Trustworthy Visual-Textual Retrieval

Yang Qin, *Graduate Student Member, IEEE*, Lifu Huang, Dezhong Peng, Bohan Jiang, Joey Tianyi Zhou, *Senior Member, IEEE*, Xi Peng, *Senior Member, IEEE*, and Peng Hu, *Member, IEEE*

***Abstract*—Visual-textual retrieval, as a link between computer vision and natural language processing, aims at jointly learning visual-semantic relevance to bridge the heterogeneity gap across visual and textual spaces. Existing methods conduct retrieval only relying on the ranking of pairwise similarities, but they cannot self-evaluate the uncertainty of retrieved results, resulting in unreliable retrieval and hindering interpretability. To address this problem, we propose a novel Trust-Consistent Learning framework (TCL) to endow visual-textual retrieval with uncertainty evaluation for trustworthy retrieval. More specifically, TCL first models the matching evidence according to cross-modal similarity to estimate the uncertainty for cross-modal uncertainty-aware learning. Second, a simple yet effective consistency module is presented to enforce the subjective opinions of bidirectional learning to be consistent for high reliability and accuracy. Finally, extensive experiments are conducted to demonstrate the superiority and generalizability of TCL on six widely-used benchmark datasets, *i.e.*, Flickr30K, MS-COCO, MSVD, MSR-VTT, ActivityNet, and DiDeMo. Furthermore, some qualitative experiments are carried out to provide comprehensive and insightful analyses for trustworthy visual-textual retrieval, verifying the reliability and interoperability of TCL. The code is available in https://github.com/QinYang79/TCL**

***Index Terms*—Visual-textual retrieval, trustworthy cross-modal learning, uncertainty learning, multimodal learning.**

Received 22 October 2024; revised 14 May 2025 and 12 June 2025; accepted 28 June 2025. Date of publication 15 July 2025; date of current version 21 July 2025. This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFB4710604; in part by NSFC under Grant 62472295, Grant 62372315, Grant 62176171, and Grant U21B2040; in part by Sichuan Science and Technology Planning Project under Grant 2024ZDZX0004, Grant 2024NSFTD0049, Grant 2024ZHCG0005, Grant 2024NSFTD0047, and Grant 2024NSFTD0038; in part by the System of Systems and Artificial Intelligence Laboratory Pioneer Fund under Grant HLJGGG20240327517-13; in part by the Fundamental Research Funds for the Central Universities under Grant CJ202403 and Grant CJ202303; and in part by Chengdu Science and Technology Project under Grant 2023-XT00-00004-GX. The associate editor coordinating the review of this article and approving it for publication was Prof. Jungong Han. *(Corresponding author: Peng Hu.)*

Yang Qin and Peng Hu are with the College of Computer Science, Sichuan University, Chengdu 610017, China (e-mail: penghu.ml@gmail.com).

Lifu Huang is with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798.

Dezhong Peng is with the College of Computer Science, Sichuan University, Chengdu 610017, China, and also with the Tianfu Jincheng Laboratory, Chengdu 610212, China.

Bohan Jiang is with the College of Software and Microelectronics, Peking University, Beijing 100871, China.

Joey Tianyi Zhou is with the Centre for Frontier AI Research (CFAR) and the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore 138634.

Xi Peng is with the College of Computer Science and the National Key Laboratory of Fundamental Algorithms and Models for Engineering Simulation, Sichuan University, Chengdu 610017, China.

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2025.3587575, provided by the authors.

Digital Object Identifier 10.1109/TIP.2025.3587575

I. Introduction

With the rapid development of multimedia technology [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], the quantity and variety of available media have exploded dramatically, particularly in the forms of images, videos, and texts. These modalities are fundamental cognitive mediums and are directly linked to the main human perceptions, thus sparking more and more interest from academic and industrial professionals. For instance, this has led to extensive studies on visual-textual retrieval [12], [13], [14], visual question-answering [15], text-based re-identification [16], visual grounding [17], and more. In this paper, we focus on the task of visual-textual retrieval, which is a fundamental task of cross/multimodal learning. It aims to learn visual-semantic similarity and ranking to retrieve the most relevant cross-modal samples from a large-scale dataset. For example, given an image query, the model should retrieve the most relevant textual descriptions, and vice versa. The primary challenge is to overcome the heterogeneity gap between visual and textual modalities to measure cross-modal relevance semantically.

To tackle the aforementioned challenge, numerous methods have been proposed in recent years, which could be roughly categorized into two groups based on the type of alignment they employ, *i.e.*, global-level methods [12], [14], [19] and local-level methods [13], [18], [20]. Global-level methods, exemplified by Visual Semantic Embedding (VSE)-based techniques, first utilize Deep Neural Networks (DNNs) to learn effective global dense representations of visual and textual samples. They then perform global coarse-grained alignments between visual and textual instances to bring semantically matched cross-modal pairs closer in the latent common space, namely, to maximize the cross-modal similarities between correlated pairs. In contrast, local-level methods focus on presenting a specific mechanism or model to explicitly learn and integrate the fine-grained relationships between image regions/frames and words for cross-modal relevance inference. Unlike global-level methods, local-level methods are dedicated to capturing the nuanced, fine-grained interplay between visual and textual features.

Although prior approaches [18], [21], [22] could achieve promising performance, they are only able to estimate visual-semantic similarities for cross-modal retrieval, wherein cross-modal pairs with high similarity are taken for granted as matched even if they actually are unmatched. For instance, in an image-text retrieval example as shown in Figure 1, model A retrieves images according to similarity rankings, where the top-1 example with the highest similarity is been seen as the most relevant result naturally, however, resulting in incorrect retrieval. Unfortunately, deterministic DNNs can

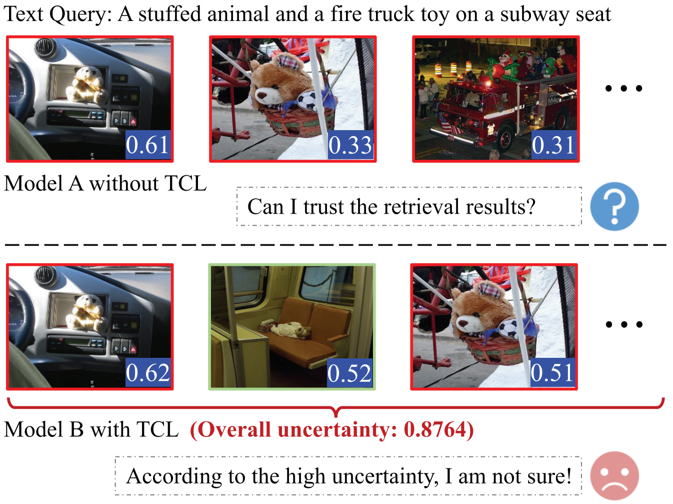

Text Query: A stuffed animal and a fire truck toy on a subway seat
0.61
0.33
0.31
...
Model A without TCL
Can I trust the retrieval results?
0.62
0.52
0.51
...
Model B with TCL
(Overall uncertainty: 0.8764)
According to the high uncertainty, I am not sure!

Fig. 1. The motivation for this paper. The first row shows the top-3 retrieved images given the text query by a traditional method (*e.g.*, SGR [18]). The second shows the results of the query retrieved by TCL-SGR. The number in the images is the corresponding cross-modal similarity, which is usually used to measure the confidence of the retrieved result. Could we trust the retrieved results only relying on the similarity? As shown in the figure, however, it cannot always correctly reflect the confidence level of the results, leading to unreliable retrieval. In this paper, we propose a trust-consistent learning (TCL) framework to endow traditional cross-modal methods with uncertainty awareness, as shown in the second row. Thanks to our TCL, the ambiguous results could be timely found for trustworthy retrieval even with high cross-modal similarity.

only infer cross-modal similarity and cannot self-evaluate retrieval reliability, thus leading to erroneous results. Since the ubiquitous uncertainty in data and models, it is inevitable to produce unreliable retrieval results. Therefore, it requires revisiting questions such as "*Is this retrieval trustworthy?*" to evaluate the uncertainty or unreliability of predictions. To this end, it is valuable and necessary to measure such uncertainty for self-evaluation.

To achieve this goal, we propose a novel trustworthy cross-modal framework as shown in Figure 2, called Trust-Consistent Learning (TCL). TCL is a general framework that can easily endow existing methods with trustworthy cross-modal learning (*e.g.*, the results of model B in Figure 1). Specifically, our TCL employs Evidential Deep Learning (EDL), which is built on the Dempster-Shafer Theory of Evidence (DST) [23] and the Subjective Logic theory (SL) [24], into visual-textual retrieval models to capture uncertainty, thus enabling the model to self-evaluate retrieval quality. We consider the pairwise similarity measured by the cross-modal model as a source of evidence and parameterize the evidence as a Dirichlet distribution, which not only models the density of matching probabilities but also the uncertainty. Unlike prior EDL methods [24], [25] that focused on trustworthy classification, our TCL focuses on instance-level visual-textual retrieval, which presents two challenges: instance-level retrieval and the conflict in bidirectional learning. To address the first task-oriented challenge, we relax instance-level retrieval to a K-way querying for Cross-Modal Uncertainty-aware Learning (CMUL), enabling uncertainty estimation via cross-modal similarities. To tackle the second challenge, we propose two-directional query models (visual-to-textual and textual-to-visual) with CMUL to learn cross-modal associations independently. However, the difference between the two tasks unavoidably leads to a gap in the uncertainty estimation. Thus, we present a simple yet effective Consistency Module (CM) to measure and minimize the difference in predicted opinions across the two task-specific models, which enforces the subjective opinions of bidirectional query models to be as consistent as possible, thus enhancing the reliability of uncertainty learning and improving performance. The main contributions and innovations of this work are summarized as follows:

- We propose a general Trust-Consistent Learning framework (TCL) to achieve trustworthy learning in visual-textual retrieval. By explicitly estimating uncertainty through minor revisions, TCL allows for self-evaluation beyond cross-modal similarity/relevance and enhances the interpretability of retrieval results, thus providing a new perspective for cross-modal retrieval.
- To achieve cross-modal trusted learning, TCL uses bidirectional inference and evidential deep learning to estimate the uncertainty of the cross-modal model at a similarity-based evidence level. An uncertainty-aware loss is proposed to achieve the same objective as the traditional ranking loss while enabling uncertainty learning.
- To address the inconsistent prediction of bidirectional inference in visual-textual retrieval, a simple yet effective consistency module is proposed to enforce models to obtain more consistent subjective opinions for high reliability, thus further improving complementarity for higher retrieval performance.
- Extensive experiments on six widely-used cross-modal benchmark datasets, namely Flickr30K, MS-COCO, MSVD, and MSR-VTT, ActivityNet, and DiDeMo, demonstrate the effectiveness of the proposed method. Notably, our TCL is used to extend nine cross-modal methods, resulting in remarkable improvements that verify its powerful generalization. In addition, comprehensive ablation studies and insightful analyses verify the reliability and practicability of TCL.

