Unified Confidence Adjustment for Robust Cross-Modal Retrieval under Test-Time Distribution Shifts

Rui Zhou[1] **Yawen Hao**[1] **Hao Zuo**[1] **Xinhang Wan**[1] **Cheng Zhu**[*1] **Yun Zhou**[*1]

[1]National Key Laboratory of Information Systems Engineering, National University of Defense Technology, Changsha, China

[*]Co-corresponding authors. {zhucheng, zhouyun}@nudt.edu.cn

Abstract

Cross-modal retrieval models often suffer substantial performance degradation under test-time distribution shifts. Existing test-time adaptation methods that are primarily based on entropy minimization tend to sharpen the similarity distribution. However, retrieval relies on feature similarities to rank candidates, and the similarity gaps between nearby-ranked candidates are often small. Oversharpening similarity gaps among candidates can distort the semantic similarity structure, leading to miscalibrated updates and unstable retrieval. To address this issue, we propose a novel *Unified Confidence Adjustment* framework that explicitly accounts for semantic similarity within the candidate set. Specifically, the *Semantic Separation Margin* is designed as a confidence prior. Building on this margin, we develop a *Confidence State Identification* mechanism and a *Unified Confidence Adjustment* strategy for adaptive confidence calibration. Extensive experiments on four benchmarks under both zero-shot transfer and natural corruption settings demonstrate that the proposed method consistently outperforms state-of-the-art test-time adaptation methods with minimal computational overhead, underscoring the effectiveness of confidence-aware regularization for robust cross-modal retrieval.

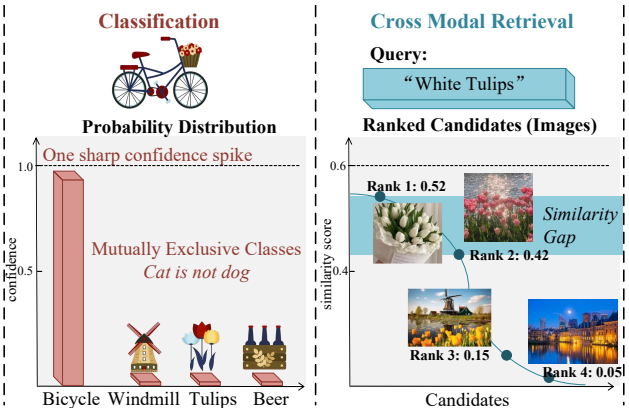

Figure 1: **Motivation.** In cross-modal retrieval, we select the best match by comparing the similarity score between the query representation and gallery representations. Unlike classification tasks where confidence is largely concentrated on the top-1 class, the candidate set for a query (especially the top-ranked candidates) often contains semantically similar items, resulting in small differences in similarity scores. Consequently, entropy-minimization-based adaptation may further sharpen already high-confidence scores and, in doing so, distort local semantic relations among near-ranked candidates. We therefore propose a *unified confidence adjustment* strategy that leverages the *similarity gap* between the best and second-best candidates to preserve the structured neighborhood within the candidate set.

1 INTRODUCTION

Cross-modal retrieval aims to learn a shared semantic space for matching heterogeneous modalities such as images and text [Wang et al., 2024]. With the rapid development of large-scale vision-language foundation models (e.g., CLIP and BLIP), cross-modal matching systems have achieved remarkable performance under standard i.i.d. settings [Li et al., 2025c]. However, in real-world applications, models are frequently deployed under distribution shifts caused by domain transfer, sensor degradation, environmental noise, or query drift [Hendrycks and Dietterich, 2019, Koh et al., 2021]. In retrieval, such shifts often manifest as unstable top-ranked neighborhoods: semantically similar candidates become difficult to separate, and small similarity perturbations can flip the top-1 ranking. Test-Time Adaptation (TTA) has emerged as an effective paradigm to alleviate this issue by updating model parameters online using unlabeled test data [Wang et al., 2021, Niu et al., 2022]. While TTA has

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:8292–8311.

shown strong performance in classification tasks, adapting it to cross-modal retrieval remains challenging due to the structured nature of similarity-based matching and the lack of confidence estimation under distribution shifts.

Existing approaches improve robustness from different perspectives. Some methods explicitly consider cross-modal discrepancy and intra-modality uniformity to maintain semantic alignment under query shifts [Li et al., 2025a]. Others adopt entropy minimization strategies, encouraging confident predictions at test time, and further incorporate uncertainty estimation or sample filtering to suppress unreliable updates [Wang et al., 2021, Niu et al., 2022]. Despite these advances, current methods largely overlook an essential aspect: confidence calibration in retrieval. In classification tasks, confidence typically corresponds to the probability assigned to a single mutually exclusive class, and entropy minimization serves as a natural proxy for certainty [Lee et al., 2024, Yang et al., 2024a]. In contrast, cross-modal retrieval operates over a ranked candidate set where semantic overlap between top candidates is common. Consequently, optimizing confidence solely for the top-ranked candidate may distort the underlying semantic relationships among candidates and impair the consistency of the ranking structure. Therefore, unlike classification, retrieval requires modeling the *relative similarity structure* among candidates. The lack of explicit modeling of over-confidence and under-confidence in this structured setting leaves an important gap that existing entropy-driven TTA methods fail to fill.

To address this limitation, we propose a **U**nified **C**onfidence **A**djustment (UCA) framework for robust cross-modal retrieval under test-time distribution shifts. Unlike classification, retrieval confidence is inherently relative: the model selects the top-ranked candidate from a semantically correlated set, where the similarity scores of near neighbors can be close. We therefore use the similarity gap between the best and second-best candidates as a structured confidence surrogate. We first estimate a reference gap, termed the *Semantic Separation Margin*, from source-domain-like query–candidate pairs that are more reliable under shift. Given an incoming test batch, we compare its empirical gap to the reference to identify over-confident and under-confident states, and introduce a *Unified Confidence Adjustment* loss $\mathcal{L}_{\mathrm{uca}}$ that steers the gap toward the reference level, stabilizing local ranking structure. UCA can be combined with lightweight regularizers for modality alignment, feature uniformity and sample re-weighted entropy minimization, yielding an efficient online adaptation procedure.

