DUQ: Dual Uncertainty Quantification for Text-Video Retrieval

Xin Liu[1], **Shibai Yin**[1*], **Jun Wang**[1*], **Jiaxin Zhu**[1], **Xingyang Wang**[1] and **Yee-Hong Yang**[2]
[1] Southwestern University of Finance and Economics
[2] University of Alberta
xliu067@163.com, {shibaiyin,wangjun1987}@swufe.edu.cn, {1764758458, 1921808558}@qq.com, herberty@ualberta.ca

Abstract

Text-video retrieval establishes accurate similarity relationships between text and video through granularity alignment and feature enhancement. However, relying solely on similarity to associate intra-pair features and distinguish inter-pair features is insufficient, *e.g.*, when querying a multi-scene video with sparse text or selecting the most relevant video from many similar candidates. In this paper, we propose a novel Dual Uncertainty Quantification (DUQ) model that separately handles uncertainties in intra-pair interaction and inter-pair exclusion. Specifically, to enhance intra-pair interaction, we propose an intra-pair similarity uncertainty module to provide similarity-based trustworthy predictions and explicitly model this uncertainty. To increase inter-pair exclusion, we propose an inter-pair distance uncertainty module to construct a distance-based diversity probability embeding, thereby widening the gap between similar features. The two components work synergistically, jointly improving the calculation of similarity between features. We evaluate our model on six benchmark datasets: MSRVTT (51.2%), DiDeMo, LSMDC, MSVD, Charades, and VATEX, achieving state-of-the-art retrieval performance.

1 Introduction

In recent years, text-video retrieval has made significant progress, focusing on finding the most relevant videos based on text queries [Liu *et al.*, 2019]. The emergence of feature representation learning [Luo *et al.*, 2022] has laid the foundation for addressing the text-video retrieval task. This method leverages powerful pre-trained models, *e.g.*, CLIP [Radford *et al.*, 2021], which project text and video into a shared latent space based on the semantic similarities of text-video pairs, enabling more efficient and accurate cross-modal retrieval. Meanwhile, coarse-grained and fine-grained feature interactions [Chen *et al.*, 2024] enhance the dense feature extractor to learn better representations. *However, is it sufficient to rely solely on the similarity between features to obtain reliable retrieval results?*

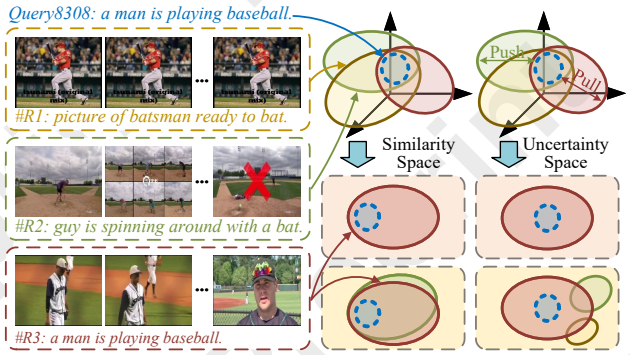


Figure 1: **Motivation.** (left) A failed text-video retrieval example based on feature-level similarity relationships, struggling with uncertainty from intra-pair interaction and inter-pair exclusion. (right) Our method models "Pull" and "Push" dynamics to reduce feature uncertainty.

Actually, relying on similarity alone is insufficient to support cross-modal interactions and distinguish within-modal differences. As illustrated in Figure 1, given the text query *"a man is playing baseball,"* the top 3 retrieval results are returned based on features similarity using the X-Pool model [Gorti *et al.*, 2022]. Although #R1 and #R2 are related to baseball, the query does not accurately match the actual description. In contrast, #R3 highlights a man playing baseball while being interviewed. However, this is a failed retrieval task because #R3 includes an irrelevant interview scene, which results in greater misalignment in feature interaction. This is a common issue in intra-pair interaction, frequently occurring in multi-scene video retrieval when using a sparse text query. Similarly, the sparse text query and weak semantic differences between highly similar candidate videos lead to higher uncertainty in inter-pair exclusion. In other words, both the interaction and the exclusion of features may involve uncertainty. Therefore, we categorize this uncertainty into two types: **(1) Intra-pair uncertainty.** Low-quality data, such as multi-scene videos, duplicate images, and non-detailed descriptions, *etc.*, is detrimental to the interaction between paired text and video, inevitably leading to unreliable retrieval results. **(2) Inter-pair uncertainty.** High-

*Corresponding author.
†Code is available at https://github.com/OPA067/DUQ

similarity data, such as same-scene videos, similar images, and consistent descriptions, *etc.*, often causes interference in the retrieval of positive samples. The root of this issue lies in the limitations of the similarity feature space, which does not provide confidence for intra-pair interaction or construct distinguishability for inter-pair exclusion. Therefore, it is essential to quantify the uncertainty in text-video pairs to enable a more reliable similarity assessment.

To address the uncertainty matching problem, existing methods for text-video retrieval focus on granularity alignment and enhanced feature learning. As for the former, UCoFiA [Wang *et al.*, 2023] proposes a unified coarse-to-fine alignment model, which combines interactive similarity aggregation and normalization strategies, effectively improving the accuracy of video-text retrieval. HBI [Jin *et al.*, 2023a] proposes the Hierarchical Banzhaf Granularity Interaction, which uses multivariate game theory to model interactions between video frames and text words, overcoming the limitations of traditional feature similarity. As for the latter, T-Mass [Wang *et al.*, 2024] addresses the inherent sparsity of text features using feature enhancement and regularization techniques, thereby mitigating uncertainty in intra-pair interaction. Although these methods alleviate feature uncertainty to some extent, they often rely on complex fine-grained alignment strategies and feature enhancement techniques. More importantly, these methods still focus on similarity-based feature spaces and far from effectively handling the inherent uncertainties in determining optimal entity combinations with appropriate granularities during text-video matching.

