Rethinking Noisy Video-Text Retrieval via Relation-aware Alignment

Huakai Lai[1] Guoxin Xiong[1] Huayu Mai[1] Xiang Liu[3] Tianzhu Zhang[1,2*]
[1]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China;
[2]National Key Laboratory of Deep Space Exploration, Deep Space Exploration Laboratory;
[3]Dongguan University of Technology
{tbhk, xgx, mai556}@mail.ustc.edu.cn, liuxiang@dgut.edu.cn, tzzhang@ustc.edu.cn

Abstract

Video-Text Retrieval (VTR) is a core task in multi-modal understanding, drawing growing attention from both academia and industry in recent years. While numerous VTR methods have achieved success, most of them assume accurate visual-text correspondences during training, which is difficult to ensure in practice due to ubiquitous noise, known as noisy correspondences (NC). In this paper, we rethink how to mitigate the NC from the perspective of representative reference features (termed agents), and propose a novel relation-aware purified consistency (RPC) network to amend direct pairwise correlation, including representative agents construction and relation-aware ranking distribution alignment. The proposed RPC enjoys several merits. First, to learn the agents well without any correspondence supervision, we customize the agents construction according to the three characteristics of reliability, representativeness, and resilience. Second, the ranking distribution-based alignment process leverages the structural information inherent in inter-pair relationships, making it more robust compared to individual comparisons. Extensive experiments on five datasets under different settings demonstrate the efficacy and robustness of our method.

1. Introduction

Video-text retrieval [5, 20, 24, 27] aims to identify videos that most accurately correspond to a textual query, and vice versa, where textual descriptions are retrieved based on video content. With the exponential growth of online videos and the increasing use of video platforms, video-text retrieval has become a critical task and benefits broad applications across various domains [2, 11, 42]. This task is fundamental to cross-modal understanding and presents significant challenges due to the substantial disparities in representation between video and text modalities.

*Corresponding author

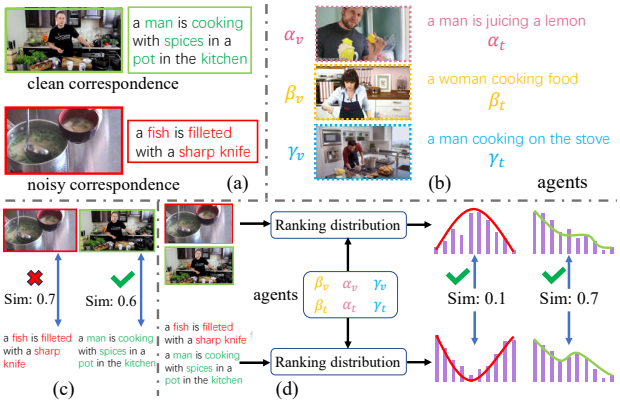

Figure 1. The motivation of our proposed method. (a) The illustration of noisy correspondence. (b) A set of representative reference pairs (termed agents). (c) False matches tend to occur under conditions where noise correspondence leads to inaccurate supervision signals. (d) The proposed ranking distribution consistency suppresses false matches and provides a more reliable supervision.

In prior literature, the dominant paradigm for video-text retrieval is to establish correspondences between the video and text modalities by global-level and local-level alignments. Specifically, global-level methods [8, 24, 25, 27] achieve semantic alignment by leveraging contrastive learning on holistic representations of videos and texts. For local-level methods [5, 15, 44, 45], cross-modal correspondences are established through more fine-grained alignments, such as frame-to-word [44] or action/entity-level alignments [5, 15]. Although these methods have shown promising results, they are based on the core assumption of accurate visual-language correspondences during training, making their performance heavily reliant on the availability of high-quality annotated data.

However, collecting such high-quality annotated data is resource-intensive and time-consuming. In fact, due to factors such as non-expert annotators and web crawling noise, it is often difficult or unrealistic to maintain the assumption that every video-text pair in a dataset is accurately

matched (see Fig. 1(a)). Therefore, the dataset collection process inevitably introduces some semantically irrelevant video-text pairs, which are erroneously treated as matched pairs, referred to as noisy correspondences (NC). Aligning these mismatched pairs during training can mislead the model and substantially degrade its matching performance.

To mitigate the negative impact of noisy data pairs, a series of noise-rectify approaches [10, 13, 49, 52] have been developed. These works typically divide the original training data into clean and noisy subsets based on the memory effect of DNNs [3] and achieve robust matching through label correction. Despite achieving some success, these methods are highly sensitive to clean sample selection, which relies heavily on the model's similarity prediction. Confident yet erroneous predictions tend to be self-perpetuating, where noisy correspondences can be strengthened and memorized in subsequent training, resulting in a feedback loop of error accumulation, especially with high noise ratios. Furthermore, erroneous supervision from noisy data can impair the model's representation ability, increasing the likelihood of mismatches between video and text. This situation further deteriorates the establishment of connections between the video and text modalities.

In this paper, we investigate the limitations of noise-rectify methods, and provide a novel viewpoint on representative reference features (termed agents) to establish more reliable supervision for robust video-text retrieval under noisy correspondences (NC). Intuitively, most methods ignore that video-text retrieval involves rich inter-pair correlation in addition to simple single-pair similarity. This highlights the potential of closer cooperation between the inter-pair correlation and single-pair similarity to reliably mitigate the effects of NC. The core point is that, for each video/text, we can acquire the agent-level correlation (i.e., a likelihood vector) via a group of representative reference agents (see Fig.1(b)) to capture the inter-pair relation. Essentially, the agent-level correlation captures the consistency across a wider scope. Therefore, it encapsulates a higher-order correlation regularization to rectify erroneously matched video-text pairs caused by NC. Building on the preceding analysis, it follows logically to enforce the regularization based on the agent-level correlation. However, this simple regularization considers each agent in isolation and is overly dependent on independent and identically distributed (i.i.d.) assumption, resulting in limited model optimization. Indeed, specific structural relations exist among the agents. For instance, as depicted in Fig. 1(b), agent α and agent γ are both 'a man', while agent β is 'a woman'. Accordingly, agent α should be more closely related to agent γ than agent β. To utilize the inter-agent structural relations, rather than treating each agent in isolation, we meticulously develop a relation-aware alignment based on ranking distribution. The main principle is to treat the agent ranking as a stochastic event instead of a fixed permutation. Given a video/text, the varying similarity among different agents can be viewed as ranking probabilities. The ranking permutation indicates the relation of the corresponding agents to video/text. By associating every ranking permutation of the agents, we build the relation-aware ranking distribution of video and text. As shown in Fig. 1(d), the consistency of the relation-aware ranking distribution between video and text provides more reliable guidance for the model optimization under NC.