Contributions. The main contributions of this work are summarized as follows:

- *Confidence Modeling for Retrieval under Distribution Shifts.* We present a systematic investigation of confidence behavior in cross-modal retrieval under test-time distribution shifts, emphasizing the structured similarity relations among near-ranked candidates that are absent in classification-style confidence assumptions.
- *Unified Confidence Adjustment Framework.* We propose a principled and efficient Unified Confidence Adjustment (UCA) framework. Specifically, we introduce the semantic separation margin as a confidence prior, and propose a confidence-state-aware objective $\mathcal{L}_{\mathrm{uca}}$ that adaptively calibrates the best/second-best similarity gap during online test-time adaptation.
- *Competitive Robustness under Diverse Shifts.* Extensive experiments on four datasets demonstrate consistent improvements over strong TTA baselines under both zero-shot transfer and natural corruption settings. In particular, under maximum-severity image corruptions (Table 1), UCA outperforms the state-of-the-art TCR by **4.6%** on Flickr30K and **1.8%** on MSCOCO, highlighting the strong robustness of UCA against severe modality degradation.

2 RELATED WORK

2.1 CROSS-MODAL RETRIEVAL

Cross-modal retrieval (CMR) learns a shared semantic space for matching heterogeneous modalities [Wang et al., 2024]. Existing studies mainly improve cross-modal alignment and representation quality through semantic alignment objectives [Li et al., 2025c, Huang et al., 2024] or hashing-based retrieval for scalable search [Li et al., 2024, Qin et al., 2025]. Recently, uncertainty-aware retrieval has been explored to quantify confidence and improve reliability [Li et al., 2025b, Duan et al., 2025], including methods that model epistemic and aleatoric uncertainty or use training-free cues such as top-1 consistency [Wang et al., 2025, Gomez, 2025]. However, these methods are typically developed under static or near-i.i.d. assumptions, and rarely address confidence calibration under test-time distribution shifts.

2.2 TEST TIME ADAPTATION

Test-Time Adaptation (TTA) updates model parameters online using online unlabeled test data to mitigate distribution shifts. Entropy minimization [Wang et al., 2021, Niu et al., 2023] is a representative strategy that encourages confident predictions and serves as a strong baseline. However, it may amplify high-confidence errors when the model is miscalibrated under shifts. To improve robustness, recent methods incorporate uncertainty estimation or sample selection, such as reweighting or filtering unreliable samples [Lee et al., 2024, Yang et al., 2024a, Niu et al., 2022]. For cross-modal retrieval, several works study adaptation under query distribution shift and reveal challenges such as reduced query discriminability and enlarged modality gaps [Yang et al., 2024b, Li et al., 2025a].

2.3 UNCERTAINTY CALIBRATION

Uncertainty calibration aims to improve model reliability in a training-free or test-time manner. Existing methods address this problem from different perspectives, including confidence-based calibration [Tan et al., 2025], feature-based calibration [Yoon et al., 2024], and attention-based calibration [Kang et al., 2025]. For example, EATA-C [Tan et al., 2025] estimates uncertainty through prediction discrepancies between the full network and sampled sub-networks; DEC [Amin and Kim, 2025] alleviates overcertainty in TTA with sample-specific temperature scaling; AdaPGC [Xu et al., 2026] models class-conditional distributions with modality- and class-specific Gaussians; C-TPT [Yoon et al., 2024] improves CLIP calibration by dispersing text features; CalibCLIP [Kang et al., 2025] reduces the influence of background and generic words; and Over the Top-1 [Gomez, 2025] estimates retrieval uncertainty via top-1 consistency.

Despite these advances, existing studies have rarely exploited inter-candidate similarity to determine whether the model is under-confident or over-confident, or to guide confidence calibration during test-time adaptation.

3 METHODOLOGY

3.1 PROBLEM SETUP

This paper addresses the *Test-Time Cross-Modal Retrieval Adaptation* problem, which considers the practical scenario where a source-domain pre-trained model encounters distribution shifts when deployed for cross-modal retrieval tasks. Let the query modality be denoted as Q and the gallery modality as G. We denote the multimodal retrieval model by ϕ, which consists of two modality-specific encoders, namely ϕ^Q and ϕ^G. Given a cross-modal retrieval dataset $\mathcal{D} = \left\{ \mathbf{X}^Q = \left\{ \mathbf{x}_i^Q \right\}_{i=1}^{N^Q}, \mathbf{X}^G = \left\{ \mathbf{x}_j^G \right\}_{j=1}^{N^G} \right\}$, where $\mathbf{x}_i^Q$ and $\mathbf{x}_j^G$ denote the i-th query sample and the j-th gallery sample. A distribution shift exists between the source domain $\mathcal{S}$ and the target domain $\mathcal{T}$, i.e., $(P_{\mathcal{S}}(\mathbf{X}^Q, \mathbf{X}^G) \neq P_{\mathcal{T}}(\mathbf{X}^Q, \mathbf{X}^G))$. Consequently, the pre-trained source model's performance degrades when deployed in the target domain. Following [Wang et al., 2021], we consider an online protocol where queries arrive in mini-batches while the gallery remains fixed, and we update only the normalization parameters of the query encoder during adaptation.

3.2 UNIFORMITY AND MODALITY GAP

In CMR, intra-modality uniformity and cross-modality discrepancy are crucial for semantic alignment, especially under distribution shifts. The formal definitions of these two properties are provided in the Appendix A. Following Li et al. [2025a], we jointly model these two factors to measure the discrepancy between sample and model distributions.

Definition 1 (Source-likeness Indicator)**.** *Given modality-specific encoder ϕ^m for $m \in \{Q, G\}$, the normalized feature is $\mathbf{z}_i^m = \frac{\phi^m(\mathbf{x}_i^m)}{\|\phi^m(\mathbf{x}_i^m)\|_2}$, and the centroid is $\boldsymbol{\mu}^m = \frac{1}{N^m} \sum_{i=1}^{N^m} \mathbf{z}_i^m$. The source-likeness indicator (SI) is defined as*

$$SI := 2\|\mathbf{z}_i^Q - \mathbf{z}_i^{G'}\|_2 - \left(\|\mathbf{z}_i^Q - \boldsymbol{\mu}^Q\|_2 + \|\mathbf{z}_i^{G'} - \boldsymbol{\mu}^{G'}\|_2 \right), \tag{1}$$

where $\mathbf{z}^{G'}$ denotes the nearest-neighbor gallery feature for the query.

