

Clover : Towards A Unified Video-Language Alignment and Fusion Model

Jingjia Huang^{◇*} Yinan Li^{♡*†} Jiashi Feng[◇] Xinglong Wu[◇] Xiaoshuai Sun^{♡‡}
 Rongrong Ji[♡]

[♡]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
 Ministry of Education of China, Xiamen University, 361005, China

[◇]ByteDance Inc, 100043, China

yinanlee@stu.xmu.edu.cn, {xssun, rrji}@xmu.edu.cn
 {huangjingjia, jshfeng, wuxinglong}@bytedance.com

Abstract

*Building a universal Video-Language model for solving various video understanding tasks (e.g., text-video retrieval, video question answering) is an open challenge to the machine learning field. Towards this goal, most recent works build the model by stacking uni-modal and cross-modal feature encoders and train it with pair-wise contrastive pre-text tasks. Though offering attractive generality, the resulted models have to compromise between efficiency and performance. They mostly adopt different architectures to deal with different downstream tasks. We find this is because the pair-wise training cannot well align and fuse features from different modalities. We then introduce **Clover**—a Correlated Video-Language pre-training method—towards a universal Video-Language model for solving multiple video understanding tasks with neither performance nor efficiency compromise. It improves cross-modal feature alignment and fusion via a novel tri-modal alignment pre-training task. Additionally, we propose to enhance the tri-modal alignment via incorporating learning from semantic masked samples and a new pair-wise ranking loss. Clover establishes new state-of-the-arts on multiple downstream tasks, including three retrieval tasks for both zero-shot and fine-tuning settings, and eight video question answering tasks. Codes and pre-trained models will be released at <https://github.com/LeeYN-43/Clover>.*

1. Introduction

Video-Language pre-training (VidL) aims to learn generalizable multi-modal models from large-scale video-text

samples so as to better solve various challenging Video-Language understanding tasks, such as text-video retrieval [1, 4, 38, 55] and video question answering [16, 47, 52]. Recent studies [9, 11, 23, 24, 49, 56, 58, 61] have shown that VidL leads to significant performance improvement and achieves state-of-the-art results on various downstream text-video retrieval and video question answering (VQA) benchmarks.

Though achieving encouraging performance, existing VidL models mostly adopt different architectures to deal with different downstream tasks. For the text-video retrieval tasks, they [2, 10, 11, 13, 31, 39] typically use two individual uni-modal encoders for processing video and text data separately, for the sake of retrieval efficiency. While for video question answering tasks, the models usually adopt the multi-modal joint encoder design to learn the association and interaction of different modalities.

Building a unified model capable of solving various Video-Language tasks is a long-standing challenge for machine learning research. A few recent works [9, 24] attempt to learn a unified VidL model for both tasks, which uses the multi-modal encoder to conduct text-video retrieval. However, the model requires an exhaustive pair-wise comparison between the query texts and gallery videos. Given N text queries and M category videos, the computation complexity of the multi-modal encoder model would be $O(NM)$, which makes it infeasible for large-scale video-text retrieval applications. Another straightforward solution is to simply combine the uni-modal and multi-modal encoders (Fig. 1 (a)), and perform the retrieval and VQA tasks through the uni-modal and multi-modal encoders respectively. Its computation complexity for retrieval tasks is only $O(N + M)$. However, the experiment results in Fig. 1 (c) show that, without a carefully designed correlating mechanism between these two types of encoders, the simple combination *i.e.*, *COMB* yields a compromised per-

*Equal Contribution in alphabetical order.

†Work done when interning at ByteDance Inc.

‡Corresponding Author.

formance compared with the models *i.e.* *IND* individually designed for retrieval and VQA.

In this work, we aim to address the above issues and build a unified pre-training model that attains high efficiency and performance simultaneously. We observe: (i) well aligning the features of the video and text from the same data pair is important for text-video matching; (ii) effectively fusing video and text features into unified representations is critical for video-text understanding. However, existing pre-training strategies that rely on either simple supervised or contrastive pre-text tasks hardly achieve promising feature alignment and fusion capability simultaneously. Motivated by these observations, we develop a new VidL method from these two aspects. Specifically, we propose Correlated Video-Language pre-training (Clover), a VidL method that not only unifies Video-Language alignment and fusion, but also makes them mutually boosted. The fused multi-modal representation contains richer context information than the uni-modal representations [11,35]. As an intermediate modality between video and text, the multi-modal representations are good anchors for cross-modality alignment. Meanwhile, keeping the fused representation closer to the uni-modal representation containing consistent semantic information and away from the inconsistent one will enhance the learning of semantic information in the fused modality. Therefore, we propose the Tri-Modal Alignment (TMA) to get Video-Language alignment and fusion mutually boosted, which takes the alignment between the multi-modal representation and text/video representations as an auxiliary objective. We note that since the tri-modal alignment is well compatible with the classical pre-training tasks [3,7,35] *e.g.*, Masked Language Modeling, its computation overhead is negligible. To help the model maintain fine-grained discriminative capability while improving its generalizability, we further introduce a pair-wise ranking loss that urges the model to be aware of the concept missing in masked samples compared to original samples.

Extensive experiments are conducted on multiple downstream tasks, including three retrieval tasks with different experimental setups (*i.e.* zero-shot and fine-tune) and eight video question answering tasks. The results demonstrate that Clover is able to get the cross-modal fusion and alignment capability mutually improved, and consistently outperforms current SOTAs on various downstream tasks. It achieves an average performance improvement of 4.9% and 8.7% Recall@10 score on the zero-shot and fine-tune settings of the three downstream retrieval datasets, while the average accuracy improvement over current SOTAs is 2.3% on the eight video question answering datasets.