In this paper, we propose a novel text-video retrieval model to tackle both intra-pair interaction and inter-pair exclusion uncertainty problems, named the **D**ual **U**ncertainty **Q**uantification (**DUQ**) Model. Figure 2 illustrates the overall framework. **First**, for intra-pair uncertainty, we propose an **I**ntra-pair **S**imilarity **U**ncertainty **M**odule **(ISUM)** to provide trustworthy predictions by quantifying the uncertainty in text-video feature interaction arising from inherent low-quality data. This method leverages the out-of-domain confidence learning problem in classification tasks, aiming not only to maximize similarity scores for in-domain matches but also to increase confidence in the presence of inherent data ambiguity. **Second**, for inter-pair uncertainty, we propose an **I**nter-pair **D**istance **U**ncertainty **M**odule **(IDUM)** to construct modality-specific differences by computing the distance in text-video probabilistic embeddings arising from inherent high-similarity data. We break the limitations of traditional single-feature approaches by proposing a local feature aggregation module to construct diversified probabilistic embeddings. Compared to traditional feature embeddings, probabilistic embeddings offer greater diversity and expressiveness, effectively distinguishing similar features within the same modality. Meanwhile, we use the maximum distance as an exclusion metric between inter-pair multiple probabilistic embeddings, which effectively reduces computational complexity and excludes similar probabilistic embeddings in an extreme boundary fashion. **Third**, the two uncertainty modules work synergistically, expanding retrieval criteria to include both feature interaction uncertainty and feature exclusion distance. Our contributions are summarized as follows:

- We propose a novel text-video retrieval framework, Dual Uncertainty Quantification, to address the uncertainty issues in text-video interactions and exclusions.
- For intra-pair interaction uncertainty caused by low-quality data, we propose an Intra-pair Similarity Uncertainty Module to provide similarity-based trustworthy predictions and explicitly model this uncertainty.
- For inter-pair exclusion uncertainty caused by high-similarity data, we propose an Inter-pair Distance Uncertainty Module to construct distance-based diversity probability embeddings, thereby increasing the gap between similar data.
- We conduct extensive experiments on six benchmark datasets: MSRVTT, DiDeMo, LSMDC, MSVD, Charades, and VATEX, achieving state-of-the-art retrieval performance (51.2%, +1.9% in R@1 on MSRVTT).

2 Related Work

Text-Video Retrieval. The text-video retrieval is a key research topic in multimodal studies. Early works [Liu *et al.*, 2019; Chen *et al.*, 2020] primarily focus on enhancing feature representations to align text and video, as well as on establishing benchmarks and foundational models. Recently, transformer-based text-video retrieval methods [Luo *et al.*, 2022; Gorti *et al.*, 2022] use cross-attention to abstract multimodal cues, achieving significant performance gains. For example, TS2Net [Liu *et al.*, 2022] employs a "token shift and selection transformer" to preserve token integrity and capture subtle actions, improving retrieval performance. With the large-scale text-image pretraining model CLIP [Radford *et al.*, 2021] achieving significant success, it has inspired improvements in retrieval tasks. For example, DiCoSA [Jin *et al.*, 2023b] enhances text-video retrieval by decoupling coarse features into semantic factors and using adaptive pooling for accurate set-to-set matching. Additionally, to better represent text features, T-Mass [Wang *et al.*, 2024] uses random text modeling and text regularization to extract effective frames and text, thereby enhancing semantic similarity between text and video. Although existing methods establish similarity alignment baselines by fine-tuning CLIP, they overlook the reliability of pairings and the impact of noisy samples. By contrast, our model goes beyond feature-level similarity, eliminating the need for complex granularity alignment strategies and focusing on addressing key issues in retrieval.

Uncertainty Model. Currently, the most influential uncertainty models include uncertainty distributions [Oh *et al.*, 2018; Song and Soleymani, 2019; Chun *et al.*, 2021] and uncertainty metrics [Lakshminarayanan *et al.*, 2017; Sensoy *et al.*, 2018; Li *et al.*, 2024]. Uncertainty distributions construct probabilistic representations to differentiate between features. For example, HIB [Oh *et al.*, 2018] proposes the Hedged Instance Embedding to handle one-to-many probabilistic correspondences for metric learning, which has been successfully applied to face recognition and 2D-to-3D pose estimation. Uncertainty metrics handle out-of-domain classification tasks and are powerful in both making accurate predictions and providing reliable uncertainty estimates. For example,

