SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

March 2026, Vol. 69, Iss. 3, 132107:1–132107:15
https://doi.org/10.1007/s11432-024-4696-2

UML: uncertainty-aware and mutual learning for noise-robust cross-lingual cross-modal retrieval

Yu LIU[1], Haipeng CHEN[1], Xun YANG[2*], Yingda LYU[3*] & Meng WANG[4,5]

[1]*College of Computer Science and Technology, Jilin University, Changchun 130012, China*
[2]*School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China*
[3]*Public Computer Education and Research Center, Jilin University, Changchun 130012, China*
[4]*School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China*
[5]*Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China*

Received 31 October 2024/Revised 7 April 2025/Accepted 10 July 2025/Published online 14 January 2026

Abstract Cross-lingual cross-modal retrieval (CCR) has recently emerged as a significant research area, focusing on aligning visual content with non-English captions without relying on human-annotated non-English cross-modal data pairs. Most CCR methods extend existing English-only datasets with other languages via machine translation (MT) to establish correspondence between vision and non-English. Regrettably, these cheaply collected datasets inevitably contain numerous mismatched vision and non-English data pairs, a.k.a noisy correspondence (NC). The presence of NC renders the supervision information unreliable, leading to a significant decline in retrieval performance. Furthermore, most existing methods attempt to improve alignment between visual and non-English representations by combining information from multiple views. However, these approaches often overlook the need for consistency across these views, capturing view-specific and task-irrelevant information, which exacerbates bias in the optimization direction. To address the issues, we propose an uncertainty-aware and mutual learning (UML) framework, which integrates a novel dual-view uncertainty-aware learning (DUL) paradigm and an efficient adaptive mutual learning (AML) loss. The DUL effectively models alignment uncertainty to assess and mitigate the effects of NC. Specifically, it employs evidential deep learning to obtain accurate cross-modal alignment uncertainty, which is then combined with labels softened by Fisher information to impose appropriate penalties for retrieval. To mitigate the exacerbation problem, we derive the AML loss, which aims to ensure effective aggregation between all modalities of a clean pair, while effectively separating the non-English representation of a noisy pair from its visual and English representations. Our UML consistently outperforms previous methods in supervised, domain generalization, and robustness settings across three challenging benchmarks.

Keywords cross-lingual cross-modal retrieval, noisy correspondence, uncertainty-based learning, mutual information

Citation Liu Y, Chen H P, Yang X, et al. UML: uncertainty-aware and mutual learning for noise-robust cross-lingual cross-modal retrieval. Sci China Inf Sci, 2026, 69(3): 132107, https://doi.org/10.1007/s11432-024-4696-2

1 Introduction

The world exhibits diversity due to its multimodal and multilingual attributes. Although multimodal research has made significant progress with the introduction of vision-language pre-training, most existing studies [1, 2] remain heavily focused on English. This is mainly due to the lack of human-annotated non-English captions. Consequently, there has been a growing interest in more general cross-lingual cross-modal retrieval (CCR), which seeks to identify visual content relevant to non-English queries without relying on human-annotated non-English cross-modal data pairs. In contrast to traditional cross-modal retrieval [3, 4], CCR boasts multilingual retrieval capabilities.

Recently, most research [5–9] in CCR employs machine translation (MT) techniques to generate pseudo-parallel data pairs, as depicted in Figure 1(a). Specifically, CCLM [5] and UC^2 [6] endeavor to achieve cross-modal alignment by developing large-scale vision-language datasets (e.g., CC3M [10]), along with the formulation of pertinent pre-training objectives. CL2CM [8] improves the alignment between vision and non-English by transferring knowledge from the cross-lingual network to the cross-modal network. Unfortunately, as shown in Figure 1(a), even with the most advanced MT technology, the translation of non-English captions may still contain various forms of noise, including spelling and grammatical errors, and even alterations in original semantics. Such noise can lead to a mismatch between vision and non-English, known as noisy correspondence [11], ultimately resulting in suboptimal performance.

* Corresponding author (email: xyang21@ustc.edu.cn, ydlv@jlu.edu.cn)

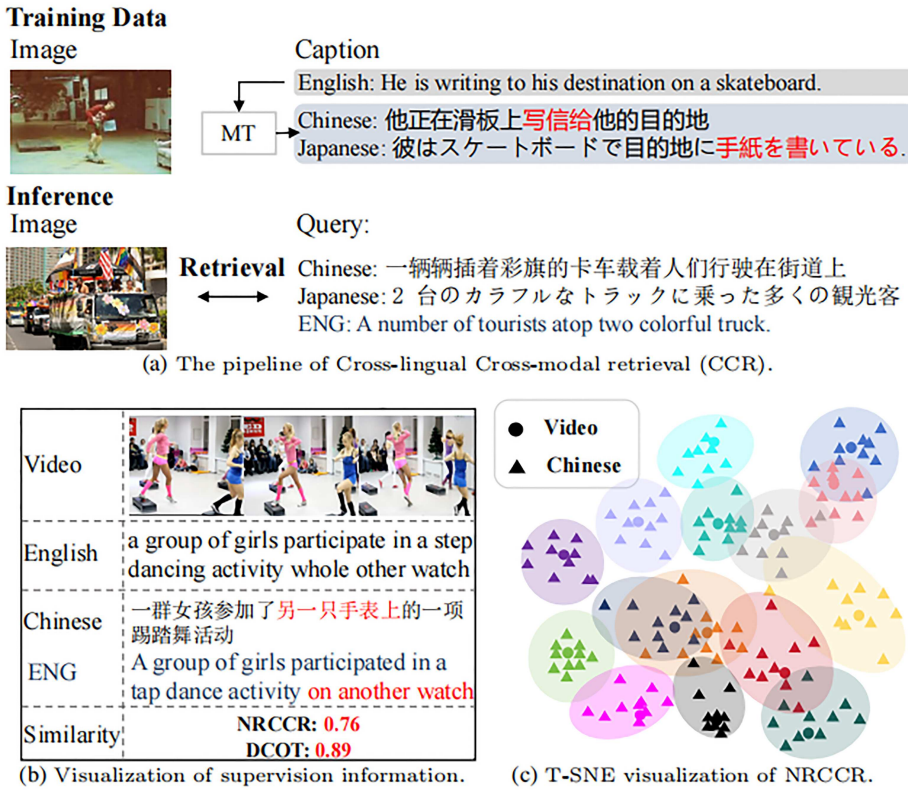

Figure 1 (Color online) (a) CCR methods are trained with noisy non-English captions caused by MT. During inference, human-labeled non-English captions are utilized to retrieve relevant visual content. Drawbacks of existing methods are as follows: (b) unreliable supervision information of NRCCR [7] and DCOT [9], and (c) scattered intra-instance representation of NRCCR [7]. The same color indicates the same instance.

To date, in the field of CCR, only a few efforts [7,9] have been made to tackle the noisy correspondence problem caused by MT. NRCCR [7] is the pioneering work to investigate this problem, which aims to learn the correct association between visual and non-English representations. Specifically, NRCCR adopts multi-view self-distillation to generate soft pseudo objects to learn noise-robust non-English representations. Additionally, NRCCR aligns visual and non-English features via a cyclic semantic consistency module and adversarial learning. Later, DCOT [9] formulates the noisy correspondence learning in CCR as an optimal transport problem to avoid overfitting to noisy pairs. However, these methods may produce unreliable supervision information because: (1) the filtered non-English representation may lose useful information; (2) DCOT relies on cumbersome parameter design. For instance, in Figure 1(b), the video and the translated Chinese query are obviously mismatched; unfortunately, NRCCR and DCOT yield high similarity scores, leading to incorrect alignment. Moreover, most CCR approaches [6–9] independently optimize the representation of different views (e.g., cross-lingual or cross-modal views), which undoubtedly introduces view-specific and task-irrelevant information. This will increase the semantic gap and amplify the optimization direction bias caused by unreliable supervision. As shown in Figure 1(c), we employ t-SNE to visualise the representations of 15 randomly selected videos and their corresponding 10 Chinese captions from the VATEX test set. It can be observed that the amplified optimization bias problem leads to very scattered representations of different modalities for the same instance. These observations and insights raise a critical research question that motivates this study: "How can we obtain reliable supervision and comprehensive semantic information to facilitate noise-robust CCR?"

To answer this, we present a novel framework called uncertainty-aware and mutual learning (UML) for noise-robust CCR. In particular, we adopt an innovative dual-view uncertainty-aware learning (DUL) paradigm to model and consider the alignment uncertainty brought by noisy correspondence, thus providing reliable supervision. Specifically, DUL first parameterizes the bidirectional evidence from two views (i.e., vision-English and vision-non-English) into a Dirichlet distribution based on cross-modal similarity to obtain alignment uncertainty. Then, we dynamically identify clean and noisy pairs based on the learned evidence, and combine the uncertainty with the label softened by Fisher information to impose appropriate penalties for both clean and noisy pairs. Additionally, to mitigate the exacerbation problem, we propose an adaptive mutual learning (AML) loss. Concretely, AML maximizes the

mutual information between multi-view representations of a clean pair, allowing the model to retain the necessary information while eliminating irrelevant distractors. For a noisy pair, AML minimizes the mutual information between its non-English representation and its visual and English representations to avoid overfitting noise. Finally, we integrate the advantages of DUL and AML to enhance the accuracy and robustness of the model. Our primary contributions are summarized as follows.