II. Related Works

A. Visual-Textual Retrieval

Visual-textual retrieval aims to retrieve the relevant samples across different modalities for a given query based on cross-modal similarity, *e.g.*, image-text retrieval and video-text retrieval. In general, most of the existing methods could be roughly divided into two groups according to the type of alignment, *i.e.*, the global-level methods represented by visual-semantic embedding [12], [21], [26], [27], [28] and the local-level methods with complex similarity inference [29], [30], [31], [32], [33]. The global-level methods mainly aim to learn good global representations from visual and textual samples with the help of a well-designed feature extraction, enhancement, or aggregation strategy, and then directly compute similarity. Although global-level methods have the advantages of high efficiency and low cost, their performance is limited due to the inability to capture the fine-grained local relationships between image and text. To learn fine-grained relevance, the local-level methods desire to learn the latent local region-word (frame-word) alignments across different modalities for more accurate similarity inference. Besides these conventional models, with the success of pre-training of the transformer-based models on large-scale data, Vision-Language Pre-training (VLP) has emerged as a popular paradigm in learning multimodal representations and building semantic relationships in visual and textual modalities. These multimodal transformer methods could be crudely classified into two categories: single-stream models [34], [35] and dual-stream models [36], [37]. The former usually concatenates visual and textual features and then inputs them into a single transformer model. Dual-stream models usually exploit independent transformer models to learn visual and textual representations, and use their correspondences to learn correct visual-semantic associations.

Although prior methods could achieve promising performance, most of them cannot infer the reliability/uncertainty of retrieved results, thus lacking the ability of self-evaluation for trustworthy retrieval. Different from most existing methods, our TCL aims to achieve the primary goal of cross-modal learning while being able to measure uncertainty for trustworthy retrieval.

B. *Uncertainty Learning*

Deep learning has made promising progress in both academic research and industrial applications, but it is hard to quantify the uncertainty of deep models directly due to deterministic network prediction. A general solution is to use Bayesian Neural Networks (BNNs) to model uncertainty by placing priors over network deterministic weights, *e.g.*, variational inference [38], approximations via dropout [39], *etc*. However, modeling uncertainty with BNNs is inevitably limited by the expensive sampling cost. Some recent works [40], [41] attempt to build a bridge between uncertainty learning by assigning prior distributions to attention weights in transformer blocks. *e.g.*, Pei et al. [41] endow Transformer with the ability to estimate uncertainty by casting the deterministic attention weights as the values sampled from a Gumbel-Softmax distribution. Unlike indirectly modeling uncertainty through model weights, Sensoy et al. [24] proposed an evidential deep learning paradigm (EDL) that combines evidence theory [42] with DNN, placing Dirichlet priors over discrete model predictions to model uncertainty at lower cost directly. EDL has been successfully applied in various tasks, *e.g.*, classification [43] and segmentation [44], [45]. However, the standard EDL cannot be effectively used for large-scale instance-level cross-modal retrieval.

Different from traditional unimodal tasks, cross-modal uncertainty learning needs to overcome uncertainty differences caused by the heterogeneity gap. To address the issue, recent work [46], [47], [48] leveraged the probabilistic embedding models to capture uncertainty for better performance on visual-textual retrieval. However, these uncertainty cross-modal methods commonly focus on the data uncertainty, while ignoring that of the cognitive level in models. But anyway, exploring recent few-shot EDL variants or probabilistic embedding-based uncertainty estimation remains an interesting direction for future work. In this paper, our TCL explicitly quantifies model uncertainty using evidence theory and achieves state-of-the-art retrieval performance in a fair setting.

III. Trust-Consistent Learning

In this section, we preview our method in Section III-A and elaborate on how to estimate evidence-based uncertainty for Cross-Modal Uncertainty-aware Learning (CMUL) in Section III-B. Moreover, we present a Consistency Module (CM) to obtain more consistent predictions on subjective opinions during CMUL in Section III-C.

A. *Overview*

To achieve trustworthy visual-textual retrieval, TCL utilizes CMUL and CM to accurately learn the cross-modal similarity/relevance for cross-modal retrieval. Additionally, unlike most standard approaches [18], [21], it can quantify the uncertainty of the cross-modal model for self-evaluation. The framework of TCL is shown in Figure 2. In the following sections, we will elaborate on the basic settings of our TCL.

1) Similarity Calculation: Given a visual-textual dataset $(\mathcal{V}, \mathcal{T})$, which contains a set of visual samples $\mathcal{I}$ and a set of texual samples $\mathcal{T}$. To calculate the cross-modal similarity between a given visual sample V and textual sample T, we first encode them into a latent common space by the modality-specific models $f_v(V; \Theta_\phi)$ and $f_t(T; \Theta_\psi)$, where Θ_ϕ and Θ_ψ are the parameters of the corresponding deep networks, respectively. In the common space, the cross-modal similarity could be measured by a similarity function as follows:

$$\mathrm{Sim}(V, T) = h(f_v(V), f_t(T); \Theta_S), \tag{1}$$

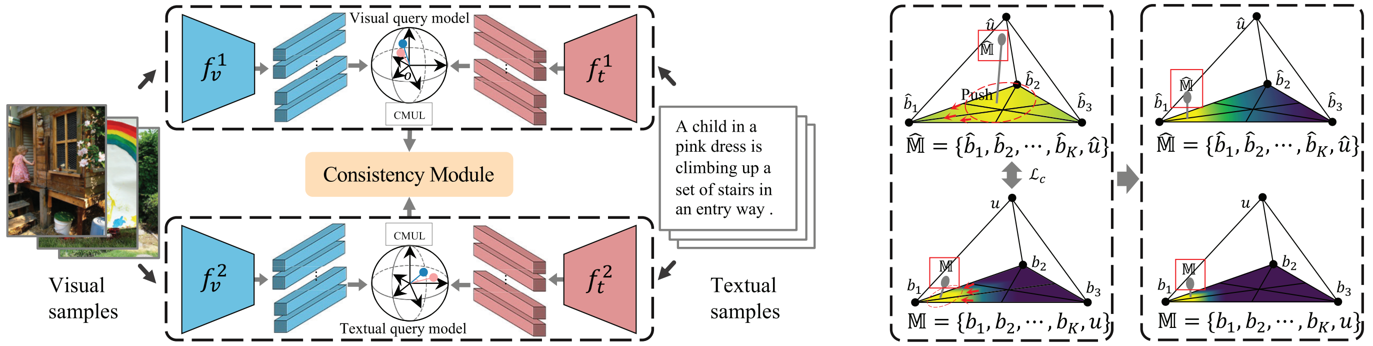
where Θ_S is the parameter set of the similarity function h. In practice, h could be a non-parametric similarity function, *e.g.*, cosine function used in [12] and [21], or a parametric similarity inference model, such as SGRAF [18], etc.

2) Cross-Modal Learning: To achieve accurate cross-modal similarity measurement, most existing deterministic methods [13], [14], [18], [21] aim to pull matched cross-modal samples closer together and push unmatched ones further apart in the shared space by using the bidirectional ranking loss [12], which is defined as:

$$\begin{aligned}\mathcal{L}(V, T) = &\left[\gamma - \mathrm{Sim}(V, T) + \mathrm{Sim}(V, \hat{T})\right]_+ \\ &+ \left[\gamma - \mathrm{Sim}(V, T) + \mathrm{Sim}(\hat{V}, T)\right]_+ ,\end{aligned} \tag{2}$$

where γ is a margin parameter, $[x]_+ \equiv \max(x, 0)$, $\hat{V}$ and $\hat{T}$ are the hardest negatives for a positive pair (V, T) in a training mini-batch.

Different from existing deterministic methods [18], [21], our TCL aims to not only achieve the same cross-modal objective but also endow the cross-modal models with the reliable capability of uncertainty estimation. Specifically, TCL conducts a two-step learning process to optimize the model. In the first step, an uncertainty-aware loss $\mathcal{L}_u$ is utilized to optimize the



(a) Trust-Consistent Learning (TCL)

(b) Consistency Module

Fig. 2. Overview of the proposed approach. (a) shows the pipeline of TCL, which consists of two independent models, and each one only learns a retrieval task, respectively, *i.e.*, the visual query model or the textual query model. Notably, each query model could be used to measure similarities for bidirectional retrieval, *i.e.*, textual retrieval given a visual query (visual-to-textual retrieval) and visual retrieval given a textual query (textual-to-visual retrieval. Each query model contains a visual encoder (f_t^1 and f_t^2) and a textual encoder (f_v^1 and f_v^2) that project the visual and textual samples into a shared common space to calculate the cross-modal similarity. Then, a Cross-Modal Uncertainty-aware Learning (CMUL) is applied in TCL to optimize the query models and capture the uncertainty lurking in the obtained similarities. Finally, our Consistency Module shown in (b) enforces the predicted subjective opinions of the two models to be more consistent for reliable uncertainty learning. Take visual-to-textual direction as an example, in (b), $\mathbb{M}$ and $\hat{\mathbb{M}}$ are the opinion assignments (red boxes) from two query model predictions in visual-to-textual direction, each assignment includes K belief scores corresponding to the K retrieval results of a visual query and an overall uncertainty score. Geometrically, the belief distribution is actually the bottom K-dimensional unit simplex plane ($K = 3$ in the figure as an example), the colors of the bottom simplex plane indicate the density of the belief masses, and the uncertainty is expressed as the height of $\mathbb{M}(\hat{\mathbb{M}})$ in the tetrahedron. In TCL, the visual query model mainly performs visual-to-textual trusted learning, but it will also generate textual-to-visual opinions due to the bi-directionality of cross-modal retrieval, as does the textual query model. Thus, the predictions from two models in the same direction (*e.g.*, visual-to-textual) should generate consistent opinions. For example, the opinions $\hat{\mathbb{M}}$ performed by the textual query model with high uncertainty gradually generate opinions with low uncertainty (close to $\mathbb{M}$ performed by the visual query model) due to consistent constraints, thus narrowing the gap.

model for cross-modal uncertainty-aware learning. The second step applies the proposed opinion-based consistency loss $\mathcal{L}_c$ to eliminate the opinion discrepancy by multifold optimization. We provide more details on the optimization process of TCL in Algorithm 1.

B. Cross-Modal Uncertainty-Aware Learning

In this section, to achieve trusted cross-modal learning, we generalize unimodal Evidential Deep Learning (EDL) [24] to model the uncertainty of visual-textual retrieval. Similar to [24], the Dempster-Shafer Theory of Evidence (DST) [23] and the theory of Subjective Logic (SL) [49] are employed to estimate uncertainty for trustworthy learning. Different from the existing EDL methods [24], [25] that mainly focus on classification, our TCL aims at enabling trustworthy instance-level cross-modal retrieval. Intuitively, the existing EDL methods are not directly applicable to cross-modal retrieval, since bidirectional instance-level retrieval is more challenging than category-level classification.