In summary, we make the following contributions: (1) we introduce Clover, a pre-training method achieving unified Video-Language alignment and fusion model that can be easily transferred to various downstream video under-

standing tasks while attaining both high efficiency and performance; (2) we propose a novel tri-modal alignment pre-training task, which correlates the uni-modal encoder and multi-modal encoder to get them mutually boosted.

2. Related Work

Video-Language pre-training for text-video retrieval.

Existing pre-training methods for text-video retrieval can be mainly divided into two categories. The first kind of method employs two individual encoders to embed video and text, and project them into a common latent space [2,10,11,13,31,39,42,50,54,59]. Then, they leverage a contrastive objective [48] to align the paired representations. Thanks to their high efficiency, this category of methods is widely used in real-world applications. The other kind of method utilizes a joint multi-modal encoder to learn comprehensive representation for a given video-text pair, and predict whether the video and text are matched or not via a binary classifier [9,23,25,45,53,63]. Despite the good performance they achieved, they need to exhaustively pair all the videos and texts, and feed them into the model during inference, making them highly inefficient.

Video-Language pre-training for video question answering.

Video question answering aims to automatically answer natural language questions given a context video, which requires a joint understanding of both the video and language. Previous works mainly focus on designing better spatio-temporal attention networks [8,16,26,29,52,62] and question-video relation networks [15,19,22,41] to extract more accurate multi-modal representations. Recently, many works leverage transformer-based models to build an additional multi-modal encoder upon the uni-modal encoders [9,24,43] or input image patches and word tokens [58,63] directly to fuse cross-modality features. The multi-modal encoder is typically driven by the pre-training tasks *e.g.* Masked Language Modeling [7] and Visual-Text Matching [46], to perform token-level cross-modal fusion.

Video-Language pre-training for multiple downstream tasks.

There are a few recent works trying to construct a unified pre-trained model for multiple downstream tasks. ActBert [63], clipBert [23], VIOLET [9] and ALPRO [24] utilize a cross-modal encoder for both retrieval and VQA tasks. They take video-text retrieval as a binary classification task and construct the classifier upon a visual-text matching head, which brings about extreme computation overload. While All-in-One [49] and HD-VILA [56] proposes a transformer-based unified backbone architecture for joint-modal representation learning, they require a large amount of data for pre-training. HERO [25] and VLM [53] design a task-agnostic model that can be finetuned on multiple downstream tasks. However, they are not end-to-end trainable and need pre-extract video features, which limits the performance of the model.

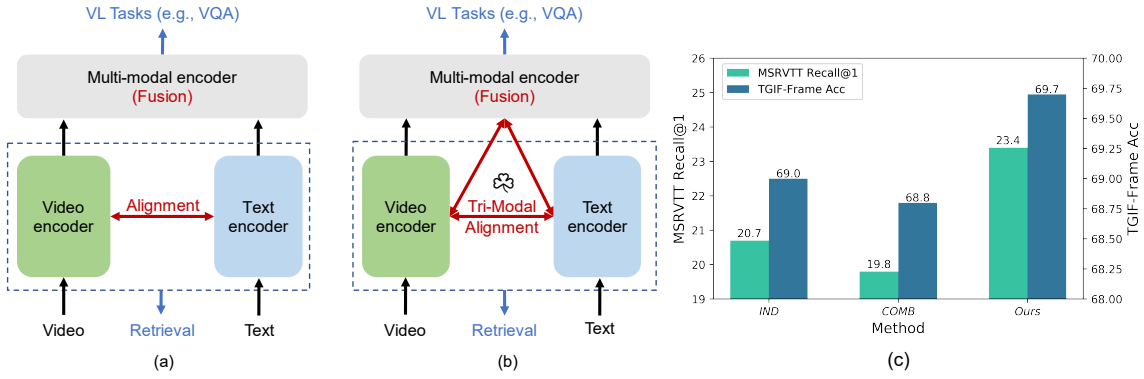


Figure 1. Different attempts on achieving a universal VidL model. (a) Naive combination by stacking the cross-modal encoder upon the uni-modal encoders with weak association (*COMB*). (b) Our proposed Clover pre-trained model that better correlates the cross-modal encoder and the uni-modal encoders via Tri-Modal Alignment. (c) Comparisons of the transfer performance on MSRVTT [60] (zero-shot) and TGIF-FrameQA [16] datasets. Benefiting from the proposed tri-modal alignment, our proposed universal model outperforms both the individually designed models (*IND*) and *COMB*.

Unlike the aforementioned methods, our Clover is a fully end-to-end VidL model that can be easily transferred to various downstream tasks while attaining both high efficiency and performance.

Note that there are some methods focus on transferring the pre-trained CLIP model to downstream tasks and achieve significant performance [6, 14, 17, 21, 30, 37, 51, 57]. The CLIP model is pretrained with 400M image-text data pairs. In contrast, our Clover focuses on presenting a better video-text pretrained model, and only leverages 5M data pairs for pretraining.

3. Method

3.1. Motivation and Overview

Recent years have witnessed a significant progress in Video-Language pre-training. Nevertheless, building a universal Video-Language model for solving various video understanding tasks (*e.g.*, text-video retrieval, video question answering) is still a challenging problem. The key to the problem lies in how to unify the representation learning for cross-modal alignment and fusion efficiently.

Cross-modal *alignment* aims to learn projection functions $f(\cdot)$ and $g(\cdot)$ that project videos and texts into a common embedding space where the similarity between the video and text with consistent semantic information is maximized and otherwise minimized. Taking a video V as the anchor, the learning objective of cross-modal alignment is formulated as:

$$\arg \max_{f,g} [s(f(V), g(T^+)) - s(f(V), g(T^-))], \quad (1)$$

where T^+ , T^- represent texts that are semantically consistent and inconsistent with V , respectively. The function $s(\cdot, \cdot)$ measures the dot-product similarity between two embeddings. A model with strong cross-modal alignment capability is desired in the video-retrieval task, which requires matching the semantically aligned videos and texts.