However, it is non-trivial to learn the agents well without any correspondence supervision. For agent mining, we meticulously design this customized for the following three characteristics. (1) **Reliability**. Since similarity calculation plays a crucial role in video-text retrieval particularly under NC, confident yet erroneous calculations can adversely affect all subsequent processes. Therefore, we design a semantic purification mechanism instead of a simple dot product. Based on this, we perform a bi-directional evaluation strategy to select agents with relatively high reliability. (2) **Representativeness**. Intuitively, the agents should represent a broad array of semantics from both video and text modality. Therefore, we adopt a cross-aggregation mechanism that allows the agents to recognize the semantics of current pair and extend to the semantics of various video-text pairs. (3) **Resilience**. Given the substantial gap between video and text, it is essential for agents to narrow this gap and increase their referability. Specifically, we attach a self-aggregation mechanism to refine agents by harmonizing the inherent resilience between video and text.

In this work, our contributions can be concluded as follows: (1) We analyze the bottlenecks exist in noise-rectify methods for amending video-text pair consistency regularization. To the best of our knowledge, this is the first work to mitigate the NC in VTR methods, from the perspective of representative reference features. (2) We develop a coherent RPC network, including construction of reliable and representative agents and a ranking distribution-based alignment to capture structure information inherent in inter-agent relationships. They can cooperate well to enable more robust matching. (3) Extensive experimental results on five challenging benchmarks show that our method achieves notable performance improvements, with particularly pronounced gains under high noise ratios.

2. Related Work

In this section, we briefly overview several lines of research in video-text retrieval and noisy correspondence learning.

2.1. Video-Text Retrieval

Building on advances in deep neural networks [22, 28, 31, 33, 39, 41, 46, 47], many video-text retrieval methods have been proposed and achieved remarkable results.

These methods broadly fall into two categories based on the alignment: global-matching [6, 8, 24, 25, 27] and local-matching methods [5, 15, 44, 45]. Global-matching methods aim to learn and align holistic video/text features within a joint embedding space. Local-matching methods establish fine-grained alignments, such as frame-word, action, and event levels, enabling more precise matching. Recently, visual-language pre-training (VLP) [14, 21, 37], particularly CLIP [37], has gained increasing attention. Benefiting from VLP, several works [9, 16, 19, 25, 27, 30, 53] have been proposed to exploit the powerful knowledge of CLIP for VTR. Despite achieving promising results, these methods typically rely on the implicit assumption that all training pairs are accurately matched. However, this assumption is often unrealistic, as real-world data is frequently affected by pervasive noise. In this paper, we tackle the unavoidable and intractable issue of noisy correspondences.

2.2. Learning with Noisy Correspondence

Noisy correspondence (NC) was first proposed in NCR [13] and is a novel paradigm in the field of noise learning. Unlike conventional noisy label which denotes incorrect category labels [23, 40], NC denotes alignment errors in paired data, i.e., semantically irrelevant videos and texts are incorrectly treated as matched. To mitigate this problem, various approaches have emerged, which are mainly classified into two categories: noise-rectify paradigm [10, 13, 29, 36, 49, 52] and robust loss functions [7, 12, 34, 35]. Noise-rectify methods typically partition the training set into clean and noisy subsets and alleviate the effects of noise through the label correction. Robust loss function methods concentrate on designing noise-tolerant loss functions to enhance the robustness of the model against NC during training. Although these methods have shown promising results in numerous tasks, the area of VTR still lacks sufficient exploration. In this paper, we offer a fresh perspective on representative reference features to construct more safe and effective supervision signals for robust video-text retrieval under NC.

3. Method

3.1. Problem Definition

Given a dataset $D = \{(v_i, t_i)\}_{i=1}^{N}$, where (v_i, t_i) is the i-th video-text pair and N is the dataset size, video-text retrieval seeks to learn a similarity function $s(\cdot)$ that assigns a high similarity score to semantically related video-text pair and a low similarity score to irrelevant pair. Formally, given a pair of text t_i with N_t words and a video v_i with N_v frames, we feed t_i and v_i into a text encoder and a video encoder, respectively, to obtain their corresponding embeddings $\mathbf{t}_i = [\mathbf{w}_i^0, \mathbf{w}_i^1, \cdots, \mathbf{w}_i^{N_t}]$ and $\mathbf{v}_i = [\mathbf{f}_i^1, \mathbf{f}_i^2, \cdots, \mathbf{f}_i^{N_v}]$. Here, $\mathbf{w}_i^0$ denotes the `[CLS]` token and $\mathbf{f}_i^n$ denotes the n-th frame feature. The similarity between video and text is calculated based on the features $\mathbf{t}_i$ and $\mathbf{v}_i$, for instance, by computing the inner product of the `[CLS]` token $\mathbf{w}_i^0$ and the average frame features $\overline{\mathbf{v}}_i = Avg([\mathbf{f}_i^1, \mathbf{f}_i^2, \cdots, \mathbf{f}_i^{N_v}])$:

$$s(v_i, t_i) = \left\langle \mathbf{w}_i^0, \overline{\mathbf{v}}_i \right\rangle . \tag{1}$$

Given a batch of B video-text pairs $\{(v_i, t_i)\}_{i=1}^{B}$, the InfoNCE loss [32] is employed to maximize the similarity between annotated video-text pairs (v_i, t_i) and minimize the similarity for irrelevant pairs $(v_i, t_j, i \neq j)$:

$$\mathcal{L}_{t2v} = -\frac{1}{B}\sum_{i}^{B}(log\frac{e^{s(t_i,v_i)/\tau}}{\sum_j^B e^{s(t_i,v_j)/\tau}}), \tag{2}$$

$$\mathcal{L}_{v2t} = -\frac{1}{B}\sum_{i}^{B}(log\frac{e^{s(v_i,t_i)/\tau}}{\sum_j^B e^{s(v_i,t_j)/\tau}}), \tag{3}$$

$$\mathcal{L}_{info} = \frac{1}{2}(\mathcal{L}_{t2v} + \mathcal{L}_{v2t}), \tag{4}$$

where τ is the temperature parameter. In reality, due to unavoidable annotation errors during the dataset collection process, some mismatched pairs are erroneously labeled as matched, resulting in noisy correspondences (NC) and consequently causing a decline in performance. Our objective is to alleviate the detrimental effects of NC for robust VTR.

