Cali-NCE: Boosting Cross-modal Video Representation Learning with Calibrated Alignment

Nanxuan Zhao
Department of Computer Science
University of Bath
nanxuanzhao@gmail.com

Jianbo Jiao
School of Computer Science
University of Birmingham
j.jiao@bham.ac.uk

Weidi Xie
Cooperative Medianet Innovation Center
Shanghai Jiaotong University
weidi@sjtu.edu.cn

Dahua Lin
Department of Information Engineering
Chinese University of Hong Kong
dhlin@ie.cuhk.edu.hk

Abstract

With the large-scale video-text datasets being collected, learning general visual-textual representation has gained increasing attention. While recent methods are designed with the assumption that the alt-text description naturally conveys the meaning and context of the video in semantics (i.e. well aligned with each other), it is unlikely to be satisfied for the Internet data, which potentially harms the quality of the learned visual-textual representation. To address this challenge, we first revisit three mainstream approaches: correspondence modeling, contrastive learning and predictive coding, demonstrating that a simple co-training strategy with these methods leads to a clear improvement in performance. To further explore the complementary nature of different training strategies, we propose a simple yet effective joint training framework that factorizes the total objective into conditional ones, termed as Cali-NCE[1]*. Our method first estimates confidence scores for measuring the correspondence between video and text descriptions, and the scores are later used to calibrate the sample weightings during contrastive training. Through extensive experiments, we show that the proposed approach achieves state-of-the-art performance on multiple downstream tasks: text-to-video retrieval, video action recognition, and video retrieval.*

1. Introduction

The highly developed intelligence of humans cannot be simply separated from vision and language [14, 15, 29]. They play key roles in our daily communication and how people understand the dynamic visual world. Modeling these means of communication becomes an important way also in the development of machine intelligence. A series of works have explored the modeling on such cross-modal tasks including text-to-video retrieval [41, 67], video captioning [31,63,74], and video question answering [3,39]. In order to learn a good representation, previous works mostly rely on a set of well-annotated pairs of video clips and text fragments for individual tasks. The expensive and tedious annotation process limits the progress of general visual-textual representations.

Taking advantage of the rich video resources from online repositories (*e.g.* YouTube and Shutterstock), recent works [9, 45] contribute datasets on large scales, showing the potential of general visual-textual representation power on several downstream tasks. For instance, the HowTo100M dataset [45] consists of over 100 million video clips and associated narration pairs; the WebVid-2M [9] dataset contains more than two million video alt-text pairs. Driven by the great success of instance discrimination on general visual representation learning [13, 28, 70], the leading methods follow a common approach to align the video and text into a shared embedding space by pulling positive visual-textual feature pairs close to each other [44, 45], while the negatives apart.

However, these methods rely on the assumption that all video and alt-text pairs are well aligned in semantics, which cannot always be true for web data. Many text descriptions are vague and inaccurate because of the uncurated collection process and underlying ambiguities. For example, as shown in Fig. 1, the scene in (a) can also appear in other cities than London, such as Hong Kong; the scene in (b) is hard to tell as a viaduct. Instead, the samples in (c) and (d)

[1] https://github.com/nanxuanzhao/Cali-NCE

are more specific and semantically aligned, which should be taken as more reliable supervision signals. This motivates us to think about whether there exists a better way to model the correlation between noisy web visual-textual pairs for representation learning.

To address this problem, we first revisit three mainstream methods for cross-modal representation learning, including correspondence modeling [4, 5, 7], contrastive learning [1,44], and predictive coding [25,56]. We find that these methods serve for learning different aspects and are complemented with each other, as a simple co-training can increase the performance clearly. Based on the observations, we propose a new method named Cali-NCE by introducing calibrated alignments in the co-training framework of correspondence modeling and contrastive learning. Cali-NCE uses the predicted confidences from correspondence modeling for calibrating the supervision signal of each visual-textual pair during training.

To validate the effectiveness of the proposed method, we conduct extensive experiments on multiple challenging downstream tasks: zero-shot text-to-video retrieval, video action recognition, and video retrieval. The proposed method achieves state-of-the-art performance on all these tasks. In summary, we make the following contributions:

- We take a step forward to investigate the problem of noisy semantic alignments for cross-modal visual-textual representation learning. We probe different approaches including correspondence modeling, contrastive learning and predictive coding, showing their potential for the target problem.
- We propose a new joint training framework with both correspondence prediction and contrastive learning. In particular, we introduce a new Calibrated Noise Contrastive Estimation (*Cali-NCE*) loss based on the predicted correspondence confidence score.
- We evaluate the quality of the learned representations and design choices through extensive experiments across different downstream tasks over several datasets with noticeable improvement, and achieve state-of-the-art performance.

2. Related Work

Learning from visual and textual cues has gained a significant amount of attention and attracted many works on related applications [3, 31, 41, 67, 74]. In this section, we focus on the major related works and the most relevant papers.

