires a more substantial alignment ability of models. Surprisingly, the results show that our model brings significant improvements on the three benchmarks compared with baselines. For the TVR validation set, **FEEL** outperforms HERO by +0.89% in R@1, +2.74% in R@10, +8.37% in R@100. For

[1]We submitted the TVR leaderboard. https://competitions.codalab.org/competitions/22780

Method	R@1	R@5	R@10
XML(2020)	2.06	-	8.96
HERO(2020)	3.01	6.33	8.57
FEEL	**3.32**	**8.34**	**10.27**

(a)

Method	IoU=0.5			IoU=0.7		
	R@1	R@5	R@10	R@1	R@5	R@10
XML(2020)	2.26	-	10.42	1.59	-	6.77
HERO(2020)	3.37	8.79	13.26	2.14	7.73	11.43
CONQUER(2021)	3.31	9.27	13.99	2.79	8.04	11.90
FEEL	**4.02**	**10.75**	**15.22**	**3.34**	**9.67**	**13.54**

(b)

Table 4: Results of text-based corpus video moment retrieval on How2R(a) and DiDeMo (b). Note that the results of HERO on How2R are from VALUE (Li et al. 2021) based on a new version of the How2R benchmark, because the old one is noisy due to short and respective textual queries. Following the suggestion, we also use the new version of the How2R benchmark to evaluate our method.

DiDeMo, **FEEL** even outperforms CONQUER (Hou, Ngo, and Chan 2021) on all three metrics. We think it is unfair to directly compare **FEEL** with methods that optimize the downstream task method while using a pre-trained model, so we grey out method CONQUER, and for methods SAN and Cascaded MPN we report the results only optimize the downstream task method without pre-training. Some methods (Lee, Oh, and Seo 2021; Gao, Liu, and Liu 2021) explore an image-text pre-trained model CLIP (Radford et al. 2021) to a well-designed video moment retrieval architecture. Besides, CUPID (Zhou et al. 2021) selects relevant pre-training data using specific downstream tasks and changes the distribution of the original pre-training dataset. Because of the use of different pre-training data, the results of these methods will not be compared. **FEEL** achieves promising results on all three benchmarks. The good results prove that our alignments are still effective in more complicated tasks.

Ablation Studies

In this section, we conduct ablation studies to verify the effectiveness of **FEEL**. To know the respective contributions of two novel techniques to the final results, We conduct ablation studies on four text-based video retrieval and text-based video corpus moment retrieval. The results are shown in Table 5.

We can see that without global modality reconstruction in the pre-training phase (MLM + MFM *vs* MLM + MFM + GMR), the results on all benchmark decreased significantly. For the MSR-VTT benchmark, without global modality reconstruction, the results drop 6.1 % in R@1, 5.5% in R@5, and 6.5% in R@10.

Besides, we can see that self-contrasting brings improvements to quantitative results (MLM + MFM + GMR + CS *vs* MLM + MFM + GMR). The results show that addressing over-alignment is helpful for for downstream tasks. For the MSR-VTT benchmark, cross-modal self-contrasting brings +3.6% in R@1, +6.1% in R@5 and +4.3% in R@10.

Discussion and Analysis

How does Model Work For a more intuitive view of how global modality reconstruction works for downstream tasks, we use the MSR-VTT benchmark for a case study. We compare some top-1 retrieval results between with and without global modality reconstruction. The results are shown in Figure 3. For example, in the first case is "somebody slices white onion with sharp a knife on the table". The model without global modality reconstruction can align some critical tokens well, such as "knife" and "onion" but fails in aligning "slices". For the second case, important tokens are well aligned, such as "soldier" and "laying down", but the final recall result is wrong. The reason is that the cross-modal information is not well aligned at the sequence level. The third case is "two child playing in the house" [2]. The model without global modality reconstruction aligns well on "child" and "house" but fails in aligning "two". Thus, it can not distinguish between one child and two children. Critical tokens are well aligned for the pre-trained model without global modality reconstruction, while the lack of alignments at the sequence level leads to incorrect semantically misaligned results. These cases show how global modality reconstruction improves the effect and the importance of alignments at the sequence level in the pre-training stage.

Figure 3: Comparison of cases with and without global modality reconstruction. The text marked in red is the query. Results with semantic alignments are on the left, while results without semantic alignments are on the right.

For cross-modal self-contrasting, We verify that our model can better identify the different information be-

[2]The original queries in the dataset have some syntax errors.