For visual-textual retrieval, the model first projects the visual and textual samples into a common space, thus making it possible to measure the similarity across different modalities. Different from existing EDL methods [24], [25], the visual-textual model does not have a nonlinear classifier to predict the evidence, thus making it difficult to quantify the uncertainty directly. To address the issue, our TCL relaxes the instance-level retrieval to a K-way querying (see Figure 3), thus the evidence could be estimated by using cross-modal similarities, *i.e.*, $\boldsymbol{e} = [g(s_1), g(s_2), \cdots, g(s_K)]$ for one *Query*, where K is the number of mutually exclusive retrieval events and $g(\cdot)$ is a function to transform similarity scores into non-negative evidence (*i.e.*, $e \in [0, +\infty)$) as shown below:

$$e = g(s) = \mathrm{ReLU}(s/\tau)\text{or } \exp(s/\tau)\text{or } \mathrm{Softplus}(s/\tau), \tag{3}$$

where s is the visual-semantic similarity computed by Equation (1), and $\tau \in (0, 1)$ is a temperature parameter [50]. Note that the impact of different evidence transformation functions is explored in the supplementary material. To model the uncertainty, the similarity-based evidence vector $\boldsymbol{e}$ could be associated with the parameters of a Dirichlet distribution $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_K]$ ($\alpha_k = e_k + 1$) built on SL theory, which provides an overall uncertainty mass u and a belief mass b_i for each event (singleton) that is one of K retrieval events (K-dimensional convex unit simplex as shown in Figure 3) for a *Qurey* in visual-textual retrieval. The $K + 1$ masses are defined as:

$$b_k = \frac{e_k}{S} = \frac{\alpha_k - 1}{S}\text{and } u = \frac{K}{S}, \tag{4}$$

where $S = \sum_{k=1}^{K} (e_k + 1) = \sum_{k=1}^{K} \alpha_k$ and $\sum_{k=1}^{K} b_k + u = 1$. The masses $\mathbb{M} = \{b_1, b_2, \cdots, b_K, u\}$ could be treated as the subjective opinions describing the retrieved results, and S is the Dirichlet strength.

Intuitively, visual-textual retrieval could be viewed as a process of retrieving counterparts with the highest matching probability from different modalities. Hence, the matching probability assignment over the retrieved samples of each query could be denoted as $\mathbf{p} = \left[p_1, p_2, \cdots, p_K\right]$, where $\sum_{j=1}^{K} p_j = 1$. By using the Dirichlet distribution to model such probability assignment, given an opinion, the expected probability of the j-th retrieval event can be written as $\mathbb{E}_{D(\mathbf{p}|\alpha)}\left[p_j\right] = \int p_j D(\mathbf{p} \mid \boldsymbol{\alpha}) d\mathbf{p} = \frac{\alpha_j}{S}$, where the Dirichlet distribution with parameters $\langle \alpha_1, \alpha_2, \cdots, \alpha_K \rangle$ are parameter-

Algorithm 1 The Pseudocode of TCL

```
Input: A well-paired subset $\{(V_i, T_i)\}_{i=1}^{N}$ of $(\mathcal{V}, \mathcal{T})$,
  the cross-modal models $\mathcal{M}_{f_v^1, f_t^1}(\cdot, \Theta_1)$ and
  $\mathcal{M}_{f_v^2, f_t^2}(\cdot, \Theta_2)$, the hyperparameters $\tau$ and $T_{\max}$, the
  maximal epoch number $N_e$;
Initialize: Initialize the model parameters $\Theta$ including
  $\Theta_1$ and $\Theta_2$;
while $n_e = 1, 2, \cdots, N_e$ do
    for $x$=$\{(V_k, T_k)\}_{k=1}^{K}$ in Batches do
        ▷ /* First step */
        $\{e_k^{v2t}\}_{k=1}^{K} \longleftarrow \mathcal{M}_{f_v^1, f_t^1}(\mathbf{x})$;
        $\{e_k^{t2v}\}_{k=1}^{K} \longleftarrow \mathcal{M}_{f_v^2, f_t^2}(\mathbf{x})$;
        Calculate corresponding Dirichlet parameters
          $\{\alpha_k^{v2t}\}_{k=1}^{K} / \{\alpha_k^{t2v}\}_{k=1}^{K}$ by $\alpha = e + 1$;
        Obtain the uncertainty-aware loss $\mathcal{L}_u$ with
          Equation (11);
        $\Theta$ = Optimizer($\mathcal{L}_u, \Theta$);
        ▷ /* Second step */
        for $t = 1, 2, \cdots, T_{max}$ do
            Calculate $\{e_k^{v2t}\}_{k=1}^{K}$ and $\{\hat{e}_k^{v2t}\}_{k=1}^{K}$ of two
              models in the same direction, respectively;
            Detach the gradients of $\{e_k^{v2t}\}_{k=1}^{K}$;
            for each visual query do
                Obtain subjective opinions $\mathbf{b}^{v2t}, \hat{\mathbf{b}}^{v2t}$
                  with Equation (4);
            end
            Calculate $\{\hat{e}_k^{t2v}\}_{k=1}^{K}$ and $\{e_k^{t2v}\}_{k=1}^{K}$ of two
              models in the same direction, respectively;
            Detach the gradients of $\{e_k^{t2v}\}_{k=1}^{K}$;
            for each textual query do
                Obtain Subjective Opinions $\hat{\mathbf{b}}^{t2v}, \mathbf{b}^{t2v}$
                  with Equation (4);
            end
            Use the formed Opinions to calculate the
              consistency loss $\mathcal{L}_c$ with Equation (13);
            $\Theta$ = Optimizer($\mathcal{L}_c, \Theta$);
        end
    end
end
Output: The learned parameters $\hat{\Theta}$
```

ized over the evidence $\langle e_1, e_2, \cdots, e_K \rangle$. $D(\mathbf{p} \mid \boldsymbol{\alpha})$ is defined as

$$D(\mathbf{p} \mid \boldsymbol{\alpha}) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \prod_{j=1}^{K} p_j^{\alpha_j - 1} & \text{for } \mathbf{p} \in \mathcal{S}_K \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

where $B(\boldsymbol{\alpha})$ is the K-dimensional multinomial *beta* function and S_K is the K-dimensional unit simplex [25]. From Proposition 1, one could be seen that $D(\mathbf{p} \mid \boldsymbol{\alpha})$ characterized by $\boldsymbol{\alpha}$ can be regarded as a prior of matching probability assignment, which models second-order probabilities (probability density function) and uncertainty [49].

Proposition 1: Given an opinion, the expected probability of the j-th retrieval event can be written as $\mathbb{E}_{D(\mathbf{p}|\boldsymbol{\alpha})}\left[p_j\right] = \frac{\alpha_j}{S}$.

Proof: See the supplementary material. □

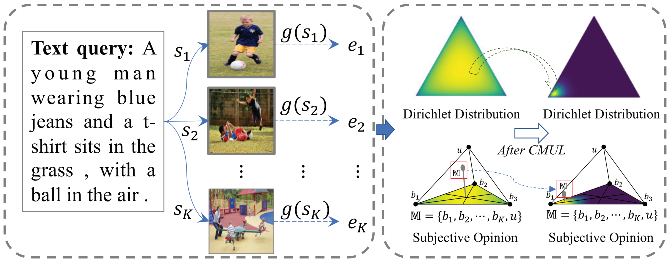


Fig. 3. Illustration of Cross-Model Uncertainty-aware Learning (CMUL). We take textual-to-visual retrieval as an example, where K deterministic similarities for the textual query are first to be viewed as the source of evidence. The evidence could be considered as a measure of the amount of support collected from data in favor of retrieval. Next, SL [49] associates the evidence with the parameters of the Dirichlet distribution, which includes the subjective opinions $\mathbb{M}$ assigning a belief mass to each retrieval and an overall uncertainty mass based on the DST [23]. After trusted learning, the $\mathbb{M}$ with high uncertainty would produce reasonable opinions with low uncertainty.

Although the K-way querying enables evidential learning for cross-modal retrieval, conducting it on the entire training set during each iteration is impractical due to the high costs of computation and storage. To overcome the issue, we draw inspiration from contrastive learning techniques [51], [52] and instead apply the K-way querying in mini-batches of size K, wherein visual and textual samples are paired one by one. Moreover, cross-modal learning aims at maximizing the similarity of positive (matched) visual-textual pairs while minimizing the similarity of negative (unmatched) pairs, *i.e.*, maximizing the corresponding matching probability. Thus, we could formulate the learning criterion for the i-th query in a mini-batch as follows:

$$\mathcal{L}_q(\mathbf{p}^{(i)}) = -\sum_{j=1}^{K} \mathbb{I}_{K(i,j)} \log\left(p_j^{(i)}\right), \tag{6}$$

where $\mathbf{p}^{(i)}$ is the matching probability assignment of the i-th query and $\mathbb{I}_K$ is an identity matrix with size of K. $\mathbb{I}_K$ could be seen as a one-hot label matrix since the visual sample V_i and textual sample T_j are relevant if and only if $i = j$ in the mini-batch.

Considering the density function $D(\mathbf{p}^{(i)} \mid \boldsymbol{\alpha}^{(i)})$ molded by the Dirichlet distribution $\boldsymbol{\alpha}^{(i)}$, the risk of $\mathcal{L}_q$ for the i-th query can be computed by

$$\begin{aligned} &\mathcal{L}_{risk}^{(i)} \\ &= \mathbb{E}_{D(\mathbf{p}^{(i)}|\boldsymbol{\alpha}^{(i)})}\left[\mathcal{L}_q(\mathbf{p}^{(i)})\right] \\ &= \int \left[-\sum_{j=1}^{K} \mathbb{I}_{K(i,j)} \log\left(p_j^{(i)}\right)\right] \frac{1}{B\left(\boldsymbol{\alpha}^{(i)}\right)} \prod_{j=1}^{K} p_j^{(i)\alpha_j^{(i)} - 1} d\mathbf{p}^{(i)} \\ &= \sum_{j=1}^{K} \mathbb{I}_{K(i,j)} \left(\psi\left(S^{(i)}\right) - \psi\left(\alpha_j^{(i)}\right)\right), \end{aligned} \tag{7}$$

where $\psi(\cdot)$ is the digamma function and $S^{(i)}$ is the Dirichlet strength for i-th query. By minimizing the risk, we could make the observations of matched cross-modal pairs generate as strong evidence as possible, ensuring a reasonable uncertainty measurement. For a training mini-batch, the loss function could be formulated as the average of the querying risks:

$$\mathcal{L}_{risk} = \frac{1}{K}\sum_{i=1}^{K}\mathcal{L}_{risk}^{(i)} = \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{D(\mathbf{p}^{(i)}|\boldsymbol{\alpha}^{(i)})}\left[\mathcal{L}_q(\mathbf{p}^{(i)})\right]. \tag{8}$$