Cross-modal *fusion* aims to integrate the correlation and interaction carried by the video and text modalities into a unified multi-modal embedding. Specifically, it can be formulated as learning a function $Fusion(\cdot, \cdot)$ that takes inputs of different modals and outputs a unified representation $M = Fusion(V, T)$, which is then used to solve downstream tasks like VQA.

Existing pre-training strategies rely on either simple supervised or contrastive pre-text tasks that hardly achieve feature alignment and fusion simultaneously. To better correlate the cross-modal alignment and fusion, we propose the Correlated Video-Language pre-training method (Clover) with three key pre-training innovations (Sec. 3.2): tri-modal alignment; pair-wise ranking loss; and semantic enhanced masked language modeling. The classical masked language modeling task is also incorporated into our method, which is able to boost the generalizability of the model as well as the interaction between visual and language.

Architecture. We briefly explain the architecture of Clover here, which is shown in Fig. 2 as well. It consists of three components: a video encoder, a text encoder and a multi-modal encoder. Following [9], we use VideoSwin Transformer [32] as our video encoder. Given a video V , it outputs a sequence of video embeddings: $V_e = \{v_1, \dots, v_K\} \subset \mathbb{R}^D$, where K is the number of flattened patches. The text encoder of Clover is a 12-layer bidirectional transformer encoder model [7]. Given an input text sentence T , the encoder outputs an embedding sequence $T_e = \{t_{cls}, t_1, \dots, t_{L-1}\} \subset \mathbb{R}^D$, where t_{cls} indicates the embedding of the [CLS] token. We employ a 3-layer bidirectional Transformer encoder to learn the fused cross-modal representation of a video-text pair. The multi-modal encoder takes the concatenation of video and text embeddings as input, and outputs the fused multi-modal embeddings $M_e = \{m_{v_1}, \dots, m_{v_K}, m_{CLS}, m_{t_1}, \dots, m_{t_{L-1}}\}$, where $m \in \mathbb{R}^D$.

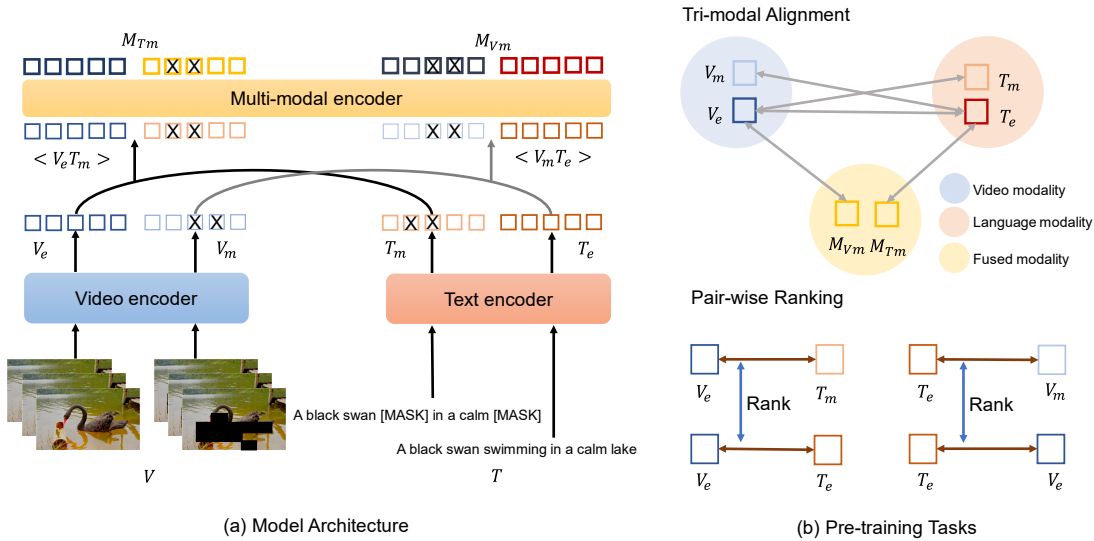


Figure 2. Overview of Clover . (a) Model architecture. (b) Tri-modal alignment pre-training task with masked samples and the pair-wise ranking. “x” indicates the embedding of the token [MASK]. For more detail on model architecture and pre-training tasks refers to Sec. 3.

3.2. Tri-Modal Alignment

Existing VidL pre-training methods typically align the embeddings of data from different modalities through pair-wise contrastive learning. Differently, we propose a Tri-Modal Alignment (TMA) learning task that not only enforces the video and text modalities to be aligned but also encourages them to well align with the third modality, *i.e.*, their fusion modality. In TMA, as an intermediate modality between video and text, the fused multi-modal representations act as anchors for cross-modality alignment, which reduces the difficulty of the direct alignment between video and text. Meanwhile, it keeps the fused representation closer to the uni-modal representation containing consistent semantic information and away from the others, so as to enhance the learning of semantic information in the fused representation.

As shown in Fig. 2 (a), given an input video-text pair $\langle V, T \rangle$ and their embeddings $\langle V_e, T_e \rangle$ from the uni-modal encoder, we mask some regions in V to make a masked video with embedding V_m , and similarly mask some tokens in T to make a masked text embedding T_m . Then, we pair the incomplete samples with the complete samples as $\langle V_m, T_e \rangle$ and $\langle V_e, T_m \rangle$. We adopt multi-modal encoder to get fused multi-modal embedding sequence M_{V_m} and M_{T_m} of $\langle V_m, T_e \rangle$ and $\langle V_e, T_m \rangle$, respectively. We denote the embedding of [CLS] token in M_{V_m} and M_{T_m} as $M_{V_m f}$ and $M_{T_m f}$.