3.2. Overview of Framework

As illustrated in Fig. 2, the proposed RPC mainly includes two procedures, i.e., 1) representative agents construction, and 2) relation-aware ranking distribution alignment. In procedure 1), we develop the semantic purification mechanism, cross-aggregation, and self-aggregation to ensure the reliability, representativeness, and resilience of the agents, respectively. In procedure 2), we conduct agent-ranking probability distributions for each video/text features on corresponding agents. The details are as follows.

3.3. Agents Construction

It is non-trivial to construct agents well-tailored for each video-text pair without any correspondence supervision. The agent construction process comprises three components, namely semantic purification, cross-aggregation, and self-aggregation, which equip the agents with reliability, representativeness, and resilience, respectively.

Semantic Purification Mechanism. Similarity calculation is crucial for video-text alignment, with further increased demands for similarity reliability in the NC. However, few noisy correspondence learning methods have explored this area. In the mainstream noise-rectify paradigm, confident yet incorrect similarity calculations can lead to errors in subsequent noisy subset partitioning and label correction, which tend to be self-perpetuating. Thus, to improve the reliability of similarity calculation, we use

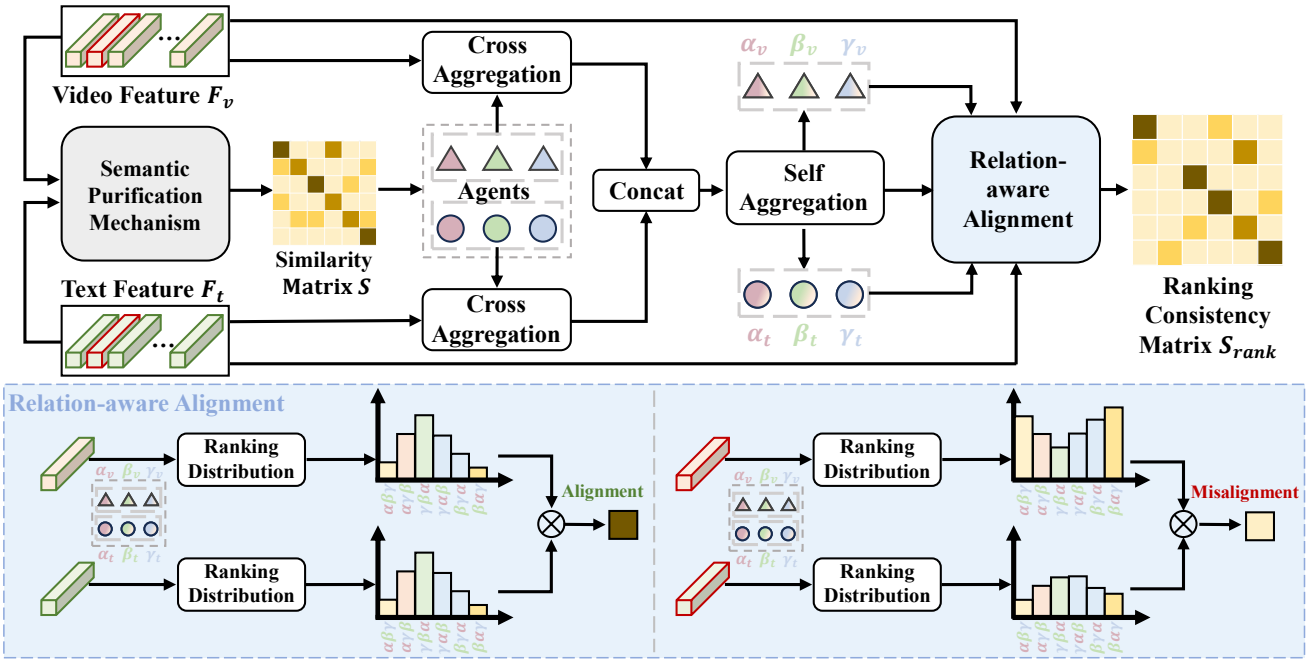

Figure 2. Framework of proposed Relation-aware Purified Consistency (RPC) Network. There are two main procedures: 1) representative agents construction and 2) relation-aware alignment based on ranking distribution, with noisy pairs in red border and clean pairs in green.

a purified maximum-correspondence interaction to consolidate the alignment. In specific, we begin by computing the token-wise similarity between frame embeddings $\mathbf{v}_p = [\mathbf{f}_p^1, \mathbf{f}_p^2, \cdots, \mathbf{f}_p^{N_v}]$ and word embeddings $\mathbf{t}_q = [\mathbf{w}_q^0, \mathbf{w}_q^1, \cdots, \mathbf{w}_q^{N_t}]$ to generate the frame-word similarity matrix $\boldsymbol{S} \in \mathbb{R}^{N_v \times N_t}$, where $(\boldsymbol{S})_{ij} = \frac{(\mathbf{f}^i)^\top \mathbf{w}^j}{\|\mathbf{f}^i\| \|\mathbf{w}^j\|}$ denotes the similarity between i-th frame and j-th word. For simplicity, we omit the subscript indicating the p-th video and q-th text.