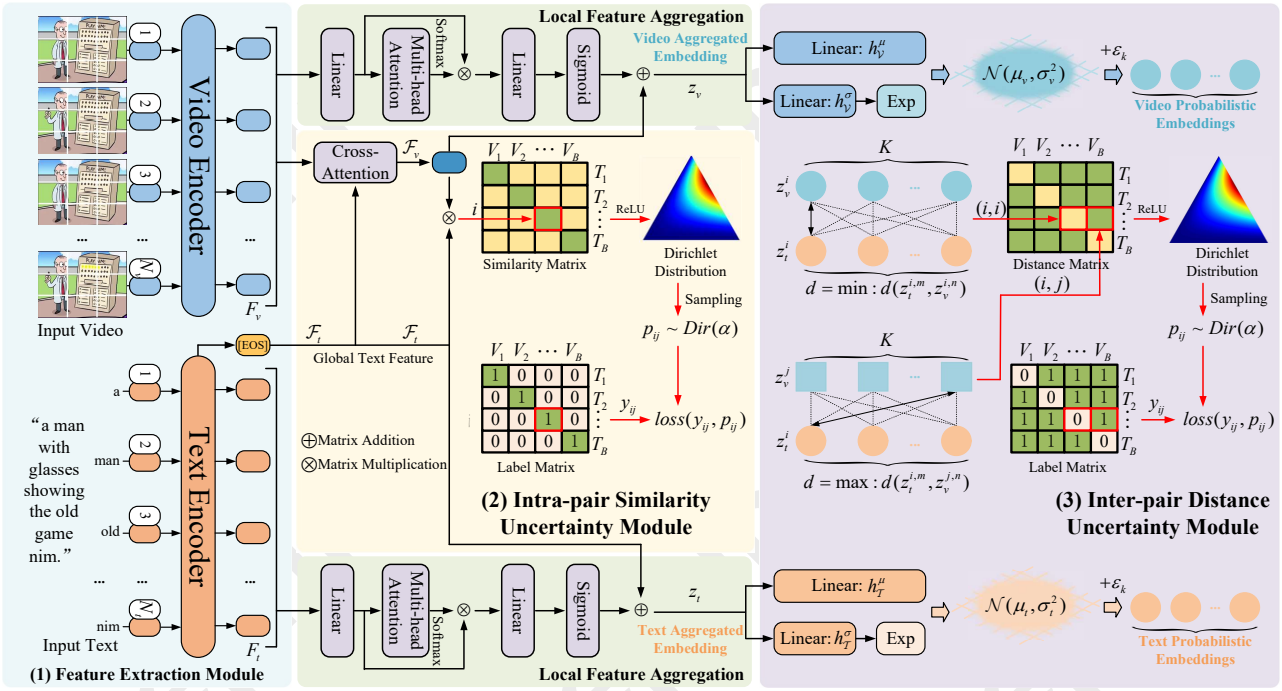


Figure 2: **Framework.** (1) The Feature Extraction Module maps text and video inputs into a joint embedding space to compute similarity. (2) The Intra-pair Similarity Uncertainty Module provides similarity-based trustworthy predictions and explicitly models intra-pair interaction uncertainty. (3) The Inter-pair Distance Uncertainty Module constructs distance-based diversity probabilistic embeddings and uses boundary distances to represent inter-pair exclusion differences.

EDL [Sensoy *et al.*, 2018] replaces the Dirichlet distribution on class probabilities and treats the predictions of a neural network as subjective opinions, learning the function that collects the evidence leading to these opinions through a deterministic neural network from data. Although uncertainty-based methods have made impressive progress in recognition and classification tasks, applying them to more complex retrieval tasks remains challenging.

3 Methodology

3.1 Preliminaries

Features Extraction. As shown in Figure 2 (1), given a text query T, we leverage the CLIP [Radford *et al.*, 2021] text encoder to output word features $F_t = [w_1, w_2, \cdots, w_{N_t}] \in \mathbb{R}^{N_t \times D}$, where N_t and D represent the number and the dimension of the word features, respectively. Then we take the representation of the `[EOS]` token as the sentence feature $\mathcal{F}_t \in \mathbb{R}^{1 \times D}$. Similar to the text encoder, we utilize the CLIP video encoder to extract visual features. Specifically, given a video V comprising multiple frames, we uniformly sample N_v frames, then feed them into the video encoder to obtain visual frame features $F_v = [f_1, f_2, \cdots, f_{N_v}] \in \mathbb{R}^{N_v \times D}$.

Features Interaction. The feature interaction lies in learning a similarity score $s(t, v)$, which aims to maximize the similarity for positive pairs (intra-pairs) and minimize it for negative pairs (inter-pairs). Typically, video frame features $F_v \in \mathbb{R}^{N_v \times D}$ are aggregated to obtain video feature $\mathcal{F}_v \in \mathbb{R}^{1 \times D}$ that can interact with the text feature $\mathcal{F}_t$, *e.g.*, through average pooling of video frames. In our work, we adopt the most effective cross-attention frame aggregation method X-Pool [Gorti *et al.*, 2022] to obtain the video feature $\mathcal{F}_v$, and compute the similarity score $s(t, v) = \frac{\mathcal{F}_t \cdot \mathcal{F}_v}{||\mathcal{F}_t|| \cdot ||\mathcal{F}_v||}$. During training, a common optimization method is to use a symmetric cross-entropy loss in both the text-to-video and video-to-text directions, as shown in Eq. (1):

$$\mathcal{L}_S = -\frac{1}{2}\left(\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{s(t_i,v_i)\cdot\lambda}}{\sum_{j=1}^{B}e^{s(t_i,v_j)\cdot\lambda}} + \frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{s(t_i,v_i)\cdot\lambda}}{\sum_{j=1}^{B}e^{s(t_j,v_i)\cdot\lambda}}\right), \tag{1}$$

where B is the batch size and λ is the temperature hyper-parameter. This loss function maximizes the similarity of intra-pairs and minimizes the similarity of inter-pairs. Although previous works optimize this loss at the feature similarity level with some success, when faced with a sparse textual query, this strategy lacks credibility in intra-pair interactions and fails to support inter-pair exclusion. Therefore, we need to break feature-level similarity and develop a more robust interaction mechanism.

3.2 Intra-pair Similarity Uncertainty

Uncertainty Theory. The Bayesian Neural Network (BNN) [Lakshminarayanan *et al.*, 2017] incorporates Bayesian statistics into traditional neural networks, producing an uncertainty distribution instead of a single value. Orthogonally to BNN, Evidential Deep Learning (EDL) [Sensoy *et al.*, 2018] uses explicit modeling based on Subjective Logic (SL) to capture prediction uncertainty by associating the belief distribution with parameters of the Dirichlet distribution. Specifically, EDL provides an evidential theoretical framework to quantify the belief masses $b = (b_1, b_2, \ldots, b_B)$ of the class probabilities $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_B)$ in a B-class classification problem, and the overall uncertainty mass u follows Eq. (2) :

$$u + \sum_{i=1}^{B} b_i = 1, \quad u \geq 0, \quad b_i \geq 0. \tag{2}$$

Let $e_i \geq 0$ denote the evidence derived from $\hat{y}_i$, *e.g.*, non-negative operation $e_i = \exp(\hat{y}_i)$. Then, the belief b_i and the uncertainty mass u are computed as:

$$b_i = \frac{e_i}{\sum_{i=1}^{B} \alpha_i}, \quad u = \frac{B}{\sum_{i=1}^{B} \alpha_i}, \tag{3}$$

where $\alpha_i = e_i + 1$ and $S = \sum_{i=1}^{B} \alpha_i$ denote the Dirichlet distribution parameters and strength, respectively. The evidence e_i serves as a measure of the support gathered from the data, indicating the likelihood of a sample being classified into a particular class. In brief, a higher u and lower e_i in target classification indicate greater ambiguity and lower confidence in the outcome. Uncertainty theory demonstrates strong robustness in classification tasks, providing uncertainty in addition to classification probabilities. However, effectively generalizing it to retrieval tasks poses many challenges, such as handling more complex cross-modal features. Therefore, it is essential to understand the differences and connections between the two in order to effectively model uncertainty in retrieval.