• We devise a novel UML framework to address a pressing and pervasive noisy correspondence caused by MT for cross-lingual cross-modal retrieval.

• A novel DUL strategy is proposed to model and consider alignment uncertainty brought by noise, which effectively improves the robustness and reliability of the model. To the best of our knowledge, our DUL is the first method that endows evidential deep learning and Fisher information with cross-lingual cross-modal retrieval.

• We design an AML loss to learn comprehensive and reliable intra-instance representations of clean pairs and mitigating noise fitting.

• Extensive experiments demonstrate the superiority and robustness of our UML on three widely-used cross-lingual cross-modal benchmark datasets.

2 Related work

2.1 Cross-lingual cross-modal retrieval

The CCR extends traditional cross-modal retrieval [3,4,12,13] to the multilingual domain, aiming to address the issue of unavailable manually labeled non-English data. Most existing approaches [5–9, 14, 15] align visual information and multilingual text by mapping them into a common semantic space. Earlier approaches [14, 16] aim to collect multilingual parallel corpora. For instance, M^{3}P [14] proposes a pre-trained model that integrates multilingual, multimodal, and multi-task learning, aiming to build a unified framework that can align visual contents with multiple languages. MMP [16] extends HowTo100M [17] into a multilingual version to enable zero-shot cross-lingual transfer of vision-language models. In addition, some methods [5,6,15] propose new optimization objectives and pre-training tasks to enable models to capture better alignment between vision and language. For example, CCR^k [15] proposes the 1-to-K contrast learning paradigm, which improves the consistency of the retrieval model. UC^2 [6] is the first to propose a pre-training model based on machine translation enhancement, which learns cross-lingual cross-modal representation by focusing primarily on images and complementing with English. Although these methods achieve promising performance, they implicitly assume that all cross-modal pairs are correctly aligned within the training data. In fact, due to the high cost of collection and annotation as well as the noise introduced by MT, collecting extremely clean large-scale data is expensive or even impossible. Even the most advanced MT tools still inevitably include spelling mistakes, grammatical errors, and altered original semantics (see Figure 1(a)). Therefore, it is important to explore robust CCR with noisy correspondence [11].

2.2 Learning with noisy correspondence

Noisy correspondence [11] refers to the situation where data pairs are incorrectly assumed to be correctly aligned despite being semantically mismatched. Noisy correspondence learning aims to mitigate the effects of this mis-correspondence and produce robust representations. It has garnered significant interest, focusing on several key areas: cross-modal retrieval [11, 18], person reidentification [19], multi-view clustering [20, 21], and so forth. These methods primarily follow a similarity-guided multi-step framework. Initially, they estimate the distribution of instance-level loss/similarity in the entire dataset. Subsequently, they select clean samples for training. To the best of our knowledge, only a few efforts [7–9] have attempted to tackle noisy correspondence caused by machine translation for cross-lingual cross-modal retrieval. NRCCR [7] and DCOT [9] respectively employ a cross-attention module and optimal transport theory to implicitly diminish the influence of noisy pairs, which may lead to the loss of semantic information and unreliable noise identification. Furthermore, these methods hope to facilitate the alignment between vision and non-English by combining complementary information from multiple views (e.g., cross-lingual or cross-modal views). However, they overlook the consistency between multiple views, which inevitably captures retrieval-irrelevant information, leading to optimization direction bias.

2.3 Uncertainty-based learning

Over the last decade, uncertainty quantification [22, 23] has attracted significant attention in deep learning. Deep neural networks (DNNs) often provide overconfident deterministic predictions and lack uncertainty estimates, which affects the credibility of the prediction results. Early study [22] uses Bayesian neural networks (BNNs) to replace the

distribution of deterministic weight parameters to measure uncertainty. Subsequently, evidential deep learning [24] is developed by adopting the Dirichlet distribution and treating output as evidence to quantify belief mass and uncertainty by jointly exploiting the Dempster-Shafer theory of evidence (DST) [25] and subjective logic [26]. Evidential deep learning (EDL) has gradually been applied to computer vision and multimodal learning [27–29]. For example, DCEL [29] utilizes EDL to alleviate uncertain cross-modal alignment caused by significant intra-class variation. Unlike these methods, UML dynamically models uncertainty via EDL and Fisher information, provides reliable supervision signals, and gradually improves cross-lingual transfer capabilities via a dynamic dual-view learning strategy.

3 Methodology

3.1 Preliminary

The CCR task aims to retrieve relevant visual content (i.e., images or videos) using non-English queries during inference, while relying solely on human-annotated vision-English pairs during training. Similar to NRCCR [7] and DCOT [9], we utilize machine translation to generate translated non-English captions for training. Formally, the cross-lingual cross-modal dataset $\mathcal{D}$ consists of N triplet sample pairs, denoted as $\mathcal{D} = \{(V_i, S_i, T_i)\}_{i=1}^{N}$, where V_i, S_i, and T_i represent the ith visual content, English caption, and non-English caption, respectively. We define F_v, F_s, and F_t as the encoders for visual, English, and non-English content, respectively. The embedded features of (V_i, S_i, T_i) are denoted as $(\mathbf{v}_i, \mathbf{s}_i, \mathbf{t}_i)$. Following previous studies [7,9], we use human-annotated non-English captions as queries during inference.

3.2 Dual-view uncertainty-aware learning

3.2.1 *Alignment uncertainty modeling*

Recently, uncertainty quantification [22,23] has attracted widespread attention in various fields, addressing overconfidence in deterministic predictions and improving the credibility of results. A notable approach is EDL [24], which integrates Dempster-Shafer theory of evidence [25] and subjective logic [26]. The fundamental concept of EDL is that models not only offer predictions but also provide evidence regarding the credibility of these predictions. The paradigm of EDL is the following. Firstly, EDL collects evidence for each class to establish prior Dirichlet distributions of class probabilities. Then, it utilizes subjective logic theory [26] to quantify predictive uncertainty.

Building upon the above EDL paradigm, we model the cross-lingual cross-modal alignment uncertainty. Specifically, for a given triplet (V_i, S_i, T_i), we take the visual content V_i as the centre, and consider its relations with English S_i and non-English T_i as two views. As illustrated in Figure 2, we model the alignment uncertainty of both views (i.e., $V \leftrightarrow T$ and $V \leftrightarrow S$ views) simultaneously. Note that we mainly address the alignment uncertainty between visual and non-English features caused by translation noise. Meanwhile, we model the uncertainty caused by the inherent semantic ambiguity between vision and English to achieve accurate alignment between them, which in turn guides the alignment between visual and non-English features. For clarity, we use the modeling of alignment uncertainty between vision and non-English (i.e., $V \leftrightarrow T$ view) as an example.

Firstly, we predict the evidence for each cross-modal alignment. Evidence measures the amount of support collected from the data, which means the degree of support for associating the retrieved cross-modal samples with a given query. An inverse relationship exists between the level of uncertainty and the amount of relevant evidence collected. Concretely, for a pair of global-level features $(\mathbf{v}_i, \mathbf{t}_j)$, we exploit the evidence extractor f_e to extract the corresponding evidence $\mathbf{e}_{ij}$, which is defined as

$$\mathbf{e}_{ij} = f_e(\mathrm{Sim}(\mathbf{v}_i, \mathbf{t}_j)) = \exp^{(\tanh(\mathrm{Sim}(\mathbf{v}_i, \mathbf{t}_j)/\tau))}, \tag{1}$$

where τ is a scaling parameter and is set to 0.1, which empirically balances expressiveness and stability. $\mathrm{Sim}(\cdot)$ represents a cosine similarity measure. Thus, we extract bidirectional evidence $\mathbf{e}_i^{vt}$, which contains vision-to-non-English evidence $\mathbf{e}_i^{v2t} = \{\mathbf{e}_{ij}\}_{j=1}^{K}$ and non-English-to-vision evidence $\mathbf{e}_i^{t2v} = \{\mathbf{e}_{ji}\}_{j=1}^{K}$. K denotes the batch size.