Next, we select the most closely aligned word (frame) for each frame (word), i.e., the highest scores in each row and column of $\boldsymbol{S}$. We then calculate the weighted average of these highest scores to obtain the overall similarity score between the video v_p and text t_q, denoted as $s(v_p, t_q)$:

$$s(v_p, t_q) = \sum_{i=1}^{N_v} \omega_v^i \max_j (\boldsymbol{S})_{ij} + \sum_{j=1}^{N_t} \omega_t^j \max_i (\boldsymbol{S})_{ij}, \quad (5)$$

$$\omega_v^i = \frac{d(\mathbf{w}^0, \mathbf{f}^i)}{\sqrt{L}}, \quad \omega_t^j = \frac{d(\overline{\mathbf{v}}, \mathbf{w}^j)}{\sqrt{L}}, \quad (6)$$

where $\mathbf{w}^0$ denotes `[CLS]` of text, $\overline{\mathbf{v}}$ is obtained by average pooling over all frame features, $d(\cdot, \cdot)$ denotes the dot product similarity in [43] and $\sqrt{L}$ is a scaling factor. Moreover, due to the presence of semantic inconsistency and noise, some frames may be semantically irrelevant to any of the words, and vice versa. Hence, it is unreasonable to include all the so-called 'most aligned' frame-word pairs in the similarity computation. To reduce the negative impact and purify similarity calculation, we measure the holistic relevance of the video and text as a threshold in a data-driven manner:

$$\omega_\star^i = \begin{cases} \omega_\star^i & \omega_\star^i > \delta \\ 0 & \omega_\star^i \leq \delta \end{cases}, \ \delta = \frac{d(\mathbf{w}^0, \overline{\mathbf{v}})}{\alpha\sqrt{L}}, \ \star \in \{v, t\}. \quad (7)$$

Finally, we use the purified similarity from Eqs. 5-7 to determine the relative reliability of video-text pairs based on bidirectional cross-modal correspondences.

$$y_i = \frac{1}{2}\left(\frac{e^{s(t_i, v_i)/\tau}}{\sum_j^B e^{s(t_i, v_j)/\tau}} + \frac{e^{s(v_i, t_i)/\tau}}{\sum_j^B e^{s(v_i, t_j)/\tau}}\right), \quad (8)$$

where $s(t_i, v_j)$ is the similarity between i-th text and j-th video in Eq. 5 and τ is the temperature parameter. For clean pairs, t_i and v_i should exhibit dominant similarity in both retrieval directions, yielding a metric value close to 1. In contrast, for noisy pairs, the similarity dominance in both directions cannot be simultaneously achieved, leading to a smaller metric value. Thus, y_i can be used to indicate cleanliness of i-th pair. The *top-K* clean pairs are selected as reference agents, denoted as $\boldsymbol{A} = \{(v_k^a, t_k^a)\}_{k=1}^K$. The corresponding video and text features are represented as $\boldsymbol{F}_v^A = [\mathbf{v}_1^a, \mathbf{v}_2^a, \cdots, \mathbf{v}_K^a]$ and $\boldsymbol{F}_t^A = [\mathbf{t}_1^a, \mathbf{t}_2^a, \cdots, \mathbf{t}_K^a]$.

Cross-aggregation Mechanism. The agents should resonate favorably with the rich semantics of different videos and texts. To enrich the semantics of the initial agents above, we employ the cross-aggregation mechanism to condense the various semantics into the corresponding agents. In specific, we derive the queries $\mathbf{Q}_\star$ from the agent features $\boldsymbol{F}_\star^A$, the keys $\mathbf{K}_\star$ for obtaining aggregation weights

from the video or text features $\boldsymbol{F}_\star$, and the values $\mathbf{V}_\star$ for feature aggregation from $\boldsymbol{F}_\star$. Here, we use $\star$ to indicate the subscript, with $\star \in \{V, T\}$ for brevity.

$$\mathbf{Q}_\star = \boldsymbol{F}_\star^A \mathbf{W}_\star^{\mathcal{Q}},\ \mathbf{K}_\star = \boldsymbol{F}_\star \mathbf{W}_\star^{\mathcal{K}},\ \mathbf{V}_\star = \boldsymbol{F}_\star \mathbf{W}_\star^{\mathcal{V}}, \tag{9}$$

$$\hat{\boldsymbol{F}}_\star^A = softmax(\frac{\mathbf{Q}_\star \mathbf{K}_\star^\mathsf{T}}{\sqrt{L}})\mathbf{V}_\star, \tag{10}$$

where $\mathbf{W}_\star^{\mathcal{Q}}, \mathbf{W}_\star^{\mathcal{K}}, \mathbf{W}_\star^{\mathcal{V}} \in \mathbb{R}^{L\times L}$ are linear projections, and $\sqrt{L}$ is scaling factor. Subsequently, a feed-forward network is employed to derive the $\hat{\boldsymbol{F}}_v^A = [\hat{\mathbf{v}}_1^a, \hat{\mathbf{v}}_2^a, \cdots, \hat{\mathbf{v}}_K^a]$ and $\hat{\boldsymbol{F}}_t^A = [\hat{\mathbf{t}}_1^a, \hat{\mathbf{t}}_2^a, \cdots, \hat{\mathbf{t}}_K^a]$, which include abundant semantics.

Self-aggregation Mechanism. The formulated $\hat{\boldsymbol{F}}_v^A$ and $\hat{\boldsymbol{F}}_t^A$ are respectively derived from the video and text features that exhibit substantial disparities due to significant modality gap. To narrow this gap, we introduce the self-aggregation mechanism to integrate complementary information. Specifically, we first concatenate the $\hat{\boldsymbol{F}}_v^A$ and $\hat{\boldsymbol{F}}_t^A$:

$$\hat{\boldsymbol{F}}^A = concat(\hat{\boldsymbol{F}}_v^A, \hat{\boldsymbol{F}}_t^A). \tag{11}$$

Next, we employ multi-head self-attention on the $\hat{\boldsymbol{F}}^A$:

$$\mathbf{Q} = \hat{\boldsymbol{F}}^A \mathbf{W}^{\mathcal{Q}},\ \mathbf{K} = \hat{\boldsymbol{F}}^A \mathbf{W}^{\mathcal{K}},\ \mathbf{V} = \hat{\boldsymbol{F}}^A \mathbf{W}^{\mathcal{V}}, \tag{12}$$

$$\tilde{\boldsymbol{F}}_\star^A = softmax(\frac{\mathbf{Q}_\star \mathbf{K}_\star^\mathsf{T}}{\sqrt{L}})\mathbf{V}_\star, \tag{13}$$

where $\mathbf{W}^{\mathcal{Q}}, \mathbf{W}^{\mathcal{K}}, \mathbf{W}^{\mathcal{V}} \in \mathbb{R}^{L\times L}$ are the linear projections. After these three processes, we establish highly reliable agents $\tilde{\boldsymbol{F}}_v^A, \tilde{\boldsymbol{F}}_t^A \in \mathbb{R}^{K\times L}$ that capture rich semantic information and exhibit cross-modal resilience.