Video self-supervised learning. Self-supervised learning aims to learn semantically meaningful representations by generating supervision signals from the videos themselves. It has gained more and more attention as a good representation can benefit many downstream tasks. At the earlier stage, a set of pretext tasks are designed artificially based on temporal or spatial information, such as identifying the odd video sequence [18], sorting sequences [37], predicting the arrow of time [68], clip order prediction [72], and rotation modeling [30]. With the impressive results generated by contrastive loss on ImageNet pretrained dataset [8, 28, 70], the contrastive learning also adapts to video unsupervised learning [54]. This kind of method pulls two views of a sampled instance together, and pushes away from the other instances.

Multimodal representation learning. Video data accompany with multiple modalities, such as visual images, audio, text, and motion. Leveraging this multi-modality nature can enhance the video representation and has been explored by many recent works. [27] learn to associate images with spoken words, and [7] directly model video, audio, and text at the same, relying on a curated annotation dataset. Instead, [7] train a multimodal versatile model without any human label. Among different multimodal modelings, video and audio is one of the biggest branches. A set of works [4, 5, 35, 50] exploit the video-audio co-occurrence for learning a good video representation. [4] design an audio-visual correspondence (AVL) task by classifying whether a pair of video and audio clips is corresponded with each other. And they further extend this idea and design a model [5] that can work for both cross-modal retrieval and sound source localization. [2] introduce a cross-modal deep clustering method by supervising one modality by the cluster of the other modality.

Vision and language. Researches in this area adopt a common approach by embedding visual and textual representation into the same space [19, 69], and use the distance (*e.g.* a dot product) to measure the semantic similarity across modalities. Learning such a joint visual-textual space has shown its effectiveness among many works [34, 51, 66]. But most of these works rely on the medium-scale well-annotated datasets, which limits the scalability of the learned representation. With the recent release of large-scale datasets collected from the Internet without much human intervention [9, 45, 57], many recent works have appeared to study visual-textural representation learning [16, 21, 38, 53, 60, 78]. Different from previous works [26, 44] studying the temporal misalignment, our work tackles the misalignment in semantics.

3. Revisiting Cross-model Representation Learning Methods

While contrastive learning is the main method used for visual-textual representation learning with large datasets, how to model cross-modal representations in other modalities has been studied for a long run [4, 5, 56]. Inspired by this, rather than directly working on contrastive learning, in

Figure 1. Video-text pairs from web data. The degrees of semantic alignment vary from vague ones (a,b) to more specific ones (c,d).

this section, we first study three general cross-modal representation learning methods, including correspondence modeling, contrastive learning, and predictive coding.

3.1. Visual-textual Correspondence Modeling

While watching movies, or instructional videos, the texts/subtitles often come with the material. To leverage this information, the visual-textual correspondence modeling is designed as a binary classification task, that is, to predict whether a video clip and the description text are corresponded or not (Fig. 2 (a)). During training, we treat the video clip and the text taken at the same time as the corresponding pair, while the video clip with texts coming from other randomly sampled videos as non-corresponding pairs. More specifically, given a video clip v_i with its corresponding text t_i, we first extract the feature embedding $f_{v_i} \in \mathbb{R}^{d_v}, f_{t_i} \in \mathbb{R}^{d_t}$ using a visual backbone and textual backbone, respectively. After concatenating these two features together, a correspondence modeling sub-network $\mathcal{C}$ is derived to obtain the correspondence confidence value $\hat{c}_{ii}$ within a range of $(0, 1)$:

$$\hat{c}_{ii} = \mathcal{C}([f_{v_i}, f_{t_i}]),$$

where $[*, *]$ is a concatenation operation, and the ground-truth value c_{ii} is set to 1. Similarly, we obtain the correspondence confidence for an unmatched pair v_i and t_j as:

$$\hat{c}_{ij} = \mathcal{C}([f_{v_i}, f_{t_j}]),$$

where the ground-truth label c_{ij} is set to 0. Then the correspondence loss is defined as a binary cross entropy loss:

$$L_{corr} = -c_{ii} \log(\hat{c}_{ii}) - (1 - c_{ij}) \log(1 - \hat{c}_{ij}). \quad (1)$$

Although the formulation is simple, learning to predict correspondence is not a trivial task. It requires a thorough understanding of the visual concepts within the videos to match with the semantic meaning conveyed by the textual descriptions. Note that there are no explicit cross-modal distance constraints imposed on the formulation, more freedom is provided to the learning process and the learned representation has the potential to be useful for more downstream tasks other than retrieval.

