Contrastive Video-Language Learning with Fine-grained Frame Sampling

Zixu Wang[1], **Yujie Zhong**[2], **Yishu Miao**[3], **Lin Ma**[2], **Lucia Specia**[1]
[1]Language and Multimodal AI Lab (LAMA), Imperial College London
[2]Meituan Inc., [3]Haiper.ai
zixu.wang@imperial.ac.uk, jaszhong@hotmail.com, yishu.miao@haiper.ai
forest.linma@gmail.com, l.specia@imperial.ac.uk

Abstract

Despite recent progress in video and language representation learning, the weak or sparse correspondence between the two modalities remains a bottleneck in the area. Most video-language models are trained via pair-level loss to predict whether a pair of video and text is aligned. However, even in paired video-text segments, only a subset of the frames are semantically relevant to the corresponding text, with the remainder representing noise; where the ratio of noisy frames is higher for longer videos. We propose **FineCo** (**Fine**-grained **Co**ntrastive Loss for Frame Sampling), an approach to better learn video and language representations with a fine-grained contrastive objective operating on video frames. It helps distil a video by selecting the frames that are semantically equivalent to the text, improving cross-modal correspondence. Building on the well established VideoCLIP model as a starting point, FineCo achieves state-of-the-art performance on YouCookII, a text-video retrieval benchmark with long videos. FineCo also achieves competitive results on text-video retrieval (MSR-VTT), and video question answering datasets (MSR-VTT QA and MSR-VTT MC) with shorter videos.

1 Introduction

Human perception is multimodal, including visual, textual, and audial information. To achieve human-level perceptional ability, intelligent systems need to understand and interpret these multimodal signals and summarise the relevant information in them. Learning from video and language data has received significant attention in recent multimodal machine learning work for downstream tasks that require joint understanding of video and textual information, including text-video retrieval (Lin et al., 2014; Liu et al., 2019; Miech et al., 2018; Wang et al., 2016; Bain et al., 2021), video question answering (Fan et al., 2019; Yang et al., 2021; Huang et al., 2020; Jiang et al., 2020; Le et al., 2020; Lei et al., 2021), and video captioning (Ging et al., 2020; Luo et al., 2020; Zhang et al., 2020b). In most of this work, contrastive learning (Gutmann and Hyvärinen, 2010) is used as training objective.

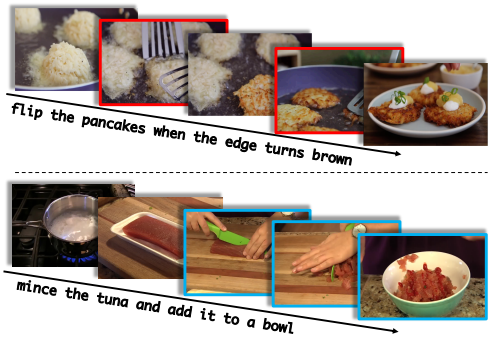

Figure 1: Illustration of the weak correspondence problem in video-language learning. Given a pair of video and its text (*e.g.* caption, instruction, or transcription), only a subset of the frames (here indicated by coloured bounding boxes) is semantically aligned to the textual content. The remaining frames represent irrelevant visual information and will not contribute to language grounding on videos.

The aim of a cross-modal contrastive loss is to maximise the similarity between an aligned video-text pair while minimising the similarity for all other pairs. One issue with standard cross-modal contrastive loss is that it focuses on pair-level alignment but ignores the negative effects of irrelevant frames that are present in a single video clip, even in a pair of aligned video and text. We define irrelevant frames as those with no or little shared semantics with the text. These irrelevant frames may negatively affect the contribution of frames that are semantically similar to the text, which further results in less informative video representation. Therefore, we posit that frame-level learning is a better strategy for video-language tasks.

In this paper, we propose FineCo, an approach that has a frame selector to sample relevant frames in a video and is trained with a fine-grained con-

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 694–705
November 20–23, 2022.

trastive loss on frame-text pairs, in order to mitigate the problem of weak correspondence in video-language representation learning. Existing video-language learning approaches (Miech et al., 2020; Xu et al., 2021) only optimise pair-level alignment but do not explicitly learn which part of a video contributes to its alignment with the text. FineCo focuses on aligning relevant frames with the text. It is inspired by the text-based temporal localisation task (Zhang et al., 2020a), however, the motivation of FineCo is different: to learn better video-level representation by adding a frame-level contrastive learning signal to the pair-level objective, with no need for temporal annotation within a video-text pair.

We hypothesise that FineCo is particularly beneficial for long videos, where each video provides more information and only a small proportion of frames will be relevant to its text counterpart, as shown in Figure 1. FineCo is able to model frame-text similarity through fine-grained contrastive learning, where the most informative frames are paired with the text as positive pairs and the remaining frames, as negatives. It then explicitly contrasts the selected informative frames against the noisy frames, without the need for frame-text annotations. This frame-level distillation provides a strong learning signal, which encourages the alignment of semantically equivalent video-text pairs. The fine-grained contrastive loss abstracts the learning signal from pair-level annotations and is trained in an end-to-end manner. This combination of pair-level learning signal and frame-level contrastive loss is novel and effective, and boosts the performance on two important video-language benchmark tasks, especially in text-video retrieval with longer videos. We devised FineCo by building on the recently proposed and well performing VideoCLIP (Xu et al., 2021), in which a video clip is represented as sequence of frame features.