3.4. Relation-aware Alignment

With the obtained agents as references, a straightforward approach is to enforce consistency constraint on the agent-level alignment ($\mathbf{r}_v$ and $\mathbf{r}_t$) between the i-th video-text pair.

$$\mathbf{r}_\star^i = softmax(\boldsymbol{F}_\star^i (\tilde{\boldsymbol{F}}_\star^A)^\mathsf{T}),\ \star \in \{v, t\}, \tag{14}$$

where $\mathbf{r}_\star \in \mathbb{R}^{1\times K}$. However, considering each agent separately increases the likelihood of misalignment, especially under noise correspondence. By further incorporating the structural information inherent in inter-agent relations, we meticulously develop the relation-aware alignment based on ranking distribution for more reliable supervision. The primary concept is to treat agent ranking as a random event instead of a fixed permutation. In other words, each permutation of the agents occurs with a certain probability, instead of a fixed order from largest to smallest. Given $\mathbf{r}$, the probability associated with a permutation $\pi \in \mathcal{P}$ ($|\mathcal{P}| = N!$) is computed as:

$$P(\pi|\mathbf{r}) = \prod_{k=1}^{K} \frac{\mathbf{r}_{\pi(k)}}{\sum_{k'=k}^{K} \mathbf{r}_{\pi(k')}}. \tag{15}$$

Here, $\pi(k)$ refers to k-th agent index in this permutation. For example, assume that for a given video-text pair, the constructed agents are α, β and γ. One possible permutation of these three agents is $\pi = (\alpha, \beta, \gamma)$. The probability of π can be obtained from the agent correlation $\mathbf{r}$:

$$P(\pi|\mathbf{r}) = \frac{\mathbf{r}(\alpha)}{\mathbf{r}(\alpha)+\mathbf{r}(\beta)+\mathbf{r}(\gamma)} \cdot \frac{\mathbf{r}(\beta)}{\mathbf{r}(\beta)+\mathbf{r}(\gamma)} \cdot \frac{\mathbf{r}(\gamma)}{\mathbf{r}(\gamma)}. \tag{16}$$

Based on the probabilities of all $|\mathcal{P}|$ permutations, we convert simple dot product alignment $\mathbf{r}$ into relation-aware ranking distribution $P(\mathcal{P}|\mathbf{r}) \in \mathbb{R}^{1\times|\mathcal{P}|}$. Under this view, ranking distribution captures the inter-agent relationship. Generally, calculating all permutations of the K agents would incur prohibitively high costs. In this regard, we observed that the *top*-4 agents account for nearly all the weight in Eq. 14. Therefore, we consider only the permutations of the *top*-4 agents for each video/text for efficiency. Note that for simplicity, we omitted the subscript denoting modality of $\mathbf{r}$ in the above process, as the principles are consistent across both modalities. Based on the agent-ranking distribution of video $P(\mathcal{P}|\mathbf{r_v})$ and text $P(\mathcal{P}|\mathbf{r_t})$, we can obtain the consistency between them as supervision:

$$S_{rank}^{ij} = cos(P(\mathcal{P}|\mathbf{r}_v^i), P(\mathcal{P}|\mathbf{r}_t^j)). \tag{17}$$

As a result, the relation-aware ranking distribution alignment regularization can be derived as:

$$\mathcal{L}_{rank} = \frac{1}{B}\sum_{i}^{B} KL(S_{rank}^{i,:}, s^{i,:}), \tag{18}$$

where B is the batch size, KL denotes KL divergence and $s^{ij} = s(v_i, t_j)$ is defined in Eq. 5. Finally, the overall loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{info} + \lambda \mathcal{L}_{rank}, \tag{19}$$

where λ is weight coefficient.

4. Experiments

4.1. Experiment Setup

Datasets. MSR-VTT [48] includes 10,000 videos, each accompanied by 20 captions. Following [52], we use two data splits: full and 1k-A [50]. **LSMDC** [38] contains 118,081 videos with equal captions from 202 movies. The 1,000 test videos are from movies disjoint with the training set. **MSVD** [4] contains 1,970 videos with 80,000 captions. We use 1,200, 100, and 670 videos to train, validate, and test, respectively. **ActivityNet** [18] contains 20,000 YouTube videos. Following [27, 51], we concatenate all descriptions of a video to form a paragraph and evaluate with video-paragraph retrieval on the 'val1' split. **DiDeMo** [1] contains

Table 1. Retrieval performance under 0%, 20% and 50% noise rates on MSR-VTT full and 1k-A. The best results are highlighted in **bold**.

Noise	Method	MSR-VTT Full						MSR-VTT 1kA					
		Text to Video			Video to Text			Text to Video			Video to Text		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
0%	CE [24]	10.0	29.0	41.2	15.6	40.9	55.2	20.9	48.8	62.4	20.6	50.3	64.0
	MMT [8]	-	-	-	-	-	-	24.6	54.0	67.1	24.4	56.0	67.8
	TT-CE+ [6]	15.0	38.5	51.7	25.3	55.6	68.6	-	-	-	-	-	-
	T2VLAD [45]	12.7	34.8	47.1	20.7	48.9	62.1	29.5	59.0	70.1	31.8	60.0	71.1
	CLIP4Clip [27]	23.6	46.4	57.1	39.6	67.5	77.5	44.5	71.4	81.6	42.7	70.9	80.6
	RVTR [52]	24.3	47.2	59.0	40.9	69.4	79.3	45.8	73.0	83.5	43.9	72.6	82.0
	Ours	**29.8**	**55.8**	**66.5**	**51.6**	**81.2**	**90.6**	**47.6**	**73.2**	**83.6**	**46.3**	**73.9**	**83.4**
20%	CE [24]	7.4	22.3	32.8	10.2	30.6	42.3	14.8	40.3	52.5	15.2	39.2	40.4
	MMT [8]	-	-	-	-	-	-	20.4	50.1	65.7	19.8	50.3	64.4
	TT-CE+ [6]	7.5	22.6	33.6	10.9	32.1	44.7	-	-	-	-	-	-
	T2VLAD [45]	-	-	-	-	-	-	21.3	44.7	56.3	22.4	45.4	59.0
	CLIP4Clip [27]	9.3	23.1	31.9	15.8	35.3	46.3	27.6	52.0	63.7	22.6	47.6	60.0
	RVTR [52]	20.8	43.1	54.7	34.6	62.7	75.0	42.3	69.7	81.0	41.0	70.8	79.3
	Ours	**28.3**	**53.7**	**64.4**	**50.0**	**80.0**	**88.1**	**46.1**	**72.3**	**82.4**	**45.5**	**72.9**	**82.6**
50%	CE [24]	4.4	14.3	21.6	3.5	13.1	20.1	9.1	23.6	31.0	6.4	19.8	27.3
	MMT [8]	-	-	-	-	-	-	14.8	28.0	39.7	13.6	26.4	36.1
	TT-CE+ [6]	5.3	17.6	27.3	7.3	23.1	34.7	-	-	-	-	-	-
	T2VLAD [45]	-	-	-	-	-	-	2.9	7.8	10.7	3.0	12.8	21.7
	CLIP4Clip [27]	11.7	28.0	37.7	8.6	21.9	31.3	24.4	48.7	59.6	23.4	47.1	58.5
	RVTR [52]	18.4	38.9	49.9	30.6	60.9	71.0	36.2	61.6	73.7	36.0	61.7	73.0
	Ours	**27.1**	**52.1**	**62.7**	**48.1**	**78.2**	**87.8**	**44.6**	**71.1**	**81.2**	**44.3**	**71.9**	**82.4**