3.2. Visual-textual Contrastive Learning

Contrastive learning [12, 49] has prevailed in representation learning in many domains, such as vision-only, visual-audio, and also visual-textual, because of its flexibility in loss design and superiority in performance. There are several variants [1, 44], but here we mainly describe the commonly used version. The goal of contrastive learning is to learn a joint embedding space $\mathcal{E}$ where the corresponded video and text features f_{v_i}, f_{t_i} are close to each other, and far away from each other if not corresponding. Besides, the semantic comparisons between the two modalities can be made by simple dot products in the embedding space [1, 44]. By assuming that the joint probability can be estimated up to a constant factor over the exponentiation of the dot product of two feature embeddings, we obtain:

$$\mathcal{E}(v_i, t_i) \propto \exp(f_{v_i}^{\intercal} f_{t_i}).$$

Accordingly, we adopt the NCE loss [1, 44] for optimization by differentiating between data obtained from the true joint distribution $\mathcal{E}$ and some artificially generated negative data. For a sample v_i, the positive training pair is the same as those generated in Section 3.1 that are taken at the same position. The negative pairs are obtained from different videos. For example, if a mini-batch contains N samples, then there will be N positive pairs and $N^2 - N$ negative pairs. Similarly, for a sample t_i, the training pairs are generated in the same manner. We thus define the NCE loss $NCE(v_i, t_i)$ and the loss for contrastive learning L_{cont} as:

$$\text{NCE}(v_i, t_i) = -\log \frac{\exp(f_{v_i}^{\intercal} f_{t_i}/\tau)}{\exp(f_{v_i}^{\intercal} f_{t_i}/\tau) + \sum_{j \in \mathcal{N}_{v_i}} exp(f_{v_i}^{\intercal} f_{t_j}/\tau)},$$

$$L_{cont} = \text{NCE}(v_i, t_i) + \text{NCE}(t_i, v_i), \quad (2)$$

where $\mathcal{N}_*$ denotes the negative index sets for the sample $*$ and τ is a temperature parameter.

3.3. Visual-textual Predictive Coding

Compared with correspondence modeling, contrastive learning can be regarded as a stricter method, with explicit

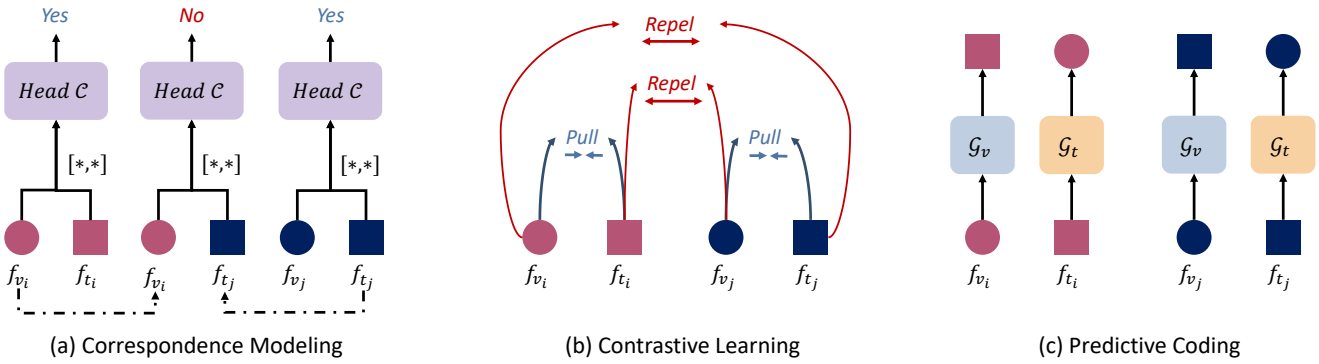

Figure 2. Cross-modal representation learning methods examined for our problem. (a) Correspondence modeling aims to classify whether a pair of video and text clip is corresponded or not, *i.e.* , binary classification. (b) Contrastive learning aims to pull positive pairs together while negative pairs apart in the embedding space. (c) Predictive coding aims to predict the embeddings of the corresponding modality A based on the embeddings of modality B ($A, B \in \{video, text\}$ in this work).

constraints imposed on the dot product of the cross-model features. To step further, in this subsection, we introduce the visual-textual predictive coding, by directly regressing the predicted coding of the other modality. This regression formulation is simpler than the above two methods, as it does not require large batches for generating negative sample pairs, and has been validated in prior works [25, 56] on representation learning of other domains. The assumption is that a good representation should maintain useful semantic features to accomplish this task.

Given a video clip v_i and its in-sync text t_i as the inputs, we treat the features f_{v_i}, f_{t_i} extracted from backbones as the coding for prediction. A small multi-layer perceptron (MLP) header is added to each backbone for the predictive coding generation, denoted as $\mathcal{G}_v, \mathcal{G}_t$ for video and text respectively. The training loss then contains two components:

$$L_{pred} = L_{v\to t}(v_i, t_i) + L_{t\to v}(v_i, t_i). \tag{3}$$

We use the mean square error (MSE) loss for predictive coding, with the detailed loss defined as:

$$L_{v\to t}(v_i, t_i) = \left\| \frac{\mathcal{G}_v(f_{v_i})}{\|\mathcal{G}_v(f_{v_i})\|_2} - sg\left[\frac{f_{t_i}}{\|f_{t_i}\|_2}\right] \right\|_2^2, \tag{4}$$

$$L_{t\to v}(v_i, t_i) = \left\| \frac{\mathcal{G}_t(f_{t_i})}{\|\mathcal{G}_t(f_{t_i})\|_2} - sg\left[\frac{f_{v_i}}{\|f_{v_i}\|_2}\right] \right\|_2^2, \tag{5}$$

where $sg[*]$ denotes the "stop gradient" operation to avoid the model collapse [23, 56].