The source-domain-like set $\mathcal{D}_{SI}$ consists of the query–candidate pairs with the lowest $P \times B$ SI values, where B is the batch size and P is a hyperparameter that controls the number of samples retained in $\mathcal{D}_{SI}$. The detailed procedure for constructing $\mathcal{D}_{SI}$ is provided in Appendix C. We propose two base objectives based on $\mathcal{D}_{SI}$ to regularize both uniformity and modality gap. The intra-modality uniformity loss is defined as:

$$\mathcal{L}_{\text{uni}} = \frac{1}{B} \sum_{i=1}^{B} \exp\left(-\frac{\|\mathbf{z}_i^Q - \overline{\mathbf{z}^Q}\|_2^2}{t} \right), \tag{2}$$

where $\overline{\mathbf{z}^Q} = \frac{1}{B} \sum_{i=1}^{B} \mathbf{z}_i^Q$ denotes the batch mean of query embeddings, and $t = 10$ controls the regularization strength. The modality gap on $\mathcal{D}_{SI}$ is estimated as:

$$\overline{\Delta g} = \left\| \frac{1}{|\mathcal{D}_{SI}|} \sum_{i=1}^{|\mathcal{D}_{SI}|} \mathbf{z}_i^Q - \frac{1}{|\mathcal{D}_{SI}|} \sum_{i=1}^{|\mathcal{D}_{SI}|} \mathbf{z}_i^{G'} \right\|_2. \tag{3}$$

Then, the modality gap learning loss is defined as:

$$\begin{aligned} \mathcal{L}_{\text{emg}} &= \left(\bar{g} - \overline{\Delta g} \right)^2 \\ &= \left(\left\| \frac{1}{B} \sum_{i=1}^{B} \mathbf{z}_i^Q - \frac{1}{B} \sum_{i=1}^{B} \mathbf{z}_i^{G'} \right\|_2 - \overline{\Delta g} \right)^2. \end{aligned} \tag{4}$$

By aligning $\bar{g}$ with $\overline{\Delta g}$, the model controls cross-modality discrepancy and prevents over-alignment.

3.3 SAMPLE RE-WEIGHTED ENTROPY MINIMIZATION

To reduce noise in query predictions, we use a sample reweighted entropy minimization technique that suppresses unreliable samples based on their estimated modality gap. Formally, for an incoming batch of B queries and the corresponding nearest-neighbor candidate features $\mathbf{Z}^{G'}$, we compute the similarity logits as $\mathbf{S} = \frac{\mathbf{Z}^Q (\mathbf{Z}^{G'})^\top}{\tau}$, where $\tau > 0$ is a temperature hyperparameter. For each query i,

we define a candidate-probability vector $\mathbf{p}_i = \mathrm{softmax}(\mathbf{s}_i)$, and the (Shannon) entropy $\mathcal{H}(\mathbf{p}_i) = -\sum_{k=1}^{B} \mathbf{p}_{i,k} \log \mathbf{p}_{i,k}$. The sample re-weighted entropy minimization target is formulated using the estimated sample-wise modality gap $\hat{g}_i = \left\| \mathbf{z}_i^Q - \mathbf{z}_i^{G'} \right\|_2$ and reference modality gap $\overline{\Delta g}$ obtained from source-domain-like samples.

$$\mathcal{L}_{\mathrm{sem}} = \frac{1}{B} \sum_{i=1}^{B} \min\left(\frac{\overline{\Delta g}}{\hat{g}_i + \epsilon}, 1 \right) \mathcal{H}(\mathbf{p}_i). \tag{5}$$

where ϵ is a small constant for numerical stability. This weighting assigns smaller weights to samples whose estimated cross-modal gap $\hat{g}_i$ exceeds the source-domain-like reference Δg, thereby suppressing unreliable high-confidence updates.

3.4 SEMANTIC SEPARATION MARGIN

The similarity between normalized feature representations guides the retrieval of relevant images in response to a given query. Let $R(\mathbf{Z}^Q, \mathbf{Z}^{G'})$ denote the retrieval ranking score list for a batch of queries, defined as

$$R(\mathbf{Z}^Q, \mathbf{Z}^{G'}) = \mathrm{sort}_{\downarrow}\left[\mathbf{Z}^Q (\mathbf{Z}^{G'})^\top \right], \tag{6}$$

where $\mathrm{sort}_{\downarrow}$ sorts similarity scores in descending order. Accordingly, $R_i^{(k)}$ denotes the k-th largest similarity score for the i-th query. For notational convenience, we use R to denote $R(\mathbf{Z}^Q, \mathbf{Z}^{G'})$ when the arguments are clear from context. Unlike classification where class labels are mutually exclusive, cross-modal retrieval operates over a ranked candidate set in which top-ranked items often exhibit semantic overlap. This observation suggests that confidence surrogates for retrieval should account for the local similarity structure among top-ranked candidates, rather than relying solely on a single best-match score.

Definition 2 (Semantic Separation Margin)**.** *For a query i, define the **similarity gap** between the best and second-best matches in the candidate set as*

$$\hat{\eta}_i = R_i^{(1)} - R_i^{(2)}, \tag{7}$$

where $R_i^{(1)}$ and $R_i^{(2)}$ are the similarity scores of the best and second-best matches, respectively. Given the source-domain-like sample set $\mathcal{D}_{SI}$, the reference (source-domain-like) Semantic Separation Margin is

$$\overline{\Delta\eta} = \frac{1}{|\mathcal{D}_{SI}|} \sum_{(\mathbf{z}_i^Q, \mathbf{z}_i^{G'}) \in \mathcal{D}_{SI}} \left(R_i^{(1)} - R_i^{(2)} \right), \tag{8}$$

where $|\mathcal{D}_{SI}|$ denotes the number of samples in $\mathcal{D}_{SI}$.

$\overline{\Delta\eta}$ measures the similarity difference between the best and second-best retrieval results for source-domain-like reliable samples, thereby characterizing the semantic separation between the two highest-ranked candidates.

Rationale. Estimating the full neighborhood similarity structure can be computationally expensive. We therefore focus on the most decision-relevant local structure: the separation between the top-1 and top-2 candidates. This choice provides a lightweight surrogate of local ranking stability while reducing computational overhead.

3.5 UNIFIED CONFIDENCE ADJUSTMENT

Since $\hat{\eta}_i$ is a noisy, sample-level surrogate affected by hard negatives and mini-batch sampling, we use the batch mean $\bar{\eta} = \frac{1}{B}\sum_{i=1}^{B} \hat{\eta}_i$ as a low-variance estimator of $\mathbb{E}[\hat{\eta}]$ to identify the confidence state. We further introduce a tolerance band to avoid unstable decisions when $\bar{\eta}$ is close to the reference margin $\overline{\Delta\eta}$.