Our contributions are summarised as follows: (1) We propose FineCo, an approach trained with fine-grained contrastive loss to mitigate the weak correspondence problem in video-text pairs; (2) We use FineCo to distil a video clip by sampling frames that are relevant to its text counterpart according to frame-text similarities; (3) On text-video retrieval and video question answering benchmarks, we show that FineCo achieves state-of-the-art performance on YouCookII and MSR-VTT MC (multiple choice).

2 Related Work

Contrastive Learning The use of contrastive loss (Gutmann and Hyvärinen, 2010) has become the dominant paradigm for learning video-language representations. The aim is to maximise the similarity of video-text pairs that are aligned to each other (positive pairs) while pushing away irrelevant (negative) pairs. However, the semantic alignment between most video-text pairs is weak, which makes it difficult to ground textual information on the videos. In order to mitigate the pair-level weak alignment issue, MIL-NCE (Miech et al., 2020) leverages multiple surrounding captions as the positive pairs and makes use of multiple instance learning (MIL) (Dietterich et al., 1997) with contrastive loss to mitigate noise in cross-modal correspondences. The main idea is to consider multiple contextual sentences for matching a video, instead of only comparing a video against a single sentence. To alleviate the issue that semantically equivalent videos and texts from different pairs may be taken as dissimilar in contrastive learning, support-set (Patrick et al., 2021) introduces a generative approach for captioning over a set of visual candidates that ensures that video-language representation does not over specialise to individual samples. MIL-NCE and support-set focus on pair-level contrastive signals to align relevant video-text pairs. However, even within a positive video-text pair, the video is likely to contain many irrelevant frames. Therefore, it can be beneficial to distil the video such that only the relevant frames, *i.e.* those which have similar content to the text, are selected for cross-modal learning.

Video-language Learning (Sun et al., 2019; Zhu and Yang, 2020; Gabeur et al., 2020; Li et al., 2020a; Miech et al., 2020; Ging et al., 2020; Luo et al., 2020) have shown promising results for video-language learning with pre-training followed by fine-tuning. This strategy has become very prominent since the release of BERT (Devlin et al., 2019) and many image-text pre-training frameworks (Tan and Bansal, 2019; Li et al., 2019, 2020b; Zhang et al., 2021; Chen et al., 2020; Zhang et al., 2019; Kim et al., 2021; Li et al., 2021, 2022). The release of datasets such as HowTo100M (Miech et al., 2019) and WebVid-2M (Bain et al., 2021) has enabled large-scale pre-training on unlabelled video-text pairs to improve representation

learning of video and language. Many approaches (Miech et al., 2020; Zhu and Yang, 2020; Patrick et al., 2021) use HowTo100M as their pre-training dataset. FiT (Bain et al., 2021) uses WebVid-2M and Google Conceptual Captions (CC3M) to take advantage of the large collection of video-text and image-text pairs for pre-training. However, large pre-training datasets rely on loosely aligned video-text pairs, without any fine-grained supervision on alignment. This makes it difficult to learn cross-modal cues present in the given video-text pairs. It is also computationally expensive to improve video-language representation learning, given that videos can contain a large number of frames, especially longer videos. ClipBERT (Lei et al., 2021) randomly samples a few frames from a video for video-language representation learning. Their motivation is to minimise memory and computation costs from processing the full sequence of frames. This sampling strategy is over simplistic and can thus be improved by better approaches to select frames based on their relevance to the paired text.

3 FineCo

3.1 Preliminaries

The most widely used objective function for video-language learning is contrastive loss, specifically the softmax version of noise-contrastive estimation (NCE) (Gutmann and Hyvärinen, 2010). It is formulated as

$$\sum_{i=1}^{n} \log \left(\frac{e^{f(x_i)^T g(y_i)}}{e^{f(x_i)^T g(y_i)} + \sum\limits_{(x',y') \in \mathcal{N}_i} e^{f(x_i')^T g(y_i')}} \right) \tag{1}$$

where x_i denotes a video clip and y_i represents the corresponding text (*e.g.* a caption, an instruction, or transcription); f and g are video encoder and text encoder respectively; $e^{f(x_i)^T g(y_i)}$ denotes the similarity of a positive video-text pair, calculated as the exponentiated dot product of the video representation $f(x_i)$ and text representation $g(y_i)$; $\mathcal{N}_i$ is a set of negative video-text pairs x_i' and y_i' that are not aligned.

This contrastive loss leverages pair-level similarity of video and text, but ignores the fact that weak video-language correspondence does not stem only from entirely negative pairs of video and text, but also from frame-level noise, which happens even when a video-text pair is aligned as a whole. Standard contrastive loss does not explicitly model frame-text relevance, *i.e.* it does not differentiate between frames that are semantically equivalent to the corresponding text and frames that are not. It can thus suffer by learning from noisy signals, particularly in long videos with various scenes.