10,000 videos with 40,000 captions. Following [20, 24], we also evaluate with video-paragraph retrieval.

Evaluation Metrics. We evaluate the retrieval performance by standard video-text retrieval metrics Recall at K (R@K, higher is better). R@K is defined as the percentage of correct videos/texts within the top K retrieved videos/texts. Following [27, 52], K was set to 1, 5, and 50 for ActivityNet, and 1, 5, and 10 for the other four datasets.

Implementation Details. Following [27, 52], visual and text encoders are initialized from CLIP with ViT-B/32 [37]. For MSR-VTT, LSMDC, and MSVD, the frame length N_v and text length N_t are set as 12 and 32. For ActivityNet and DiDeMo, N_v and N_t are 64. Feature dimension L is set as 512 for all benchmarks, and α is set as 5. All frames are resized into 224×224. Our model is optimized with Adam [17] and the cosine learning rate schedule [26] is employed. The initial learning rate is set as $1e$-7 for CLIP encoders and $1e$-4 for others. The batch size is set as 128. The number of agent K is 10 and weight coefficient λ is 0.2.

4.2. Comparison with State-of-the-arts

We compared our method with previous state-of-the-art methods on five popular benchmarks. To comprehensively evaluate the robustness of our method, we simulate two levels of noisy correspondences, namely 20%, and 50% by randomly shuffling the captions like [52].

Results on MSR-VTT. As shown in Tab. 1, our method consistently outperforms previous advanced methods across various scenarios, clearly demonstrating its effectiveness. Specifically, we analyze the results from the following perspectives. (1) In the synthetic NC settings, our method significantly exceeds all methods under all noise ratios. Notably, at 50% noise, our method achieves R@1 gains of 8.7, and 8.4 over the best baseline RVTR [52] across full and 1k-A splits. (2) As the noise rate rises, other methods experience significant performance degradation, while ours decreases only slightly. For example, on 1k-A split, as the noise rate rises from 20% to 50%, the R@1 of the RVTR decreases by 6.1, whereas our method only decreases by 1.5. (3) In the NC-free setting, our method also outperforms state-of-the-art approaches, further demonstrating the effectiveness and advantage of our method.

Results on Other Four Datasets. Tabs. 2 and 3 demonstrate the performance comparison on MSVD, LSMDC, ActivityNet, and Didemo. It can be found that our method constantly surpasses other methods across all metrics by a significant margin. These results underscore the powerful reliable semantic alignment capability under the noisy correspondence setting of our method.

Stability Comparison. To further highlight the stability of our method, we plot the text-to-video R@1 curves for each method in MSR-VTT 1k-A at increasing noise ratios in Fig. 3. It can be observed that our method exceeds all other methods across all noise ratios. Meanwhile, as the noise ratio increases, the performance degradation of our method is much smaller compared to other methods. The results show that our method is highly noise-resistant.

Table 2. Retrieval performance under 50% noise rates on MSVD and LSMDC. The best results are highlighted in **bold**.

Method	MSVD						LSMDC					
	Text to Video			Video to Text			Text to Video			Video to Text		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CE [24]	9.9	27.8	40.6	11.8	27.9	38.7	4.6	13.9	22.2	5.2	14.8	21.7
MMT [8]	-	-	-	-	-	-	5.1	13.9	20.1	5.1	13.7	19.9
TT-CE+ [6]	10.5	32.5	47.4	12.1	31.9	43.7	6.2	20.5	27.3	8.8	18.0	24.7
CLIP4Clip [27]	8.4	22.3	33.4	6.2	17.4	29.8	14.3	32.3	41.7	13.9	29.8	40.4
RVTR [52]	28.7	57.0	70.3	30.0	64.9	80.0	19.2	38.0	47.0	19.9	38.1	46.5
Ours	**38.5**	**69.9**	**80.9**	**48.1**	**71.8**	**81.3**	**22.8**	**42.3**	**52.4**	**22.0**	**39.8**	**50.9**

Table 3. Retrieval performance under 50% noise rates on DiDeMo and ActivityNet. The best results are highlighted in **bold**.

Method	DiDeMo						ActivityNet					
	Text to Video			Video to Text			Text to Video			Video to Text		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@50	R@1	R@5	R@50
CE [24]	5.7	18.1	28.0	6.5	19.3	27.6	5.1	15.8	55.4	4.2	15.4	61.0
MMT [8]	-	-	-	-	-	-	19.1	48.0	87.0	19.8	48.1	86.6
TT-CE+ [6]	12.5	31.5	42.5	11.8	31.1	42.2	7.3	24.1	69.1	4.9	15.3	51.1
CLIP4Clip [27]	14.5	36.2	48.3	15.6	37.7	50.7	8.4	25.8	38.3	4.2	16.3	26.8
RVTR [52]	26.9	49.5	59.9	27.0	50.1	59.5	31.2	61.3	95.9	34.2	62.8	94.9
Ours	**34.7**	**63.4**	**74.4**	**33.4**	**61.8**	**73.2**	**33.0**	**63.2**	**96.1**	**35.9**	**65.9**	**96.4**

Table 4. Ablation on different agent initialization strategies.