For the convenience of description, we use superscripts i and j to index the data sample, and introduce TMA from the perspective of video and language, respectively. Firstly, for a video representation V_e^i , besides its associated text embedding T_e^i and T_m^i , we also consider its associated fusion embedding $M_{V_m f}^i$ as its positive pair—that should be pushed closer within the embedding space. To align these tri-modal representations, we propose a novel exclusive-NCE loss computed within a batch of B samples as follows:

$$L_v = - \sum_{i=1}^B \left[\log \frac{e^{s(V_e^i, T_e^i)}/\tau}{e^{s(V_e^i, T_e^i)}/\tau + Z} + \log \frac{e^{s(V_e^i, T_m^i)}/\tau}{e^{s(V_e^i, T_m^i)}/\tau + Z} + \log \frac{e^{s(V_e^i, M_{V_m f}^i)}/\tau}{e^{s(V_e^i, M_{V_m f}^i)}/\tau + Z} \right], \text{ where}$$

$$Z = \sum_{j \neq i}^B \left[e^{s(V_e^i, T_e^j)}/\tau + e^{s(V_e^i, T_m^j)}/\tau + e^{s(V_e^i, M_{V_m f}^j)}/\tau \right]. \quad (2)$$

Here τ is a temperature scalar, and $s(\cdot, \cdot)$ is the dot-product similarity function to measure the degree of alignment between different modalities. In each term of Eq. (2), we exclude the other positive pairs when performing contrastive learning on one positive pair. Thus the intra-suppression among positive pairs is effectively avoided, and the video-/text-/fusion-modal representations are better aligned. For example, in the first term of Eq. (2), $e^{s(V_e^i, M_{V_m f}^i)}/\tau$ and $e^{s(V_e^i, T_m^i)}/\tau$ are excluded from the denominator, which prevents the model from yielding the sub-optimal solution that gets T_e^i closer to V_e^i while pushes T_m^i and $M_{V_m f}^i$ away from V_e^i . Meanwhile, to align the video to the text and fusion modalities, we have the following objective:

$$L_{v'} = - \sum_{i=1}^B \left[\log \frac{e^{s(T_e^i, V_e^i)}/\tau}{\sum_{j=1}^B e^{s(T_e^i, V_e^j)}/\tau} + \log \frac{e^{s(T_m^i, V_e^i)}/\tau}{\sum_{j=1}^B e^{s(T_m^i, V_e^j)}/\tau} + \log \frac{e^{s(M_{V_m f}^i, V_e^i)}/\tau}{\sum_{j=1}^B e^{s(M_{V_m f}^i, V_e^j)}/\tau} \right]. \quad (3)$$

The above two losses (2) and (3) are summed up to give the tri-modal alignment objective w.r.t. the visual modality: $L_V = L_v + L_{v'}$. Similarly, from the perspective of text modality, we define the tri-modal alignment objective $L_T = L_t + L_{t'}$ w.r.t. the text modality.¹ Then, we obtain the final loss of tri-modal alignment $L_{TMA} = L_V + L_T$.

¹For the detail of L_T , please refer to the supplementary material.

To get the masked samples for TMA, we conduct masking over the videos and texts to form masked samples using the following two strategies.

Video-block masking strategy. Objects in a video often appear at similar spatial locations across consecutive frames, which compose a tube area in the spatial-temporal domain. In order to construct incomplete video samples with less information leakage, we extend the block-wise mask [3] to video via performing a block-wise mask at the same position in all video frames. In this paper, we randomly replace 20% patches with a learnable mask token to obtain V_m .

Semantic text masking strategy. The detailed semantic information in the text basically lies in the verb, noun and adjective words. To facilitate learning their representations, we construct incomplete text samples by randomly masking some verbs, nouns and adjectives in the sentence. Specifically, given a sentence, we use a part-of-speech tagger [33] to tag each word. Then we pick the verb phrases and nouns, and replace 30% of them with a special [MASK] token. Note that to avoid drastic semantic changes, we do not mask the auxiliary verb like *have*, *should*, *will*, *would*, etc. With the masking strategy, when some key elements are masked from the texts, it would form an expression carrying partial information rather than completely changing the information expressed.

3.3. Training Objective

Pair-wise ranking. In TMA, we take the masked pairs $\langle V_e, T_m \rangle$ and $\langle V_m, T_e \rangle$ as positive pairs, considering that masked samples would still carry partial meaningful information though some others are missed. For instance, as shown in Fig. 2 (a), given the text “A black swan swimming in a calm lake”, “swimming” and “lake” are masked with [MASK] tokens, generating the masked pair $\langle V_e, T_m \rangle$. Although some information is removed, the masked sentence still contains other concepts that make it partially matched with the video, such as the “black swan”. Therefore, in contrastive learning, V_m and T_m can be considered pseudo positive candidates for T and V . However, compared to the original pair $\langle V_e, T_e \rangle$, we consider that the semantic consistency in the masked pairs should be weaker. Based on this prior, we further propose a pair-wise ranking loss:

$$L_{rank} = \max(0, -(\text{sim}(V_e, T_e)/\tau - \text{sim}(V_e, T_m)/\tau) + \lambda) + \max(0, -(\text{sim}(V_e, T_e)/\tau - \text{sim}(V_m, T_e)/\tau) + \lambda), \quad (4)$$

where $\lambda > 0$ is a margin hyper-parameter. Eq. (4) urges the model to be aware of the gap of semantic consistency between $\langle V_e, T_m \rangle / \langle V_m, T_e \rangle$ and $\langle V_e, T_e \rangle$ brought about by concepts missing in the masked pairs. With the pair-wise ranking objective, our model is able to maintain fine-grained perceptual capability while improving its generalizability.