Similarity Uncertainty Modeling. For a B-class classification task, the probability of classifying x_i into the i^{th} class is $\hat{y}_i$, with the label $y_i = 1$. For a B-size retrieval task, the similarity between the text t_i and the video feature v_j is s_{ij}, with no label y_{ij}. In fact, if we provide label support and reduce the dimension of the similarity matrix, retrieval tasks can be transformed into classification tasks. Therefore, we propose an **I**ntra-pair **S**imilarity **U**ncertainty **M**odule **(ISUM)** to guide the model in learning uncertainty from text-video modality interactions in Figure 2 (2). Intuitively, we expect the similarity matrix $S^{B\times B}$ output by the model to be as close as possible to the identity matrix $E^{B\times B}$. At this point, we first obtain the Dirichlet distribution $\mathrm{Dir}(\alpha_{ij})$, where $\alpha_{ij} = e_{ij} + 1 = \mathrm{ReLU}(S_{ij}) + 1$, and the distribution strength $S_i = \sum_{j=1}^{B} \alpha_{ij}$. Given the known labels $y_{ij} \sim E^{B\times B}$ and the Dirichlet distribution $p_{ij} \sim \mathrm{Dir}(\alpha_{ij})$, we use a generalized Mean-Squared Error loss to supervise the model in learning how to generate a similarity matrix that approaches the identity matrix:

$$\mathbb{E}_{p_{ij}\sim \mathrm{Dir}(\alpha_{ij})} ||y_{ij} - p_{ij}||_2^2. \tag{4}$$

For the specific i^{th} text to retrieve all videos in the entire batch, we can derive:

$$\begin{aligned} \mathcal{L}_i^U &= \sum_{j=1}^{B} \mathbb{E}\left[y_{ij}^2 - 2y_{ij}p_{ij} + p_{ij}^2\right] \\ &= \sum_{j=1}^{B} \left(y_{ij}^2 - 2y_{ij}\mathbb{E}[p_{ij}] + \mathbb{E}[p_{ij}^2]\right), \end{aligned} \tag{5}$$

where $\mathbb{E}[p_{ij}] = \alpha_{ij}/S_i$, $\mathbb{E}[p_{ij}^2] = \mathrm{Var}[p_{ij}] + \mathbb{E}^2[p_{ij}]$, and $\mathrm{Var}[p_{ij}] = \frac{\alpha_{ij}(S_i - \alpha_{ij})}{S_i^2(S_i+1)}$. Letting $\hat{p}_{ij} = \mathbb{E}[p_{ij}] = \alpha_{ij}/S_i$, Eq. (5) can then be further simplified to:

$$\mathcal{L}_i^U = \sum_{j=1}^{B} (y_{ij} - \hat{p}_{ij})^2 + \frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{S_i + 1}. \tag{6}$$

Similarly, for the video $\mathcal{L}_j^U$ can also be computed, and the total evidence loss function is given by Eq. (7):

$$\mathcal{L}_S^U = \frac{1}{B}\left(\sum_{i=1}^{B} \mathcal{L}_i^U + \sum_{j=1}^{B} \mathcal{L}_j^U\right). \tag{7}$$

It is important to note that we elevate the metric for measuring retrieval pair relationships from similarity s to confidence e, and cleverly integrate the identity matrix E to provide a theoretical foundation for building reliable pairings.

3.3 Inter-pair Distance Uncertainty

Uncertainty Probabilistic Construction. Retrieval relationships based solely on feature-level similarity are often unreliable due to false positives (highly similar candidate data) in inter-pair exclusion. To mitigate semantic interference within the same modality, PCME [Chun *et al.*, 2021] constructs probability distributions from text and images, enabling diverse embeddings and overcoming single-feature limitations. However, PCME focus on individual images and fail to address complex video scenarios. Therefore, we propose a Local Feature Aggregation Module to generate aggregated embeddings for text and video in Figure 2 (top). This module includes local feature aggregation via multi-head attention and global feature fusion via concatenation; the former captures fine-grained local details, while the latter provides contextual support for local features. However, the aggregated embeddings (z_t, z_v) obtained through this module do not show significant differences compared to traditional features. Therefore, it is necessary to re-represent the aggregated embeddings to capture the diversity of probabilistic embeddings.

$$\begin{aligned} p(z \mid t) &\sim \mathcal{N}\big(h_{\mathcal{T}}^{\mu}(z_t), e^{h_{\mathcal{T}}^{\sigma}(z_t)}\big), \\ p(z \mid v) &\sim \mathcal{N}\big(h_{\mathcal{V}}^{\mu}(z_v), e^{h_{\mathcal{V}}^{\sigma}(z_v)}\big), \end{aligned} \tag{8}$$

where h_*^{μ} and h_*^{σ} denote linear operations, and $p(z \mid *)$ denotes a Gaussian distribution. To enable stable training and facilitate sampling, we introduce a standard Gaussian distribution $\epsilon_k \sim \mathcal{N}(0, I)$ to generate probabilistic embedding z_t^k:

$$\begin{aligned} z_t^k &= h_{\mathcal{T}}^{\mu}(z_t) + e^{h_{\mathcal{T}}^{\sigma}(z_t)} \cdot \epsilon_k, \\ z_v^k &= h_{\mathcal{V}}^{\mu}(z_v) + e^{h_{\mathcal{V}}^{\sigma}(z_v)} \cdot \epsilon_k, \end{aligned} \tag{9}$$

where $k = 1, 2, \ldots, K$, $z_*^k \in \mathbb{R}^{1\times D}$, and $z_* \in \mathbb{R}^{K\times D}$.