3.4. Analyses

Implementation details. For the video branch, we use the standard S3D implementation following previous works [44, 71] throughout all the experiments. The video clip is sampled at 5 fps with 16 frames (*i.e.* 3.2 seconds). Each frame is resized to a resolution of 224 × 224. For the text branch, we use the word2vec embedding ($d = 300$) pretrained on Google News in a self-supervised manner [46] followed by two linear layers with a max-pooling layer in between [44]. For each text input, the maximum number of words is set to 30. The dimensions of both video and text features in the embedding space are 512. For the correspondence subnet $\mathcal{C}$, we follow the work [4] using a two-layer MLP and output a 2-dimensional softmax vector. For the coding subnet $\mathcal{G}$, we use a linear layer with the same dimension on input and output for prediction. We train our model on WebVid-2M [9] which contains 2.5M video-text pairs scraped from the web. The dataset includes a variety of styles for description text and is one of the most recently released web datasets. We compare the performance on MSR-VTT zero-shot text-to-video retrieval task to evaluate the learned cross-modal representation.

Table 1. Analysis of different methods on MSR-VTT zero-shot text-to-video retrieval task. **R@K**: Recall@K. **MedR**: Median Rank. **Corr.**: Correspondence modeling. **Contra.**: Contrastive learning. **Coding.**: Predictive Coding.

Method	**R@1**↑	**R@5**↑	**R@10**↑	**MedR** ↓
Random	0.01	0.05	0.1	500
Coding.	0.4	1.9	4.8	180.0
Corr.	2.4	7.3	11.0	141.0
Contra.	10.1	22.9	31.5	36.0
Contra. + Corr.	11.1	26.0	35.9	28.0
Contra. + Coding.	11.6	26.0	34.8	28.0
Contra. + Corr. + Coding.	11.5	26.2	34.8	27.0

The result is shown in Table 1. Though correspondence modeling has demonstrated its efficiency in visual-audio representation learning, training only with correspondence loss degrades the performance a lot, especially com-

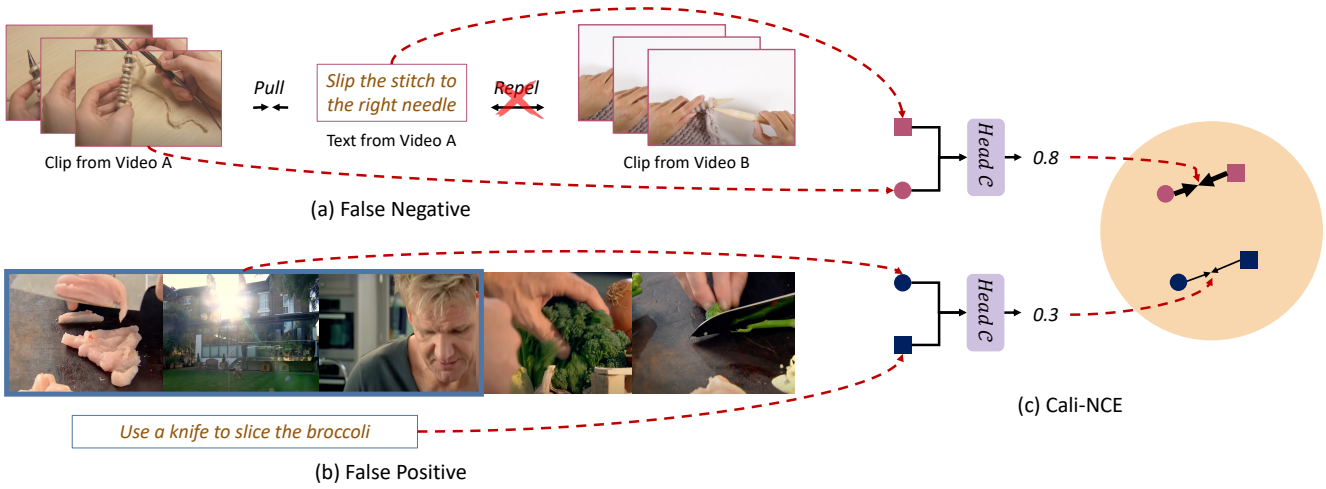

Figure 3. Illustration of the proposed approach. (a) and (b) are common problematic patterns in contrastive learning. False negative (a) happens when two different videos share the content under same semantics. And false positive (b) happens when misalignment exists between sampled clip and in-sync text (frames marked by dark blue box are the sampled clip). (c) An illustration of the proposed Cali-NCE loss, by re-weighting NCE with predicted correspondence confidence. The width of arrow indicates the level of confidence (*i.e.* correspondence score).

pared with the model trained with contrastive loss. We find that the training loss of correspondence classification converges very quickly. This may be due to the task of judging whether a video-text pair is corresponding is relatively easier. In this way, the model may overfit to the pretrained task without learning a more useful representation for the downstream task. Instead, by contrasting with a set of negative pairs and adding constraints on the dot production across modalities, contrastive learning has shown its superiority. This also aligns with the conclusion drawn from previous works [1, 9, 45]. For predictive coding, the performance is even worse than correspondence modeling. One possible reason is that there still has a certain level of collapse although we adopt the stop gradient during training. We encourage future research to further explore this direction.