	Text-to-Video			Video-to-Text		
Method	R@1	R@5	R@10	R@1	R@5	R@10
All	43.1	70.2	80.5	42.9	70.8	81.1
Learnable	42.9	69.8	80.6	43.3	71.1	80.2
Random	42.8	70.0	79.7	43.1	70.6	80.5
Ours	**44.6**	**71.1**	**81.2**	**44.3**	**71.9**	**82.4**

Table 5. Ablation on agents construction.

Agents Construction			Text-to-Video			Video-to-Text		
SPM	Cross	Self	R@1	R@5	R@10	R@1	R@5	R@10
			40.4	66.1	77.4	40.1	67.6	76.8
✓			42.9	69.8	79.5	42.6	70.8	80.5
✓	✓		43.7	70.3	81.1	43.3	71.6	81.7
✓		✓	43.3	70.4	80.4	43.5	71.7	81.4
✓	✓	✓	**44.6**	**71.1**	**81.2**	**44.3**	**71.9**	**82.4**

Table 6. Ablation on relation-aware correlation consistency.

	Text-to-Video			Video-to-Text		
Method	R@1	R@5	R@10	R@1	R@5	R@10
Pair-wise	39.3	65.5	76.1	40.1	66.7	75.3
Agent-level	42.6	70.6	78.7	42.9	70.0	80.4
Ours	**44.6**	**71.1**	**81.2**	**44.3**	**71.9**	**82.4**

4.3. Ablation Study

To examine the effectiveness of our method, we conduct extensive ablation studies on MSR-VTT 1k-A at 50% noise.

Effectiveness of Agent Construction. As shown in Tab. 5, we conduct diagnostic experiments progressively to verify the effectiveness of each components in the agents construction process. Note that the first line indicates the result of applying relation-aware alignment in the absence of proposed agent construction process. We can observe that this result has already surpassed the previous state-of-the-art methods, which fully demonstrates the effectiveness of relation-aware alignment. The 2-nd row shows the introduction of semantic purification mechanism (SPM) yields a substantial performance lift, i.e., 2.5 on R@1. The gains can primarily be attributed to our more reliable similarity calculation method, which effectively suppresses false matches. By further integrating the cross-aggregation, performance improved by 0.8 R@1. This can be mainly ascribed to various semantic cues absorbed during cross-attention. Finally, when the self-aggregation is adopted, the performance increases by 0.9 R@1. This proves that refined agents after self-aggregation bridge the modality gap.

Ablation on Agent Initialization. To explore the effectiveness of different agent selection strategies, we compare the performance of several intuitive initialization ways in Tab. 4. In these methods, *'All'* indicates that all video-text pairs are used as agents. *'Random'* refers to randomly selecting k video-text pairs as agents. *'Learnable'* means that agents are randomly initialized trainable parameters. It can be found that our bidirectional selection consistently exceeds other methods. This can be attributed to the fact that

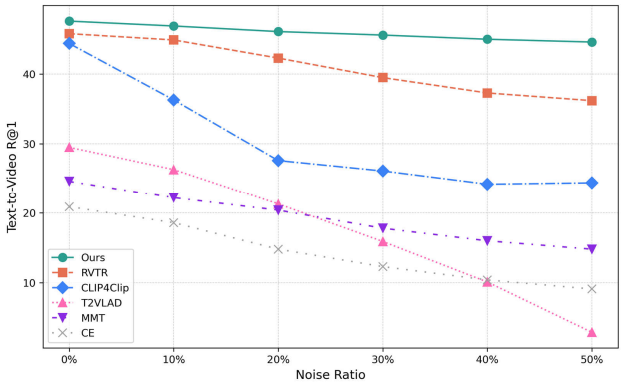

Figure 3. R@1 comparison for each method at various noise ratios.

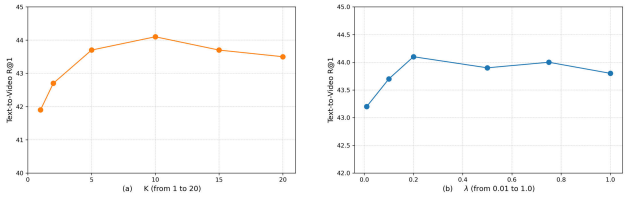

Figure 4. Hyperparameter experiments on the number of agents and weight for the regularization in Eq. 19.

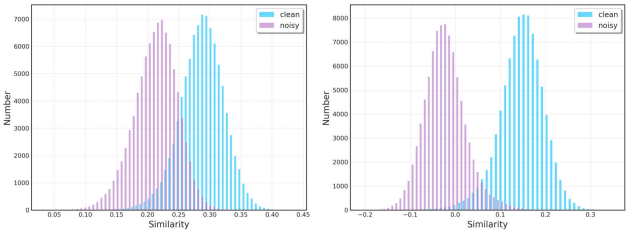

Figure 5. The similarity distribution of clean pairs and noisy pairs before and after training on MSR-VTT with a 50% noise rate.

Figure 6. Instances of NC in MSR-VTT training set with 50% noise, where noisy texts are highlighted in red, and for clarity, the corresponding ground truth is displayed below in green. The proposed ranking consistency more accurately reflects the true degree of correspondence compared to pair-wise similarity.

our method selects more reliable agents as references, while effectively alleviating the interference from noisy pairs.

Effectiveness of Relation-aware Alignment. As shown in Tab. 6, we compare the performance of different correlation schemes. *'Pair-wise'* denotes directly calculating one-to-one similarity between video and text. *'Agent-level'* refers to measuring the similarity of agent-level correlation in Eq. 14 via the dot product or L2 distance. The former may lead to considerable false matches, while the latter overlooks the inherent relationships between agents. The proposed relation-aware alignment yields optimal results, underscoring that modeling inter-agents relationships provides more effective and reliable supervision.

Hyperparameter Evaluations. Quantitative experiments are conducted to explore the appropriate number of agents and weight for the ranking regularization. As illustrated in Fig. 4(a), the performance continues to grow until the number reaches 10, beyond which it starts to decline. It is reasonable as too few agents cannot represent diverse semantic information while too many agents increase the risk of interference from noisy samples. Similarly, in Fig. 4(b), both too small or too large for λ will lead to suboptimal results, and we select 0.2 in our experiments. Overall, our method exhibits robustness to variations in hyperparameters.