By minimizing $\mathcal{L}_{risk}$, TCL encourages the cross-modal model to generate as strong evidence as possible for positive pairs, which guarantees that evidence of positive pairs is higher than that of negative pairs. Furthermore, to further extreme the predicted evidence, we introduce *Kullback-Leibler (KL) divergence* to enforce the evidence of negative pairs to be zero. The penalization loss could be formulated as:

$$\begin{aligned}\mathcal{L}_{kl} &= \frac{1}{K}\sum_{i=1}^{K}\mathbf{KL}\left[D\left(\mathbf{p}^{(i)} \mid \tilde{\alpha}^{(i)}\right) \| D\left(\mathbf{p}^{(i)} \mid \langle 1, 1, \cdots, 1\rangle\right)\right] \\ &= \frac{1}{K}\sum_{i=1}^{K}\left[\log\left(\frac{\Gamma\left(\sum_{j=1}^{K}\tilde{\alpha}_j^{(i)}\right)}{\Gamma(K)\prod_{j=1}^{K}\Gamma(\tilde{\alpha}_j^{(i)})}\right)\right. \\ &\quad \left. + \sum_{j=1}^{K}\left(\tilde{\alpha}_j^{(i)} - 1\right)\left(\psi\left(\tilde{\alpha}_j^{(i)}\right) - \psi\left(\tilde{S}^{(i)}\right)\right)\right],\end{aligned} \tag{9}$$

where $\tilde{S}^{(i)} = \sum_{j=1}^{K}\tilde{\alpha}_j^{(i)}$, $\tilde{\boldsymbol{\alpha}}^{(i)} = \mathbb{I}_{K(i,:)} + \left(1 - \mathbb{I}_{K(i,:)}\right) \odot \boldsymbol{\alpha}^{(i)}$, $\Gamma(\cdot)$ is the gamma function, and $\psi(\cdot)$ is the digamma function. Thus, the uncertainty-aware loss of one query model (*e.g.*, textual retrieval given a visual query) is given by

$$\mathcal{L}_u^{v2t} = \mathcal{L}_{risk}^{v2t} + \lambda\mathcal{L}_{kl}^{v2t}, \tag{10}$$

where λ is a balance factor that dynamically increases with the number of epochs, *i.e.*, $\lambda = \min(1, 0.005 * n_e)$. n_e is the current epoch. The dynamical strategy prevents the optimizer from overemphasizing the KL divergence at the beginning of training, otherwise, the optimizer will be misled by immature opinions, leading to performance degradation. Finally, to simultaneously consider the bidirectional learning of cross-modal retrieval, we jointly optimize the two query models as below:

$$\mathcal{L}_u = \mathcal{L}_u^{v2t} + \mathcal{L}_u^{t2v}, \tag{11}$$

where $\mathcal{L}_u^{t2v}$ is the uncertainty-aware loss of textual query model, which could be computed like Equations (8) to (10).

C. Consistency Module

In our TCL framework, two independent learning models are designed to focus on different specific learning directions *w.r.t.* distinct retrieval tasks, *i.e.*, one for visual-to-textual and another for textual-to-visual. However, the predicted subjective opinions from models can be divergent or conflicting due to random initialization or noisy data, leading to cognitive bias or uncertainty. That is, the discrepancy between different directions will inevitably lead to inconsistent uncertainty estimates from different models in the same direction. More specifically, given one query, one model produces a prediction of low uncertainty, whereas the uncertainty of another model might be higher, as shown in Figure 2b. Thus, we introduce a consistency regularization to enforce the two query models to produce consistent predictions on subjective opinions, *i.e.*, utilizing the L1 norm to measure and minimize the difference in predicted subjective opinions across the two task-specific query models. To simplify presentation without losing generality, we only elaborate on the consistency loss of one direction (visual-to-textual) as follows:

$$\mathcal{L}_c^{v2t}\left(\mathbf{b}^{v2t}, \hat{\mathbf{b}}^{v2t}\right) = \frac{1}{K}\sum_{j=1}^{K}\left|b_j^{v2t} - \hat{b}_j^{v2t}\right|, \tag{12}$$

where $\mathbf{b}^{v2t} = \{b_j^{v2t}\}_{j=1}^{K}$ and $\hat{\mathbf{b}}^{v2t} = \{\hat{b}_j^{v2t}\}_{j=1}^{K}$ are obtained from visual and textual query models with Equation (4), respectively. Similarly, we could easily obtain the consistency loss $\mathcal{L}_c^{t2v}$ in another direction (textual-to-visual). Finally, the consistency loss $\mathcal{L}_c$ of our TCL could be formulated as:

$$\mathcal{L}_c = \frac{1}{K}\sum_{i=1}^{K}\left[\mathcal{L}_c^{v2t}\left(\mathbf{b}_i^{v2t}, \hat{\mathbf{b}}_i^{v2t}\right) + \mathcal{L}_c^{t2v}\left(\hat{\mathbf{b}}_i^{t2v}, \mathbf{b}_i^{t2v}\right)\right]. \tag{13}$$

To explain why the above consistency loss works, we take Equation (12) as an example and obtain its lower bound:

$$\begin{aligned}&\mathcal{L}_c^{v2t}\left(\mathbf{b}^{v2t}, \hat{\mathbf{b}}^{v2t}\right) \\ &\quad = \frac{1}{K}\left\|\mathbf{b}^{v2t} - \hat{\mathbf{b}}^{v2t}\right\|_1 \\ &\quad \geq \frac{1}{K}\left|\left\|\mathbf{b}^{v2t}\right\|_1 - \left\|\hat{\mathbf{b}}^{v2t}\right\|_1\right| = \frac{1}{K}\left|\hat{u}^{v2t} - u^{v2t}\right|,\end{aligned} \tag{14}$$

where u^{v2t} and $\hat{u}^{v2t}$ are uncertainty estimation from visual and textual query models at visual-to-textual direction. Obviously, minimizing Equation (12) (Equation (13)) can ultimately make the uncertainty difference smaller.

IV. Experiment

To evaluate our TCL, we conduct extensive comparison experiments with state-of-the-art methods on four widely used benchmark datasets for visual-textual retrieval, including image-text retrieval and video-text retrieval. Following [13], we measure the performance of image/video-to-text and text-to-image/video retrieval by Recall@K ($K = 1, 5, 10$), which is defined as the proportion of items correctly retrieved in the top K samples of the query. In addition, like most of methods [22], [50], [53], we adopt the sum of all Recall results (rSum) to evaluate the overall performance.

A. Datasets and Implementation Details

1) Datasets: The benchmark datasets used in our experiments are Flickr30K [54], MS-COCO [55], MSVD [56], and MSR-VTT [57]. The first two are used for image-text retrieval, and the latter two are used for text-video retrieval. See the supplementary material for more details.

2) Implementation Detail: To fully verify the superiority and generalizability of our TCL, we apply our TCL to 12 methods for visual-textual retrieval, including nine methods (VSE++ [12], VSRN [19], IMRAM [66], GSMN [67], VSE∞ [21], SAF [18], SGR [18], DivE [62], and RVSE++ [60]) for image-text retrieval and three methods (DE [14], CE+ [68], TT-CE++ [68]) for video-text retrieval. More specifically, for the TCL variants, we follow the basic setups of the corresponding standard methods, and the corresponding bidirectional sorting loss is replaced by $\mathcal{L}_u$. If the standard method has a specially-designed loss function, we keep it in the first step of TCL and optimize the model parameters together with $\mathcal{L}_u$. Among these variants, we report TCL-VSE∞ and TCL-SGR in our comparison experiments. TCL-VSE∞ uses the same encoder networks as VSE∞ [21] to project the local region features and word embeddings into the shared common space with the dimensionality of 1024. Then, a Generalized Pooling Operator is used to aggregate local features. In TCL-SGR, we use the same image-text model settings as those of the original SGR [18]. As TCL-SGR utilizes SGR [18] to infer the similarity score, we replace the activation function Sigmoid of the final fully-connected layer with the activation function Tanh to limit the similarity score to $(-1, 1)$ for loss calculation. For an intuitive and fair comparison, we normalized similarity scores in all qualitative experiments to $(0, 1)$. For a fair comparison, for all image-text retrieval methods, we utilize a Faster R-CNN detection model (ResNet 101) [69] to extract local-level BUTD features of salient regions for each image as input, like [13]. The detection model could extract the region proposals with top-36 confidence scores and then project each region into a 2,048-dimensional feature vector. For each text, the Bi-GRU or pre-trained language model BERT [70] encodes the word tokens into the same dimensional semantic vector space, whose dimensionality is the same as that of image representation. For video-text retrieval, the TCL variants all followed the data processing of the standard methods, and the hyperparameters are also the same as those of the standard methods, except for TCL-specific hyperparameters.

B. Comparisons With State-of-the-Art Methods

1) Image-Text Retrieval: For a comprehensive evaluation, we compare our TCL with 16 baselines, including ten global-level methods: VSE∞ [21], VSRN++ [26], DivE [62], HREM [27], ESA [22], CORA [65], DBL [63], FEM [33], IMEB [64], and RVSE++ [60]; Six local-level methods: SGRAF [18], CMCAN [58], NAAF [53], BCAN [61], RCL [52], and CHAN [32]. Like [27] and [53], we report the ensemble retrieval performance (*i.e.*, TCL$_{i\&t}$) of our TCL by averaging the similarities computed by two query models for a fair comparison, *i.e.*, image query model (TCL$_i$) and text query model (TCL$_t$). Moreover, the experiments are divided into three groups with different settings of textual backbone, namely, a randomly initialized Bi-GRU network is used to train from scratch, a Bi-GRU network initialized by pre-trained Glove vectors [22], [61] is used to train, and the Transformer-based model BERT is utilized to fine-tune.

To verify the superiority of TCL, we carry out experiments on the image-text datasets mentioned above, *i.e.*, Filckr30K and MS-COCO. Note that, all baselines and our TCL are trained and tested on the same data partition as [12] for a fair comparison. Specifically, in Table I, we report the performance on the Flickr30K 1K test set and MS-COCO 5-fold 1K test set. In Table II, following the protocol used in [12] and [13], we report the performance on the MS-COCO 5K test set. From the results, one can see that our TCL methods showed obvious performance advantages on the two widely used benchmark datasets. More specifically, **(1)** Compared with original methods, our TCL brings significant performance improvements. For example, in the group of BUTD + Bi-GRU, TCL-VSE∞ surpasses the global-level baseline VSE∞ by more than **5%** in terms of rSum on the Flickr30K 1K test set. **(2)** Compared with the local-level methods, our TCL-SGR outperforms almost all local-level baselines on the three test sets. Notably, thanks to TCL, SGR achieves even better performance than SGRAF. For example, in the group of BUTD + Bi-GRU, TCL-SGR surpasses SGRAF by about **18%** and **5%** in terms of rSum on Flickr30K 1K test set and MS-COCO 5-fold 1K test set, which indicates the effectiveness of our TCL for performance improvement. Not only that, in group BUTD + BERT, our TCL-SGR shows amazing performance with the best overall performance (rSum) on the three test sets, *i.e.*, 535.0%, 538.7%, and 458.7%, respectively. **(3)** Better textual encoding can achieve better performance. For instance, VSRN++, NAAF, and ESA could achieve better performance by using the BERT model and the pre-trained GloVe vector [71] in the textual encoder, respectively. Similarly, TCL-VSE∞/SGR could achieve better performance by initializing Bi-GRU with the GloVe vector and replacing the textual encoder (Bi-GRU) with the pre-trained BERT model. For example, TCL-SGR delivers the best performance (*i.e.*, rSum = 458.7%) on the MS-COCO 5K test set, which is not only the best but also exceeds the rSum of TCL using Bi-GRU by more than **10%** absolutely.