We then try on different combinations of methods. The loss function is defined as $L = \theta_1 L_{corr} + \theta_2 L_{cont} + \theta_3 L_{pred}$. We set $\theta_1 = \theta_2 = 1$ and $\theta_3 = 200$ to balance the losses if it is used in the training. For example, *Contra. + Corr.* means we jointly train correspondence and contrastive together with $L_{corr} + L_{cont}$. We treat contrastive learning as a basic model because of its superiority in performance. It works surprisingly well when combining either correspondence classification or coding prediction with contrastive learning. This indicates that other methods could complement with contrastive learning, and further enhance the representation quality. Whereas adding all of these methods together does not explicitly boost the performance. We thus choose correspondence modeling to complement with contrastive learning and propose our Cali-NCE next.

4. Cali-NCE with Correspondence Modeling

In this section, we further explore the complementary information between correspondence modeling and contrastive learning. According to the instance discrimination assumption, samples extracted from the other videos are all treated as negative ones. However, among these negatives, some of them can be false negatives. For example, there may exist two different videos (*i.e.* v_a, v_b) both describing "*slip the stitch to the right needle*" (*i.e.* for texts t_a, t_b, and $t_a = t_b$ in this case), and contrastive learning will treat v_a, t_b as negative pairs, which is apparently not the reliable supervision signal (Fig. 3 (a)). Without posing any constraints on negative samples, joint training with correspondence tasks can alleviate this false negative issue. The reason behind this is that it is much easier for the model to pull such two videos together in the embedding space to fit the correspondence modeling goal (*i.e.* binary classification) than repelling them apart.

On the other hand, false positives also exist in the data samples, especially for the pretraining data that are not carefully curated/annotated. Except the examples shown in Fig. 1, the sample video clip and the text taken at the same time may not be aligned well [44]. For example, when watching an NBA game, the description text comes out only after the goal. In Fig. 3 (b), the misalignment comes out because of the shooting way, and the video producer may add some irrelevant clips for better storytelling and aesthetic pleasure. Directly increasing the dot product similarity among video and text pairs is inappropriate. Instead, the correspondence loss does not add on dot product

and provides model flexibility for training. Correspondence can be considered as a weak label, which is a more accurate supervision signal in this case.

Simply joint training with the correspondence modeling task is a very straightforward way to mitigate the above problems, and bring the performance gain (Section 3.4). But this is hard-coded and the affinity between modalities is uncontrollable. Here, based on the findings, in addition to joint training, we further propose a more adaptive solution **Cali-NCE**, by leveraging the correspondence prediction confidence. For each positive pair (v_i, t_i), we first estimate its correspondence confidence value $\hat{c_{ii}}$ via the correspondence modeling (Section 3.1). If the pair is more correlated, the confidence is larger (with an upper bound of 1). We then calibrate the contrastive loss by re-weighting with $\hat{c_{ii}}$ as follows:

$$L_{CaliNCE} = \lambda_1[\hat{c}_{ii} * (\text{NCE}(v_i, t_i) + \text{NCE}(t_i, v_i))] + \lambda_2 L_{corr}, \quad (6)$$

where λ_i is the weight used to balance the objective terms, and we set $\lambda_1 = \lambda_2 = 1$ empirically throughout our experiments. Note that L_{corr} is necessary in *Cali-NCE* for estimating $\hat{c}_{ii}$ in a data-driven and adaptive manner.

5. Experiments

Implementation details. We mainly follow the details mentioned in Section 3.4. Except for taking the convolutional neural network (CNN) as backbones, we also examine the Transformer-based backbone, which has shown great performance for representation learning tasks. We replace the standard S3D and MLP backbones with CLIP (ViT-B/32) [55] and extract features in 1 fps across 8 seconds. We freeze all the parameters from the CLIP model and use the features only. We build two Transformer layers with 512 hidden units and 8 heads on top of CLIP features for final evaluation on the downstream task.

5.1. Downstream Tasks

To show the effectiveness of our model and the generality of the learned representations, we conduct evaluations on three diverse downstream tasks: text-to-video retrieval, video action recognition, and video-to-video retrieval. We first introduce the downstream tasks with the corresponding datasets and required metrics below.