Definition 3 (Confidence State Identification)**.** *Given a batch of queries, we define the empirical average similarity gap as $\bar{\eta}$:*

$$\bar{\eta} = \frac{1}{B} \sum_{i=1}^{B} \hat{\eta}_i. \tag{9}$$

The confidence state of the batch is determined by comparing $\bar{\eta}$ with the reference margin $\overline{\Delta\eta}$:

$$\textit{state} = \begin{cases} \textit{over-confident}, & \bar{\eta} > \overline{\Delta\eta} + \gamma, \\ \textit{under-confident}, & \bar{\eta} < \overline{\Delta\eta} - \gamma, \\ \textit{well-calibrated}, & |\bar{\eta} - \overline{\Delta\eta}| \le \gamma, \end{cases} \tag{10}$$

where γ controls decision stability (we set $\gamma = \kappa\widehat{\sigma}_\eta/\sqrt{B}$), with $\widehat{\sigma}_\eta$ being the sample standard deviation of $\{\hat{\eta}_i\}_{i=1}^{B}$. Statistical interpretation for Definition 3 is detailed in Appendix D. When $\bar{\eta} > \overline{\Delta\eta} + \gamma$, the model exhibits an over-confident state, i.e., it assigns overly high confidence to the best match relative to the predicted semantic margin. In this scenario, the similarity gap between the best and second-best candidates should be narrowed to avoid over-confidence. In contrast, when $\bar{\eta} < \overline{\Delta\eta} - \gamma$, the model is under-confident, assigning insufficient confidence to the best match or unduly high confidence to the second-best candidate. In this case, the similarity gap should be enlarged to enhance discriminability. To calibrate the margin toward the source-domain-like level, we introduce a unified confidence regulation loss:

$$\mathcal{L}_{\mathrm{uca}} = \max\left(0, \left| \frac{1}{B} \sum_{i=1}^{B} \hat{\eta}_i - \overline{\Delta\eta} \right| - \gamma \right), \tag{11}$$

which dynamically steers the similarity margin toward the reference confidence level only when the empirical gap falls outside the stable zone.

Gradient analysis. Let $\bar{\eta} = \frac{1}{B}\sum_{i=1}^{B}\hat{\eta}_i$ and $d = \bar{\eta} - \overline{\Delta\eta}$. The loss becomes $\mathcal{L}_{\text{uca}} = \max(0, |d| - \gamma)$. For $|d| > \gamma$, we have

$$\frac{\partial \mathcal{L}_{\text{uca}}}{\partial \bar{\eta}} = \text{sign}(d), \qquad \frac{\partial \bar{\eta}}{\partial \hat{\eta}_i} = \frac{1}{B}. \tag{12}$$

Incorporating the indicator function $\mathbb{I}(|d| > \gamma)$, we have $\frac{\partial \mathcal{L}_{\text{uca}}}{\partial \hat{\eta}_i} = \frac{1}{B}\text{sign}(d)\mathbb{I}(|d| > \gamma)$. Because $\hat{\eta}_i = R_i^{(1)} - R_i^{(2)}$, the gradients w.r.t. the best and second-best similarities are

$$\begin{aligned} \frac{\partial \mathcal{L}_{\text{uca}}}{\partial R_i^{(1)}} &= \frac{1}{B}\text{sign}(d)\mathbb{I}(|d| > \gamma), \\ \frac{\partial \mathcal{L}_{\text{uca}}}{\partial R_i^{(2)}} &= -\frac{1}{B}\text{sign}(d)\mathbb{I}(|d| > \gamma). \end{aligned} \tag{13}$$

Over-Confident State. When the model is over-confident, i.e., $\bar{\eta} > \overline{\Delta\eta} + \gamma$ (equivalently $d > \gamma$), we have $\text{sign}(d) = +1$ and the indicator is active:

$$\frac{\partial \mathcal{L}_{\text{uca}}}{\partial R_i^{(1)}} = \frac{1}{B} > 0, \qquad \frac{\partial \mathcal{L}_{\text{uca}}}{\partial R_i^{(2)}} = -\frac{1}{B} < 0. \tag{14}$$

Under gradient descent, $R_i^{(1)}$ is decreased and $R_i^{(2)}$ is increased, so the margin $\hat{\eta}_i = R_i^{(1)} - R_i^{(2)}$ is reduced, suppressing excessive confidence.

Under-Confident State. When the model is under-confident, i.e., $\bar{\eta} < \overline{\Delta\eta} - \gamma$ (equivalently $d < -\gamma$), we have $\text{sign}(d) = -1$ and the indicator is active:

$$\frac{\partial \mathcal{L}_{\text{uca}}}{\partial R_i^{(1)}} = -\frac{1}{B} < 0, \qquad \frac{\partial \mathcal{L}_{\text{uca}}}{\partial R_i^{(2)}} = \frac{1}{B} > 0. \tag{15}$$

Under gradient descent, $R_i^{(1)}$ is increased and $R_i^{(2)}$ is decreased, so the margin $\hat{\eta}_i$ is enlarged, enhancing discriminability when confidence is insufficient.

Well-Calibrated State. When the model is well-calibrated ($|d| \leq \gamma$), the gradients are exactly 0, preventing unnecessary and potentially destabilizing updates. Overall, $\mathcal{L}_{\text{uca}}$ functions as a batch-wise semantic margin stabilizer, adaptively calibrating confidence by driving the average similarity gap toward $\overline{\Delta\eta}$ while maintaining retrieval discriminability.

3.6 MODEL OPTIMIZATION

Following the standard online test-time adaptation protocol, we adapt the model in a sequential, batch-wise manner. During adaptation, we freeze all parameters except parameters in the normalization layers of the query encoder, and update them by minimizing the following objective on each incoming batch:

$$\mathcal{L} = \lambda_{\text{uca}}\mathcal{L}_{\text{uca}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{uni}}\mathcal{L}_{\text{uni}} + \lambda_{\text{emg}}\mathcal{L}_{\text{emg}}, \tag{16}$$

where λ_{uca}, λ_{sem}, λ_{uni}, and λ_{emg} are trade-off hyperparameters that balance the contributions of different loss terms.

4 EXPERIMENTS

We evaluate the proposed UCA under two test-time distribution shift scenarios: (i) cross-dataset zero-shot transfer and (ii) natural corruptions, and report results for both image-to-text and text-to-image retrieval on four benchmarks, using CLIP and BLIP as backbones. Analyses on confidence calibration and ablations emphasize the contribution of UCA.