Method	video corpus moment retrieval						video retrieval					
	DeDimo			How2R			TVR			MSR-VTT		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
MLM + MFM	1.91	6.13	9.47	2.47	5.10	7.03	22.65	44.28	54.98	15.8	41.9	55.3
MLM + MFM + GMR	2.71	8.39	12.61	3.01	8.19	10.04	30.50	53.24	62.92	21.9	47.4	61.8
MLM + MFM + GMR + CS	3.34	9.67	13.54	3.32	8.34	10.27	31.15	55.01	64.66	25.5	53.5	66.1

Table 5: Ablations on two video retrieval tasks and two video corpus moment retrieval tasks. GMR denotes global modality reconstruction, and CS means cross-modal self-contrasting.

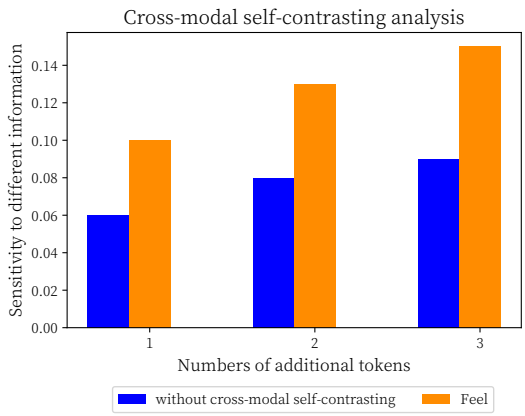

Figure 4: Comparison of sensitivity to different information with and without cross-modal self-contrasting.

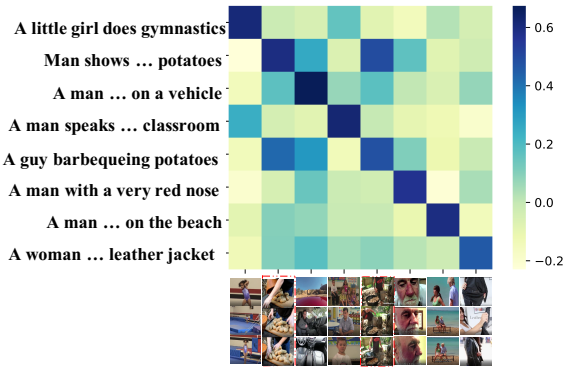

Figure 5: Distribution of similarity scores for video-language pairs on the MSR-VTT benchmark. The score is calculated based on cosine similarity and the range is [-1,1].

tween the two modalities through a specific experiment on MSR-VTT. We sample 100 cases from the testing set. We randomly add some unrelated tokens to the queries and construct hard negative queries (strongly related to origin queries but semantically inconsistent). Intuitively, the scores between hard negative queries and origin correct videos will be reduced in the process of inference with the same model. If the scores decrease significantly, the model is sensitive to different information between the two modalities, which is important for cross-modal retrieval tasks. As shown in Figure 4, after adding unrelated tokens, compared with the pre-trained model without cross-modal self-contrasting, our method significantly change the similarity score between queries and videos. The results show that the new method is more sensitive to misaligned information than the pre-trained model without cross-modal self-contrasting and can better identify the different information.

Visualization For visualization, we illustrate the similarity score distribution of video-text pairs on the MSR-VTT benchmark in Figure 5. Each row represents a query, and each column stands for a video. Three frames are extracted from the corresponding video to show the visual contents. Block in row i, and column j represents the similarity score between the i-th query and the j-th video. The i-th query and the i-th video are matched in the testing dataset. Our model gives the highest scores to the diagonal of the matrix. In other words, our model can align the queries to the videos correctly, even though there are some hard samples like the 2-nd video and the 5-th video. The two videos are both about potatoes and are difficult to distinguish. The results demonstrate the effectiveness of our model on video retrieval.

Discussion on CycleGAN and MirrorGAN Our cross-modal self-contrasting loss is different from CycleGan (Zhu et al. 2017) and MirrorGan (Qiao et al. 2019) in two aspects. First, in terms of goals, CycleGan and MirrorGan are introduced for the image-to-image translation task and text-to-image generation task respectively, while our method is devised to enhance video-language pre-training for two text-to-video retrieval tasks. Second, in terms of model architecture, CycleGan and MirrorGAN are based on the generative adversarial network, while our method is mainly composed of multi-head self-attention blocks and cross-modal transformation is in the latent semantic space.

Conclusion

In this paper, we proposed **FEEL** to learn semantic alignments by video-language pre-training. We introduced the global modality reconstruction to learn the alignments at the sequence level. Besides, cross-modal self-contrasting was applied to handle the misaligned information in the pre-training video-language pairs. We conducted experiments on 5 benchmarks to evaluate the cross-modal alignments. The results showed that our pre-trained model maintained promising performances on text-based video retrieval and text-based video corpus moment retrieval.