Semantic enhanced masked language modeling. Masked

language modeling (MLM) is a classical pre-training task in VidL, which promotes the interaction between different modalities in the cross-modal encoder. Combining the classical MLM with our semantic text masking strategy, we present the Semantic Enhanced Masked Language Modeling task, which facilitates the representation learning on the key concepts, *i.e.*, verb, noun and adjective words. Besides, considering the class-imbalance between different words, we use focal loss [28] instead of the traditional cross-entropy loss to improve the MLM loss. We apply the MLM loss to the reconstruction of the masked text tokens *i.e.*, $m_{t_j} \in M_{T_m}$, where $t_j = [Mask]$. The MLM loss is defined as:

$$L_{mlm} = -\frac{1}{B} \sum_{i=1}^B \sum_{m_{t_j} \in M_{T_m}^i} [(1 - p_{t_j}^i)^\gamma p_{t_j}^i] \quad (5)$$

where $p_{t_j}^i$ denote the predicted probability distribution of the masked token t_j in i th sentence in the batch, and γ is a hyper-parameter. Finally, the overall pre-training objective of Clover is:

$$L = L_{TMA} + L_{rank} + L_{mlm}. \quad (6)$$

4. Experiments

4.1. Experiment Setup

Pre-training datasets. Following recent work [2, 11, 24], we jointly pre-train our Clover on a video dataset We-Vid2M [2] with 2.5M video-text pairs and an image dataset Google Conceptual Captions (CC3M) [44] with 3.3M image-text pairs (we only obtain 2.8M image-text pairs in CC3M due to image url broken). During pre-training, we treat image data as one frame video data.

Downstream tasks. We evaluate our proposed models on the following downstream tasks. (a) **Text-to-Video Retrieval** on MSRVT [55], LSMDC [38] and DiDeMo [1]. For DiDeMo, we follow [23, 31] and evaluate Clover on the paragraph-to-video retrieval, where text sentences for each video are concatenated together as one text query. We do not use the ground-truth proposal for fair comparison with previous works; (b) **Multiple-choice QA** on TGIF-Action [16], TGIF-Transition [16], MSRVT-MC [60] and LSMDC-MC [47]; (c) **Open-Ended QA** on TGIF-Frame [16], MSRVT-QA [52], MSVD-QA [52] and LSMDC-FiB [38]. For retrieval tasks, we only use the two uni-modal encoders of Clover for fine-tuning and inference. We adopt the two encoders to get the video and text embeddings, and calculate their cosine similarity for retrieval. For QA tasks, we use all the modules for fine-tuning and inference. More details on these datasets and their evaluation usage are provided in the supplementary material.

Method	Pre-training dataset	DiDeMo				MSRVTT				LSMDC			
		R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR
Fine-tune													
clipBert [23]	COCO [5], VG [20]	20.4	48.0	60.8	6	22.0	46.8	59.9	6	-	-	-	-
Frozen [2]	W2M+C3M	31.0	59.8	72.4	3	31.0	59.5	70.5	3	15.0	30.8	39.8	20
VIOLET [9]	W2M+C3M+Y180M	32.6	62.8	74.7	-	34.5	63.0	73.4	-	16.1	36.6	41.2	-
HD-VILA [56]	HDV100M	28.8	57.4	69.1	4	35.6	65.3	78.0	3	17.4	34.1	44.1	15
ALPRO [24]	W2M+C3M	35.9	67.5	78.8	3	33.9	60.7	73.2	3	-	-	-	-
TMVM [27]	W2M+C3M	36.5	64.9	75.4	3	36.2	64.2	75.7	3	17.8	37.1	45.9	13.5
All-in-1 [49]	W2M+H100M	32.7	61.4	73.5	-	37.9	68.1	77.1	-	-	-	-	-
OA-Trans [50]	W2M+C3M	34.8	64.4	75.1	3	35.8	63.4	76.5	3	18.2	34.3	43.7	18.5
MILES [12]	W2M+C3M	36.6	63.9	74.0	3	37.7	63.6	73.8	3	17.8	35.6	44.1	15.5
MCQ [11]	W2M+C3M	37.0	62.2	73.9	3	37.6	64.8	75.1	3	17.9	35.4	44.5	15
Clover (ours)	W2M+C3M	50.1	76.7	85.6	1	40.5	69.8	79.4	2	24.8	44.0	54.5	8
Zero-shot													
MIL-NCE [39]	H100M	-	-	-	-	9.9	24.0	32.4	29.6	-	-	-	-
Frozen [2]	W2M+C3M	21.1	46.0	56.2	7	18.7	39.6	51.6	10	9.3	22.0	30.1	51.0
VIOLET [9]	W2M+C3M+Y180M	23.5	49.8	59.8	-	25.9	49.5	59.7	-	-	-	-	-
ALPRO [24]	W2M+C3M	23.8	47.3	57.9	6	24.1	44.7	55.4	8	-	-	-	-
OA-Trans [50]	W2M+C3M	23.5	50.4	59.8	6	23.4	47.5	55.6	8	-	-	-	-
MILES [12]	W2M+C3M	27.2	50.3	63.6	5	26.1	47.2	56.9	7	11.1	24.7	30.6	50.7
MCQ [11]	W2M+C3M	25.6	50.6	61.1	5	26.0	46.4	56.4	7	12.2	25.9	32.2	42
Clover (ours)	W2M+C3M	29.5	55.2	66.3	4	26.4	49.5	60.0	6	14.7	29.2	38.2	24

Table 1. Text-to-video retrieval performance comparison under fine-tune and zero-shot setups. Here higher R@k (Recall K) and lower MedR (Median Recall) indicate better performance. W2M, C3M, H100M, HDV100M, Y180M are short for WebVid2M [2], CC3M [44], HowTo100M [40], HD-VILA-100M [56], YT-Temporal-180M [61], respectively.