4.1 EXPERIMENT SETUP

Datasets. We conduct experiments on four standard benchmarks for cross-modal retrieval: 1) ***MSCOCO*** [Lin et al., 2014], a large-scale dataset widely used for cross-modal retrieval. We evaluate on the COCO 2014 test split, which contains 5,000 images, each associated with five captions; 2) ***Flickr30K*** [Plummer et al., 2015], a cross-modal retrieval dataset collected from natural scenes. We use the test split containing 1,000 images, each paired with five descriptive captions; 3) ***Fashion-Gen*** [Rostamzadeh et al., 2018], a cross-modal retrieval dataset from the e-commerce domain. We use the test split containing 32,528 image-text pairs; 4) ***NoCaps*** [Agrawal et al., 2019], a benchmark derived from the OpenImages dataset with images spanning 600 object classes. We evaluate on the test set consisting of 648 *in-domain (ID)* images, 2,938 *near-domain (ND)* images, and 914 *out-of-domain (OD)* images; each image is paired with 10 captions. We adopt two testing protocols, namely, image-to-text retrieval (a.k.a. TR) and text-to-image retrieval (a.k.a. IR). We primarily report Recall@1 (R@1; denoted as TR@1 and IR@1 for image-to-text and text-to-image retrieval, respectively)

Test-time Settings. We consider two distribution-shift settings commonly used in cross-modal retrieval: 1) ***Zero-Shot Transfer.*** We directly deploy a pretrained backbone on a target dataset without additional training, and perform online adaptation on the test stream. We denote this protocol as *B2DATASET* (e.g., B2COCO), indicating a backbone transferred to dataset X; 2) ***Natural Corruption.*** We first finetune the backbone on the in-domain training split and then apply corruptions at test time. For images, we use 16 corruption types with severity $s_{\text{img}} \in \{1, 2, 3, 4, 5\}$. For texts, we use 15 corruption types involving character-level, word-level and sentence-level perturbations, with fixed severities $s_{\text{char}} = 7$, $s_{\text{word}} = 2$ and $s_{\text{sen}} = 1$.

Baselines. We compare UCA with representative test-time adaptation methods, including entropy-minimization approaches (TENT [Wang et al., 2021], EATA [Niu et al., 2022]), pseudo-labeling methods (PL [Lee et al., 2013], SHOT [Liang et al., 2020]), augmentation-driven methods (SAR [Niu et al., 2023], DeYO [Lee et al., 2024]), and retrieval-specific TTA methods (READ [Yang et al., 2024b], TCR [Li et al., 2025a]).

Table 1: **Comparison results under image natural corruptions** with maximum severity level ($s_{\text{img}} = 5$) regarding the Recall@1 metric. All methods are evaluated with BLIP ViT-B/16 on Flickr30K and MSCOCO. The best and second-best results are highlighted in **bold** and underline, respectively.

Dataset	Method	Noise				Blur				Weather				Digital				Avg.
		Gauss.	Shot	Impul.	Speckle	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	
Flickr	ViT-B/16	49.8	56.6	50.3	71.6	53.1	84.5	47.4	15.5	66.4	80.4	79.5	85.5	60.6	53.3	35.1	80.3	60.6
	Tent	54.9	54.9	54.3	73.1	53.3	85.3	47.9	1.6	67.2	80.9	79.6	86.8	63.6	53.4	35.4	81.4	60.9
	PL	51.2	60.3	49.2	73.7	29.9	86.2	28.9	2.3	69.7	83.1	81.9	87.4	66.1	57.8	38.4	81.4	59.2
	SHOT	54	63.3	56.2	76.6	39	87.8	21.7	1.5	72.6	84.9	84	88.9	70.6	63.1	32.9	82.6	61.2
	EATA	55.5	60.5	55.8	75.8	64.6	86.2	52.2	8.5	72	83.7	82.5	87.9	68.4	60.1	45.9	81.6	65.1
	SAR	54.8	62.5	55.6	75.2	48.3	87.2	34.8	15.5	71.9	83.1	82.2	87.9	68.2	60.3	42.2	81.4	63.2
	READ	50.1	58.2	52.2	74.8	63.7	87.0	55.1	2.2	71.7	83.8	81.9	87.7	67.4	62.3	42.5	81.4	63.9
	DeYO	55.4	62	56.3	76.2	63.8	86.3	50.3	3.2	73.1	84.1	83.2	88.6	70.1	63.1	46.8	81.3	65.2
	TCR	62.0	66.6	61.4	80.0	68.1	87.9	65.2	39.9	78.2	85.2	85.7	89.5	75.1	73.1	56.8	83.3	72.4
	Ours	**67.4**	**71.4**	**67.7**	**83.0**	**72.5**	**88.8**	**72.8**	**52.9**	**82.6**	**87.4**	**87.0**	**90.0**	**79.8**	**83.1**	**60.2**	**84.8**	**77.0**
COCO	ViT-B/16	43.4	46.3	43.2	57.3	43.3	68.0	39.7	8.4	32.3	52.2	57.0	66.8	36.0	41.3	20.6	63.7	45.0
	Tent	41.6	40.5	37.9	54	44.7	65.1	39.6	8.3	31.9	48.7	56.3	66.5	31.8	40.3	19.2	62.3	43.0
	PL	43.2	44.9	29.9	57.3	13.3	71.8	31	1.1	25	50.5	57.9	70.3	14.4	38.5	10.4	65.8	39.1
	SHOT	47.2	42.9	34.4	63.8	10.8	71.9	6.8	0.5	8.3	52.2	51.8	70.6	7.9	39.8	4.7	67.7	36.3
	EATA	41.4	50.3	35.7	63.1	49.8	72.2	46.2	6.9	45.6	56.7	62.5	71.4	43.6	51.3	25.6	67.0	49.3
	SAR	42.3	51.5	37.5	61.8	40.3	71.5	32.8	6.2	38	56.2	59.1	70.6	31.1	53.5	17.5	66.4	46.0
	READ	45.8	48.4	37.2	59.9	44.5	71.8	46.6	11.5	39.9	49.9	58.4	70.3	35.8	45	18.8	66.2	46.9
	DeYO	47.9	53.5	46.8	63.4	42.9	72.1	36.7	3.2	37.5	59.7	66.4	71.2	40.3	49.0	13.1	67.6	48.2
	TCR	53.2	56.2	54.8	64.4	58.0	73.0	56.4	32.2	56.5	64.1	71.0	**72.9**	57.9	63.7	41.8	68.4	59.0
	Ours	**55.2**	**57.4**	**55.5**	**64.7**	**58.1**	**73.0**	**58.5**	**38.0**	**59.1**	**65.1**	**71.6**	**72.9**	**61.5**	**65.7**	**48.1**	**68.7**	**60.8**