Acknowledgments

This research is supported by National Natural Science Foundation of China (Grant No.62276154 and 62011540405), AMiner.Shenzhen SciBrain Fund, Beijing Academy of Artificial Intelligence (BAAI), the Natural Science Foundation of Guangdong Province (Grant No. 2021A1515012640), Basic Research Fund of Shenzhen City (Grant No. JCYJ20210324120012033 and JSGG20210802154402007), and Overseas Cooperation Research Fund of Tsinghua Shenzhen International Graduate School (Grant No. HW2021008).

References

Amrani, E.; Ben-Ari, R.; Rotman, D.; and Bronstein, A. 2021. Noise estimation using density estimation for self-supervised multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 6644–6652.

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, 5803–5812.

Cai, G.; Ge, Y.; Wang, A. J.; Yan, R.; Lin, X.; Shan, Y.; He, L.; Qie, X.; Wu, J.; and Shou, M. Z. 2022. Revitalize Region Feature for Democratizing Video-Language Pre-training. *arXiv preprint arXiv:2203.07720*.

Cao, M.; Yang, T.; Weng, J.; Zhang, C.; Wang, J.; and Zou, Y. 2022. Locvtp: Video-text pre-training for temporal localization. In *European Conference on Computer Vision*, 38–56. Springer.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6202–6211.

Fu, T.-J.; Li, L.; Gan, Z.; Lin, K.; Wang, W. Y.; Wang, L.; and Liu, Z. 2021. VIOLET: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*.

Gabeur, V.; Nagrani, A.; Sun, C.; Alahari, K.; and Schmid, C. 2022. Masking modalities for cross-modal video retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1766–1775.

Gabeur, V.; Sun, C.; Alahari, K.; and Schmid, C. 2020. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, 214–229. Springer.

Gao, Z.; Liu, H.; and Liu, J. 2021. Coarse to Fine: Video Retrieval before Moment Localization. *arXiv preprint arXiv:2110.07201*.

Ge, Y.; Ge, Y.; Liu, X.; Li, D.; Shan, Y.; Qie, X.; and Luo, P. 2022. Bridging Video-Text Retrieval With Multiple Choice Questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16167–16176.

Ging, S.; Zolfaghari, M.; Pirsiavash, H.; and Brox, T. 2020. Coot: Cooperative hierarchical transformer for video-text representation learning. *Advances in neural information processing systems*, 33: 22605–22618.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hendricks, L. A.; and Nematzadeh, A. 2021. Probing Image-Language Transformers for Verb Understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3635–3644.

Hou, Z.; Ngo, C.-W.; and Chan, W. K. 2021. Conquer: Contextual query-aware ranking for video corpus moment retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3900–3908.

Huang, J.; Li, Y.; Feng, J.; Sun, X.; and Ji, R. 2022. Clover: Towards A Unified Video-Language Alignment and Fusion Model. *arXiv preprint arXiv:2207.07885*.

Kim, D.; Yoon, S.; Hong, J. W.; and Yoo, C. D. 2022. Semantic Association Network for Video Corpus Moment Retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1720–1724. IEEE.

Lee, A. S.; Oh, H.; and Seo, M. 2021. ViSeRet: A simple yet effective approach to moment retrieval via fine-grained video segmentation. *arXiv preprint arXiv:2110.05146*.

Lei, C.; Luo, S.; Liu, Y.; He, W.; Wang, J.; Wang, G.; Tang, H.; Miao, C.; and Li, H. 2021a. Understanding chinese video and language via contrastive multimodal pre-training. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2567–2576.

Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T. L.; Bansal, M.; and Liu, J. 2021b. 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.

Lei, J.; Yu, L.; Berg, T. L.; and Bansal, M. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*, 447–463. Springer.

Li, D.; Li, J.; Li, H.; Niebles, J. C.; and Hoi, S. C. 2022. Align and Prompt: Video-and-Language Pre-training with Entity Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4953–4963.

Li, L.; Chen, Y.-C.; Cheng, Y.; Gan, Z.; Yu, L.; and Liu, J. 2020. HERO: Hierarchical Encoder for Video+ Language Omni-representation Pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2046–2065.

Li, L.; Lei, J.; Gan, Z.; Yu, L.; Chen, Y.-C.; Pillai, R.; Cheng, Y.; Zhou, L.; Wang, X.; Wang, W. Y.; Wang, W. Y.; Berg, T. L.; Bansal, M.; Liu, J.; Wang, L.; and Liu, Z. 2021. VALUE: A Multi-Task Benchmark for Video-and-Language Understanding Evaluation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luo, H.; Ji, L.; Shi, B.; Huang, H.; Duan, N.; Li, T.; Li, J.; Bharti, T.; and Zhou, M. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

Luo, J.; Li, Y.; Pan, Y.; Yao, T.; Chao, H.; and Mei, T. 2021. CoCo-BERT: Improving video-language pre-training with contrastive cross-modal matching and denoising. In *Proceedings of the 29th ACM International Conference on Multimedia*, 5600–5608.