Text-to-video retrieval. This task aims to retrieve the best-matched video according to the input text. We focus on the zero-shot text-to-video retrieval as we want to directly evaluate the quality of the learned representation without further finetuning. And we use the dot product to measure the similarity across modalities. We test this task on two datasets: 1) **MSR-VTT [73]** is a dataset that contains 200K unique video clip-caption pairs covering 20 different categories. We use the same test split with 1K constructed by a previous work [75] that is commonly used in the literature [9, 10, 44]. 2) **MSVD [11]** contains 1,970 videos with 80K descriptions. We use the standard split with 670 videos in the test dataset [42]. We report the results on recall metrics ($R@K, K = 1, 5, 10$) which measure the percentage of correctly retrieved clips at the top K. We also report the median rank (MedR) of videos to be retrieved.

Table 2. Comparison to SOTA results on MSR-VTT for text-to-video retrieval. Numbers of previous works for non zero-shot methods trained on MSR-VTT are copied from the work [45].

Method	**R@1↑**	**R@5↑**	**R@10↑**	**MedR ↓**
Random	0.01	0.05	0.1	500
C+LSTM+SA+FC7 [62]	4.2	12.9	19.9	55.0
VSE-LSTM [33]	3.8	12.7	17.1	66.0
SNUVL [76]	3.5	15.9	23.8	44.0
CT-SAN [77]	4.4	16.6	22.3	35.0
JSFusion [75]	10.2	31.2	43.2	13.0
Zero-shot				
HowTo100M [45]	7.5	21.2	29.6	38.0
MIL-NCE [44]	9.9	24.0	32.4	29.5
SupportSet [53]	12.7	27.5	36.2	**24.0**
Baseline	10.1	22.9	31.5	36.0
Ours	**13.1**	**27.6**	**36.8**	26.0

Video action recognition. This task targets to classify the action class of each given video. We follow the released protocol [44] by training directly on pre-extracted features with a linear SVM. We test on both HMDB-51 [36] and UCF-101 [58] datasets and report the average accuracy over three splits.

Video-to-video retrieval. In this task, we evaluate the learned video representation in a zero-shot manner through video retrieval. Similarly, we examine on the most commonly used datasets: HMDB-51 [36] and UCF-101 [58]. Given a video, we uniformly sample 10 clips, each with 16 frames [72]. We obtain the final feature by averaging over these 10 clips. Then after extracting the feature from our model, we use cosine similarity to measure the distance between videos from the test set and those from the training split. The video retrieval performance is measured by querying the top K-nearest neighbors (NN) on the test split, where K is set to 1, 5, 10, 20, and 50. The retrieval result is considered as success if the class label of the test clip is within the top K-NN.

5.2. Comparison with State-of-the-art Methods

To better validate the effectiveness of the proposed Cali-NCE in a more controlled manner, here we design a baseline by training with vanilla NCE and keeping all the other settings the same as ours.

Zero-shot text-to-video retrieval. In Table 2, we evalu-

Table 3. Comparison to SOTA results on MSVD for text-to-video retrieval. Numbers of previous works for non zero-shot methods are copied from [9].

Method	R@1↑	R@5↑	R@10↑	MedR ↓
VSE-LSTM [33]	12.3	30.1	42.3	14.0
VSE++ [17]	15.4	39.6	53.0	9.0
Multi. Cues [48]	20.3	47.8	61.1	6.0
CE [42]	19.8	49.0	63.8	6.0
SupportSet [53]	23.0	52.8	65.8	5.0
Zero-shot				
SupportSet [53]	21.4	46.2	57.7	6.0
MIL-NCE [44]	32.6	59.8	73.2	3.5
Baseline	26.6	54.6	66.9	5.0
Ours	**32.8**	**63.0**	**76.4**	**3.0**

ate our learned representation on MSR-VTT. We achieve the state-of-the-art (SOTA) performance and outperforms contrastive learning-based models [44, 45] by a large margin. Our off-the-shelf model without finetuning even outperforms models directly trained on MSR-VTT [33, 62, 75–77]. Additional results on MSVD can be found in Table 3. Though the performance gain is not as high as that on MSR-VTT, our model still achieves better performance than other methods. We attribute this to that MSVD is a rather simple task with little room for improvement. The above shown performance over different datasets validates the effectiveness of our learned visual-textual representations.

Video retrieval. To evaluate the learned representation quality from our video branch, we conduct a zero-shot video retrieval experiment as in Table 4, where we report the average recalls over three splits on both HMDB-51 and UCF-101. A baseline method as well as other SOTA video representation learning methods are included for the comparison (including both single-modal and multi-modal approaches). The proposed approach achieves comparable performance to SOTAs on HMDB-51 and outperforms them on UCF-101 by a large margin.

Video action recognition. We show the results of our model compared to baseline, and other alternative methods on the task of action recognition in Table 5. The accuracy is averaged over three splits on both HMDB-51 and UCF0-101 datasets. We can see that our model surpasses baseline and other SOTAs on both datasets, putting more evidence on the effectiveness of our video-only representation learned from the cross-modal task. Though our model only achieves comparable results on video retrieval with HMDB-51, after a linear SVM, we surpass the SOTA methods with a notable gain, suggesting that our learned embedding is semantically-aware and separable.

Table 4. Comparison to SOTA results on zero-shot video retrieval on HMDB51 and UCF101.