Implementation Details. Following Qiu et al. [2024], we adopt two widely used foundation models for cross-modal tasks, namely CLIP [Radford et al., 2021] and BLIP [Li et al., 2022], as the source models. During adaptation, UCA optimizes its objective for each incoming mini-batch of queries, with the batch size set to 64. For TR, we use a unified learning rate of 7×10^{-4} for natural corruptions and 5×10^{-4} for zero-shot transfer. For IR, we set the learning rate to 5×10^{-4} when using CLIP and 3×10^{-5} when using BLIP. UCA updates only the parameters in the normalization layers of the query-specific encoder using the AdamW optimizer. The temperature hyperparameter τ is fixed to 0.02. Unless stated otherwise, λ_{sem}, λ_{uni}, and λ_{emg}, and λ_{uca} are set to 1. All experiments are conducted on NVIDIA RTX 4090 GPUs with 24GB memory. The proposed UCA method is implemented using PyTorch 1.13.1 with CUDA 11.7. The code is available at UAI-UCA.

4.2 BENCHMARK RESULTS

Natural Corruptions. Table 1 reports results under image natural corruptions at the maximum severity level on Flickr30K and MSCOCO. Our method achieves the highest average R@1 across all 16 corruption types on both datasets. Specifically, it outperforms READ and DeYO by 13.1% and 11.8% on Flickr30K and by 14.0% and 12.7% on MSCOCO, respectively. Compared to the recent state-of-the-art method TCR, our approach consistently yields higher R@1, with gains of 4.6% on Flickr30K and 1.8% on MSCOCO, further demonstrating the effectiveness of the proposed method. Table A10 presents results under text natural corruptions at the maximum severity level. We observe that UCA achieves the best average Recall@1 on both Flickr30K (64.0%) and MSCOCO (42.6%), providing additional evidence of its robustness.

Zero-Shot Transfer. We report the zero-shot transfer results in Table 2. We evaluate our method using CLIP ViT-B/16 and BLIP ViT-B/16 across four benchmarks. The proposed approach achieves the highest average R@1 over six subtasks. Specifically, it outperforms READ and DeYO by 6.2% and 3.9% on image-to-text retrieval and by 2.2% and 2.1% on text-to-image retrieval, respectively. These results demonstrate the robustness of our method under zero-shot distribution shifts.

4.3 CONFIDENCE ADJUSTMENT ANALYSIS

To verify whether our method adjusts model confidence as intended, we visualize reliability for the cross-modal matching task in Fig. 2. We use the similarity gap between the best and second-best matches as a proxy for confidence. The gray bars denote the expected values, computed from the best/second-best gaps on source-domain-like samples, which serve as the target calibration level we aim to achieve. The colored bars denote the predicted values produced by UCA, computed from the best/second-best gaps on all samples, reflecting the model's current confidence behavior. Our central objective is to make the predicted similarity gap in each confidence bin align with the expected one as closely as possible, thereby realizing *unified confidence adjustment*.

Image to Text (Columns 1 and 3). Before calibration (first row), many bins with high expected confidence contain almost no corresponding predicted bars, indicating pro-

Table 2: **Comparison results under zero-shot transfer regarding the Recall@1 metric.** All methods are evaluated with CLIP ViT-B/16 and BLIP ViT-B/16 on four benchmarks.

Query	Method	B2Flickr		B2COCO		B2Fashion		B2Nocaps(ID)		B2Nocaps(ND)		B2Nocaps(OD)		Avg.
		CLIP	BLIP	CLIP	BLIP	CLIP	BLIP	CLIP	BLIP	CLIP	BLIP	CLIP	BLIP	
TR@1	ViT-B/16	80.2	70.0	52.5	59.3	8.5	19.9	84.9	88.2	75.4	79.3	73.8	81.9	64.5
	Tent	81.4	81.9	48.8	61.7	5.6	14.1	85.1	88.5	74.6	82.6	71.8	82.7	64.9
	EATA	80.4	82.3	52.1	64.2	8.1	12.8	84.7	87.8	75.1	82.8	74.1	81.5	65.5
	SAR	80.3	81.7	51.8	63.5	8.0	17.9	84.7	88.2	75.4	81	73.7	81.2	65.6
	READ	80.6	80	46.0	62.1	5.8	5.6	85.1	87.3	75.0	80.6	73.5	80.7	63.5
	DeYO	80.1	83.5	51.5	65.0	6.9	12.2	84.4	89.2	75.1	83.7	73.2	84.3	65.8
	TCR	82.4	86.8	52.9	68.9	8.9	23.6	85.1	89.7	75.7	86.3	74.4	87.2	68.5
	Ours	**83.3**	**89.8**	**54.9**	**69.7**	**10.2**	**24.5**	**85.9**	**91.1**	**77.4**	**87.3**	**75.7**	**87.5**	**69.8**
IR@1	ViT-B/16	61.5	68.3	33.0	45.4	13.2	26.1	61.4	74.9	49.2	63.6	55.8	67.8	51.7
	Tent	64.0	68.5	27.6	41.7	10.7	26.1	61.7	75.4	48.6	64.1	56.1	68.9	51.1
	EATA	63.4	69.4	34.8	47.9	12.0	25.2	62.0	75.1	52.3	63.9	56.9	67.9	52.6
	SAR	62.2	68.3	33.9	46.6	13.3	26.1	61.3	75.6	51.3	65.4	56.1	69.3	52.5
	READ	64.4	69.9	35.7	46.4	11.2	24.1	63.0	75.1	52.1	63.9	57.0	67.9	52.6
	DeYO	64.0	69.9	33.4	47.3	10.9	24.1	62.2	75.6	52.0	65.7	57.3	69.4	52.7
	TCR	64.8	70.3	36.5	48.9	14.0	30.3	63.5	76.0	54.0	66.1	58.0	69.5	54.3
	Ours	**66.2**	**70.5**	**36.8**	**48.9**	**14.1**	**30.9**	**65.4**	**76.2**	**54.5**	**66.2**	**58.6**	**69.8**	**54.8**