Miech, A.; Alayrac, J.-B.; Smaira, L.; Laptev, I.; Sivic, J.; and Zisserman, A. 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.

Miech, A.; Laptev, I.; and Sivic, J. 2018. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*.

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 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.

Patrick, M.; Huang, P.-Y.; Asano, Y.; Metze, F.; Hauptmann, A. G.; Henriques, J. F.; and Vedaldi, A. 2020. Support-set bottlenecks for video-text representation learning. In *International Conference on Learning Representations*.

Qiao, T.; Zhang, J.; Xu, D.; and Tao, D. 2019. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1505–1514.

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

Rouditchenko, A.; Boggust, A.; Harwath, D.; Chen, B.; Joshi, D.; Thomas, S.; Audhkhasi, K.; Kuehne, H.; Panda, R.; Feris, R.; et al. 2020. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*.

Ruan, L.; and Jin, Q. 2022. Survey: Transformer based video-language pre-training. *AI Open*, 3: 1–13.

Salin, E.; Farah, B.; Ayache, S.; and Favre, B. 2022. Are Vision-Language Transformers Learning Multimodal Representations? A Probing Perspective. In *AAAI 2022*.

Shvetsova, N.; Chen, B.; Rouditchenko, A.; Thomas, S.; Kingsbury, B.; Feris, R. S.; Harwath, D.; Glass, J.; and Kuehne, H. 2022. Everything at Once-Multi-Modal Fusion Transformer for Video Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20020–20029.

Sun, C.; Baradel, F.; Murphy, K.; and Schmid, C. 2019a. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*.

Sun, C.; Myers, A.; Vondrick, C.; Murphy, K.; and Schmid, C. 2019b. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7464–7473.

Tang, Z.; Lei, J.; and Bansal, M. 2021. Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In *Proceedings of NAACL-HLT*, 2415–2426.

Thomee, B.; Shamma, D. A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; and Li, L.-J. 2016. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2): 64–73.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, A. J.; Ge, Y.; Yan, R.; Ge, Y.; Lin, X.; Cai, G.; Wu, J.; Shan, Y.; Qie, X.; and Shou, M. Z. 2022a. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*.

Wang, J.; Ge, Y.; Cai, G.; Yan, R.; Lin, X.; Shan, Y.; Qie, X.; and Shou, M. Z. 2022b. Object-aware Video-language Pre-training for Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3313–3322.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Arora, P.; Aminzadeh, M.; Feichtenhofer, C.; Metze, F.; and Zettlemoyer, L. 2021a. VLM: Task-agnostic Video-Language Model Pre-training for Video Understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4227–4239.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021b. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6787–6800.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 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.

Xue, H.; Hang, T.; Zeng, Y.; Sun, Y.; Liu, B.; Yang, H.; Fu, J.; and Guo, B. 2022. Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5036–5045.

Yan, R.; Shou, M. Z.; Ge, Y.; Wang, A. J.; Lin, X.; Cai, G.; and Tang, J. 2021. Video-Text Pre-training with Learned Regions. *arXiv preprint arXiv:2112.01194*.

Yang, J.; Bisk, Y.; and Gao, J. 2021. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11562–11572.

Yoon, S.; Kim, D.; Kim, J.; and Yoo, C. D. 2022. Cascaded MPN: Cascaded Moment Proposal Network for Video Corpus Moment Retrieval. *IEEE Access*, 10: 64560–64568.

Yu, Y.; Kim, J.; and Kim, G. 2018. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision*, 471–487.

Zellers, R.; Lu, X.; Hessel, J.; Yu, Y.; Park, J. S.; Cao, J.; Farhadi, A.; and Choi, Y. 2021. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34: 23634–23651.

Zhang, B.; Hu, H.; Lee, J.; Zhao, M.; Chammas, S.; Jain, V.; Ie, E.; and Sha, F. 2020. A hierarchical multi-modal encoder for moment localization in video corpus. *arXiv preprint arXiv:2011.09046*.

Zhang, H.; Sun, A.; Jing, W.; Nan, G.; Zhen, L.; Zhou, J. T.; and Goh, R. S. M. 2021. Video corpus moment retrieval with contrastive learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 685–695.

Zhou, L.; Liu, J.; Cheng, Y.; Gan, Z.; and Zhang, L. 2021. Cupid: Adaptive curation of pre-training data for video-and-language representation learning. *arXiv preprint arXiv:2104.00285*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

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