	Method	R@1	R@5	R@10	R@20	R@50
HMDB51	Pace [65]	12.9	31.6	43.2	58.0	77.1
	CoCLR [25]	26.1	45.8	-	69.7	-
	SeLaVi [6]	24.8	47.6	-	**75.5**	-
	GDT [52]	25.4	51.4	-	75.0	-
	MIL-NCE [44]	32.9	**52.3**	**61.1**	70.5	**81.4**
	Baseline	22.2	39.5	48.3	58.2	70.5
	Ours	**34.4**	**52.2**	**60.9**	68.9	78.5
UCF101	Pace [65]	25.6	42.7	51.3	61.3	74.0
	CoCLR [25]	55.9	70.8	-	82.5	-
	SeLaVi [6]	52.0	68.6	-	84.5	-
	GDT [52]	57.4	73.4	-	88.1	-
	MIL-NCE [44]	51.7	64.1	70.5	77.1	87.9
	Baseline	44.8	58.8	66.1	73.5	84.9
	Ours	**64.2**	**78.4**	**83.8**	**88.8**	**94.0**

Table 5. Comparison to SOTA results on video action recognition on HMDB51 and UCF101.

Method	HMDB51-Acc.	UCF101-Acc.
OPN [37]	23.8	59.6
ShuffleLearn [47]	35.8	68.7
MAS [64]	33.4	61.2
CMC [61]	26.7	59.1
Geometry [22]	23.3	55.1
Fernando *et al.* [18]	32.5	60.3
ClipOrder [72]	30.9	72.4
3DRotNet [30]	40.0	75.3
DPC [24]	35.7	75.7
Pace [65]	36.6	77.1
3D STPuzzle [32]	33.7	65.8
CBT [59]	44.6	79.5
MIL-NCE [44]	47.0	80.0
Baseline	46.1	76.3
Ours	**49.6**	**81.4**

5.3. Ablation Studies

The effect of backbone. Because of the great success of using Transformers [9, 38], we compare our method with SOTAs by using CLIP as the backbone. The results are shown in Table 6, and we copy the values for direct retrieval with CLIP features from the previous paper [43]. As can be seen, our model still works with a transformer-based backbone. As our method is agnostic to the underlying backbone, it has the potential to combine with more advanced

Figure 4. Example text-to-video retrieval results on MSR-VTT using our trained joint embedding. The input text is shown on the top of each group and the correctly retrieved video (GT) is in an orange box. The candidate retrieval set is the whole test set.

Table 6. Comparison to SOTA results using Transformer-based backbone on MSR-VTT zero-shot text-to-video retrieval task.

Method	R@1↑	R@5↑	R@10↑	MedR ↓
Frozen [9]	18.7	39.5	51.6	10.0
ALPRO [40]	24.1	44.7	55.4	8.0
VIOLET [20]	25.9	49.5	59.7	-
CLIP [55]	31.2	53.7	64.2	4.0
Baseline	32.2	55.9	64.1	4.0
Ours	**32.9**	**56.5**	**66.8**	**4.0**

Table 7. The robustness of Cali-NCE with noisy training data. We show the random shuffle noisy rate on the left (*e.g.* , 10% indicates that this amount of training pairs is unmatched).

	Method	R@1↑	R@5↑	R@10↑	MedR ↓
10%	Baseline	31.0	55.2	65.0	4.0
	Ours	31.0	55.8	65.4	4.0
30%	Baseline	30.9	54.1	64.7	4.0
	Ours	32.0	55.8	65.3	4.0
50%	Baseline	31.4	53.2	63.0	5.0
	Ours	31.6	54.3	64.2	4.0

network architectures to further enhance performance.

The robustness of Cali-NCE. To verify the robustness of our method in dealing with noisy training pairs, we conduct the experiment by deliberately injecting noises into the training data. We randomly shuffle a ratio of pairs in training data to make them unmatched. We run this experiment with the CLIP-based backbone same as the above on a randomly selected $1/6$ WebVid-2M dataset. The results are shown in Table 7, testing on MSR-VTT zero-shot text-to-video retrieval task. With the help of Cali-NCE, the performance drop incurred by the noisy training data has been mitigated even for a random shuffle ratio of 50%, demonstrating the effectiveness of our Cali-NCE for dealing with noisy input data.

Qualitative results. Example results of text-to-video retrieval on MSR-VTT are shown in Fig. 4, and we regard vanilla NCE as the baseline. Note how our model can retrieve matched videos with the input text. Compared to vanilla NCE, rather than only focusing on a single concept in the input text (*e.g.* "car" in Fig. 4), our Cali-NCE pays attention to more fine-grained details (*e.g.* "car" and "cartoon"), supporting a deeper understanding of text semantics.

6. Conclusion

In this work, the main question we would like to address, is the noisy pairwise data used for learning the visual-textual representation. We have examined several methods, and found that the methods like correspondence modeling and predictive coding, can complement contrastive learning. Based on the findings, we proposed a new optimization solution called Cali-NCE, and showed its effectiveness over extensive quantitative and qualitative experiments. We hope that our work can inspire more following-up research on exploring semantic misalignment problems for cross-modal representation learning.