2) Video-Text Retrieval: To verify the versatility and effectiveness of TCL on different retrieval tasks, we conduct comparison experiments for video-text retrieval on the MSVD [56] and MSR-VTT [57] datasets. Specifically, we compared our extended methods (TCL-DE, TCL-CE+, and TCL-TTCE+) with the original versions (DE [14], CE+ [68] and TTCE+ [68]). All bidirectional retrieval results are reported in table III. From the table, one can see that our method achieves better performance for video-text retrieval. More specifically, **(1)** Thanks to our TCL, TCL-DE, TCL-CE+, and TCL-TTCE+ achieved better performance on most of the metrics compared to the original versions, especially ensemble results. Specifically, there is an improvement of about 20% in terms of rSum for all extensions on the MSVD and MSR-VTT datasets. **(2)** In addition to the ensemble results, almost all single models performed better than the original methods, which indicates that our TCL could not only boost their performance with the dual-modal ensemble but also directly improve the single cross-modal models. In addition to these two datasets, we also perform additional experiments on the DiDeMo [72] and ActivityNet [73] datasets, and the results are reported in the supplementary material.

In brief, our TCL-(VSE∞, SGR, DE, CE+, TTCE+) achieves considerable improvements in terms of overall performance (rSum) compared with the baselines. This remarkable success serves as compelling evidence to demonstrate the effectiveness and superiority of our TCL, which is owed to the proposed cross-modal uncertainty-aware learning and consistency module, *e.g.*, unlike CORA [65] and FEM [33] that focus on relation composition and adaptive feature aggregation, respectively, our TCL introduces consistency-based opinion alignment between bidirectional query branches, which is

TABLE I

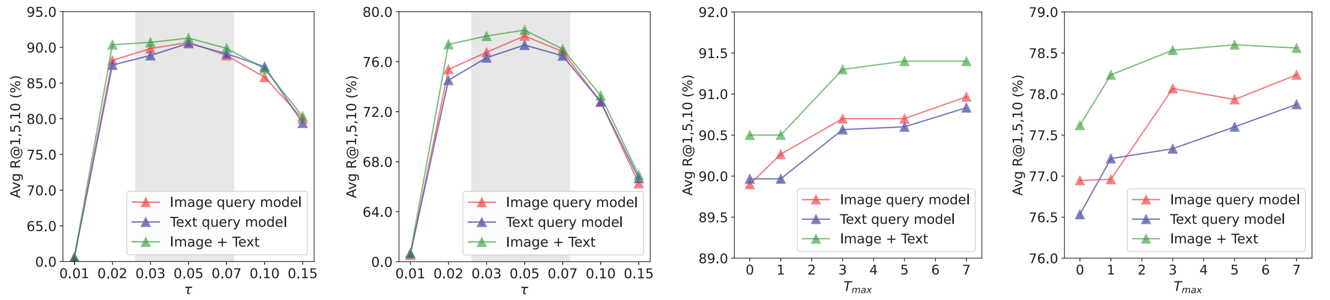
Comparison of Retrieval Performance on Flickr30K 1K Test Set and MS-COCO 5-Fold 1K Test Set. The Best Result Is bolded and the Second-Best One Is underlined. "*" Indicates the Ensemble Results. "G/L" Means the Global/local-Level Methods

Method	Ref.	Type	Flickr30K 1K test							MS-COCO 5-fold 1K test						
			Image⟶Text			Text⟶Image				Image⟶Text			Text⟶Image			
			R@1	R@5	R@10	R@1	R@5	R@10	rSum	R@1	R@5	R@10	R@1	R@5	R@10	rSum
BUTD + Bi-GRU																
VSE∞ [21]	CVPR'21	G	76.5	94.2	97.7	56.4	83.4	89.9	498.1	78.5	96.0	98.7	61.7	90.3	95.6	520.8
SGRAF* [18]	AAAI'21	L	77.8	94.1	97.4	58.5	83.0	88.8	499.6	79.6	96.2	98.5	63.2	90.7	96.1	524.3
NAAF* [53]	CVPR'22	L	78.3	94.1	97.7	58.9	83.3	89.0	501.3	78.9	96.0	98.7	63.1	**91.4**	**96.5**	524.6
CMCAN [58]	AAAI'22	L	79.5	95.6	97.6	60.9	84.3	89.9	507.8	78.6	**96.5**	**98.9**	63.9	90.7	96.2	524.8
RCL* [52]	TPAMI'23	L	79.9	**96.1**	97.8	61.1	85.4	90.3	510.6	80.4	96.4	98.7	64.3	90.8	96.0	526.6
BiCro* [59]	CVPR'23	L	81.7	95.3	**98.4**	61.6	85.6	90.8	513.4	79.1	96.4	98.6	63.8	90.4	96.0	524.5
RVSE++* [60]	TCSVT'25	G	78.2	95.4	97.8	58.5	84.6	90.9	505.5	78.6	96.3	98.8	62.8	90.7	96.0	523.3
TCL$_i$-VSE∞	–	G	79.4	95.4	97.3	58.9	84.4	90.9	506.3	78.3	96.3	98.6	63.0	90.3	95.8	522.3
TCL$_t$-VSE∞	–	G	79.4	94.8	97.5	58.1	83.7	90.2	503.7	79.1	96.2	98.6	63.4	90.4	95.8	523.5
TCL$_{i\&t}$-VSE∞*	–	G	80.8	95.6	97.5	59.3	85.0	91.3	509.5	79.9	96.4	98.7	63.9	90.8	96.0	525.7
TCL$_i$-SGR	–	L	81.2	94.8	97.5	61.2	85.2	90.5	510.4	79.8	96.2	98.7	64.5	90.6	95.9	525.7
TCL$_t$-SGR	–	L	81.2	95.0	97.8	61.6	85.3	91.0	511.9	80.3	96.2	98.7	64.6	90.7	95.9	526.4
TCL$_{i\&t}$-SGR*	–	L	**83.3**	95.4	97.7	**62.7**	**86.5**	**91.7**	**517.3**	**81.3**	96.4	98.7	**65.7**	91.2	96.2	**529.5**
BUTD + Bi-GRU (Glove)																
BCAN* [61]	TNNLS'23	L	78.8	94.8	98.1	56.2	83.1	89.4	499.6	79.2	**96.9**	**99.2**	63.9	91.1	96.4	526.8
DivE [62]	CVPR'23	G	77.8	94.0	97.5	57.5	84.0	90.0	500.8	79.8	96.2	98.6	63.6	90.7	95.7	524.6
CHAN [32]	CVPR'23	L	79.7	94.5	97.3	60.2	85.3	90.7	507.8	79.7	96.7	98.7	63.8	90.4	95.8	525.0
NAAF* [53]	CVPR'22	L	81.9	96.1	98.3	61.0	85.3	90.6	513.2	80.5	96.5	98.8	64.1	90.7	**96.5**	527.2
HREM* [27]	CVPR'23	G	81.4	**96.5**	98.5	60.9	85.6	91.3	514.3	81.2	96.5	98.9	63.7	90.7	96.0	527.1
ESA* [22]	TCSVT'23	G	83.1	96.3	**98.7**	62.4	87.2	**92.5**	520.2	80.4	96.5	98.8	64.2	**91.3**	96.3	527.6
DBL$_{\mathrm{ESA}}$ [63]	TIP'24	G	83.2	96.2	-	62.2	86.5	-	517.5	80.1	96.5	-	63.8	91.2	-	526.7
IMEB [64]	TCSVT'24	G	80.0	96.0	98.1	60.0	85.9	91.5	511.5	81.0	96.6	98.8	64.1	90.8	95.9	527.1
TCL$_i$-VSE∞	–	G	79.3	94.9	98.0	58.8	85.4	91.1	507.5	79.4	96.2	98.7	63.3	90.6	95.8	524.0
TCL$_t$-VSE∞	–	G	80.2	95.4	97.8	60.1	85.0	91.1	509.6	80.0	96.4	98.7	63.2	90.5	95.8	524.6
TCL$_{i\&t}$-VSE∞*	–	G	80.7	95.9	98.3	60.5	86.1	91.8	513.3	80.6	96.5	98.8	63.8	91.0	96.1	526.8
TCL$_i$-SGR	–	L	82.7	96.4	98.4	62.6	86.3	91.8	518.2	80.8	96.4	98.5	65.1	91.0	96.1	527.9
TCL$_t$-SGR	–	L	81.1	95.5	98.1	62.2	86.7	91.6	515.2	80.1	96.4	98.7	64.6	90.7	96.1	526.6
TCL$_{i\&t}$-SGR*	–	L	**84.1**	96.3	98.4	**64.1**	**87.5**	**92.5**	**522.9**	**81.8**	96.6	98.7	**65.5**	91.1	96.3	**530.0**
BUTD + BERT																
VSE∞ [21]	CVPR'21	G	81.7	95.4	97.6	61.4	85.9	91.5	513.5	79.7	96.4	98.9	64.8	91.4	96.3	527.5
VSRN++* [26]	TPAMI'22	G	79.2	94.6	97.5	60.6	85.6	91.4	508.9	77.9	96.0	98.5	64.1	91.0	96.1	523.6
CHAN [32]	CVPR'23	L	80.6	96.1	97.8	63.9	87.5	92.6	518.5	81.4	96.9	98.9	66.5	92.1	96.7	532.6
HREM* [27]	CVPR'23	G	84.0	96.1	98.6	64.4	88.0	93.1	524.2	82.9	96.9	**99.0**	67.1	92.0	96.6	534.6
ESA* [22]	TCSVT'23	G	**84.6**	96.6	98.6	66.3	88.8	93.1	528.0	81.0	96.9	98.9	66.4	92.2	96.5	531.9
FEM [33]	ICASSP'24	G	81.8	95.8	98.1	59.9	84.9	91.2	510.9	80.1	96.3	98.7	64.0	90.9	96.0	526.1
CORA* [65]	CVPR'24	G	83.4	95.9	98.6	64.1	88.1	93.1	523.3	82.4	96.8	98.8	66.2	91.9	96.6	532.7
IMEB [64]	TCSVT'24	G	84.2	96.7	98.4	64.0	88.0	92.8	524.1	82.4	96.9	**99.0**	66.7	91.7	96.6	533.5
RVSE++* [60]	TCSVT'25	G	83.6	96.5	98.6	64.3	88.2	93.0	524.2	81.6	96.6	98.8	66.6	92.1	96.6	532.4
TCL$_i$-VSE∞	–	G	82.1	97.0	98.5	62.4	87.0	91.9	518.9	82.0	96.6	98.8	65.9	91.5	96.4	531.2
TCL$_t$-VSE∞	–	G	83.3	96.5	98.7	63.3	87.0	92.1	520.9	81.8	96.5	98.8	65.9	91.8	96.6	531.4
TCL$_{i\&t}$-VSE∞*	–	G	83.6	**97.3**	98.7	64.3	87.7	92.8	524.4	82.8	96.7	98.7	66.6	92.0	96.7	533.5
TCL$_i$-SGR	–	L	82.5	96.7	98.7	66.8	89.5	93.4	527.6	82.9	96.8	98.7	68.1	92.3	96.8	535.6
TCL$_t$-SGR	–	L	83.2	96.1	98.6	67.6	89.1	93.6	528.2	82.9	96.8	98.8	68.2	92.4	97.0	536.1
TCL$_{i\&t}$-SGR*	–	L	84.3	97.1	**98.9**	**69.7**	**90.5**	**94.5**	**535.0**	**83.9**	**97.2**	98.8	**68.9**	**92.8**	**97.1**	**538.7**

orthogonal and complementary to these strategies. In the following sections, we will conduct an in-depth analysis to verify the rationality of TCL.