3.2 Fine-grained Contrastive Learning

A video consists of a sequence of frames. For video-language learning, the video is paired with a text which describes/refers to some of the content of the video. For most tasks, only some of the visual information has an equivalent textual signal, *e.g.* a video description is only a summary of the visual information. To sample and optimise for the relevant visual information from a video, we propose a fine-grained contrastive loss to distil each video-text pair.

Formally, a video-text pair is denoted as (x, y), where x is a video clip consisting of a sequence of N video frames $\{x_1, x_2, \ldots, x_K\}$ where K is the number of frames in the video clip, and y is the paired text. We assume that a video x contains a set of C positive frames $\mathcal{P}(x)$ and a set of $(K - C)$ negative frames $\mathcal{N}(x)$, where positive frames contains relevant information to the text while negative frames are noisy/irrelevant ones. The aim is to maximise the joint probability of relevant frame-text pairs (x_k, y) by exponentiating the similarity of the two representations:

$$p(x_k, y) = h(f(x_k), g(y)) \propto e^{\text{sim}(f(x_k), g(y))} \tag{2}$$

3.2.1 Objective Function

Given n pairs of video representation $f(x)$ and text representation $g(y)$, the ith pair is denoted as $f(x_i) = \{f(x_{i_1}), f(x_{i_2}), \ldots, f(x_{i_K})\}$ and $g(y_i)$, our fine-grained contrastive loss $\mathcal{L}$ is defined as:

$$\begin{aligned} \mathcal{A}_i &= \sum_{x_{i_k} \in \mathcal{P}(x_i)} e^{\text{sim}(f(x_{i_k}), g(y_i))} \\ \mathcal{B}_i &= \sum_{x'_{i_k} \in \mathcal{N}(x_i)} e^{\text{sim}(f(x'_{i_k}), g(y_i))} \\ \mathcal{L} &= \sum_{i=1}^{n} \log \left(\frac{\mathcal{A}_i}{\mathcal{A}_i + \mathcal{B}_i} \right) \end{aligned} \tag{3}$$

where $\mathcal{P}(x_i)$ contains the positive frames in a video that have higher similarities to the text representation $g(y_i)$, and $\mathcal{N}(x_i)$ is the set of remaining frames in the same video, which refers to the negative frames. The similarity is calculated by our frame selector ($\mathcal{FS}$) (Section 3.2.2) with the frame

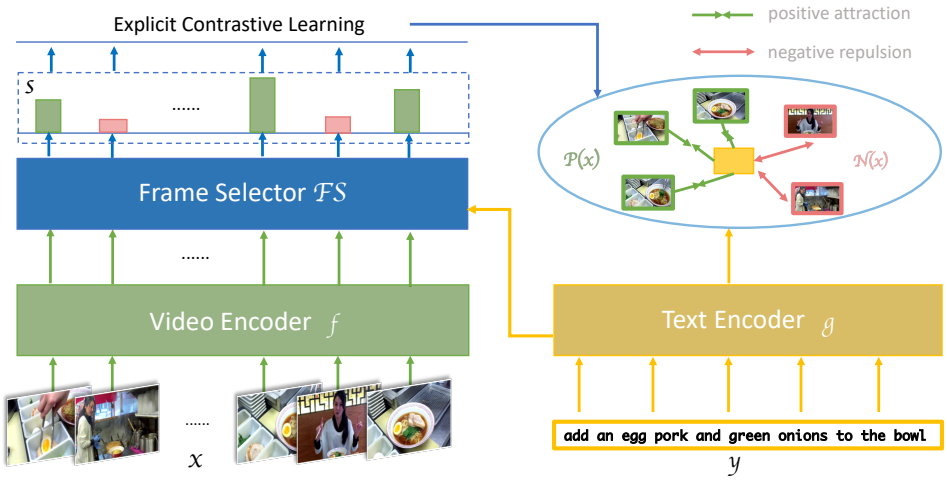

Figure 2: FineCo architecture. Given a sequence of frames in a video clip x, the video encoder f transforms them into a sequence of video features. The corresponding sentence y is fed into the text encoder g to get the text representation. The frame selector FS takes the text representation and the sequence of video features as inputs and outputs the similarities (probabilities of each frame being relevant). The top k frames are then used as the positive candidates and the remaining ones as negative, both of which are combined with the text representation to compute the fine-grained contrastive loss.

x_{i_k} and text representations y_i as inputs. $\mathcal{A}_i$ and $\mathcal{B}_i$ represent the sum of similarity scores for positive and negative frames, respectively. This objective function aims to maximise the similarity between the positive frames and the text, while increasing the dissimilarity between the negative frames and the text. Therefore, the sampled relevant frames can directly contribute to the cross-modal learning of video-text alignments.