Then, we employ subjective logic theory [26] to assign a belief mass b_{ij} to each query and an overall uncertainty mass u_i based on the collected cross-modal evidence as follows:

$$b_{ij} = \frac{e_{ij}}{D_i} = \frac{\alpha_{ij} - 1}{D_i} \quad \text{and} \quad u_i = \frac{K}{D_i}, \tag{2}$$

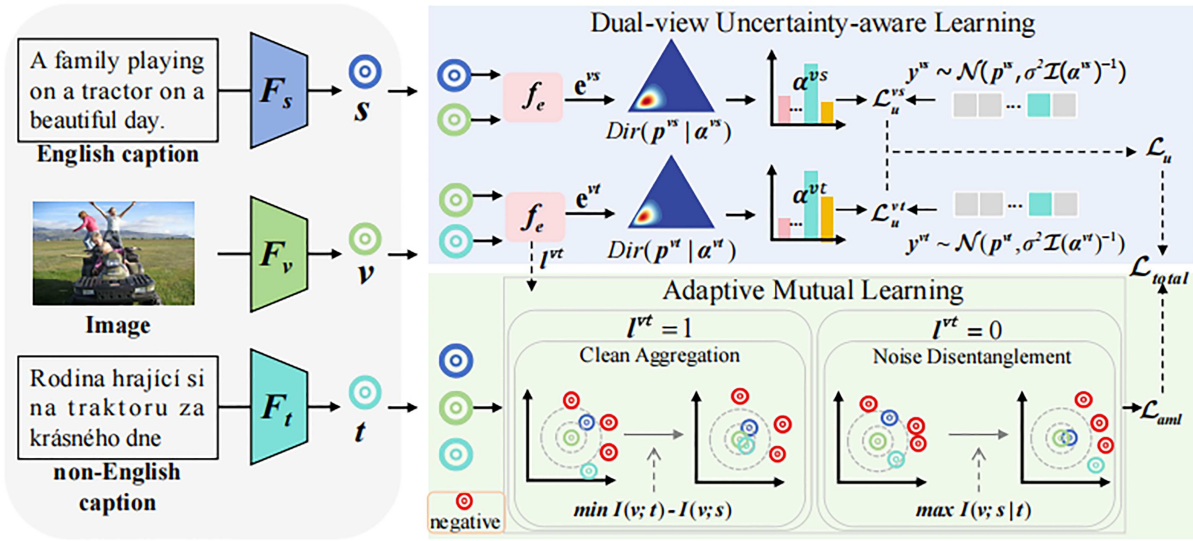

Figure 2 (Color online) Overview of the proposed UML. The images, English captions, and non-English captions are first encoded as feature representations. Then, the evidence extractor f_e is used to collect the dual-view bidirectional evidence $\mathbf{e}^{vs}$ and $\mathbf{e}^{vt}$ parameterized into Dirichlet distributions $\boldsymbol{\alpha}^{vs}$ and $\boldsymbol{\alpha}^{vt}$, respectively. By making the probability of cross-modal alignment $\boldsymbol{p}^{vs}$ ($\boldsymbol{p}^{vt}$) approximate to $\boldsymbol{y}^{vs}$ ($\boldsymbol{y}^{vt}$) softened by Fisher information, an appropriate penalty is assigned to each pair. Besides, we devise an AML loss to further strengthen the comprehensiveness and robustness of features.

where $D_i = \sum_{j=1}^{K} \alpha_{ij}$ and $u_i = 1 - \sum_{j=1}^{K} b_{ij}$. D_i can be regarded as intensity of Dirichlet distribution, and the belief mass assignment $\boldsymbol{b}_i = \{b_{ij}\}_{j=1}^{K}$ represents subjective opinions corresponding to the Dirichlet distribution with parameters $\boldsymbol{\alpha}_i = \{\alpha_{ij}\}_{j=1}^{K}$, where $\alpha_{ij} = e_{ij} + 1$.

Intuitively, retrieval between vision and non-English is analogous to classifying instances, the query similarity is equivalent to probability alignment. By employing the Dirichlet distribution parameterized over evidence, we define the density of each probability assignment, allowing the modeling of second-order probabilities and alignment uncertainty between vision and non-English [26]. Essentially, the Dirichlet distribution serves as a probability density function modeling the potential values of alignment probabilities. The density function is parameterized by $\boldsymbol{\alpha}_i$, defined as follows:

$$\mathrm{Dir}(\boldsymbol{p}_i|\boldsymbol{\alpha}_i) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1}, & \text{for } \boldsymbol{p}_i \in S_K, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where $\boldsymbol{p}_i \in S_K$ are the alignment probabilities, $B(\boldsymbol{\alpha}_i)$ represents the K-dimensional beta function, and S_K is the K-dimensional unit simplex [30].

3.2.2 *Uncertainty learning for CCR*

Previous methods [7,28,31] usually regard the matching between different modalities of the same sample as a simple K-way classification task, where a query is assigned a hard one-hot label corresponding to its positive cross-modal counterpart. This assumes that the ground-truth labels are independent, neglecting any potential correlations between unpaired instances. However, the essence of cross-modal retrieval is to capture the semantic relationships between different modalities, extending beyond mere classification. Simplifying the CCR task into a single-label classification problem may ignore the potentially valuable inter-modal relationships, resulting in the alignment between vision and non-English more challenging.

To address this, we introduce the Fisher information matrix to soften the hard one-hot labels, thereby providing a softer target that facilitates the establishment of comprehensive relationships between modalities. Concretely, given a pair (V_i, T_i), its definitive hard one-hot label l_i can be derived based on the collected bidirectional evidence (i.e., $\mathbf{e}_i^{v2t}$ and $\mathbf{e}_i^{t2v}$) as follows:

$$l_i = \begin{cases} 1, & \text{if } i = \arg\max(\mathbf{e}_i^{v2t} + \mathbf{e}_i^{t2v}), \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Later, we soften the hard one-hot label l_i to the target variable $\boldsymbol{y}_i$. Following [32], the Fisher information matrix can measure the amount of information that the alignment probabilities $\boldsymbol{p}_i$ carry about the concentration parameters

$\boldsymbol{\alpha}_i$ of a Dirichlet distribution that models $\boldsymbol{p}_i$. The class label with higher evidence is associated with lower Fisher information. Hence, we employ the inverse of the Fisher information matrix ($\mathcal{I}(\boldsymbol{\alpha_i})^{-1}$) as the variance of the generative distribution of $\boldsymbol{y}_i$. Intuitively, a class label that has higher evidence is assigned a larger variance, so that more contextual information can be preserved. Thus, we assume that the alignment target variable $\boldsymbol{y}_i$ follows a multivariate Gaussian distribution:

$$\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{p}_i, \sigma^2 \mathcal{I}(\boldsymbol{\alpha}_i)^{-1}), \tag{5}$$

where $\boldsymbol{p}_i \sim \mathrm{Dir}(\boldsymbol{\alpha}_i)$, σ^2 is the scalar used to adjust covariance value, and $\mathcal{I}(\boldsymbol{\alpha}_i)$ is referred to as the Fisher information matrix for $\mathrm{Dir}(\boldsymbol{\alpha}_i)$. Ultimately, the mean square error loss (MSE) is exploited to make the alignment probabilities $\boldsymbol{p}_i$ approach the softened ground-truth $\boldsymbol{y}_i$. The density of $\boldsymbol{p}_i$ conforms to the parameterized Dirichlet distribution $\boldsymbol{\alpha}_i$. The loss $\mathcal{L}_i^e$ can be formulated as follows:

$$\mathcal{L}_i^e\left(\boldsymbol{\alpha}_i, \boldsymbol{y}_i\right) = \mathcal{L}_i^{\mathcal{I}\text{-}MSE} - \varphi \cdot \mathcal{L}_i^{|\mathcal{I}|}, \tag{6}$$

where

$$\mathcal{L}_i^{\mathcal{I}\text{-}MSE} = \sum_{j=1}^{K} \left(\left(Y_{ij} - \frac{\alpha_{ij}}{D_i} \right)^2 + \frac{\alpha_{ij}\left(D_i - \alpha_{ij}\right)}{D_i^2\left(D_i + 1\right)} \right) \psi^{(1)}(\alpha_{ij}), \tag{7}$$

$$\mathcal{L}_i^{|\mathcal{I}|} = \sum_{j=1}^{K} \log \psi^{(1)}(\alpha_{ij}) + \log\left(1 - \sum_{j=1}^{K} \frac{\psi^{(1)}(D_i)}{\psi^{(1)}(\alpha_{ij})} \right), \tag{8}$$

and $\varphi = 0.01$. $\psi^{(1)}(\cdot)$ is a trigamma function with $\psi^{(1)}(x) = \frac{d}{dx}\psi(x)$. Therefore, the uncertainty loss $\mathcal{L}_u^{vt}$ between vision and non-English can be expressed as

$$\mathcal{L}_u^{vt}(\mathbf{v}_i, \mathbf{t}_i, l_i^{vt}) = \mathcal{L}_i^e\left(\boldsymbol{\alpha}_i^{v2t}, l_i^{vt}\right) + \mathcal{L}_i^e\left(\boldsymbol{\alpha}_i^{t2v}, l_i^{vt}\right). \tag{9}$$

Similarly, we can compute the uncertainty loss $\mathcal{L}_u^{vs}$ between vision and English by

$$\mathcal{L}_u^{vs}(\mathbf{v}_i, \mathbf{s}_i, l_i^{vs}) = \mathcal{L}_i^e\left(\boldsymbol{\alpha}_i^{v2s}, l_i^{vs}\right) + \mathcal{L}_i^e\left(\boldsymbol{\alpha}_i^{s2v}, l_i^{vs}\right). \tag{10}$$

Furthermore, to address the model's limited ability to recognize noisy correspondence during the initial stages of training, we devise a dynamic dual-view learning strategy. Specifically, in the early phases of training, we prioritize the vision-English view, which serves as a corrective guide for aligning vision with non-English. As training advances, we progressively shift focus towards enhancing the alignment between vision and non-English, thereby enhancing the model's cross-lingual transfer capabilities. Finally, the dual-view uncertainty-aware loss $\mathcal{L}_u$ can be formulated as

$$\mathcal{L}_u(\mathbf{v}_i, \mathbf{s}_i, \mathbf{t}_i, l_i^{vs}, l_i^{vt}) = \sigma\left(t\right) \cdot \mathcal{L}_u^{vs}(\mathbf{v}_i, \mathbf{s}_i, l_i^{vs}) + \left(1 - \sigma\left(t\right)\right) \cdot \mathcal{L}_u^{vt}(\mathbf{v}_i, \mathbf{t}_i, l_i^{vt}), \tag{11}$$

where $\sigma\left(t\right) = \max(\gamma, 1 - \lambda \cdot \frac{t}{E})$. Here, t and E denote current and total epoch, respectively. γ and λ are hyper-parameters. By combining EDL and the Fisher information matrix, DUL quantifies the alignment uncertainty between modalities and dynamically adjusts label weights to improve the robustness.