Uncertainty Probabilistic Distance. Most previous methods adopt Euclidean distance [Oh *et al.*, 2018] and Multi-Instance InfoNCE loss [Chun *et al.*, 2021] to achieve effective alignment between single probabilistic embedding. However, these methods are often constrained by computational efficiency and embedding selection when handling multiple probabilistic embeddings ($K > 1$). To address this issue and support inter-pair exclusion, we propose an **I**nter-pair **D**istance **U**ncertainty **M**odule (**IDUM**), which elegantly resolves the challenge of aligning probabilistic embeddings across multiple instances in Figure 2 (3). Intuitively, to achieve effective inter-pair exclusion, we only need to focus on specific boundary information, namely the maximum distance between inter-pair probabilistic embeddings. This approach not only reduces the number of probabilistic embeddings involved in alignment computation, thereby lowering the computational cost, but also improves the model's sensitivity to embedding boundaries. Specifically, we define the distance between probabilistic embeddings as:

$$d(z_t^{i,m}, z_v^{j,n}) = 1 - s(z_t^{i,m}, z_v^{j,n}), \tag{10}$$

where $i, j = 1, 2, \ldots, B$, $m, n = 1, 2, \ldots, K$, and $z_t^{i,m}$ and $z_v^{j,n}$ represent the m^{th} probabilistic embedding of the i^{th} text feature and the n^{th} probabilistic embedding of the j^{th} video feature, respectively. Based on effective boundary selection, we choose the maximum value for inter-pair exclusion ($i \neq j$) and the minimum value for intra-pair interaction ($i = j$):

$$d(z_t^i, z_v^j) = \begin{cases} \max\ d(z_t^{i,m}, z_v^{j,n}), & \forall n, m,\ i \neq j, \\ \min\ d(z_t^{i,m}, z_v^{j,n}), & \forall n, m,\ i = j. \end{cases} \tag{11}$$

Similar to the feature similarity loss in Eq. (1), we can directly obtain the probabilistic embedding distance loss:

$$\begin{aligned} \mathcal{L}_D = \frac{1}{2}\Bigg(& \frac{1}{B}\sum_{i=1}^{B} \log \frac{e^{d(z_t^i, z_v^i)\cdot\lambda}}{\sum_{j=1}^{B} e^{d(z_t^i, z_v^j)\cdot\lambda}} \\ & + \frac{1}{B}\sum_{i=1}^{B} \log \frac{e^{d(z_t^i, z_v^i)\cdot\lambda}}{\sum_{j=1}^{B} e^{d(z_t^j, z_v^i)\cdot\lambda}} \Bigg). \end{aligned} \tag{12}$$

Additionally, the distance matrix D in Eq. (11) between probabilistic embeddings can directly adopt the method used for processing the similarity matrix S in Sec. 3.2, with the labels $y_{ij} \sim (1 - \mathbb{E}^{B\times B})$:

$$\mathcal{L}_D^U = \frac{1}{B}\Bigg(\sum_{i=1}^{B} \mathcal{L}_i^U + \sum_{j=1}^{B} \mathcal{L}_j^U \Bigg). \tag{13}$$

3.4 Training and Inference

Training. Following [Oh *et al.*, 2018], we also introduce an additional KL divergence loss between the probabilistic embeddings to prevent them from collapsing to zero:

$$\mathcal{L}_{KL} = KL(p(z|t)||\mathcal{N}(0, I)) + KL(p(z|v)||\mathcal{N}(0, I)). \tag{14}$$

Therefore, based on feature similarity (Eq. (1) and (7)) and probabilistic distance (Eq. (12) and (13)) losses, the total training objective can be defined as:

$$\mathcal{L}_{total} = \mathcal{L}_S + \mathcal{L}_S^U + \alpha(\mathcal{L}_D + \mathcal{L}_D^U) + \beta\mathcal{L}_{KL}, \tag{15}$$

where α and β are weight hyper-parameters.

Inference. After training, in addition to using the regular feature similarity s in Eq. (1) for sampling, the similarity uncertainty u_s in Eq. (3), the probabilistic distance d in Eq. (11), and the distance uncertainty u_d in Eq. (3) can also be used to update the similarity calculation:

$$s' = \exp(-\gamma_1 u_d) \cdot (1 - d)) \circ (\exp(-\gamma_2 u_s) \cdot s, \tag{16}$$

where γ_1 and γ_2 are hyper-parameters used to control the impact of re-ranking on prediction.

4 Experiments

4.1 Experimental Setting

Datasets. We adopt six benchmark datasets for evaluation: **(1) MSRVTT** [Xu *et al.*, 2016] consists of 10K videos, each paired with 20 captions. We follow the training and testing splits from [Yu *et al.*, 2018]. **(2) DiDeMo** [Anne Hendricks *et al.*, 2017] contains 10,642 video clips and 40,543 captions. We use the training and testing protocols from [Gabeur *et al.*, 2020]. **(3) LSMDC** [Rohrbach *et al.*, 2015] includes 118,081 video clips from 202 movies. We use the split from [Torabi *et al.*, 2016], with 1,000 videos reserved for testing. **(4) MSVD** [Liu *et al.*, 2019] includes 1,970 videos and over 80K captions, with training, validation, and test sets containing 1,200, 100, and 670 videos, respectively. **(5) Charades** [Sigurdsson *et al.*, 2016] consists of 9,848 video clips, and we adopt the split protocol from [Lin *et al.*, 2022]. **(6) VATEX** [Wang *et al.*, 2019] contains 34K video clips. We follow the train-test split from [Chen *et al.*, 2020].