3.2.2 Assignment of Positives and Negatives

Inspired by MIL-NCE (Miech et al., 2020), which makes use of multiple sentences for matching a video and its corresponding text, we extract multiple positive frames from the complete set according to the similarity score between each frame and the text. Consider an example (x, y) with K frames $\{x_1, x_2, \ldots, x_K\}$, we introduce a frame selector $\mathcal{FS}$, a cross-modal module which takes video and text representation as the input and outputs the similarity scores between each frame and the text, denoted as:

$$\text{sim}_k = \mathcal{FS}(f(x_k), g(y)); x_k \in \{x_1, x_2, \ldots, x_K\} \quad (4)$$

where $f(x_k)$ is the representation of the kth frame; $g(y)$ is the representation that encodes the meaning of the complete text sequence, which is used to find semantically similar frames in the corresponding video x; sim_k is the similarity score between the kth frame and the text y.

By ranking the similarity scores of K frames, we choose top C frames to form the positive set and the remaining $(K - C)$ as the negative set. This is an explicit sampling strategy which extracts the relevant frames in a video. There is no constraint on the architecture of $\mathcal{FS}$. In this work, we use a multi-layer perceptron (MLP) with a softmax layer to compute the similarity scores.

3.3 Model Architecture

As our methodology focuses on fine-grained contrastive learning signal for a single pair of video and its text, it makes no assumptions on the encoder architectures and can work with pre-training frameworks with different video and text backbones. In our experiments, we use Transformer (Vaswani et al., 2017) as both the video encoder and the text encoder, as we detail below.

3.3.1 Text Encoder

We use BERT (Devlin et al., 2019) as the text encoder g to get text representation $g(y)$. The text encoder is trained together with the video encoder to learn better text representations. Following VideoCLIP (Xu et al., 2021), we use average pooling (instead of using the `[CLS]` token) as the final text encoding. The text representation is used as the guiding element and anchor to calculate the frame-text similarity scores and to sample the most semantically similar frames in a video clip.

3.3.2 Video Encoder

Our video encoder f is composed of an S3D (Xie et al., 2018; Miech et al., 2020) and a Transformer (Vaswani et al., 2017), following VideoCLIP (Xu et al., 2021). To speed up training, we use a S3D pre-trained on HowTo100M (Miech et al., 2019) to extract pre-trained video features, where the video feature of a video clip is represented by a sequence of video frames. The output from the S3D is formulated as $x = [x_1, x_2, \ldots, x_K]$, where x is the representation of a sequence of video frames. We extract the frames at a rate of one frame per second, so the number of video frames equals the number of seconds. x is concatenated with learnable tokens `[CLS]` and `[SEP]` at the beginning and the end of the sequence, respectively. We then train the Transformer using the pre-extracted video representation as the input, to obtain the last hidden states as the representation of the sequence of video frames.

3.4 Training

Training with the pair-level contrastive loss is challenging due to the intractability of computing the normalisation constant over all possible pairs of videos and texts. It is however more feasible in our fine-grained contrastive loss as the number of possible frames in a single video clip is limited. The normalising constant is computationally tractable and can be directly computed by summing over exponentiated similarity scores across all the frame-text pairs. The overall training objective ($\mathcal{L}$) is defined by combining our fine-grained contrastive loss ($\mathcal{L}_1$) and task-specific losses ($\mathcal{L}_2$), denoted by $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$; where in text-video retrieval, the task loss $\mathcal{L}_2$ is pair-level contrastive loss and in video question answering, it is cross-entropy.

3.5 Inference

For text-video retrieval, there is no cross-modal fusion module at inference time. It requires only video and text representations which are first projected to a common dimension via linear layers. The similarity between a video-text pair is calculated by performing the exponentiated dot product between the two projected embeddings. This ensures retrieval inference is of trivial cost, since it is indexable and scalable to large-scale retrieval at inference time. For video question answering, we follow the pipeline in Figure 2, where we concatenate the video and text representations, and feed it into an MLP module to obtain the final representation for answer prediction.

4 Experiments

In this section, we describe the tasks and datasets used in our experiments with FineCo.

4.1 Datasets and Metrics

FineCo is mainly beneficial for long videos, therefore we focus our evaluation on YouCookII (Zhou et al., 2018) - a text-video retrieval dataset with long videos. **YouCookII** consists of 2K cooking videos with 14K video clips. The videos are of a total duration of 176 hours with average **5.26 minutes** per video. Each video clip is annotated with one sentence on a cooking instruction. It is collected from YouTube and contains 89 types of recipes. We split the dataset according to Miech et al. (2020) where 9.6k video-text pairs are used for training and 3.3k pairs for validation.