3.3 Adaptive mutual learning

Existing methods [8, 9] independently optimize different pairing views (e.g., cross-lingual ($T \leftrightarrow S$) or cross-modal ($V \leftrightarrow T$) views) of a given (V_i, S_i, T_i) triplet, while ignoring the consistency of the different views. This can amplify the optimization direction bias issue, especially in the presence of noisy correspondence. To solve this problem, we propose an AML loss, as illustrated in Figure 2. It optimises clean and noisy pairs separately, leveraging information theory to achieve comprehensive semantic aggregation and reduce the semantic gap. Overall, AML consists of two key components: the clean aggregation module and the noise disentanglement module. The clean aggregation module aims to remove views discrepancies to obtain a comprehensive representation, while the noise disentanglement module aims to push away incorrect alignment. Specifically, as depicted in Figure 2, given a (V_i, S_i, T_i) triplet, we can first obtain the predicted correspondence label l_i^{vt} corresponding to the (V_i, T_i) pair according to (4). If l_i^{vt} is equal to 1, the triplet will be judged as belonging to the clean subset $\mathcal{D}_c$, otherwise it is judged as belonging to the noise subset $\mathcal{D}_n$. The division can be formulated as

$$\begin{cases} \mathcal{D}_c \supseteq (V_i, S_i, T_i), & l_i^{vt} = 1, \\ \mathcal{D}_n \supseteq (V_i, S_i, T_i), & l_i^{vt} = 0. \end{cases} \tag{12}$$

So far, we refine sample filtration based on more accurate evidence, which can be used to adaptively adjust sample contributions in subsequent training.

3.3.1 *Clean aggregation module*

For clean pairs, our objective is to integrate the complementary features of multiple views while minimizing view-specific and task-irrelevant information. To achieve this, we focus on ensuring the semantic consistency of the two cross-modal views (i.e., $V \leftrightarrow T$ and $V \leftrightarrow S$), which can eliminate view-specific information and obtain comprehensive task-relevant information. Specifically, for a triplet $(V_i, S_i, T_i) \in \mathcal{D}_c$, we consider $\mathbf{s}_i$ and $\mathbf{t}_i$ as two observations of $\mathbf{v}_i$ from different viewpoints and define consistency from the perspective of information theory as

$$\min I(\mathbf{v}_i; \mathbf{t}_i) - I(\mathbf{v}_i; \mathbf{s}_i), \tag{13}$$

where $I(*)$ represents mutual information. $I(\mathbf{v}_i; \mathbf{t}_i)$ and $I(\mathbf{v}_i; \mathbf{s}_i)$ indicate the visual representation $\mathbf{v}_i$ (i.e., current task-related information) contained in the non-English representation $\mathbf{t}_i$ and the English representation $\mathbf{s}_i$, respectively. To minimize the disparity between $I(\mathbf{v}_i; \mathbf{t}_i)$ and $I(\mathbf{v}_i; \mathbf{s}_i)$, we introduce the multi-view consistency loss $\mathcal{L}_c$, leveraging variational mutual-learning [33] to equate (13) to

$$\mathcal{L}_c(\mathbf{v}_i, \mathbf{s}_i, \mathbf{t}_i) = \frac{1}{2}[\mathrm{JS}[q_i^{v2t} || q_i^{v2s}] + \mathrm{JS}[q_i^{t2v} || q_i^{s2v}]], \tag{14}$$

where JS denotes the Jensen-Shannon divergence. The q_i^{v2t} (q_i^{t2v}) and q_i^{v2s} (q_i^{s2v}) represent the vision to non-English (non-English to vision) and vision to English (English to vision) softmax-normalized similarity, respectively. In practice, Eq. (14) encourages the two cross-modal views to learn from each other, thereby reducing view-specific redundant information and enhancing robustness against view changes.

3.3.2 *Noise disentanglement module*

If a triplet $(V_i, S_i, T_i) \in \mathcal{D}_n$, it indicates the presence of noisy correspondence between T_i and V_i. Consequently, during the optimization of the representation, the noisy non-English text T_i must be moved away from the vision V_i and English text S_i to prevent fitting the noise. To this end, we propose a disentanglement loss $\mathcal{L}_n$ for noisy pairs based on conditional mutual information [34, 35], which can be defined as

$$\max I\left(\mathbf{v}_i; \mathbf{s}_i \mid \mathbf{t}_i\right), \tag{15}$$

where $I\left(\mathbf{v}_i; \mathbf{s}_i \mid \mathbf{t}_i\right)$ denotes the amount of visual-relevant information in the English feature $\mathbf{s}_i$, excluding information from the noisy non-English feature $\mathbf{t}_i$. Intuitively, we retain only the information that the correspondence is correct. However, directly estimating (15) is typically impractical. Previous studies [34, 36] have highlighted significant challenges in estimating mutual information, primarily due to the curse of dimensionality. Thus, we first factorize (15) as follows:

$$I\left(\mathbf{v}_i; \mathbf{s}_i \mid \mathbf{t}_i\right) = I\left(\mathbf{v}_i; \mathbf{s}_i\right) - I\left(\mathbf{s}_i; \mathbf{t}_i\right) + I(\mathbf{s}_i; \mathbf{t}_i | \mathbf{v}_i), \tag{16}$$

where $I\left(\mathbf{v}_i; \mathbf{s}_i\right)$ measures the relevance of the visual feature $\mathbf{v}_i$ and English feature $\mathbf{s}_i$, $I\left(\mathbf{s}_i; \mathbf{t}_i\right)$ indicates the dependence between English feature $\mathbf{s}_i$ and non-English feature $\mathbf{t}_i$, and $I(\mathbf{s}_i; \mathbf{t}_i | \mathbf{v}_i)$ represents the task-irrelevant information in both $\mathbf{s}_i$ and $\mathbf{t}_i$. Heuristically, optimizing for the task objective typically results in task-specific information overshadowing the irrelevant. Therefore, we can assume that task-irrelevant information will become negligible upon sufficient training [37, 38]. This simplifies (16) to

$$I\left(\mathbf{v}_i; \mathbf{s}_i \mid \mathbf{t}_i\right) \to I\left(\mathbf{v}_i; \mathbf{s}_i\right) - I\left(\mathbf{s}_i; \mathbf{t}_i\right). \tag{17}$$

In our experiments, we employ the variational self-distillation [33] to estimate (17), thus the disentanglement loss $\mathcal{L}_n$ can be expressed as

$$\mathcal{L}_n(\mathbf{v}_i, \mathbf{s}_i, \mathbf{t}_i) = \frac{1}{2}[\mathrm{KL}[q_i^{v2s} || q_i^{s2t}] + \mathrm{KL}[q_i^{s2v} || q_i^{t2s}]], \tag{18}$$

in which KL is the Kullback-Leibler divergence. $\mathcal{L}_n$ can preserve the correct correspondence while avoiding overfitting noisy correspondence. Finally, the learning objective for the adaptive mutual information loss $\mathcal{L}_{aml}$ can be expressed as

$$\mathcal{L}_{aml}(\mathbf{v}_i, \mathbf{s}_i, \mathbf{t}_i, l_i^{vt}) = l_i^{vt} \cdot \mathcal{L}_c(\mathbf{v}_i, \mathbf{s}_i, \mathbf{t}_i) - (1 - l_i^{vt}) \cdot \mathcal{L}_n(\mathbf{v}_i, \mathbf{s}_i, \mathbf{t}_i). \tag{19}$$

By integrating $\mathcal{L}_c$ and $\mathcal{L}_n$, $\mathcal{L}_{aml}$ significantly enhances the model's ability to capture correct matches while effectively suppressing noise interference.

3.4 Training objective

Our model is trained by minimizing the combination of the dual-view uncertainty-aware loss $\mathcal{L}_u$ in (11) and the adaptive mutual information loss $\mathcal{L}_{aml}$ in (19). To sum up, the total loss function is defined as

$$\mathcal{L}_{total} = \frac{1}{K}\sum_{i=1}^{K}(\mathcal{L}_u + \beta \cdot \mathcal{L}_{aml}), \tag{20}$$

where β is a trade-off parameter. Overall, the optimization steps of the proposed UML are summarized in Algorithm 1.