C. Ablation Study

In this section, extensive ablation studies are carried out on the Flickr30K dataset to verify the contribution of each component to cross-modal retrieval. The experimental results, all performed by TCL-VSE∞, are presented in Table IV. The first column indicates whether the consistency module ($\mathcal{L}_c$) is utilized to obtain consistent predictions. The task-specific query models used for performance evaluation are *v2t* and *t2v*, which represent the visual query model and textual query model, respectively. From Table IV, we can draw the following observations. **(1) Effectiveness.** To verify the effectiveness of our CMUL applied in TCL, we replace the proposed uncertainty-aware loss $\mathcal{L}_u$ with the widely-used bidirectional ranking loss [12] (Equation (2)) to optimize the cross-modal model, *i.e.*, #7. From Table IV, one could see that other variants with CMUL (*i.e.*, #1–6) achieve better bidirectional retrieval performance than that of training with the bidirectional ranking loss, which indicates that the existing cross-modal model endowed with CMUL could remarkably improve performance by capturing the uncertainty. Moreover, our consistency module could further improve the retrieval performance of the two query models, even using only one model for inference. More specifically, the module could improve the performance by 1.5% (#1 vs. #2), 1.4% (#3 vs. #4), and



(a) Visual-to-textual (b) Textual-to-visual (c) Visual-to-textual (d) Textual-to-visual

Fig. 4. The performance with different settings of two TCL hyperparameters (*i.e.*, τ and $T_{\max}$) for parametric analysis on Flickr30K. (a) is the visualization of the parametric experiments for τ in Equation (3), and (b) is that of the parametric experiments for $T_{\max}$ in Algorithm 1.

TABLE II

Comparison of the Bidirectional Retrieval Performance (R@K %) on MS-COCO 5K Test. The Best Result Is bolded and the Second Best One Is Underlined. " * ": Ensemble Results

Mthods	Ref.	Type	Image⟶Text R@1	R@5	R@10	Text ⟶Image R@1	R@5	R@10	rSum
BUTD + Bi-GRU									
SGRAF* [18]	AAAI'21	L	50.4	82.2	90.0	38.6	69.3	80.4	410.9
VSE∞ [21]	CVPR'21	G	56.6	83.6	91.4	39.3	69.9	81.1	421.9
NAAF [53]	CVPR'22	L	57.3	84.3	91.9	41.6	70.5	81.7	427.3
BiCro* [59]	CVPR'23	L	59.0	84.4	91.7	42.4	71.2	81.7	430.4
RCL* [52]	TPAMI'23	L	60.0	85.5	91.8	43.0	72.0	82.2	434.5
RVSE++* [60]	TCSVT'25	G	56.6	84.7	91.6	40.4	70.9	81.9	426.1
TCL$_{i\&t}$-VSE∞*	–	G	59.6	85.9	92.3	41.8	71.5	82.2	433.3
TCL$_{i\&t}$-SGR*	–	L	**61.9**	**86.5**	**92.5**	**44.1**	**73.0**	**82.7**	**440.7**
BUTD + Bi-GRU (Glove)									
ESA [22]	TCSVT'23	G	58.2	84.8	91.8	41.2	71.4	82.2	429.6
NAAF [53]	CVPR'22	L	58.9	85.2	92.0	42.5	70.9	81.4	430.9
DivE [62]	CVPR'23	G	58.8	84.9	91.5	41.1	72.0	82.4	430.7
CHAN [32]	CVPR'23	L	60.2	85.9	92.4	41.7	71.5	81.7	433.4
HREM* [27]	CVPR'23	G	60.6	**86.4**	92.5	41.3	71.9	82.4	435.1
DBL$_{\text{ESA}}$ [63]	TIP'24	G	58.8	85.2	-	41.6	72.0	-	431.8
IMEB [64]	TCSVT'24	G	60.4	86.3	**92.6**	41.8	72.1	82.2	435.4
TCL$_{i\&t}$-VSE∞*	–	G	60.5	86.1	92.4	42.1	71.6	82.1	434.8
TCL$_{i\&t}$-SGR*	–	L	**62.7**	86.3	**92.6**	**44.3**	**73.0**	**82.5**	**441.4**
BUTD + BERT									
VSE∞ [21]	CVPR'21	G	58.3	85.3	92.3	42.4	72.7	83.2	434.3
VSRN++* [19]	TPAMI'22	G	54.7	82.9	90.9	42.0	72.2	82.7	425.4
ESA* [22]	TCSVT'23	G	61.1	86.6	92.9	43.9	74.1	84.4	443.0
CHAN [32]	CVPR'23	L	59.8	87.2	93.3	44.9	74.5	84.2	443.9
HREM* [27]	CVPR'23	G	64.0	**88.5**	93.7	45.4	75.1	84.3	450.9
FEM [33]	ICASSP'24	G	59.3	85.9	92.3	41.4	72.0	82.3	433.2
CORA* [65]	CVPR'24	G	64.3	87.5	93.6	45.4	74.7	84.6	450.1
IMEB [64]	TCSVT'24	G	62.8	87.8	93.5	44.9	74.6	84.0	447.6
RVSE++* [60]	TCSVT'25	G	60.6	86.4	92.8	44.5	74.5	84.5	443.4
TCL$_{i\&t}$-VSE∞*	–	G	64.2	87.3	93.2	44.9	74.3	84.1	448.0
TCL$_{i\&t}$-SGR*	–	L	**66.5**	**88.5**	**93.9**	**47.9**	**76.3**	**85.6**	**458.7**

0.9% (#5 vs. #6), and in terms of R@1 for sentence retrieval, respectively. By fusing the two query models, our TCL could achieve further improvement, *e.g.*, the full version of our TCL (#1) could improve the version of one query model #3 and #5 by 1.4% and 1.4% in terms of R@1 for sentence retrieval, respectively. **(2) Complementarity.** Two query models are exploited to focus on different retrieval tasks, *i.e.*, image-to-text and text-to-image retrieval. Due to the difference, ensembling the two query models will take advantage of their complementary information, embracing further improvement. Specifically, the variants with the ensemble (*i.e.*, #1 and #2) achieve better performance compared to the variants with single models (*i.e.*, #3-6). **(3) Consistency.** Thanks to our consistency module, our TCL has shown improvement not only in ensemble models but also in single query models, *e.g.*, #3 vs. #4, and #5 vs. #6. That is to say, our consistency module could mutually promote the performance of different branches by eliminating the prediction discrepancy across different branches. Furthermore, our full version of TCL (#1) could achieve the best retrieval performance, which indicates that our consistency module not only mutually promotes the performance of each branch but also captures complementary information from different branches.

D. Parametric Analysis

TCL has two key hyperparameters: the temperature parameter τ in Equation (3) and $T_{\max}$ in Algorithm 1. To investigate the influence of the hyperparameters, we conduct extensive parameter analysis experiments on the Flickr30K dataset as shown in Figure 4, wherein "Image + Text" indicates the ensemble results. From Figures 4a and 4b, one could observe that our TCL demonstrates stable performance in a suitable range of τ, *i.e.*, 0.03 ~ 0.07. However, if τ is too large or too small, the retrieval performance of our TCL gets degraded. Specifically, a very small value of τ will make the cross-modal model hard to optimize, thus leading to poor performance. Moreover, the performance of TCL gradually decreases with increasing τ after about 0.07. Therefore, we recommend setting τ for TCL within 0.03 ~ 0.07 to obtain stable performance. For another hyperparameter $T_{\max}$, the experiments are shown in Figures 4c and 4d. The figures show that as the number of times the consistency module is performed increases, our TCL will achieve better performance. This is due to the fact that a larger $T_{\max}$ produces more consistent predictions. Furthermore, it can be observed from the figure that TCL achieves relatively stable performance when $T_{\max}$ is set in 3 ~ 7. However, each CM execution will incur additional costs, so we recommend setting $T_{\max}$ to 1 ~ 3.