Metrics and Implementation. We report R@k (k = 1, 5, 10), median rank (MdR), and mean rank (MnR) as evaluation metrics, where higher R@k and lower MdR/MnR indicate better performance. Following previous methods [Gorti *et al.*, 2022], we use CLIP as our backbone model. The batch size is set to 32, and the model is trained for 5 epochs across different datasets. We sample an average of $F = 12$ frames from each video clip, resizing them to 224×224 pixels for all datasets. The hyper-parameters are set as $\alpha = 1 \times 10^{-1}$, $\beta = 1 \times 10^{-4}$, $\gamma_1 = \gamma_2 = 1 \times 10^{-1}$, and the number of probabilistic embeddings is $K = 7$.

4.2 Comparisons with State-of-the-art Methods

Table 1 provides the retrieval results on the MSRVTT, and DUQ outperforms recently proposed SOTA methods on both text-to-video (+0.6% in R@1) and video-to-text (+1.4% in R@1) retrieval tasks. Meanwhile, Table 2 provides text-to-video retrieval results on other datasets. DUQ demonstrates consistent improvements across multiple datasets, including the long-video DiDeMo (+10.9% in R@1) and the sparse-text LSMDC (+11.3% in R@1), highlighting the effectiveness of our method. See Appendix A for more results.

4.3 Ablation Study

To evaluate the impact of each component on the DUQ model, we conduct an ablation study in Table 3: **(1) Baseline**: We adopt the conventional feature similarity loss as the baseline. **(2) Similarity Uncertainty**: Comparing #1 and #2,

Methods	MSRVTT (Text-to-Video)					MSRVTT (Video-to-Text)				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
CLIP-Non										
CE [Liu *et al.*, 2019]	20.9	48.8	62.4	6.0	28.2	20.6	50.3	64.0	5.3	25.1
MMT [Gabeur *et al.*, 2020]	26.6	57.1	69.6	4.0	24.0	27.0	57.5	69.7	3.7	21.3
SST [Patrick *et al.*, 2020]	27.4	56.3	67.7	3.0	-	26.6	55.1	67.5	3.0	-
CLIP-ViT-B/32										
EMCL-Net [Jin *et al.*, 2022]	46.8	73.1	83.1	2.0	-	46.5	73.5	83.5	2.0	-
X-Pool [Gorti *et al.*, 2022]	46.9	72.8	82.2	2.0	14.3	44.4	73.3	84.0	2.0	9.0
DiCoSA [Jin *et al.*, 2023b]	47.5	74.7	83.8	2.0	13.2	46.7	75.2	84.3	2.0	8.9
UATVR [Fang *et al.*, 2023]	47.5	73.9	83.5	2.0	12.3	46.9	73.8	83.8	2.0	8.6
HBI [Jin *et al.*, 2023a]	48.6	74.6	83.4	2.0	12.0	46.8	74.3	84.3	2.0	8.9
Cap4Video [Wu *et al.*, 2023]	49.3	74.3	83.8	2.0	12.0	47.1	73.7	84.3	2.0	8.7
T-Mass [Wang *et al.*, 2024]	50.2	75.3	85.1	**1.0**	11.9	47.7	78.0	86.3	2.0	8.0
DUQ (Ours)	**51.2**	**77.3**	**86.1**	**1.0**	**10.8**	**50.4**	**79.2**	**87.5**	**1.0**	**6.4**
CLIP-ViT-B/16										
X-Pool [Gorti *et al.*, 2022]	48.2	73.7	82.6	2.0	12.7	46.4	73.9	84.1	2.0	8.4
UATVR [Fang *et al.*, 2023]	50.8	76.3	85.5	**1.0**	12.4	48.1	76.3	85.4	2.0	8.0
Cap4Video [Wu *et al.*, 2023]	51.4	75.7	83.9	**1.0**	12.4	49.0	75.2	85.0	2.0	8.0
T-Mass [Wang *et al.*, 2024]	52.7	77.1	85.6	**1.0**	10.5	50.9	80.2	88.0	**1.0**	7.4
DUQ (Ours)	**55.9**	**81.0**	**88.6**	**1.0**	**8.4**	**54.6**	**82.4**	**89.9**	**1.0**	**5.3**

Table 1: Retrieval performance on the MSRVTT-1K dataset. "↑" means that higher is better. "↓" means that lower is better.

Methods	DiDeMo (Text-to-Video)					LSMDC (Text-to-Video)				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
CE [Liu *et al.*, 2019]	16.1	41.1	-	8.3	43.7	11.2	26.9	34.8	25.3	-
EMCL-Net [Jin *et al.*, 2022]	45.3	74.2	82.3	2.0	12.3	23.9	46.4	53.7	8.0	-
TS2-Net [Liu *et al.*, 2022]	41.8	71.6	82.0	2.0	14.8	23.4	42.3	50.9	9.0	56.9
X-Pool [Gorti *et al.*, 2022]	44.6	73.2	82.0	2.0	15.4	25.2	43.7	53.5	8.0	53.2
CLIP-VIP [Xue *et al.*, 2022]	48.6	77.1	84.4	2.0	-	25.6	45.3	54.4	8.0	-
DiCoSA [Jin *et al.*, 2023b]	45.7	74.6	83.5	2.0	11.7	25.4	43.6	54.0	8.0	41.9
DiffusionRet [Jin *et al.*, 2023c]	46.7	74.7	82.7	2.0	14.3	24.4	43.1	54.3	8.0	40.7
UATVR [Fang *et al.*, 2023]	43.1	71.8	82.3	2.0	15.1	-	-	-	-	-
DUQ (Ours)	**51.8**	**77.9**	**86.5**	**1.0**	**10.6**	**28.5**	**48.2**	**58.0**	**6.0**	**41.2**
Methods	Charades (Text-to-Video)					VATEX (Text-to-Video)				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Support-Set [Patrick *et al.*, 2020]	-	-	-	-	-	44.9	82.1	89.7	**1.0**	-
CLIP4Clip [Luo *et al.*, 2022]	9.9	27.1	36.8	21.0	85.4	-	-	-	-	-
X-Pool [Gorti *et al.*, 2022]	11.2	28.3	38.8	20.0	82.7	60.0	90.0	95.0	**1.0**	3.8
UATVR [Fang *et al.*, 2023]	-	-	-	-	-	61.3	91.0	95.6	**1.0**	3.3
T-Mass [Wang *et al.*, 2024]	14.2	36.2	48.3	12.0	54.8	63.0	92.3	96.4	**1.0**	3.2
DUQ (Ours)	**28.5**	**55.0**	**66.8**	**4.0**	**22.9**	**80.0**	**97.4**	**99.0**	**1.0**	**1.6**