Algorithm 1 Optimization algorithm for UML.

Input: Noisy training dataset $\mathcal{D} = \{(V_i, S_i, T_i)\}_{i=1}^{N}$, batch size K, max epochs E, learning rate η, hyper-parameters γ, λ, β.
Output: Trained cross-modal matching model $\mathcal{M}$.

for epoch $t \leftarrow 1$ **to** E **do**
  Sample minibatch $\mathcal{B} \subset \mathcal{D}$ with size K;
  Learn the common representations of vision (i.e., images or videos) modality **v**, non-English text modality **t**, and English text modality **s**;
  //Subsection 3.2: dual-view uncertainty-aware learning
  Compute the dual-view uncertainty-aware loss using (11);
  //Subsection 3.3: adaptive mutual learning
  Use (12) to split $\mathcal{D}$ into clean pairs $\mathcal{D}_c$ and noisy ones $\mathcal{D}_n$;
  //Subsection 3.3.1: clean aggregation module
  Compute the multi-view consistency loss $\mathcal{L}_c$ by (14) for $\mathcal{D}_c$;
  //Subsection 3.3.2: noise disentanglement module
  Compute the disentanglement loss $\mathcal{L}_n$ by (18) for $\mathcal{D}_n$;
  Compute the adaptive mutual information loss $\mathcal{L}_{aml}$ by (19);
  //Subsection 3.4: training objective
  Compute the total loss $\mathcal{L}_{total}$ according to (20);
  Update $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}_{total}$ via Adam optimizer;
end for
Return $\mathcal{M}$ with optimized parameters θ^*.

4 Experiments

In this section, we conduct a comprehensive comparative analysis with prior methods. We perform ablation studies to assess the efficacy of our proposed components and provide visualization results to illustrate the reliability and interpretability of our approach. Our extensive experimental investigations aim to address the following research questions.

- **RQ1:** To what extent does the proposed UML alleviate the issue of noisy correspondence (NC)?
- **RQ2:** What are the roles and impacts on performance of the different components in UML?
- **RQ3:** What are the learning patterns and insights of UML?

4.1 Experimental settings

Datasets. We perform experiments on two widely-used multilingual image-text matching datasets (i.e., Multi-MSCOCO [9] and Multi30K [39]), as well as a video-text retrieval dataset (i.e., VATEX [40]). Consistent with previous methods [7,9], during training we utilize annotated visual-English pairs and non-English captions generated by Google Translate, while manually labeled non-English queries are evaluated during inference. UML learns a unified representation across modalities, enabling it to handle missing data during training. If one modality is absent, UML can process the available data and infer the missing one, maintaining retrieval performance. Below, we provide a description of the datasets.

- **Multi-MSCOCO** [9] is a multilingual extension of MSCOCO [41]. It consists of 123287 images, and each image has 5 English captions. In addition, each image contains 5 Chinese (ch) and Japanese (ja) captions obtained by machine translation. To establish the dataset partition, we adopt the methodology proposed in [7].
- **Multi30K** [39] represents a multilingual version of Flickr30K [42], comprising 31000 images. Each image is associated with 5 captions in English (en) and German (de), and a single caption in French (fr) and Czech (cs). Following previous work [39], we partition all data into training, validation, and test sets in 29000/1000/1000.

• **VATEX** [40] contains more than 41250 videos, each with 10 English and 10 Chinese sentences. Consistent with previous work [9], we exclusively utilize annotated English captions from the training set and employ machine translation to generate their corresponding Chinese captions. We follow the similar data partition as [43].

Evaluation protocol. (1) For cross-lingual video-text retrieval, we adopt the same evaluation metrics as [8], which include the recall rate at K (R@K), the sum of all recalls (SumR) for both video-to-text and text-to-video retrieval, and mean average precision (mAP). Here, R@K (K = 1, 5, 10) measures the proportion of correctly retrieved items among the top K items most similar to the query. (2) For cross-lingual image-text retrieval, we only report SumR for both image-to-text and text-to-image retrieval.

Implementation details. Following [7], we utilize CLIP (ViT-B/32) as the image encoder [44]. For video representations, we employ I3D video features [45]. The text encoder is derived from mBERT [46] to generate text representations. We train the model for a total of 40 epochs with a batchsize of 128. The optimizer is Adam, with an initial learning rate of 2.5e−5 and a cosine decay scheduler. During training, for Multi30K and Multi-MSCOCO, we set the hyper-parameters γ, λ, and β to 0.2, 4, and 0.6, respectively. For VATEX, we fix these hyper-parameters to 0.25, 3, and 0.4, respectively. During inference, we apply the same similarity calculation method as NRCCR [7] and DCOT [9].

4.2 Comparisons with SOTA methods (RQ1)

In this section, we comprehensively evaluate our UML approach on two widely-used image-text datasets (i.e., Multi-MSCOCO [9] and Multi30K [39]) and a video-text retrieval dataset (i.e., VATEX [40]). The methods against which we compare can be broadly categorized into two groups: (1) noise-robust methods, including NRCCR [7], DCOT [9], and CL2CM [8], and (2) methods that leverage pre-trained models on large-scale datasets, such as M^{3}P [14], UC^2 [6], MURAL [47], MLA [48] and CCLM [5]. To ensure a fair comparison, we directly utilize the results reported in the respective papers and retrain the baseline models according to the recommended settings to obtain results not reported in these studies.

4.2.1 *Cross-lingual image-text retrieval*

Results on Multi30K. According to the comparison results on Multi30K reported in Table 1, we observe the following. (1) Our UML approach achieves the highest performance compared to other noise-robust methods (i.e., NRCCR, DCOT, and CL2CM). Specifically, UML significantly outperforms the strongest baseline, CL2CM, with an improvement of +1.5% to 3.0% in SumR across all languages, underscoring the suitability of our uncertainty modeling for learning the noisy correspondence. (2) When applying our uncertainty-aware and mutual learning strategies to powerful backbones (i.e., SwinTransformer for image encoding and XLM-R for text encoding), UML‡ consistently leads to improvements across all languages. This demonstrates the strong extensibility of UML. Overall, these observations demonstrate the effectiveness and scalability of our approach for cross-lingual image-text retrieval, making it a powerful tool for managing real-world complexities in multilingual, multimodal retrieval applications.

Results on Multi-MSCOCO. Table 1 presents the results on Multi-MSCOCO. UML demonstrates a significant performance advantage over large-scale pre-trained models that do not address noisy correspondence issues. When compared with the leading baseline CLCM, UML maintains a substantial edge, showing respective improvements of 2.4% and 3.0% in terms of SumR. It is important to note that, unlike German and French in Multi30K, Chinese and Japanese in Multi-MSCOCO exhibit notable structural differences from English, rendering them more susceptible to noise during the translation process. Consequently, Multi-MSCOCO presents greater challenges than Multi30K. UML's superior performance on Multi-MSCOCO compared to Multi30K further attests to its robustness in handling noise.

4.2.2 *Cross-lingual video-text retrieval*

For cross-lingual video-text retrieval, we compare our model with four state-of-the-art (SOTA) methods: MMP, NRCCR, DCOT, and CL2CM. MMP* indicates that MMP is pre-trained on Multi-HowTo100M [16], while NRCCR, DCOT, and CL2CM employ robust learning techniques to resist noise from machine translation. Our results, shown in Table 2, reveal that UML outperforms MMP* by a substantial 4% in R@1, without the need for pre-trained datasets, highlighting UML's cost-effectiveness. Furthermore, UML exceeds the best-performing baseline, CL2CM, by 2.1% in SumR. These findings underscore UML's exceptional ability to identify noise and achieve more accurate alignment, surpassing other methods in both robustness and precision.

Table 1 Performance comparison (SumR) of cross-lingual image-text retrieval on Multi30K and Multi-MSCOCO. "en", "de", "fr", "cs", "zh", and "ja" indicate the English, German, French, Czech, Chinese, and Japanese, respectively. The ⋆ denotes models pre-trained on large-scale datasets, e.g., CC3M [49]. The ‡ denotes models employing the same initialization parameters in the backbone with CCLM [5]. "–" means that the result on the dataset is not reported by the paper or its model is unavailable.