References

[1] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 2(6):7, 2020. 2, 3, 5

[2] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667*, 2019. 2

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015. 1, 2

[4] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, pages 609–617, 2017. 2, 4

[5] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *ECCV*, pages 435–451, 2018. 2

[6] Yuki Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. *NeurIPS*, 33:4660–4671, 2020. 7

[7] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017. 2

[8] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019. 2

[9] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. *arXiv preprint arXiv:2104.00650*, 2021. 1, 2, 4, 5, 6, 7, 8

[10] Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, et al. Multimodal clustering networks for self-supervised learning from unlabeled videos. *arXiv preprint arXiv:2104.12671*, 2021. 6

[11] 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. 6

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 3

[13] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1

[14] Banchiamlack Dessalegn and Barbara Landau. More than meets the eye: The role of language in binding and maintaining feature conjunctions. *Psychological science*, 19(2):189–195, 2008. 1

[15] Banchiamlack Dessalegn and Barbara Landau. Interaction between language and vision: It's momentary, abstract, and it develops. *Cognition*, 127(3):331–344, 2013. 1

[16] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *T-PAMI*, 2021. 2

[17] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 7

[18] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *CVPR*, pages 3636–3645, 2017. 2, 7

[19] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *NeurIPS*, 26, 2013. 2

[20] 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. 8

[21] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, pages 214–229, 2020. 2

[22] Chuang Gan, Boqing Gong, Kun Liu, Hao Su, and Leonidas J Guibas. Geometry guided convolutional neural networks for self-supervised video representation learning. In *CVPR*, pages 5589–5597, 2018. 7

[23] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 4

[24] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *ICCV Workshops*, pages 0–0, 2019. 7

[25] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *arXiv preprint arXiv:2010.09709*, 2020. 2, 4, 7

[26] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2906–2916, 2022. 2

[27] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *ECCV*, pages 649–665, 2018. 2

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 1, 2

[29] Ray Jackendoff. On beyond zebra: The relation of linguistic and visual information. *Cognition*, 26(2):89–114, 1987. 1

[30] Longlong Jing and Yingli Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. *arXiv preprint arXiv:1811.11387*, 2(7):8, 2018. 2, 7

[31] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, pages 4565–4574, 2016. 1, 2

[32] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *AAAI*, volume 33, pages 8545–8552, 2019. 7

[33] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 6, 7

[34] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, pages 4437–4446, 2015. 2

[35] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *arXiv preprint arXiv:1807.00230*, 2018. 2

[36] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011. 6

[37] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *ICCV*, pages 667–676, 2017. 2, 7

[38] 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 *CVPR*, pages 7331–7341, 2021. 2, 7

[39] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 1

[40] 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. 8

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 1, 2

[42] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019. 6, 7

[43] 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. 7

[44] 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 *CVPR*, pages 9879–9889, 2020. 1, 2, 3, 4, 5, 6, 7

[45] 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 *ICCV*, pages 2630–2640, 2019. 1, 2, 5, 6, 7

[46] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 4

[47] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, pages 527–544. Springer, 2016. 7

[48] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27, 2018. 7

[49] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[50] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, pages 631–648, 2018. 2

[51] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, pages 4594–4602, 2016. 2

[52] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, Joao F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298*, 2020. 7

[53] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020. 2, 6, 7

[54] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *CVPR*, pages 6964–6974, 2021. 2

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 6, 8

[56] Adrià Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Patraucean, Florent Altché, Michal Valko, et al. Broaden your views for self-supervised video learning. *arXiv preprint arXiv:2103.16559*, 2021. 2, 4

[57] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*, 2018. 2

[58] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6

[59] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. *URL http://arxiv. org/abs*, 2019. 7

[60] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, pages 7464–7473, 2019. 2

[61] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, pages 776–794. Springer, 2020. 7

[62] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124*, 2016. 6, 7

[63] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *T-PAMI*, 39(4):652–663, 2016. 1

[64] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *CVPR*, pages 4006–4015, 2019. 7

[65] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *ECCV*, pages 504–521. Springer, 2020. 7

[66] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *T-PAMI*, 41(2):394–407, 2018. 2

[67] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, pages 5005–5013, 2016. 1, 2

[68] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *CVPR*, pages 8052–8060, 2018. 2

[69] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011. 2

[70] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018. 1, 2

[71] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, pages 305–321, 2018. 4

[72] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *CVPR*, pages 10334–10343, 2019. 2, 6, 7

[73] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016. 6

[74] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, pages 4651–4659, 2016. 1, 2

[75] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, pages 471–487, 2018. 6, 7

[76] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. Video captioning and retrieval models with semantic attention. *arXiv preprint arXiv:1610.02947*, 6(7), 2016. 6, 7

[77] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *CVPR*, pages 3165–3173, 2017. 6, 7

[78] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, pages 8746–8755, 2020. 2