4.4. Qualitative Results

To further analyze and understand the proposed RPC, we show several qualitative examples in MSR-VTT under 50% noise. In Fig. 5, it can be observed that after training, the mean of the clean sample distribution becomes distinctly separated from that of the noisy sample distribution. In Fig. 6, it can be noticed that noisy samples incorrectly learn relatively high similarity due to erroneous supervision, whereas our ranking consistency accurately and reliably reflects the true semantic relevance. Note that 0.3 is already a relatively high similarity (see Fig. 5). For more results, please refer to Supplementary Material.

5. Conclusion

In this paper, we present a novel and coherent RPC network, including reliable, representative, and resilient agents construction and relation-aware ranking distribution alignment, to tackle the challenging video-text retrieval under NC. We not only achieve consistent performance gains over state-of-the-art methods on five datasets under different settings, but also opens up new avenues from the perspective of representative reference features. We hope this paper could help advance the issue in future research.

Acknowledgements

This work was supported by National Defense Science and Technology Foundation Strengthening Program Funding (Grant 2023-JCJQ-JJ-0219).

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 5

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[3] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017. 2

[4] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 5

[5] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10635–10644, 2020. 1, 3

[6] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11583–11593, 2021. 3, 6, 7

[7] Zhuohang Dang, Minnan Luo, Chengyou Jia, Guang Dai, Xiaojun Chang, and Jingdong Wang. Noisy correspondence learning with self-reinforcing errors mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1463–1471, 2024. 3

[8] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal Transformer for Video Retrieval. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 3, 6, 7

[9] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5006–5015, 2022. 3

[10] Haochen Han, Kaiyao Miao, Qinghua Zheng, and Minnan Luo. Noisy correspondence learning with meta similarity correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7517–7526, 2023. 2, 3

[11] 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. 1

[12] Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9595–9610, 2023. 3

[13] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419, 2021. 2, 3

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3

[15] Peng Jin, Jinfa Huang, Pengfei Xiong, Shangxuan Tian, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2472–2482, 2023. 1, 3

[16] Peng Jin, Hao Li, Zesen Cheng, Jinfa Huang, Zhennan Wang, Li Yuan, Chang Liu, and Jie Chen. Text-video retrieval with disentangled conceptualization and set-to-set alignment. *arXiv preprint arXiv:2305.12218*, 2023. 3

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[18] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 5

[19] Huakai Lai, Wenfei Yang, Tianzhu Zhang, and Yongdong Zhang. Reliable phrase feature mining for hierarchical video-text retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 3

[20] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 1, 6

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3

[22] Zhaoyang Li, Yuan Wang, Wangkai Li, Rui Sun, and Tianzhu Zhang. Localization and expansion: A decoupled framework for point cloud few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 18–34. Springer, 2025. 2

[23] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015. 3

[24] 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. 1, 3, 6, 7

[25] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *European Conference on Computer Vision*, pages 319–335. Springer, 2022. 1, 3

[26] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[27] 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 and captioning. *Neurocomputing*, 508:293–304, 2022. 1, 3, 5, 6, 7

[28] Naisong Luo, Yuan Wang, Rui Sun, Guoxin Xiong, Tianzhu Zhang, and Feng Wu. Exploring the better correlation for few-shot video object segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(3):2133–2146, 2025. 2

[29] Xinran Ma, Mouxing Yang, Yunfan Li, Peng Hu, Jiancheng Lv, and Xi Peng. Cross-modal retrieval with noisy correspondence via consistency refining and mining. *IEEE Transactions on Image Processing*, 2024. 3

[30] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022. 3

[31] Huayu Mai, Rui Sun, Yuan Wang, Tianzhu Zhang, and Feng Wu. Pay attention to target: Relation-aware temporal consistency for domain adaptive video semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4162–4170, 2024. 2

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

[33] Yuwen Pan, Rui Sun, Naisong Luo, Tianzhu Zhang, and Yongdong Zhang. Exploring reliable matching with phase enhancement for night-time semantic segmentation. In *European Conference on Computer Vision*, pages 408–424. Springer, 2024. 2

[34] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4948–4956, 2022. 3

[35] Yang Qin, Yuan Sun, Dezhong Peng, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Cross-modal active complementary learning with self-refining correspondence. *Advances in Neural Information Processing Systems*, 36:24829–24840, 2023. 3

[36] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27197–27206, 2024. 3

[37] 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. 3, 6

[38] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123:94–120, 2017. 5

[39] Zhangxiang Shi, Tianzhu Zhang, Xi Wei, Feng Wu, and Yongdong Zhang. Decoupled cross-modal phrase-attention network for image-sentence matching. *IEEE Transactions on Image Processing*, 33:1326–1337, 2022. 2

[40] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11):8135–8153, 2022. 3

[41] Rui Sun, Yihao Li, Tianzhu Zhang, Zhendong Mao, Feng Wu, and Yongdong Zhang. Lesion-aware transformers for diabetic retinopathy grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10938–10947, 2021. 2

[42] Mingkang Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li. Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4858–4862, 2021. 1

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[44] Qiang Wang, Yanhao Zhang, Yun Zheng, Pan Pan, and Xian-Sheng Hua. Disentangled representation learning for text-video retrieval. *arXiv preprint arXiv:2203.07111*, 2022. 1, 3

[45] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5079–5088, 2021. 1, 3, 6

[46] Yuan Wang, Rui Sun, Naisong Luo, Yuwen Pan, and Tianzhu Zhang. Image-to-image matching via foundation models: A new perspective for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3952–3963, 2024. 2

[47] Guoxin Xiong, Meng Meng, Tianzhu Zhang, Dongming Zhang, and Yongdong Zhang. Reference-aware adaptive network for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[48] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 5

[49] Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19883–19892, 2023. 2, 3

[50] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European conference on computer vision (ECCV)*, pages 471–487, 2018. 5

[51] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *Proceedings of the european conference on computer vision (ECCV)*, pages 374–390, 2018. 5

[52] Huaiwen Zhang, Yang Yang, Fan Qi, Shengsheng Qian, and Changsheng Xu. Robust video-text retrieval via noisy pair calibration. *IEEE Transactions on Multimedia*, 25:8632–8645, 2023. 2, 3, 5, 6, 7

[53] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. Centerclip: Token clustering for efficient text-video retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 970–981, 2022. 3