Method	Backbone (#parameters)	Multi30K			Multi-MSCOCO	
		en2de	en2fr	en2cs	en2zh	en2ja
NRCCR [7]	mBERT (170M)	480.6	482.1	467.1	512.4	507.0
DCOT [9]	mBERT (170M)	494.9	495.3	481.8	521.5	515.3
CL2CM [8]	mBERT (170M)	498.0	499.7	485.3	522.0	515.9
UML (ours)	mBERT (170M)	505.6	507.9	493.8	534.7	531.6
M^3P [14]⋆	XLMR-Large (560M)	351.0	276.0	220.8	332.8	336.0
UC^2 [6]⋆	XLMR-Base (278M)	449.4	444.0	407.4	492.0	430.2
MURAL [47]⋆	XLMR-Large (560M)	456.0	454.2	409.2	–	435.0
MLA [48]⋆	CLIP (–)	495.6	510.0	457.2	–	482.4
CCLM [5]⋆	XLMR-Large (560M)	503.4	490.6	481.6	511.2	496.4
DCOT [9]‡	XLMR-Large (560M)	515.2	518.7	512.1	535.6	536.2
CL2CM [8]‡	XLMR-Large (560M)	530.4	536.0	526.3	544.3	546.2
UML‡ (ours)	XLMR-Large (560M)	536.4	536.1	530.3	548.6	549.1

Table 2 Performance comparison (R@K, mAP and SumR) of cross-lingual video-text retrieval on VATEX (en2zh). The ⋆ indicates that the model is pre-trained on a large-scale dataset Multi-HowTo100M [16].

Method	T2V				V2T				SumR
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	
MMP [16]	23.9	55.1	67.8	–	–	–	–	–	–
MMP [16]⋆	29.7	63.2	75.5	–	–	–	–	–	–
NRCCR [7]	30.4	65.0	75.1	45.64	40.6	72.7	80.9	32.40	364.8
DCOT [9]	31.4	66.3	76.8	–	46.0	76.3	84.8	–	381.8
CL2CM [8]	32.1	66.7	77.3	47.49	48.2	77.1	85.5	35.77	386.9
UML (ours)	33.7	67.8	78.5	49.33	49.1	78.3	87.5	36.85	394.9

Table 3 Performance comparison (SumR) of zero-shot retrieval on Multi30K. "CC3M-MT" indicates the multilingual version of CC3M [49]. "en", "de", "fr", and "cs" denote the English, German, French, and Czech, respectively.

Method	Training data	en2de	en2fr	en2cs
M^3P [14]	CC3M + Wikipedia	220.8	162.6	122.4
UC^2 [6]	CC3M-MT	375.0	362.4	330.6
CCLM [5]	CC3M-MT	409.5	384.4	375.3
NRCCR [7]	Multi-MSCOCO	448.7	433.8	411.2
DCOT [9]	Multi-MSCOCO	458.9	445.3	424.2
CL2CM [8]	Multi-MSCOCO	461.2	447.0	428.9
UML (ours)	Multi-MSCOCO	463.7	450.5	433.1

4.2.3 *Generalization analysis*

To evaluate the generalization ability of UML, we present the results of cross-lingual image-text retrieval under a zero-shot setting, as shown in Table 3. Specifically, UML is trained on the Multi-MSCOCO dataset and is evaluated on the Multi30K dataset, allowing us to examine its performance when directly applied to a different domain without fine-tuning. Compared to large-scale pre-trained models like M3P, UC2, and CCLM, UML achieves superior results with less training data. These findings suggest that noise-robust learning can reduce reliance on large-scale datasets, emphasizing its importance. Furthermore, UML consistently outperforms other noise-robust methods (i.e., NRCCR, DCOT, and CL2CM) by a significant margin, using the same training data. This is largely due to UML's ability to identify accurate cross-modal correspondences, leading to improved generalization.

4.2.4 *Robustness analysis*

To verify the robustness of the UML against noisy correspondence, we perform experiments with different noise ratios on Multi30K. Since Multi30K is a well-annotated dataset, we artificially generate synthetic noisy correspondence by randomly disrupting the correspondence between vision and non-English captions like [9] for a specific percentage (i.e., 20%, 40%, and 60%). In Figure 3, UML consistently outperforms the robust baseline models NRCCR and

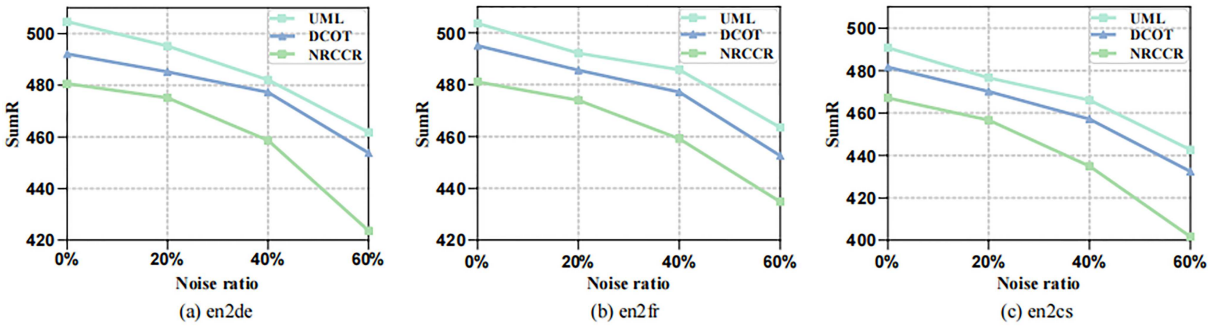

Figure 3 (Color online) Performance comparison (SumR) with different noise ratios on Multi30K. The noise (i.e., 20%, 40%, and 60%) is introduced by artificially switching the correspondence of (V, T) pairs, where "0%" indicates that no artificial noise is added. "en", "de", "fr", and "cs" indicate the English, German, French, and Czech, respectively.

Table 4 Ablation studies for UML's components on Multi30K and Multi-MSCOCO. "DUL" and "AML" denote the dual-view uncertainty-aware learning and adaptive mutual learning, respectively.

No.	DUL		AML		Multi30K			Multi-MSCOCO	
	$\mathcal{L}^{\mathcal{I}\text{-}MSE}$	$\mathcal{L}^{\lvert\mathcal{I}\rvert}$	$\mathcal{L}_c$	$\mathcal{L}_n$	en2de	en2fr	en2cs	en2zh	en2ja
0					475.2	481.9	473.4	510.4	507.2
1	✓				493.8	494.0	481.7	520.9	515.6
2	✓	✓			494.8	496.2	485.4	523.7	519.8
3	✓		✓		497.5	498.3	485.9	526.3	523.6
4	✓	✓	✓		501.6	499.3	487.2	529.5	527.6
5	✓	✓		✓	504.2	505.1	489.6	532.7	529.1
6	✓	✓	✓	✓	505.6	507.9	493.8	534.7	531.6

DCOT across different synthetic noise ratios. This demonstrates the superior robustness of UML in handling noisy correspondence. In addition, even with a high noise ratio, UML maintains a strong performance due to its effective noise mitigation through explicit partitioning.

4.3 In-depth studies of UML (RQ2)

4.3.1 *Contributions of the UML's components*

To comprehensively understand UML, we examine its structure with careful scrutiny. Specifically, we explore the effectiveness of the proposed DUL module and AML loss by analyzing their performance with different backbones on Multi30K. We report the corresponding performances in Table 4 and summarize our observations as follows.

• **Effectiveness of DUL.** As demonstrated in Table 4, the initial row (i.e., No. 0) showcases the performance of the baseline method, which is solely trained using the triplet ranking loss. No. 0 assumes that all pairs are correctly related and does not incorporate any noise-robust design. In comparison to No. 0, DUL (i.e., No. 2) notably enhances the model's performance (+2.5%–4.1%). This observation indicates that DUL offers more reliable cross-modal supervision by capturing and learning from uncertainty.

• **Effectiveness of AML.** In Nos. 4 and 5, we investigate the clean aggregation module and the noise disentanglement module mentioned in AML. The results demonstrate that better performance can be achieved by using either of the two. Furthermore, we conduct experiments to analyze the interplay between DUL and AML, with the findings indicating that their combined use leads to significant performance improvements (+4.3%–6.4%). This is attributed to AML's ability to not only improve the semantic consistency of multiple views of clean samples but also to prevent fitting to noisy samples. Overall, the results underscore that the combined utilization of DUL and AML can enhance visual representation and improve alignment between visual features and non-English features.

4.3.2 *Study of hyper-parameters*

The parameters γ and λ in (11) control the intensity of attention in the dynamic dual-view learning strategy across different views. Meanwhile, the parameter β in (20) balances $\mathcal{L}_u$ and $\mathcal{L}_{aml}$. The results of ablation studies on these hyperparameters are presented in Figure 4. We investigate their impact on UML's performance in cross-lingual image-text retrieval using the Multi-MSCOCO dataset and in cross-lingual video-text retrieval using the VATEX dataset. Across a wide range of hyperparameter values, UML exhibits minimal fluctuations. Notably, variations in

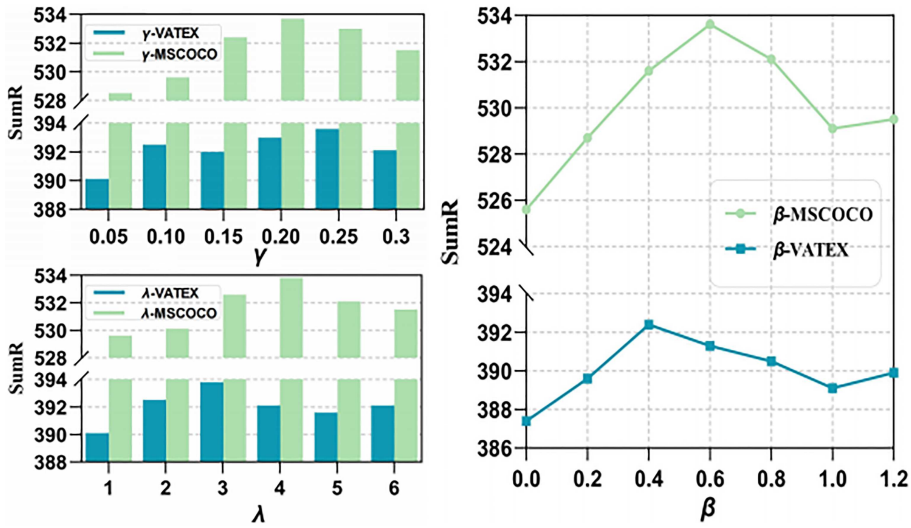

Figure 4 (Color online) Study of the three hyper-parameters (i.e., γ and λ in (11), and β in (20)).