We further evaluate FineCo on other benchmark datasets for text-video retrieval and video question answering with shorter videos. **MSR-VTT** (Xu et al., 2016) is another popular benchmark dataset for text-video retrieval. It contains 10K YouTube videos (an average **20 seconds** per video) with 200K captions. We report the results on the **1k** test split and use the remaining 9k videos for training. **MSVD** (Chen and Dolan, 2011) consists of 80K captions for 1,970 videos from YouTube, with each video containing 40 sentences. We use the standard split of 1200, 100, and 670 videos for training, validation, and testing as in (Liu et al., 2019; Patrick et al., 2021). **DiDeMo** (Hendricks et al., 2018) contains 10K Flickr videos with 40K sentences. Following (Liu et al., 2019; Lei et al., 2021), we evaluate paragraph-to-video retrieval, where all sentence descriptions from a video are concatenated into a single query. **MSR-VTT QA** contains 10K videos and 243K open-ended questions, which is created using the videos and captions from original MSR-VTT. We use 1500 most frequent answers as the answer vocabulary, which covers over 93% samples. **MSR-VTT MC** (multiple choice) is also created from original MSR-VTT. Multiple choice QA is formulated as a video-text retrieval task where the videos are the questions and captions are the answers.

Evaluation Metrics Following the standard evaluation protocols as described in most video-language work (Miech et al., 2019; Zhang et al., 2018; Mithun et al., 2018; Miech et al., 2018, 2020),

we report the text-video retrieval performance using recall-based metrics: Recall at rank K (R@K) which measures the rate at which the correct video is retrieved amongst the top ranked results, and Median Rank (MdR) which calculates the median of a list of indices representing the rank of the ground truth video; where the higher R@K and lower median rank indicate better performance. For MSR-VTT QA and MSR-VTT MC, accuracy is reported, as in Xu et al. (2021).

4.2 Training Details

To minimise computation costs, we use S3D (Xie et al., 2018) for video feature extraction, which is pre-trained on HowTo100M (Miech et al., 2019) following MIL-NCE (Miech et al., 2020). The feature dimensionality is 512 (*e.g.* given a 10-second video, the shape of the video feature extracted is $[10, 512]$). We apply video feature pre-extraction to all the downstream datasets in our experiments. We follow the pre-training steps as in VideoCLIP (Xu et al., 2021) where pre-training is done using HowTo100M, which contains uncurated instructional videos. A total of 1.1M videos are used for pre-training after cleaning and filtering.

For the video Transformer encoder, we use 6 attention blocks, while for the text Transformer encoder, we use 12 blocks. The weights for both encoders are initialised with *bert-base-uncased*. The maximum length of a video is 32; for text inputs it is 64. Before feeding video and text inputs into their respective encoders, [CLS] and [SEP] tokens are concatenated to the beginning and end of each modality. All the models are trained on one NVIDIA Tesla V100 GPU with 32 GB of RAM memory for 15 epochs, with fp16 precision for 2-3 hours. We select the final checkpoint according to the loss on the validation set. Optimisation is performed using Adam (Kingma and Ba, 2015) with a learning rate of 5e-5. The model takes 1000 steps for warm-up, and we use a learning rate schedule with polynomial decay.

5 Results

In this section, we describe the experimental results and compare FineCo with state-of-the-art approaches (Section 5.1). We further explore different sampling strategies to select positive frames (Section 5.2), and fine-grained word sampling (Section 5.3). We also provide examples of the frames selected by FineCo (Section 5.4).

YouCookII	R@1	R@5	R@10	MedR
HowTo100M (Miech et al., 2019)	8.2	24.5	35.3	24.0
MIL-NCE (Miech et al., 2020)	15.1	38.0	51.2	10.0
COOT (Ging et al., 2020)	16.7	40.2	52.3	9.0
UniVL (Luo et al., 2020)	28.9	57.6	70.0	4.0
VideoCLIP (Xu et al., 2021)	32.2	62.6	75.0	**3.0**
Ours w/o DS	35.7	65.9	77.5	**3.0**
Ours w DS	**37.6**	**66.6**	**78.2**	**3.0**

Table 1: YouCookII Retrieval Results. DS denotes Dual Softmax.

5.1 Comparison to State-of-the-art

Overall, as we detail below, FineCo outperforms its base model VideoCLIP across all benchmark datasets. Additionally, it achieves state-of-the-art performance on YouCookII and MSR-VTT MC.

5.1.1 Text-video Retrieval

We start by evaluating on YoucookII, which contains longer videos than other text-video benchmarks, and is therefore more challenging for video-language representation learning. As shown in Table 1, FineCo outperforms all previous approaches by a large margin. We report results w/ and w/o Dual Softmax (DS) following Cheng et al. (2021) and Gao et al. (2021). In Dual Softmax, given a similarity matrix in text-video retrieval, a prior probability is calculated in the cross direction, which is then multiplied with the original similarity matrix as an efficient regulariser. FineCo surpasses previous state-of-the-art with fine-grained contrastive loss (3.5% gains for R@1). Dual Softmax further improves the results (1.6% for R@1) and achieves an even higher state-of-the-art (37.3% R@1).

We provide additional results on text-video retrieval across MSR-VTT [1] (Table 2), MSVD (Table 3), and DiDeMo (Table 4). Our reported scores of VideoCLIP on MSVD and DiDeMo are from our implementation as their paper does not test on the datasets. As FineCo builds on VideoCLIP (Xu et al., 2021), our results are directly comparable with the scores reported in VideoCLIP. [2] From

[1] We omit the results of text-video retrieval on MSR-VTT from CLIP (Radford et al., 2021) models (Cheng et al., 2021; Luo et al., 2021; Fang et al., 2021; Gao et al., 2021) as it would not be a fair comparison since CLIP-based models benefit mainly from large-scale image-text pre-training, which we do not use.