Method	Pre-training dataset	TGIF			MSRVTT		LSMDC		MSVD
		Action	Transition	Frame	MC	QA	MC	FiB	QA
clipBert [23]	COCO, VG	82.8	87.8	60.3	88.2	37.4	-	-	-
JuskAsk [58]	HTVQA69M [58]	-	-	-	-	41.5	-	-	46.3
ALPRO [24]	W2M+C3M	-	-	-	-	42.1	-	-	45.9
All-in-1 [49]	W2M+H100M	92.7	94.3	64.2	92.0	42.9	83.1	-	47.9
VIOLET [9]	W2M+C3M+Y180M	92.5	95.7	68.9	91.9	43.9	82.8	53.7	47.9
MERLOT [61]	Y180M	94.0	96.2	69.5	90.9	43.1	81.7	52.9	-
Clover (ours)	W2M+C3M	95.0	98.2	71.6	95.2	44.1	83.7	54.1	52.4

Table 2. Performance comparison on transferring to downstream video question answering tasks.

Implementation details. Following [9], we initialize the video encoder with a VideoSwin-Base [32] model pre-trained on Kinetics-400 [18]. The text encoder is initialized from pre-trained Bert-Base [7]. The multi-modal encoder is initialized from the first three layers of the pre-trained Bert-Base model. We train Clover end-to-end during both pre-training and fine-tuning. We pre-train Clover for 40 epochs, using a batch size of 1024 on 64 NVIDIA A100 GPUs. We use AdamW [34] optimizer with a weight decay 0.005 and betas (0.9, 0.98). The learning rate is first warmed-up by 4 epochs to $5e-5$ and then decays following a cosine annealing decay schedule. We resize all the video frames to 224×224 and split each frame into patches with a size of 32×32 . We choose hyper-parameter $\tau = 0.05$, $\lambda = 5$ and $\gamma = 2$. We set dimension $D = 768$ in our model. For each video, we randomly sample 8 frames while preserving their order in-between. For fine-tuning on retrieval tasks, we only fine-

tune the uni-modal encoders of Clover with InfoNCE [48] loss. For fine-tuning on video QA task, we add a simple MLPs that takes the multi-modal [CLS] embedding as input for classification, and optimize the whole Clover model with cross-entropy loss. During fine-tuning, we follow the conventional set-up in [9, 36], and all the fine-tuning experiments are performed on 8 NVIDIA A100 GPUs.

4.2. Comparing to State-of-the-art

Text-to-video retrieval. Tab. 1 shows the retrieval results on the DiDeMo [1], MSRVTT [55] and LSMDC [38] datasets with both zero-shot and fine-tuning settings. Our Clover surpasses previous works on all the datasets by a large margin. The significant performance gain under zero-shot evaluation demonstrates its stronger generalization ability. Specifically, Clover brings 4.9% gain in R@10 over MCQ [11], the current SOTA method on re-

Method	DiDeMo				MSRVTT				LSMDC-MC	TGIF-Frame
	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	Acc	Acc
Baseline	22.6	47.9	58.2	7	19.8	41.7	51.1	10	78.8	68.9
+ TMA	24.7	49.7	60.0	6	22.7	42.2	52.2	9	80.0	69.3
+ TMA+SM	25.3	49.5	60.5	6	22.5	42.7	52.0	9	80.5	69.4
+ TMA+SM+RankL	26.4	51.1	61.3	5	23.4	43.3	52.4	9	80.7	69.7

Table 3. Effects of our pre-training tasks for Clover. We report *zero-shot* text-video retrieval performance on DiDeMo and MSRVTT, and *fine-tune* video QA performance on LSMDC-MC and TGIF-Frame. TMA, SM and RankL: tri-modal alignment, semantic masking strategy and pair-wise ranking loss.

Method	Averaged similarity scores of positive pairs	Averaged similarity margin between positive and negative samples
w/o TMA	0.56	0.30
w TMA	0.65	0.35

Table 4. Affect of TMA on the video and text representations in learned embedding space. The experiment is conducted on MSRVTT-1kA test set under zero-shot setting.

retrieval tasks with the same pre-train data, averaged on the three datasets. Moreover, Clover outperforms VIOLET [9], though VIOLET adopts a considerably larger pre-training dataset (YTT180M [61] of 180M visual-text pairs vs. WebVid2M+CC3M of only 5.5M samples). Clover also presents superior performance when fine-tuned on the downstream datasets, compared with the SOTAs. On all three datasets, Clover surpasses previous works by a large margin across all the metrics and obtains an average improvement of 8.7% on R@10. Moreover, different from some previous work [9, 23, 24] that adopts a joint multi-modal encoder and requires exhaustively pairing every video and text during retrieval, Clover directly deploys the video encoder and text encoder for retrieval tasks, making it much more efficient.