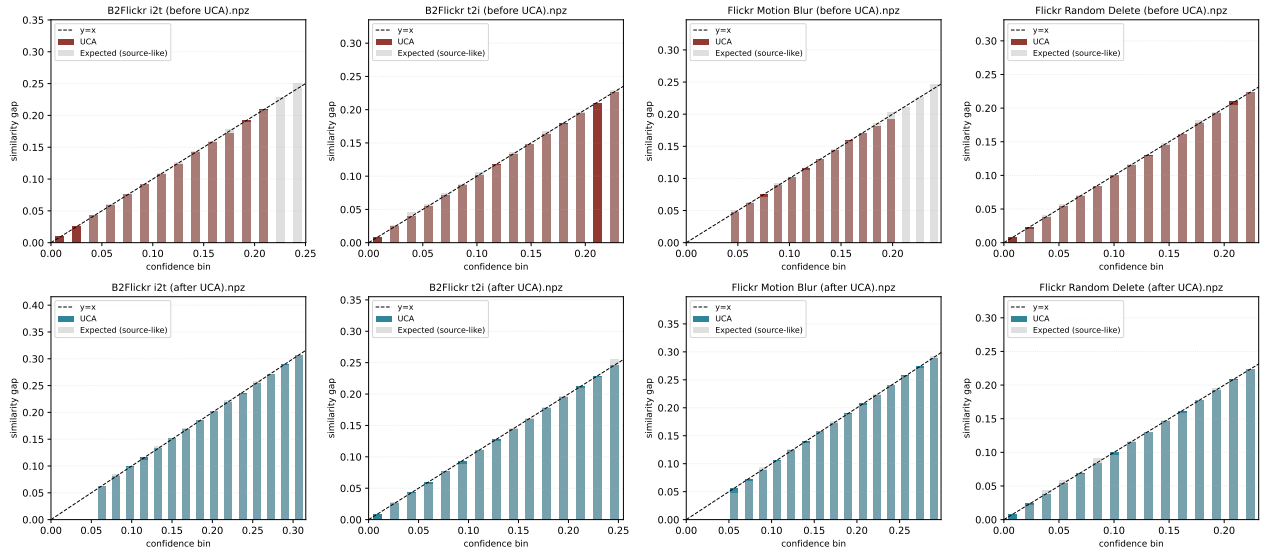

Figure 2: **Reliability diagrams for cross-modal retrieval before and after UCA** on four benchmarks: zero-shot transfer to B2Flickr for image-to-text (i2t) and text-to-image (t2i), Motion Blur image corruption, and Random Delete text corruption. Gray bars denote the Semantic Separation Margin, while red and green bars indicate the similarity gaps before and after UCA, respectively. The $y = x$ line represents ideal calibration.

nounced under-confidence and insufficient separation. After applying UCA (third row), the colored bars shift toward the gray expected distribution across bins. The improvement is most evident in the medium-to-high confidence region, where the gap is effectively increased and becomes more consistent with the expected values. These results suggest that UCA mitigates under-confidence in i2t and aligns the matching separation with the target calibration level.

Text to Image (Columns 2 and 4). Before calibration (first row), the model already follows the expected distribution more closely, implying that its native confidence is relatively usable. This observation is consistent with our empirical finding that t2i tasks offer less room for improvement and benefit less from entropy/confidence-based optimization than i2t tasks. Nevertheless, we still observe a few bins where predicted exceeds expected, revealing localized over-confidence (overly large gaps). After calibration, such deviations are suppressed, and the overlap between predicted and expected increases, indicating that UCA effectively constrains over-confident regions.

Overall, UCA consistently moves the predicted similarity-gap distribution toward the source-like expectation, showing stable performance across retrieval directions and distribution shifts.

Table 3: Ablation study of different loss components. The best and second-best results are highlighted in bold and underlined, respectively.

$\mathcal{L}_{uni}$	$\mathcal{L}_{emg}$	$\mathcal{L}_{uca}$	$\mathcal{L}_{sem}$	TR@1	MG	CG	MD	IR@1	MG	CG	MD	Avg. R@1	ΔR@1
✓	✓	✓	✓	**69.6**	**0.591**	0.237	**0.955**	**48.9**	**0.632**	0.163	**0.961**	**59.3**	—
✗	✓	✓	✓	69.4	0.584	0.236	0.951	48.3	0.634	0.162	0.956	58.8	0.444
✗	✗	✓	✓	69.3	0.692	**0.239**	0.951	47.6	0.757	**0.164**	0.956	58.4	0.386
✗	✗	✗	✓	64.1	0.686	0.171	0.837	46.9	0.751	0.137	0.904	55.5	2.920

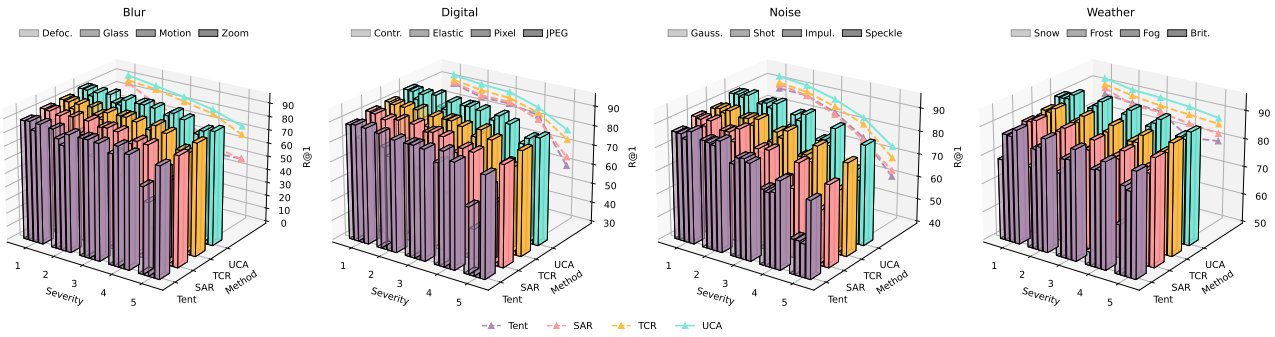

Figure 3: Image-to-text retrieval performance on the Flickr benchmark under natural image corruptions across all severity levels $s_{\text{img}} \in \{1, 2, 3, 4, 5\}$, evaluated using the R@1 metric. The 16 corruption types are grouped into four categories: *blur*, *digital*, *noise*, and *weather*. Within each category, the four sub-corruptions are visualized using bar charts with varying transparency to indicate their respective R@1 performance. The average performance of each method across the five severity levels within each category is represented by colored lines.