[2] We also implemented FineCo in FiT (Bain et al., 2021), however the improvements are not obvious as in VideoCLIP.

MSR-VTT 1k	R@1	R@5	R@10	MedR
JSFusion (Yu et al., 2018)	10.2	31.2	43.2	13.0
HowTo100M (Miech et al., 2019)	14.9	40.2	52.8	9.0
ClipBERT (Lei et al., 2021)	22.0	46.8	59.9	6.0
Support-set (Patrick et al., 2021)	30.1	58.5	69.3	3.0
FiT (Bain et al., 2021)	32.5	61.5	71.2	3.0
VideoCLIP (Xu et al., 2021)	30.9	55.4	66.8	4.0
Ours	**32.6**	**62.1**	**71.4**	**3.0**

Table 2: MSR-VTT Results - 1k

MSVD	R@1	R@5	R@10	MedR
VSE (Kiros et al., 2014)	12.3	30.1	42.3	14.0
VSE ++ (Faghri et al., 2018)	15.4	39.6	53.0	9.0
CE (Liu et al., 2019)	19.8	49.0	63.8	6.0
Support-set (Patrick et al., 2021)	28.4	60.0	72.9	4.0
FiT (Bain et al., 2021)	**33.7**	**64.7**	**76.3**	**3.0**
VideoCLIP (Xu et al., 2021)	26.4	52.2	63.3	5.0
Ours	27.2	54.0	64.0	5.0

Table 3: MSVD Results

DiDeMo	R@1	R@5	R@10	MedR
S2VT (Venugopalan et al., 2015)	11.9	33.6	-	13.0
FSE (Zhang et al., 2018)	13.9	36.0	-	11.0
CE (Liu et al., 2019)	16.1	41.1	-	8.3
ClipBERT (Lei et al., 2021)	20.4	44.5	56.7	7.0
FiT (Bain et al., 2021)	**31.0**	**59.8**	**72.4**	**3.0**
VideoCLIP (Xu et al., 2021)	16.6	46.9	-	-
Ours	19.5	48.8	55.9	7.0

Table 4: DiDeMo Results

MSR-VTT QA	Accuracy
AMU (Xu et al., 2017)	32.5
HME (Fan et al., 2019)	33.0
HCRN (Le et al., 2020)	35.6
ClipBERT (Lei et al., 2021)	**37.4**
VideoCLIP (Xu et al., 2021)	35.9
Ours	**37.4**

Table 5: MSR-VTT QA Results

the additional results, it can be seen that FineCo outperforms VideoCLIP on all text-video retrieval datasets by a large margin. This shows that FineCo is generalisable to various types of text-video retrieval data. The smaller improvements (*e.g.*, 30.9% $\rightarrow$ 32.6% R@1 on MSR-VTT 1k in Table 2) compared to those on YouCookII (32.2% $\rightarrow$ 37.6% R@1) might be due to the less varied scenes in shorter videos of MSR-VTT, which makes it challenging to distinguish among intra-video frames in a short video.

Note that video-text pairs in these downstream datasets are constructed to be aligned in order to provide strong supervision learning signals to video-language representation learning. FineCo distils aligned video-text pairs and achieves noticeable improvements over approaches without any frame sampling, which corroborates our hypothesis that there are irrelevant or less useful frames in a video even if it is annotated as aligned to its text counterpart.

The reason might be the difference of video encoding in VideoCLIP and FiT. FineCo contributes more to complete frame features where a video is encoded into a long sequence of video features with more temporally contextual information, rather than only a few visual frames in ViT (Dosovitskiy et al., 2021) and Timesformer (Bertasius et al., 2021).

5.1.2 Video Question Answering

Tables 5 and 8 show the results on video question answering (VideoQA) for MSR-VTT QA and MSR-VTT MC, respectively. For both datasets, FineCo improves over VideoCLIP. For MSR-VTT MC, it achieves a new state-of-the-art (92.7% accuracy). This further shows the generalisation ability of FineCo across different video-language tasks and datasets.

For MST-VTT QA, the score reported for VideoCLIP is from our implementation as their paper does not test on this dataset. For MSR-VTT MC, the score reported is from the original paper. For VideoQA, we note that ClipBERT also achieves good results, which might be because it employs a multimodal Transformer encoder after two separate encoders for the video and the question to learn better cross-modal relationships. The improvement is particularly noticeable on MSR-VTT MC, which quantitatively suggests that FineCo can distil question-relevant frames to improve answer accuracy. We speculate that this is because a question only needs partial information in some frames of a video clip to be answered, which is addressed by FineCo.

5.2 Decision on Number of Frames

Given a pair of video clip and text, we choose the positive frames according to the similarities be-

Table 6: Comparison of different sampling strategies for positive frames.