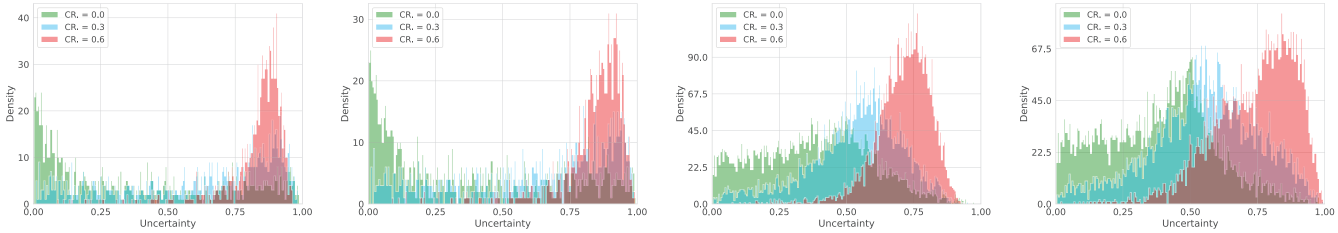
E. Generalization Study

In this section, we perform comprehensive experiments to further demonstrate the effectiveness and generalizability of TCL. More specifically, we apply our TCL into nine instance-level image-text retrieval methods (VSE++ [12], VSRN [19], IMRAM [66], GSMN [67], VSE∞ [21], and SGRAF (SAF and SGR) [18]) and three video-text retrieval methods (DE

TABLE III

COMPARISON OF THE BIDIRECTIONAL RETRIEVAL PERFORMANCE (R@K %) ON THE MSVD AND MSR-VTT DATASETS. THE BEST RESULT IS BOLDED AND THE SECOND BEST ONE IS UNDERLINED

Mthods	MSVD							MSR-VTT						
	Text⟶Video			Video⟶Text				Text⟶Video			Video⟶Text			
	R@1	R@5	R@10	R@1	R@5	R@10	rSum	R@1	R@5	R@10	R@1	R@5	R@10	rSum
DE [14]	12.7	34.5	46.4	16.1	32.1	41.5	183.3	7.7	22.0	31.8	13.0	30.8	43.3	148.6
TCL-DE (Text)	14.3	36.8	50.5	18.2	33.7	43.9	197.4	8.2	24.1	34.6	13.9	33.9	45.8	160.5
TCL-DE (Video)	14.1	37.5	50.6	17.5	34.8	44.0	198.5	8.3	24.3	34.6	14.4	34.5	46.3	162.4
TCL-DE (Text + Video)	**14.7**	**38.2**	**51.7**	**18.7**	**35.8**	**45.2**	**204.3**	**8.6**	**25.0**	**35.5**	**15.4**	**35.5**	**47.5**	**167.5**
CE+ [68]	25.1	56.5	70.9	26.3	54.3	66.8	299.9	13.8	36.5	49.4	22.1	51.5	64.8	238.1
TCL-CE+ (Text)	27.0	59.7	73.4	**29.6**	**59.6**	**70.3**	**319.6**	15.1	38.8	51.5	24.5	55.1	**68.6**	253.6
TCL-CE+ (Video)	24.7	56.9	71.4	24.0	52.2	63.4	292.6	14.7	38.2	50.9	23.5	53.0	65.8	246.1
TCL-CE+ (Text + Video)	**27.6**	**60.5**	**74.4**	28.5	59.3	68.4	318.7	**15.2**	**38.9**	**51.7**	**24.8**	**55.3**	68.3	**254.2**
TT-CE+ [68]	25.1	56.5	70.9	26.3	54.3	66.8	299.9	13.8	36.5	49.4	22.1	51.5	64.8	238.1
TCL-TTCE+ (Text)	25.6	56.9	70.3	22.1	49.6	60.1	284.6	14.9	37.9	50.5	21.2	51.2	65.7	241.4
TCL-TTCE+ (Video)	27.3	60.0	74.0	29.1	**59.7**	**70.1**	320.2	15.8	39.0	51.5	25.6	57.0	69.6	258.5
TCL-TTCE+ (Text + Video)	**28.3**	**61.3**	**74.8**	**29.4**	57.2	69.4	**320.4**	**16.1**	**40.3**	**52.7**	**25.9**	**57.5**	**70.1**	**262.6**



(a) Flickr30K* (b) Flickr30K # (c) MS-COCO 5K* (d) MS-COCO 5K#

Fig. 5. The visualization of the estimated uncertainty on Flickr30K 1K and MS-COCO 5K test sets. “ * ” means the sentence retrieval given image query and “ # ” expresses the image retrieval given sentence query. “CR” means corruption ratio.

TABLE IV

THE IMPACT OF DIFFERENT TCL CONFIGURATIONS. “$^\diamond$” INDICATES THAT THE RESULTS COME FROM THE ORIGINAL PAPER. THE BEST RESULT IS BOLDED AND THE SECOND BEST ONE IS UNDERLINED

No.	$\mathcal{L}_c$	$v2t$	$t2v$	Image → Text			Text → Image			
				R@1	R@5	R@10	R@1	R@5	R@10	rSum
#1	✓	✓	✓	**80.8**	**95.6**	97.5	**59.3**	**85.0**	**91.3**	**509.5**
#2		✓	✓	79.3	94.7	97.5	58.1	84.2	90.5	504.3
#3	✓	✓		79.4	95.4	97.3	58.9	84.4	90.9	506.3
#4		✓		78.0	94.3	97.4	56.9	83.8	90.1	500.5
#5	✓		✓	79.4	94.8	97.5	58.1	83.7	90.2	503.7
#6			✓	78.5	94.1	97.3	56.5	83.6	90.0	500.0
#7	VSE∞$^\diamond$			76.5	94.2	**97.7**	56.4	83.4	89.9	498.1

[14], CE+ [68] and TTCE+ [68]) to improve their retrieval performance. All TCL variants adopt the standard settings of the original methods, except for TCL-specific hyperparameters. Among them, VSE++ is reproduced by Chen et al. [21], which uses the BUTD visual features and adopts mean-pooling to aggregate local features for embedding. The comparison results between TCL variants and original methods are reported in Tables III and V. From the results of image-text retrieval, our TCL significantly improves the bidirectional retrieval performances of the original methods, for example, TCL improves the performance of SAF and SGR by more than 10% in terms of rSum. Besides, compared with the ensemble results, *i.e.*, VSRN, IMRAM, and SGRAF, our TCL also shows an obvious advantage, which fully verifies the effectiveness of TCL. In addition to image-text retrieval, TCL also improves the performance of existing two video-text retrieval methods, *e.g.*, TCL has improved the performance of DE by more than 10% in terms of rSum. which further proves its effectiveness and generalizability.

F. Visualization of Uncertainty

To visually illustrate the estimation of uncertainty, we plotted the distribution diagrams of uncertainty obtained from the test sets of Flickr30K and MS-COCO. However, since the intrinsic perturbations are uncontrollable and inconspicuous, it is difficult to quantitatively evaluate the estimated uncertainty. To address this, we manually corrupted the test data to increase the likelihood of unreliable retrievals for easier observation. Specifically, for the image, we randomly masked the extracted regions in a given proportion. Meanwhile, for the text, we randomly masked, replaced, and deleted the words of the text in the same proportion as the image. For the convenience of presentation, the corruption proportion of the image and text is denoted as “corruption ratio (CR)”. In the experiment, we investigated the uncertainty distribution quantified by our TCL under three CRs (*i.e.*, 0.0, 0.3, 0.6) as shown in Figure 5. The results indicate that most retrievals under low CRs had low uncertainty and clustered on the left, while the uncertainty of the retrievals gradually increased as the CR increased and gathered to the right, as shown in Figures 5a to 5d. That is to say, as the CR increased, the correlation between image-text pairs degraded, leading to increased retrieval uncertainty, which is consistent with the notion that data disturbance increases unreliability and uncertainty. In conclusion, our method effectively captures the uncertainty in models and data, thereby enabling the self-evaluation of retrieval quality.

TABLE V

COMPARISON OF TCL VARIANTS ON FLICKR30K 1K TEST SET. THE BEST RESULT IS BOLDED AND THE SECOND BEST ONE IS UNDERLINED. " * " INDICATES THE ENSEMBLE RESULTS OF TWO MODELS

	Image → Text			Text → Image			
Method	R@1	R@5	R@10	R@1	R@5	R@10	rSum
VSE++ [12]	63.4	87.2	92.7	45.6	76.4	84.4	449.7
TCL$_i$-VSE++	63.2	87.4	93.2	47.0	76.0	84.6	451.4
TCL$_t$-VSE++	63.4	87.8	93.8	47.1	76.3	84.7	453.1
TCL$_{i\&t}$-VSE++*	**65.2**	**88.5**	**93.9**	**47.7**	**76.8**	**85.2**	**457.3**
VSRN [19]	66.8	90.5	95.2	51.5	78.7	86.1	468.8
VSRN* [19]	71.3	90.6	**96.0**	54.7	81.8	88.2	482.6
TCL$_i$-VSRN	68.9	88.7	93.7	52.8	80.0	86.8	470.9
TCL$_t$-VSRN	70.9	91.2	94.7	53.3	80.3	87.3	477.7
TCL$_{i\&t}$-VSRN*	**72.6**	**91.5**	95.3	**56.0**	**82.5**	**88.7**	**486.6**
IMRAM [66]	68.8	91.6	95.6	53.0	79.0	87.1	475.5
TCL$_i$-IMRAM	71.2	92.1	95.6	53.9	77.6	85.7	476.1
TCL$_t$-IMRAM	71.9	92.6	96.4	53.2	78.6	85.9	478.6
TCL$_{i\&t}$-IMRAM*	**74.5**	**93.0**	**96.5**	**55.6**	**80.**2	**87.2**	**486.9**
GSMN [67]	72.6	**93.5**	96.8	53.7	80.0	87.0	483.6
TCL$_i$-GSMN	74.0	93.2	95.8	56.5	81.3	88.0	488.8
TCL$_t$-GSMN	73.7	92.6	96.4	57.2	81.7	88.3	489.9
TCL$_{i\&t}$-GSMN*	**76.7**	93.2	**96.9**	**59.1**	**83.4**	**89.3**	**498.5**
VSE∞ [21]	76.5	94.2	**97.7**	56.4	83.4	89.9	498.1
TCL$_i$-VSE∞	79.4	95.4	97.3	58.9	84.4	90.9	506.3
TCL$_t$-VSE∞	79.4	94.8	97.5	58.1	83.7	90.2	503.7
TCL$_{i\&t}$-VSE∞*	**80.8**	**95.6**	97.5	**59.3**	**85.0**	**91.3**	**509.5**
SAF [18]	73.7	93.3	96.3	56.1	81.5	88.0	489.5
SGRAF* [18]	77.8	94.1	97.4	58.5	83.0	88.8	499.6
TCL$_i$-SAF	78.1	94.4	96.7	60.2	84.9	90.4	504.8
TCL$_t$-SAF	77.4	94.3	97.3	60.5	84.3	90.3	504.1
TCL$_{i\&t}$-SAF*	**79.8**	**94.6**	**97.7**	**62.5**	**85.9**	**91.4**	**511.9**
SGR [18]	75.2	93.3	96.6	56.2	81.0	86.5	488.8
SGRAF* [18]	77.8	94.1	97.4	58.5	83.0	88.8	499.6
TCL$_i$-SGR	81.2	94.8	97.5	61.2	85.2	90.5	510.4
TCL$_t$-SGR	81.2	95.0	**97.8**	61.6	85.3	91.0	511.9
TCL$_{i\&t}$-SGR*	**83.3**	**95.4**	97.7	**62.7**	**86.5**	**91.7**	**517.3**
DivE [62]	77.8	94.0	97.5	57.5	84.0	90.0	500.8
TCL$_i$-DivE	79.3	93.5	97.5	59.2	84.3	90.5	504.3
TCL$_t$-DivE	79.7	94.0	**97.6**	59.8	84.8	91.1	507.0
TCL$_{i\&t}$-DivE*	**81.1**	**94.1**	97.4	**61.1**	**85.8**	**91.8**	**511.3**
RVSE++ [60]	77.2	94.3	97.6	57.3	83.4	89.8	499.6
RVSE++* [60]	78.2	95.4	97.8	58.5	84.6	**90.9**	505.5
TCL$_i$-RVSE++	77.6	94.6	97.7	58.1	84.2	90.2	502.4
TCL$_t$-RVSE++	78.8	95.3	97.7	58.8	84.3	89.8	504.7
TCL$_{i\&t}$-RVSE++*	**79.6**	**95.4**	**97.9**	**59.5**	**84.9**	**90.9**	**508.2**

V. CONCLUSION

In this paper, we revisit a practicable and meaningful problem in visual-textual retrieval, *i.e.*, "*Is this retrieval trustworthy?*". To this end, we present a general Trust-Consistent Learning framework (TCL) that performs uncertainty learning, thus endowing the cross-modal models with the ability to self-evaluate the retrieval quality. Specifically, first, cross-modal uncertainty-aware learning is proposed to capture the accurate uncertainty of cross-modal retrieval. Second, a consistency module is presented to enforce the subjective opinions of distinct query models to be consistent for high reliability. Finally, we apply TCL to nine existing visual-textual methods to verify its generalizability. Besides, we conduct extensive experiments and analyses to verify the effectiveness and self-evaluation of TCL.