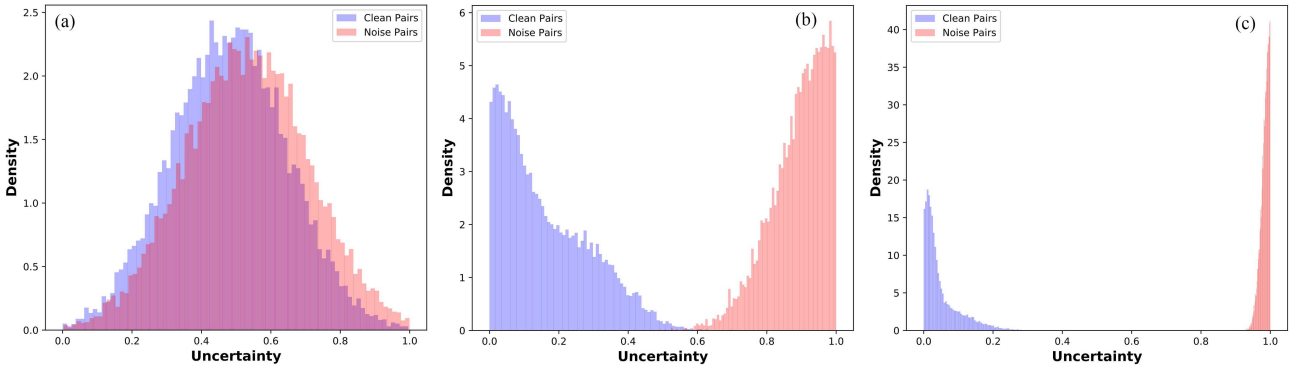

Figure 5 (Color online) We visualize the uncertainty distribution of clean and noisy pairs at different training stages of our UML, which is conducted on Multi30K under 40% noise. Thanks to our UML, the uncertainty of clean pairs gradually approaches the left (low) and the uncertainty of noisy pairs tightly gathers to the right (high). (a) Initial distribution; (b) epoch 20; (c) epoch 40.

these hyperparameters do not lead to significant performance degradation, further demonstrating the robustness of UML.

4.4 Qualitative analysis (RQ3)

4.4.1 *Uncertainty visualization*

To visually analyze the evolution of uncertainty during training, as shown in Figure 5, we conduct experiments under 40% noise ratios on French of Multi30K. The results demonstrate that the uncertainty of clean pairs decreases gradually as training progresses, while the uncertainty of noisy pairs increases, indicating a clear polarization trend. This pattern confirms the effectiveness of uncertainty estimation in identifying and handling noisy correspondence. Notably, leveraging uncertainty provides a natural mechanism to distinguish between clean and noisy pairs, thereby enhancing the model's robustness.

4.4.2 *Representation visualization*

In Figure 6, we utilize t-SNE to visualize the images and non-English representations of NRCCR, DCOT, and UML. We randomly select 20 images along with their corresponding 5 German captions from the Multi-MSCOCO test set, assigning the same color to indicate the same instance. We observe that UML achieves more precise and aggregated intra-instance cross-modal alignment compared to NRCCR and DCOT. This discovery indicates that UML demonstrates superior performance in capturing and integrating precise cross-modal representations.

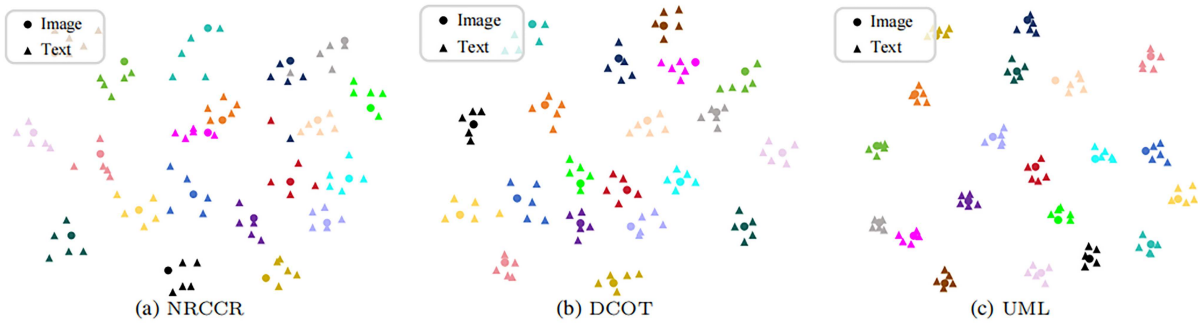

Figure 6 (Color online) A t-SNE visualization is conducted to represent 20 images alongside their corresponding 5 German sentence representations on the Multi30K dataset. The same color indicates the same instance.

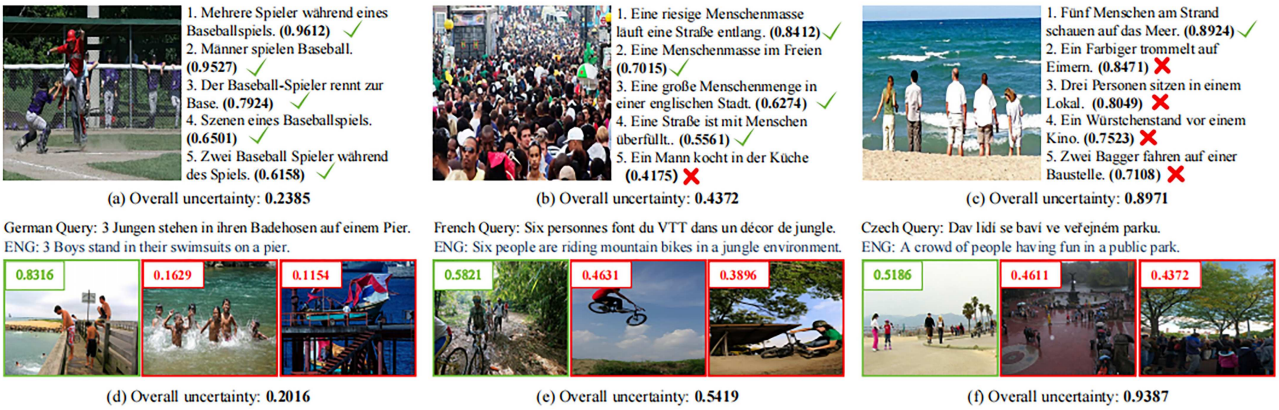

Figure 7 (Color online) Several retrieved examples of cross-lingual image-text retrieval on Multi30K under 40% noise. We present the top-5 ranked German sentences for each image query, identified as (a)–(c). The correctly matched texts are denoted by a green tick, while incorrectly matched texts are marked with a red cross. Similarly, for each sentence query, we display the top-3 ranked images from left to right, denoted as (d)–(f). Correctly matched images are highlighted in green boxes, while incorrectly matched ones are indicated in red boxes. Estimated uncertainty and similarity (i.e., bold font with bracket in sentences, and green or red font with white background in images) are given in sub-captions and exemplars, respectively. "ENG" denotes English translations.

4.4.3 *Retrieval visualization*

To demonstrate the effectiveness of UML, we visualize six representative retrieval cases from Multi30K in Figure 7. UML retrieves cross-modal samples while estimating the uncertainty of each result. The examples show that uncertainty is generally inversely correlated with retrieval quality-more and higher-ranked correct matches correspond to lower uncertainty, as illustrated in Figures 7(a), (d) and Figures 7(c), (f). Notably, high similarity does not always indicate correct retrieval, whereas uncertainty offers a more reliable confidence measure, as shown in Figure 7(c). By quantifying uncertainty, UML enhances the interpretability.

In addition, Figure 8(a) shows the successful cases of UML on the VATEX dataset. UML provides a more reasonable estimate of the similarity of the sample pairs. For example (from left to right), as the translation quality decreases, the similarity given by UML also decreases accordingly. However, the similarity estimates of NRCCR and DCOT are not accurate. Figure 8(b) shows the failure cases of UML. For the semantic ambiguity caused by translation errors in the first two cases, UML overestimates the similarity, indicating that it may overfocus on local semantic alignment and ignore overall semantic learning. Although the translation of the last case is accurate, the video semantics evolve significantly over time, and UML also makes a wrong judgment. This is mainly because UML focuses on global feature optimization and has inherent limitations in understanding the temporal characteristics of videos. Future work will focus on fine-grained semantic alignment to solve the hidden noisy correspondence problem.