Strategy	fixed-k ($k = 1, 10, 30, 50, 100, 256$)						median	ratio (30%, 50%, 80%)		random
R@1	26.90	30.44	37.17	**37.32**	37.04	34.80	**37.62**	**37.29**	36.99	36.85	30.08

fixed-k	1	5	10	15	20	25	32
MSR-VTT QA	35.5	36.3	36.2	36.8	**37.4**	37.2	35.9
MSR-VTT MC	90.3	92.3	92.6	92.4	92.6	**92.7**	92.1

Table 7: Effect of different number of positive frames on MSR-VTT QA and MSR-VTT MC. When $k = 32$, FineCo equals VideoCLIP.

MSR-VTT MC	Accuracy
MLB (Kim et al., 2016)	76.1
JSFusion (Yu et al., 2018)	83.4
ActBERT (Zhu and Yang, 2020)	85.7
ClipBERT (Lei et al., 2021)	88.2
VideoCLIP (Xu et al., 2021)	92.1
Ours	**92.7**

Table 8: MSR-VTT MC Results

tween each frame and the text. The number of positive frames k is the key factor, deciding the set of frames to be treated as positive, and hence the extent of the contribution of the fine-grained contrastive learning signal. We propose four strategies to choose positive frames in a video clip.

Fixed-k: We select a fixed number of positive frames which have the highest similarities to the text. We experiment with $k = [1, 10, 30, 50, 100, 256]$ as the number of positive frames, with 256 as the maximum number of frames (one frame per second).[3] **Median:** We use the averaged similarity medians in a mini-batch as the thresholds for each video: in a sequence of video frames, the ones with higher similarities than the median are used as the positive frames. The number of positive frames will vary across different mini-batches, depending on the distribution of similarities. **Ratio:** We apply 30%, 50%, and 80% of the original video length (without padding or trimming) as the positive frames. Note that different video clips have different lengths, so the number of sampled frames will differ from video to video.

[3]We set the maximum length of a video sequence to 256 frames for YouCookII, but 32 frames for other datasets with much shorter videos.

Random: We randomly sample $k = 50$ frames in a video clip as the positives.

We show the performance of the four strategies on YouCookII in Table 6. **Median** has the best performance (37.62), which is followed by **fixed-k** with $k = 50$ ($\approx 20\%$ of the data) (37.32), and similarly to **ratio** with 30% (37.29). This indicates that on average only $\approx 20\% - 30\%$ frames in the long videos from YouCookII are informative for the retrieval task. **Fixed-k** with $k = 1$ has the lowest score, which makes sense given that the entire videos are summarised by the one most similar frame to be used as the positive candidate. This mistakenly treats many other possibly relevant frames as negative frames, hence degrading the performance significantly. The best number 50 indicates that for most video-text pairs in YouCookII, 50 frames (=50 seconds as we extract video features at a rate of one feature per second, so the length of the extracted video features is the same as the number of seconds) ($\approx 20\%$) are the most relevant and sufficient. For **random**, we choose $k = 50$ as this was the best number according to the fixed-k analysis. The comparison between **random** and **fixed-k** clearly shows that sampling positive pairs based on their similarity to the text is an effective strategy to improve performance on the downstream task: on the same number of positive frames, **fixed-k** improves over **random** by 7.24%.

We also compare the performance of **fixed-k** on MSR-VTT QA and MSR-VTT MC. In Table 7, we show that FineCo has the best performance on MSR-VTT QA with $k = 20$ and on MSR-VTT MC with $k = 25$, where both have a sequence with maximum number of 32 frames. The ratio of positive frames ($\approx 70\% - 80\%$) is higher than in YouCookII. This corroborates our hypothesis that fine-grained sampling is more applicable to longer videos, which tend to contain more varied scenes and where there is more scope to filter out noisy or irrelevant frames. Therefore, in video-language datasets with shorter videos, a higher proportion of frames is needed as positive frames for effective contrastive learning. As the number of informative frames k in a video clip varies across different types

of videos, we recommend that this is treated as hyperparameter that is tuned for each new dataset, following our **fixed-k** strategy to select the number k on a development set.

5.3 Fine-grained Word Sampling

Given the improvements of FineCo with fine-grained frame sampling, we were curious about potential improvements if applying the same strategy to the text instead of the video, *i.e.* sampling most relevant words. Therefore, we experiment with this idea over a sequence of words to sample the most informative words as those with the highest similarity to the entire video clip in YouCookII. The text-video retrieval results in this setup are {R@1-32.1, R@5-62.6, R@10-75.5}. These figures are similar to those obtained by VideoCLIP {R@1-32.2, R@5-62.6, R@10-75.0}, but substantially lower than our results from FineCo in Table 1. The reason is intuitive: by removing certain words, the meaning of the sentence or paragraph can be substantially compromised, and having an understanding of the meaning of the complete text is important for video-language tasks. Video frames, on the other hand, can be more redundant or contribute less to the complete video understanding, and therefore fine-grained sampling from frames proves more effective.