REFERENCES

[1] J. Wen, G. Xu, Z. Tang, W. Wang, L. Fei, and Y. Xu, "Graph regularized and feature aware matrix factorization for robust incomplete multi-view clustering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3728–3741, May 2024.

[2] C. Liu, J. Wen, Z. Wu, X. Luo, C. Huang, and Y. Xu, "Information recovery-driven deep incomplete multiview clustering network," *IEEE Trans. Neural Netw. Learn. Syst.*vol. 35, no. 11, pp. 15442–15452, 2023.

[3] J. Wen et al., "Adaptive graph completion based incomplete multi-view clustering," *IEEE Trans. Multimedia*, vol. 23, pp. 2493–2504, 2021.

[4] Y. Qin, N. Pu, N. Sebe, and G. Feng, "Latent space learning-based ensemble clustering," *IEEE Trans. Image Process.*, vol. 34, pp. 1259–1270, 2025.

[5] Y. Qin, C. Qin, X. Zhang, and G. Feng, "Dual consensus anchor learning for fast multi-view clustering," *IEEE Trans. Image Process.*, vol. 33, pp. 5298–5311, 2024.

[6] Y. Sun, Y. Qin, Y. Li, D. Peng, X. Peng, and P. Hu, "Robust multi-view clustering with noisy correspondence," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 12, pp. 9150–9162, Dec. 2024.

[7] J. Cai, Y. Zhang, S. Wang, J. Fan, and W. Guo, "Wasserstein embedding learning for deep clustering: A generative approach," *IEEE Trans. Multimedia*, vol. 26, pp. 7567–7580, 2024.

[8] Y. Zhang, J. Cai, Z. Wu, P. Wang, and S.-K. Ng, "Mixture of experts as representation learner for deep multi-view clustering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, Apr. 2025, pp. 22704–22713.

[9] Y. Liu, C. Li, S. Xu, and J. Han, "Part-whole relational fusion towards multi-modal scene understanding," *Int. J. Comput. Vis.*, vol. 133, no. 7, pp. 4483–4503, Jul. 2025.

[10] T. Zhang, Q. Zhang, K. Debattista, and J. Han, "Cross-modality distillation for multi-modal tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 7, pp. 5847–5865, Jul. 2025.

[11] H. Duan, S. Shao, B. Zhai, T. Shah, J. Han, and R. Ranjan, "Parameter efficient fine-tuning for multi-modal generative vision models with Möbius-inspired transformation," *Int. J. Comput. Vis.*, vol. 133, pp. 4590–4603, Mar. 2025.

[12] F. Fartash, D. Fleet, J. Kiros, and S. Fidler, "VSE+: Improved visual semantic embeddings with hard negatives," in *Proc. Brit. Mach. Vis. Conf. (BMCV)*, 2018, pp. 1–14.

[13] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comp. Vis.*, 2018, pp. 201–216.

[14] J. Dong et al., "Dual encoding for zero-example video retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9346–9355.

[15] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1–9.

[16] Y. Qin, Y. Chen, D. Peng, X. Peng, J. T. Zhou, and P. Hu, "Noisy-correspondence learning for text-to-image person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 27187–27196.

[17] L. Yang, Y. Xu, C. Yuan, W. Liu, B. Li, and W. Hu, "Improving visual grounding with visual-linguistic verification and iterative reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9489–9498.

[18] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2021, vol. 35, no. 2, pp. 1218–1226.

[19] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4654–4662.

[20] S. Chen, Y. Zhao, Q. Jin, and Q. Wu, "Fine-grained video-text retrieval with hierarchical graph reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10638–10647.

[21] J. Chen, H. Hu, H. Wu, Y. Jiang, and C. Wang, "Learning the best pooling strategy for visual semantic embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15789–15798.

[22] H. Zhu, C. Zhang, Y. Wei, S. Huang, and Y. Zhao, "ESA: External space attention aggregation for image-text retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 10, pp. 6131–6143, Oct. 2023.

[23] R. R. Yager and L. Liu, *Classic Works of the Dempster-Shafer Theory of Belief Functions*, vol. 219. Cham, Switzerland: Springer, 2008.

[24] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, S. Bengio, H. Wallach, K. Grauman, N. Cesa-Bianch, and R. Garnett, Eds., Curran Associates, 2018.

[25] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[26] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Image-text embedding learning via visual and textual semantic reasoning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 641–656, Jan. 2023.

[27] Z. Fu, Z. Mao, Y. Song, and Y. Zhang, "Learning semantic relationship among instances for image-text matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15159–15168.

[28] X. Ma, M. Yang, Y. Li, P. Hu, J. Lv, and X. Peng, "Cross-modal retrieval with noisy correspondence via consistency refining and mining," *IEEE Trans. Image Process.*, vol. 33, pp. 2587–2598, 2024.

[29] H. Diao, Y. Zhang, W. Liu, X. Ruan, and H. Lu, "Plug-and-play regulators for image-text matching," *IEEE Trans. Image Process.*, vol. 32, pp. 2322–2334, 2023.

[30] Z. Fu, L. Zhang, H. Xia, and Z. Mao, "Linguistic-aware patch slimming framework for fine-grained cross-modal alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 26297–26306.

[31] Y. Yang, L. Wang, E. Yang, and C. Deng, "Robust noisy correspondence learning with equivariant similarity consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 17700–17709.

[32] Z. Pan, F. Wu, and B. Zhang, "Fine-grained image-text matching by cross-modal hard aligning network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19275–19284.

[33] Z. Wang, Y. Yin, and I. V. Ramakrishnan, "Enhancing image-text matching with adaptive feature aggregation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 8245–8249.

[34] J. Li, R. R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2021, pp. 9694–9705.

[35] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 5583–5594.

[36] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, Jul. 2021, pp. 4904–4916.

[37] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[38] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, Dec. 2015, pp. 2575–2583.

[39] Y. Gal, J. Hron, and A. Kendall, "Concrete dropout," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jan. 2017, pp. 3581–3590.

[40] F. Yang et al., "Uncertainty-guided transformer reasoning for camouflaged object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4146–4155.

[41] J. Pei, C. Wang, and G. Szarvas, "Transformer uncertainty estimation with hierarchical stochastic attention," in *Proc. Conf. Artif. Intell. (AAAI)*, Jun. 2022, vol. 36, no. 10, pp. 11147–11155.

[42] A. P.Dempster,*Upper and Lower Probabilities Inducedby a Multivalued Mapping*. Berlin, Germany: Springer, 2008, pp. 57–72.

[43] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2551–2566, Feb. 2023.

[44] Y. Chen et al., "Evidence-based uncertainty-aware semi-supervised medical image segmentation," *Comput. Biol. Med.*, vol. 170, Mar. 2024, Art. no. 108004.

[45] Y. Chen, Z. Yang, C. Shen, Z. Wang, Y. Qin, and Y. Zhang, "EVIL: Evidential inference learning for trustworthy semi-supervised medical image segmentation," in *Proc. IEEE 20th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2023, pp. 1–5.

[46] S. Chun, S. J. Oh, R. S. De Rezende, Y. Kalantidis, and D. Larlus, "Probabilistic embeddings for cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 8415–8424.

[47] H. Zhang, Y. Yang, F. Qi, S. Qian, and C. Xu, "Robust video-text retrieval via noisy pair calibration," *IEEE Trans. Multimedia*, vol. 25, pp. 8632–8645, 2023.

[48] Y. Ji et al., "MAP: Multimodal uncertainty-aware vision-language pre-training model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23262–23271.

[49] A. Jsang, *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Cham, Switzerland: Springer, 2016.

[50] Y. Qin, D. Peng, X. Peng, X. Wang, and P. Hu, "Deep evidential learning with noisy correspondence for cross-modal retrieval," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 4948–4956.

[51] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[52] P. Hu, Z. Huang, D. Peng, X. Wang, and X. Peng, "Cross-modal retrieval with partially mismatched pairs," *IEEE Trans. Pattern Anal. Mach. Intell.*vol. 45, no. 8, pp. 9595–9610, 2023.

[53] K. Zhang, Z. Mao, Q. Wang, and Y. Zhang, "Negative-aware attention framework for image-text matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15661–15670.

[54] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, Dec. 2014.

[55] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[56] D. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2011, pp. 190–200.

[57] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5288–5296.

[58] H. Zhang, Z. Mao, K. Zhang, and Y. Zhang, "Show your faith: Cross-modal confidence-aware network for image-text matching," in *Proc. Conf. AAAI*, 2022, pp. 3262–3270.

[59] S. Yang et al., "BiCro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19883–19892.

[60] Z. Li, C. Guo, X. Wang, Z. Feng, and Z. Du, "Selectively hard negative mining for alleviating gradient vanishing in image-text matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 2, pp. 1921–1935, Feb. 2025.

[61] Y. Liu, H. Liu, H. Wang, F. Meng, and M. Liu, "BCAN: Bidirectional correct attention network for cross-modal retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, pp. 14247–14258, Oct. 2024.

[62] D. Kim, N. Kim, and S. Kwak, "Improving cross-modal retrieval with set of diverse embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23422–23431.

[63] H. Diao, Y. Zhang, S. Gao, X. Ruan, and H. Lu, "Deep boosting learning: A brand-new cooperative approach for image-text matching," *IEEE Trans. Image Process.*, vol. 33, pp. 3341–3352, 2024.

[64] Z. Li, L. Zhang, K. Zhang, Y. Zhang, and Z. Mao, "Fast, accurate, and lightweight memory-enhanced embedding learning framework for image-text retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 6542–6558, Jul. 2024.

[65] K. Pham, C. Huynh, S.-N. Lim, and A. Shrivastava, "Composing object relations and attributes for image-text matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 14354–14363.

[66] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, "IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12655–12663.

[67] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, "Graph structured network for image-text matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10921–10930.

[68] I. Croitoru et al., "TeachText: CrossModal generalized distillation for text-video retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11563–11573.

[69] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[70] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jun. 2019, pp. 4171–4186.

[71] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[72] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5803–5812.

[73] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 706–715.