Table 2: Retrieval performance on other datasets. "↑" means that higher is better. "↓" means that lower is better.

the latter additionally incorporates the intra-similarity uncertainty module, leading to a noticeable improvement in retrieval performance (+1.3% in R@1), demonstrating the effectiveness of modeling intra-similarity uncertainty. **(3) Distance Metric**: Comparing #1 and #3 (or #2 and #4), the latter addresses the limitations of single-feature representations by constructing distance-based diversity probabilistic embeddings. This approach increases inter-modal differences and improves retrieval performance (+2.8% in R@1). **(4) Distance Uncertainty**: Comparing #4 and #6, the latter directly integrates the similarity uncertainty from (2), transforming it into distance-based uncertainty. This integration similarly enhances the model's retrieval performance (+2.9% in R@1). See Appendix B for more ablation studies.

4.4 Out-of Domain Tasks

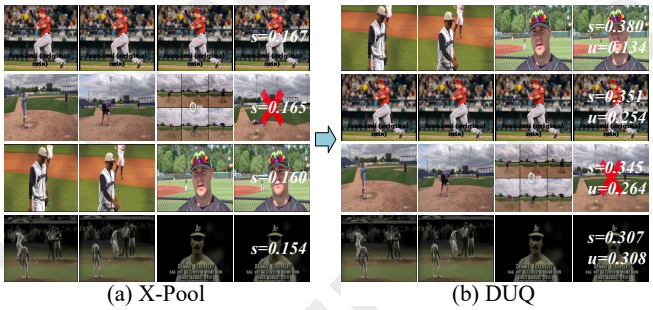
To evaluate the model's generalization on unseen data, we compare the out-of-domain text-to-video retrieval in Table

Baseline	DUQ			MSRVTT (Text-to-Video)					MSRVTT (Video-to-Text)				
$N/\mathcal{L}_S$	$\mathcal{L}_S^U$	$\mathcal{L}_D$	$\mathcal{L}_D^U$	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
1/✓				46.9	74.5	82.2	2.0	14.3	44.4	73.3	84.0	2.0	9.0
2/✓	✓			47.5	73.9	83.5	2.0	12.3	46.9	73.8	83.8	2.0	8.6
3/✓		✓		48.2	74.6	83.4	2.0	12.2	45.2	73.7	84.8	2.0	8.6
4/✓	✓	✓		49.2	**77.6**	85.4	2.0	11.2	49.2	77.7	84.8	2.0	8.1
5/✓	✓		✓	48.9	76.3	84.6	2.0	12.6	48.5	74.5	84.3	2.0	8.1
6/✓	✓	✓	✓	**51.2**	77.3	**86.1**	**1.0**	**10.8**	**50.4**	**79.2**	**87.5**	**1.0**	**6.4**

Table 3: Ablation study of different Components. We use the symbol $\mathcal{L}_X^Y$ to represent the corresponding components.

Methods	MSRVTT			MSRVTT>DiDeMo			MSRVTT>LSMDC		
	R@1↑	R@Sum↑	MdR↓	R@1↑	R@Sum↑	MdR↓	R@1↑	R@Sum↑	MdR↓
CLIP4Clip [Luo *et al.*, 2022]	43.8	195.8	2.0	31.8	154.9	4.0	15.3	87.1	21.0
X-Pool [Gorti *et al.*, 2022]	46.9	201.9	2.0	35.3	168.5	3.0	16.4	93.5	18.0
DiffusionRet [Jin *et al.*, 2023c]	49.0	206.9	2.0	33.2	160.9	3.0	17.1	90.5	21.0
T-Mass [Wang *et al.*, 2024]	50.2	210.6	**1.0**	39.5	178.2	**2.0**	19.7	102.5	14.0
DUQ (Ours)	**51.2**	**214.6**	**1.0**	**43.0**	**188.6**	**2.0**	**21.4**	**110.0**	**11.0**

Table 4: Out-of-domain text-to-video retrieval performance. $X > Y$, where X denotes the training data and Y denotes the test data.



(a) X-Pool (b) DUQ

Figure 3: Visualization of re-ranked retrieval results.

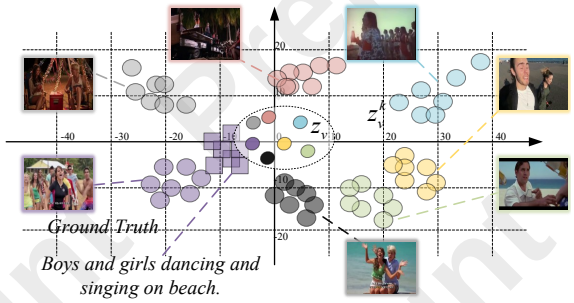


Figure 4: Visualization of probabilistic embeddings.

4. We first train on the "source" dataset MSRVTT and then test on the "target" datasets DiDeMo and LSMDC. We observe that models achieving strong in-domain performance often experience significant drops when generalized to out-of-domain data. Compared to other methods, our approach demonstrates excellent retrieval performance both in-domain and out-of-domain. Additionally, we provide the performance comparison for video question answering in Table 5.