5.4 Qualitative Examples

To further elaborate the contribution of FineCo and understand the effect of fine-grained contrastive loss, we show two examples where FineCo improves over VideoCLIP in Figure 3.[4] As we can observe from the examples, some of the information in each video clip can be considered irrelevant, given the meaning of the text. For example, in the first case, the long video (82 seconds) describes the cooking instruction *"brush the circles with egg washa and sprinkle with sesame seeds"* but there are only two frames delivering this meaning. This is a common feature in the YouCookII dataset, hence the positive results from sampling subsets of frames. In the third example we show a failure case where FineCo does not distinguish between similar videos hence a similar but incorrect video retrieved. We also observed failure cases where the video is either relatively short or less dynamic. FineCo might not effectively distil these

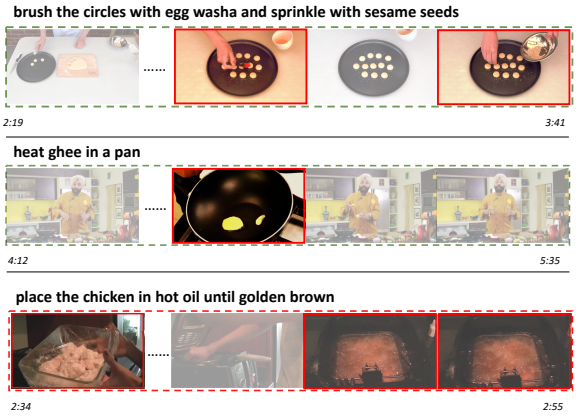

Figure 3: Qualitative examples. FineCo makes correct retrieval predictions on the frist two examples from YouCookII dataset. We calculate the frame-text similarities and highlight the frames with the highest scores.

types of videos to find the most informative frames. The issues could be potentially mitigated by incorporating FineCo into large-scale video-language pre-training to learn from more dynamic videos of various lengths.

6 Conclusions

We propose FineCo, an approach with a fine-grained contrastive loss to mitigate the weak correspondence problem in video-language representation learning. Experiments conducted on text-video retrieval and video question answering datasets suggest that FineCo can distil video frames that are relevant to its corresponding text and contribute to significant gains in performance, especially on the text-video retrieval dataset YouCookII with long videos. FineCo achieves state-of-the-art on YouCookII and MSR-VTT MC, and for text-video retrieval datasets with shorter videos, it substantially improves over the base model. Ablation studies analyse the key factors in FineCo including number of positive frames and word sampling. Our strategy for frame selection is simple and can generalise to different video-language frameworks, as long as they are based on contrastive learning, which is standard in this area. In addition, we posit that FineCo can be useful for video-language *pre-training* on large loosely or misaligned video-text datasets. We hope that our work will draw attention to the need for frame-level alignment to improve video-language representation learning.

[4] We only show a subset of informative and irrelevant frames for each example due to space limitations.

References

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. 2021. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas G. Dietterich, Richard H. Lathrop, and TomÃ¡s Lozano-PÃ©rez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31–71.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives.

Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In *CVPR*.

Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. 2021. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal Transformer for Video Retrieval. In *European Conference on Computer Vision (ECCV)*.

Zijian Gao, Jingyu Liu, Sheng Chen, Dedan Chang, Hao Zhang, and Jinwei Yuan. 2021. Clip2tv: An empirical study on transformer-based methods for video-text retrieval. *arXiv preprint arXiv:2111.05610*.

Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. 2020. Coot: Cooperative hierarchical transformer for video-text representation learning. In *Advances on Neural Information Processing Systems (NeurIPS)*.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy. PMLR.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2018. Localizing moments in video with temporal language. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. 2020. Location-aware graph convolutional networks for video question answering. In *AAAI*.

Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. 2020. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11101–11108.

Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *ArXiv*, abs/1411.2539.

Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020. Hierarchical conditional relation networks for video question answering. *arXiv preprint arXiv:2002.10698*.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learningvia sparse sampling. In *CVPR*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020a. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. In *Arxiv*.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman. 2019. Use what you have: Video retrieval using representations from collaborative experts. In *arXiv preprint arxiv:1907.13487*.

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

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. CLIP4Clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*.

Antoine Miech, Ivan Laptev, and Josef Sivic. 2018. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv*, abs/1804.02516.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K. Roy-Chowdhury. 2018. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ICMR '18, page 19–27, New York, NY, USA. Association for Computing Machinery.

Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F. Henriques, and Andrea Vedaldi. 2021. Support-set bottlenecks for video-text representation learning. In *International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

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

Hao Tan and Mohit Bansal. 2019. 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, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1494–1504, Denver, Colorado. Association for Computational Linguistics.

Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR).

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1686–1697.

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

Bowen Zhang, Hexiang Hu, and Fei Sha. 2018. Cross-modal and hierarchical modeling of video and text. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, pages 385–401.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Zhu Zhang, Zhou Zhao, Zhijie Lin, jieming zhu, and Xiuqiang He. 2020a. Counterfactual contrastive learning for weakly-supervised vision-language grounding. In *Advances in Neural Information Processing Systems*, volume 33, pages 18123–18134. Curran Associates, Inc.

Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. 2020b. Object relational graph with teacher-recommended learning for video captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*.

Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.