5 Conclusion

This paper investigates the challenges of cross-lingual cross-modal retrieval with NC. To address this problem, we present a novel UML framework to mitigate the adverse impacts caused by NC. Specifically, the proposed DUL

Figure 8 (Color online) Several qualitative results of cross-lingual video-text retrieval on VATEX. "ENG" denotes English translations. (a) The successful cases of UML; (b) the failure cases of UML.

measures and mitigates the impact of unreliable supervision by considering alignment uncertainty modeling. Moreover, we design an AML loss to obtain comprehensive and well-aligned cross-modal representations. Experimental results demonstrate the effectiveness and robustness of our proposed UML against noisy correspondence. In our future work, we plan to explore additional strategies to further improve retrieval performance and extend the UML to a wider range of multimodal analysis tasks [50–52].

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62276112, U22A2094, 62272435).

References

1 Yang X, Feng F, Ji W, et al. Deconfounded video moment retrieval with causal intervention. In: Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021. 1–10
2 Dong J, Li X, Xu C, et al. Dual encoding for video retrieval by text. IEEE Trans Pattern Anal Mach Intell, 2021, 44: 4065–4080
3 Yang X, Dong J, Cao Y, et al. Tree-augmented cross-modal encoding for complex-query video retrieval. In: Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020. 1339–1348
4 Liu Y, Qin G, Chen H, et al. Causality-inspired invariant representation learning for text-based person retrieval. In: Proceedings of AAAI Conference on Artificial Intelligence, 2024. 14052–14060
5 Zeng Y, Zhou W, Luo A, et al. Cross-view language modeling: towards unified cross-lingual cross-modal pre-training. 2022. ArXiv:2206.00621
6 Zhou M, Zhou L, Wang S, et al. Uc2: universal cross-lingual cross-modal vision-and-language pre-training. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2021. 4155–4165
7 Wang Y, Dong J, Liang T, et al. Cross-lingual cross-modal retrieval with noise-robust learning. In: Proceedings of ACM International Conference on Multimedia, 2022. 422–433
8 Wang Y, Wang F, Dong J, et al. Cl2cm: improving cross-lingual cross-modal retrieval via cross-lingual knowledge transfer. In: Proceedings of AAAI Conference on Artificial Intelligence, 2024. 5651–5659
9 Wang Y, Wang S, Luo H, et al. Dual-view curricular optimal transport for cross-lingual cross-modal retrieval. IEEE Trans Image Process, 2024, 33: 1522–1533
10 Sharma P, Ding N, Goodman S, et al. Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of Annual Meeting of the Association for Computational Linguistics, 2018. 2556–2565
11 Huang Z, Niu G, Liu X, et al. Learning with noisy correspondence for cross-modal matching. In: Proceedings of Conference on Neural Information Processing Systems, 2021. 29406–29419
12 Shen X, Zhang X, Yang X, et al. Semantics-enriched cross-modal alignment for complex-query video moment retrieval. In: Proceedings of ACM International Conference on Multimedia, 2023. 4109–4118
13 Yang X, Chang T, Zhang T, et al. Learning hierarchical visual transformation for domain generalizable visual matching and recognition. Int J Comput Vis, 2024, 132: 4823–4849
14 Ni M, Huang H, Su L, et al. M3p: learning universal representations via multitask multilingual multimodal pre-training. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2021. 3977–3986
15 Nie Z, Zhang R, Feng Z, et al. Improving the consistency in cross-lingual cross-modal retrieval with 1-to-K contrastive learning. In: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024. 2272–2283
16 Huang P Y, Patrick M, Hu J, et al. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. 2021. ArXiv:2103.08849
17 Miech A, Zhukov D, Alayrac J B, et al. Howto100m: learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2019. 2630–2640

18 Liu Y, Chen H, Qin G, et al. Bias mitigation and representation optimization for noise-robust cross-modal retrieval. ACM Trans Multimed Comput Commun Appl, 2024, 21: 310
19 Qin Y, Chen Y, Peng D, et al. Noisy-correspondence learning for text-to-image person re-identification. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2024. 27197–27206
20 Sun Y, Qin Y, Li Y, et al. Robust multi-view clustering with noisy correspondence. IEEE Trans Knowl Data Eng, 2024, 36: 9150–9162
21 Lu Y, Lin Y, Yang M, et al. Decoupled contrastive multi-view clustering with high-order random walks. In: Proceedings of AAAI Conference on Artificial Intelligence, 2024. 14193–14201
22 Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? In: Proceedings of Conference on Neural Information Processing Systems, 2017
23 Malinin A, Gales M. Predictive uncertainty estimation via prior networks. In: Proceedings of Conference on Neural Information Processing Systems, 2018
24 Sensoy M, Kaplan L, Kandemir M. Evidential deep learning to quantify classification uncertainty. In: Proceedings of Conference on Neural Information Processing Systems, 2018
25 Liu L, Yager R R. Classic Works of the Dempster-Shafer Theory of Belief Functions: An Introduction. Berlin: Springer, 2008. 1–34
26 Jsang A. Subjective Logic: A Formalism for Reasoning Under Uncertainty. Berlin: Springer, 2018
27 Bao W, Yu Q, Kong Y. Evidential deep learning for open set action recognition. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2021. 13349–13358
28 Qin Y, Peng D, Peng X, et al. Deep evidential learning with noisy correspondence for cross-modal retrieval. In: Proceedings of ACM International Conference on Multimedia, 2022. 4948–4956
29 Li S, Xu X, Yang Y, et al. Dcel: deep cross-modal evidential learning for text-based person retrieval. In: Proceedings of ACM International Conference on Multimedia, 2023. 6292–6300
30 Han Z, Zhang C, Fu H, et al. Trusted multi-view classification with dynamic evidential fusion. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 2551–2566
31 Li S, He C, Xu X, et al. Adaptive uncertainty-based learning for text-based person retrieval. In: Proceedings of AAAI Conference on Artificial Intelligence, 2024. 3172–3180
32 Deng D, Chen G, Yu Y, et al. Uncertainty estimation by Fisher information-based evidential deep learning. In: Proceedings of International Conference on Machine Learning, 2023. 7596–7616
33 Tian X, Zhang Z, Lin S, et al. Farewell to mutual information: variational distillation for cross-modal person re-identification. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2021. 1522–1531
34 Yang Y, Chen H, Liu Z, et al. Action recognition with multi-stream motion modeling and mutual information maximization. 2023. ArXiv:2306.07576
35 Wu S, Chen H, Yin Y, et al. Joint-motion mutual learning for pose estimation in videos. 2024. ArXiv:2408.02285
36 Hjelm R D, Fedorov A, Lavoie-Marchildon S, et al. Learning deep representations by mutual information estimation and maximization. 2018. ArXiv:1808.06670
37 Federici M, Dutta A, Forré P, et al. Learning robust representations via multi-view information bottleneck. 2020. ArXiv:2002.07017
38 Zhao L, Wang Y, Zhao J, et al. Learning view-disentangled human pose representation by contrastive cross-view mutual information maximization. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2021. 12793–12802
39 Elliott D, Frank S, Sima'an K, et al. Multi30k: multilingual English-German image descriptions. 2016. ArXiv:1605.00459
40 Wang X, Wu J, Chen J, et al. Vatex: a large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2019. 4581–4591
41 Chen X, Fang H, Lin T Y, et al. Microsoft coco captions: data collection and evaluation server. 2015. ArXiv:1504.00325
42 Young P, Lai A, Hodosh M, et al. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Trans Assoc Comput Linguist, 2014, 2: 67–78
43 Chen S, Zhao Y, Jin Q, et al. Fine-grained video-text retrieval with hierarchical graph reasoning. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2020. 10638–10647
44 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of Conference on Machine Learning, 2021. 8748–8763
45 Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2017. 6299–6308
46 Devlin J. Bert: pre-training of deep bidirectional transformers for language understanding. 2018. ArXiv:1810.04805
47 Jain A, Guo M, Srinivasan K, et al. Mural: multimodal, multitask retrieval across languages. 2021. ArXiv:2109.05125
48 Zhang L, Hu A, Jin Q. Generalizing multimodal pre-training into multilingual via language acquisition. 2022. ArXiv:2206.11091
49 Sharma P, Ding N, Goodman S, et al. Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018. 2556–2565
50 Yang X, Wang S, Dong J, et al. Video moment retrieval with cross-modal neural architecture search. IEEE Trans Image Process, 2022, 31: 1204–1216
51 Chang T, Yang X, Luo X, et al. Learning style-invariant robust representation for generalizable visual instance retrieval. In: Proceedings of ACM International Conference on Multimedia, 2023. 6171–6180
52 Han N, Yang X, Lim E P, et al. Efficient cross-modal video retrieval with meta-optimized frames. IEEE Trans Multimedia, 2024, 26: 10924–10936