Methods	Accuracy(%)↑
VQA-T [Yang *et al.*, 2021]	41.5
MERLOT [Zeng *et al.*, 2022]	43.1
Co-Tokenization [Piergiovanni *et al.*, 2022]	45.7
EMCL-QA [Jin *et al.*, 2022]	45.8
HBI [Jin *et al.*, 2023a]	46.2
TG-VQA [Li *et al.*, 2023]	46.3
DUQ (Ours)	**47.8**

Table 5: Video question answering performance.

4.5 Visualization Results

Figure 3 illustrates a failed retrieval case discussed in the introduction. By incorporating the similarity uncertainty module into feature interactions, the weak semantic interaction capability under sparse queries is effectively enhanced. Figure 4 visualizes the distances between probabilistic embeddings. The central circle represents the aggregated embedding, while the others represent probabilistic embeddings.

5 Conclusion

In this paper, we propose a novel Dual Uncertainty Quantification (DUQ) framework to provide trustworthy predictions by quantifying the confidence of intra-pair interactions and inter-pair exclusions induced by data uncertainty. Accurate text-video retrieval requires reliable intra-pair associations and inter-pair exclusions to ensure the model effectively captures semantic consistency between text and video. Extensive experiments on six benchmark datasets demonstrate that our approach achieves new SOTA retrieval performance. We hope this work inspires further applications in other domains.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) (No: 72471197), the Natural Science Foundation of Xinjiang Uygur Autonomous Region (No: 2024D01A18), the Natural Sciences and Engineering Research Council of Canada, the Sichuan Provincial Philosophy and Social Science Fund General Project(No: SCJJ24ND133), the Sichuan Provincial International Science and Technology Innovation Project (No: 25GJHZ0147) and the Sichuan Science and Technology Program (No: MZGC20240146).

References

[Anne Hendricks *et al.*, 2017] 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.

[Chen *et al.*, 2020] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10638–10647, 2020.

[Chen *et al.*, 2024] Lei Chen, Zhen Deng, Libo Liu, and Shibai Yin. Multilevel semantic interaction alignment for video–text cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[Chun *et al.*, 2021] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424, 2021.

[Fang *et al.*, 2023] Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Yuxin Song, Weiping Wang, Xiangbo Shu, Xiangyang Ji, and Jingdong Wang. Uatvr: Uncertainty-adaptive text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13723–13733, 2023.

[Gabeur *et al.*, 2020] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 214–229. Springer, 2020.

[Gorti *et al.*, 2022] 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.

[Jin *et al.*, 2022] Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David Clifton, and Jie Chen. Expectation-maximization contrastive learning for compact video-and-language representations. *Advances in neural information processing systems*, 35:30291–30306, 2022.

[Jin *et al.*, 2023a] 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.

[Jin *et al.*, 2023b] 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. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 938–946. International Joint Conferences on Artificial Intelligence Organization, 2023.

[Jin *et al.*, 2023c] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Xiangyang Ji, Chang Liu, Li Yuan, and Jie Chen. Diffusionret: Generative text-video retrieval with diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2470–2481, 2023.

[Lakshminarayanan *et al.*, 2017] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[Li *et al.*, 2023] Hao Li, Peng Jin, Zesen Cheng, Songyang Zhang, Kai Chen, Zhennan Wang, Chang Liu, and Jie Chen. Tg-vqa: ternary game of video question answering. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 1044–1052, 2023.

[Li *et al.*, 2024] Hao Li, Jingkuan Song, Lianli Gao, Xiaosu Zhu, and Hengtao Shen. Prototype-based aleatoric uncertainty quantification for cross-modal retrieval. *Advances in Neural Information Processing Systems*, 36, 2024.

[Lin *et al.*, 2022] Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. Eclipse: Efficient long-range video retrieval using sight and sound. In *European Conference on Computer Vision*, pages 413–430. Springer, 2022.

[Liu *et al.*, 2019] 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.

[Liu *et al.*, 2022] 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.

[Luo *et al.*, 2022] 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.

[Oh *et al.*, 2018] Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew Gallagher. Modeling uncertainty with hedged instance embedding. *arXiv preprint arXiv:1810.00319*, 2018.

[Patrick *et al.*, 2020] 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.

[Piergiovanni *et al.*, 2022] AJ Piergiovanni, Kairo Morton, Weicheng Kuo, Michael S Ryoo, and Anelia Angelova. Video question answering with iterative video-text co-tokenization. In *European Conference on Computer Vision*, pages 76–94. Springer, 2022.

[Radford *et al.*, 2021] 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.

[Rohrbach *et al.*, 2015] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015.

[Sensoy *et al.*, 2018] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[Sigurdsson *et al.*, 2016] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016.

[Song and Soleymani, 2019] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019.

[Tang *et al.*, 2025] Haoran Tang, Meng Cao, Jinfa Huang, Ruyang Liu, Peng Jin, Ge Li, and Xiaodan Liang. Muse: Mamba is efficient multi-scale learner for text-video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7238–7246, 2025.

[Torabi *et al.*, 2016] Atousa Torabi, Niket Tandon, and Leon Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv:1609.08124*, 2016.

[Wang *et al.*, 2019] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4581–4591, 2019.

[Wang *et al.*, 2023] Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Unified coarse-to-fine alignment for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2816–2827, 2023.

[Wang *et al.*, 2024] Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Text is mass: Modeling as stochastic embedding for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16551–16560, 2024.

[Wu *et al.*, 2023] Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4video: What can auxiliary captions do for text-video retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10704–10713, 2023.

[Xu *et al.*, 2016] 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.

[Xue *et al.*, 2022] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022.

[Yang *et al.*, 2021] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1697, 2021.

[Yu *et al.*, 2018] 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.

[Zeng *et al.*, 2022] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11101–11111, 2022.