Video question answering. Tab. 2 reports the results of Clover and current SOTAs on four video QA datasets. Though using much less pre-training data, Clover outperforms another unified VidL model *i.e.*, VIOLET [9] significantly. Moreover, compared to the current SOTA method MERLOT [61], Clover brings improvements of 1.0% on TGIF-Action, 2.0% on TGIF-Transition, 2.1% on TGIF-Frame, 4.3% on MSRVTT-MC, 1.0% on MSRVTT-QA, 2.0% on LSMDC-MC and 1.2% on LSMDC-FiB. Note that JustAsk [58] and MERLOT [61] are specifically designed for video QA and trained on orders of magnitude larger datasets (*e.g.*, HTVQA69M [58], YTT180M [61]). The experiment results again confirm the superiority of Clover.

4.3. Analysis

We conduct ablation experiments on two retrieval datasets (DiDeMo and MSRVTT) and two Video QA datasets (TGIF-Frame and LSMDC-MC). Due to the computational resource limit, we randomly sample 1 million video-text pairs from WebVid2M [2] to build the WebVid-1M subset for model pre-training under all the ablation stud-

ies. We keep the model architecture unchanged but pre-train the model with only MLM and InfoNCE losses as a baseline, which is similar to the baseline adopted in VIOLET [9]. The only difference is that we replace the ITM applied to the cross-modal encoder outputs in VIOLET [9] with InfoNCE applied to the outputs of uni-modal encoders and separate the text encoder from the cross-modal encoder. All the ablation experiments are conducted with batch size 1024 on 32 NVIDIA A100 GPUs.

Effect of tri-modal alignment. Tri-modal alignment (TMA) aims to better correlate cross-modal alignment and fusion. To evaluate the contribution of TMA, we remove the pair-wise ranking objective and semantic masking strategy, and employ the classical MLM task as in the baseline method. Compared with the baseline, TMA explicitly associates the outputs of uni-modal encoders with the outputs of the multi-modal encoder. As shown in Tab. 3, the model trained with TMA outperforms the baseline on both retrieval and video QA, demonstrating its effectiveness. For a better understanding of the effect of TMA, we further report averaged cosine similarity of videos and texts in MSRVTT dataset (in zero-shot setting) calculated by models trained with/without TMA, respectively. For a fair comparison, we select all video queries (197 in total) that were successfully responded to by both models. We recall 100 results for each query, and take the ground-truth pairs as positive and others as negative. As shown in Tab. 4, the model armed with tri-modal alignment assigns higher similarity scores to positive pairs compared to the model without it. Meanwhile, with tri-modal alignment, the margin between the positive and negative samples is also increased. It reveals that with the help of tri-modal alignment, the distance between video and text that contain consistent semantics is reduced, while the margin between the misaligned samples is increased. The results demonstrate that with fused multi-modal representation as anchors, tri-modal alignment helps the video and language modalities to be better aligned.

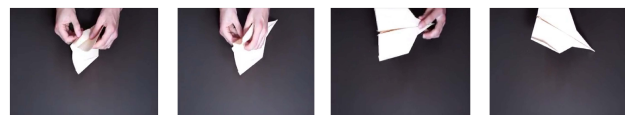
Effect of semantic masking strategy. We incorporate a semantic masking strategy to form the masked pairs in tri-modal alignment task. As shown in Tab. 3, on more diverse evaluation benchmarks, *e.g.* DiDemo, where each video is paired with 5 different ground-truth texts, using masked samples brings more gains. The result reveals that the pro-

Method	Model Architecture		Training Objectives			MSRVTT	DiDeMo	TGIF-Frame	LSMDC-MC
	Two uni-modal encoder	Cross-modal encoder	InfoNCE	MLM	TMA + Rankloss	R@1	R@1	Acc	Acc
IND-A	✓		✓			20.7	22.9	-	-
IND-F	✓	✓		✓		-	-	69.0	79.0
COMB	✓	✓	✓	✓		19.8	22.6	68.9	78.8
Clover	✓	✓		✓	✓	23.4	26.4	69.7	80.7

Table 5. Comparisons with task-independently trained models (*IND*) and naive combination of the uni- and cross- modal encoder (*COMB*)



Baseline: a man is showing the interior of a car
Clover: inside of a car with large space and safety



Baseline: a man is folding pieces of paper
Clover: a person folds a paper airplane

Figure 3. Qualitative results of zero-shot video to text retrieval results on MSRVTT [55].

posed semantic masking strategy facilitates the model to capture key semantic information in more complex scenes and makes it achieve better results.

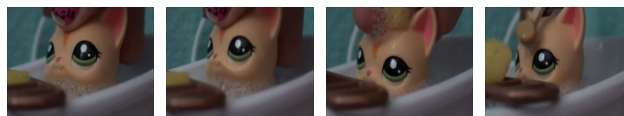
Effect of pair-wise ranking. To make the model capture the semantic information difference between the masked pairs and complete pairs, we adopt the pair-wise rank loss. As shown in Tab. 3, with pair-wise ranking loss, the performance of the model is further improved. We also show the visualization results below to further illustrate the effect of pair-wise ranking.

Clover makes cross-modal alignment and fusion mutually improving. We compare Clover with the ones that use the same architecture but employ different pre-training tasks in Tab. 5. *IND-A* represents the model only trained with InfoNCE loss for cross-modal alignment. *IND-F* indicates the model trained with MLM loss for cross-modal fusion. We also report results achieved by the model (*i.e.*, *COMB*) that simply combines the uni-modal encoders and multi-modal encoder as well as the training objectives. We can see that the simple combination hurts the performance, while our Clover achieves superior performance than the task-independently trained *IND-A* and *IND-F*. It reveals that Clover is able to get the model’s cross-modal alignment and fusion capability mutually enhanced.

Qualitative analysis. In Fig. 3 and Fig. 4, we show the qualitative results of Clover and the baseline method on zero-shot video-text retrieval and VQA tasks. Specifically, we present the query videos with the matched texts, and the video-question pair with the answer. Clover empowers the model with stronger video-text understanding capabil-



Question: What is a guy hugging?
Baseline: **Animal** Clover: **Lion**



Question: What does a person wash with a brush?
Baseline: **Face** Clover: **Toy**

Figure 4. Qualitative results of video question answering results on MSRVTT-QA [52].

ity. For example, for the second video in Fig. 3, Clover returns the text with “paper airplane”, which is a more accurate description of the video content; for the first video in Fig. 4, Clover generates a more accurate answer “Lion” than the baseline answer “Animal”. The results demonstrate the superiority of Clover.

5. Conclusion

In this paper, we present Clover, a new end-to-end Video-Language pre-training method for both high-efficiency video-text retrieval and video question answering. Clover introduces a novel Tri-modal Alignment task to better align the representations from visual, text and fused modalities, which explicitly correlate the uni-modal encoder and multi-modal encoder. It also introduces semantic masking strategy and pair-wise ranking loss to further improve the cross-modality modal training. Extensive experiments conducted on the three retrieval datasets and eight video QA datasets clearly demonstrated, as a general video-text model, its consistent superiority for video-text understanding.

Acknowledgements. This work was supported by National Key R&D Program of China (No.2022ZD0118201), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), and the Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001).

References

- [1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 1, 5, 6
- [2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 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*, pages 1728–1738, 2021. 1, 2, 5, 6, 7
- [3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Bert: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2021. 2, 5
- [4] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 1
- [5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 6
- [6] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021. 3
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. 2, 3, 6
- [8] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019. 2
- [9] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 1, 2, 3, 6, 7
- [10] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229. Springer, 2020. 1, 2
- [11] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16167–16176, 2022. 1, 2, 5, 6
- [12] Yuying Ge, Yixiao Ge, Xihui Liu, Alex Jinpeng Wang, Jianping Wu, Ying Shan, Xiaohu Qie, and Ping Luo. Miles: Visual bert pre-training with injected language semantics for video-text retrieval. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 6
- [13] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. *Advances in neural information processing systems*, 33:22605–22618, 2020. 1, 2
- [14] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5006–5015, 2022. 3
- [15] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11021–11028, 2020. 2
- [16] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 1, 2, 3, 5
- [17] Jie Jiang, Shaobo Min, Weijie Kong, Hongfa Wang, Zhifeng Li, and Wei Liu. Tencent text-video retrieval: Hierarchical cross-modal interactions with multi-level representations. *IEEE Access*, 2022. 3
- [18] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6
- [19] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. Progressive attention memory network for movie story question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8337–8346, 2019. 2
- [20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 6
- [21] Alexander Kunitsyn, Maksim Kalashnikov, Maksim Dzabraev, and Andrei Ivaniuta. Mdmmt-2: Multidomain multimodal transformer for video retrieval, one more step towards generalization. *arXiv preprint arXiv:2203.07086*, 2022. 3
- [22] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981, 2020. 2
- [23] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 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*, pages 7331–7341, 2021. 1, 2, 5, 6, 7

- [24] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963, 2022. 1, 2, 5, 6, 7
- [25] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, 2020. 2
- [26] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8658–8665, 2019. 2
- [27] Chengzhi Lin, Ancong Wu, Junwei Liang, Jun Zhang, Wenhong Ge, Wei-Shi Zheng, and Chunhua Shen. Text-adaptive multiple visual prototype matching for video-text retrieval. *Advances in Neural Information Processing Systems*, 2022. 6
- [28] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5
- [29] Fei Liu, Jing Liu, Weining Wang, and Hanqing Lu. Hair: Hierarchical visual-semantic relational reasoning for video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1698–1707, 2021. 2
- [30] Ruyang Liu, Jingjia Huang, Ge Li, Jiashi Feng, Xinglong Wu, and Thomas H Li. Revisiting temporal modeling for clip-based image-to-video knowledge transferring. *arXiv preprint arXiv:2301.11116*, 2023. 3
- [31] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, 2019. 1, 2, 5
- [32] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 3, 6
- [33] Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002. 5
- [34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 6
- [35] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 2
- [36] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 6
- [37] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022. 3
- [38] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6884–6893, 2017. 1, 5, 6
- [39] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 1, 2, 6
- [40] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 6
- [41] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bridge to answer: Structure-aware graph interaction network for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15526–15535, 2021. 2
- [42] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, et al. Avlnet: Learning audio-visual language representations from instructional videos. In *Annual Conference of the International Speech Communication Association*. International Speech Communication Association, 2021. 2
- [43] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16877–16887, 2021. 2
- [44] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernamed, 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)*, pages 2556–2565, 2018. 5, 6
- [45] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019. 2
- [46] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019. 2

- [47] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124*, 2016. **1, 5**
- [48] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv-1807, 2018. **2, 6**
- [49] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022. **1, 2, 6**
- [50] Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Object-aware video-language pre-training for retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3313–3322, 2022. **2, 6**
- [51] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. **3**
- [52] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. **1, 2, 5, 8**
- [53] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. Vlm: Task-agnostic video-language model pre-training for video understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4227–4239, 2021. **2**
- [54] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, 2021. **2**
- [55] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. **1, 5, 6, 8**
- [56] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5036–5045, June 2022. **1, 2, 6**
- [57] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022. **3**
- [58] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021. **1, 2, 6, 7**
- [59] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11562–11572, 2021. **2**
- [60] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018. **3, 5**
- [61] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34, 2021. **1, 6, 7**
- [62] Zheng-Jun Zha, Jiawei Liu, Tianhao Yang, and Yongdong Zhang. Spatiotemporal-textual co-attention network for video question answering. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(2s):1–18, 2019. **2**
- [63] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020. **2**