4.4 FURTHER ANALYSIS

Trade-off Analysis. We analyze the impact of the four loss weights on performance on the Flickr dataset. All weights are swept over $(0, 1]$ with a step size of 0.1. We group the hyperparameters as $(\lambda_{\text{uca}}, \lambda_{\text{sem}})$ and $(\lambda_{\text{uni}}, \lambda_{\text{emg}})$, and visualize the results with 3D heatmaps. The results show that emphasizing confidence-aware regularization (larger λ_{uca}) while keeping the uniformity constraints balanced yields consistently strong performance. Detailed analysis can be found in Appendix B.1.

Ablation Study. We conduct ablation experiments on both image-to-text and text-to-image retrieval to analyze the contribution of each loss term. Four evaluation metrics are reported in Table 3. Overall, the effects of different loss terms are consistent across text and image retrieval. Removing UCA leads to a substantial performance drop, demonstrating the contribution of the proposed UCA. Detailed discussions are provided in the Appendix B.2.

Results on All Severity Settings. We conduct additional experiments which encompasses natural image corruptions across all severity levels. The results shown in Fig. 3 demonstrate that UCA consistently maintains strong performance under varying degrees of query distribution shifts, further confirming the effectiveness of UCA in handling diverse and progressively intensified natural perturbations.

Efficiency Comparisons. We conduct additional experiments to analyze the efficiency of UCA. We adopt the pre-trained CLIP model as the source model and evaluate zero-shot transfer settings across six datasets. The results are shown in Table A11. We measure the GPU runtime during the test-time adaptation stage. Experimental results show that UCA achieves significantly higher adaptation efficiency than augmentation-based methods, including DeYO and SAR, reducing the runtime by 18 seconds and 25 seconds, respectively. Compared to entropy-minimization-based approaches such as TENT and TCR, UCA achieves consistently better robustness while incurring only a negligible additional time overhead of approximately 2 seconds.

5 CONCLUSION

To the best of our knowledge, this work presents the first systematic investigation of confidence behavior in cross-modal retrieval under test-time distribution shifts. By incorporating the semantic neighborhood structure within the candidates, we propose a *confidence state identification* criterion and develop a *unified confidence adjustment* strategy based on the score margin between the best and second-best matches. The proposed joint optimization framework achieves state-of-the-art performance under both zero-shot transfer and naturally corrupted test-time adaptation settings, highlighting the valuable role of confidence-aware adjustment.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62276272) and the Training Program for Excellent Young Innovators of Changsha (Grant No. KQ2009009).

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2019.

Fin Amin and Jung-Eun Kim. The over-certainty phenomenon in modern test-time adaptation algorithms. *Transactions on Machine Learning Research*, 2025.

George Casella and Roger Berger. *Statistical Inference*. Chapman and Hall/CRC, 2024.

Siyuan Duan, Yuan Sun, Dezhong Peng, Zheng Liu, Xiaomin Song, and Peng Hu. Fuzzy multimodal learning for trusted cross-modal retrieval. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 20747–20756, 2025.

Lluis Gomez. Over the top-1: Uncertainty-aware cross-modal retrieval with clip. In Silvia Chiappa and Sara Magliacane, editors, *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, volume 286, pages 1521–1532, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of the International Conference on Learning Representations*, 2019.

Hailang Huang, Zhijie Nie, Ziqiao Wang, and Ziyu Shang. Cross-modal and uni-modal soft-label alignment for image-text retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18298–18306, 2024.

Bin Kang, Bin Chen, Junjie Wang, Yulin Li, Junzhi Zhao, Junle Wang, and Zhuotao Tian. CalibCLIP: Contextual calibration of dominant semantics for text-driven image retrieval. In *Proceedings of the ACM International Conference on Multimedia*, pages 5140–5149, 2025.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the International Conference on Machine Learning*, pages 5637–5664, 2021.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Proceedings of the International Conference on Machine Learning Workshop*, 2013.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *Proceedings of the International Conference on Learning Representations*, 2024.

Erich Leo Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer, 2005.

Haobin Li, Peng Hu, Qianjun Zhang, Xi Peng, Mouxing Yang, et al. Test-time adaptation for cross-modal retrieval with query shift. In *Proceedings of the International Conference on Learning Representations*, 2025a.

Jiaxing Li, Wai Keung Wong, Lin Jiang, Xiaozhao Fang, Shengli Xie, and Yong Xu. CKDH: CLIP-based knowledge distillation hashing for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6530–6541, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the International Conference on Machine Learning*, volume 162, pages 12888–12900, 2022.

Yuan Li, Liangli Zhen, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Deep evidential hashing for trustworthy cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18566–18574, 2025b.

Zheng Li, Caili Guo, Xin Wang, Hao Zhang, and Lin Hu. Multi-view visual semantic embedding for cross-modal image–text retrieval. *Pattern Recognition*, 159:111088, 2025c.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the International Conference on Machine Learning*, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 2014.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the International Conference on Machine Learning*, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *Proceedings of the International Conference on Learning Representations*, 2023.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the International Conference on Computer Vision*, 2015.

Qibing Qin, Lei Wu, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep semantic-consistent penalizing hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 27:4613–4626, 2025.

Jielin Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. Benchmarking robustness of multimodal image-text models under distribution shift. *Journal of Data-centric Machine Learning Research*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the International Conference on Machine Learning*, volume 139, pages 8748–8763, 2021.

Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018.

Mingkui Tan, Guohao Chen, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Peilin Zhao, and Shuaicheng Niu. Uncertainty-calibrated test-time model adaptation without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6274–6289, 2025.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *Proceedings of the International Conference on Learning Representations*, 2021.

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Cross-modal retrieval: A systematic review of methods and future directions. *Proceedings of the IEEE*, 112(11):1716–1754, 2024.

Yijing Wang, Xu Tang, Jingjing Ma, Xiangrong Zhang, Fang Liu, and Licheng Jiao. Cross-modal remote sensing image–text retrieval via context and uncertainty-aware prompt. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):11384–11398, 2025.

Jinglin Xu, Yi Li, Chuxiong Sun, Xiao Xu, Jiangmeng Li, and Fanjiang Xu. Multi-modal test-time adaptation via adaptive probabilistic gaussian calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 30268–30277, 2026.

Hao Yang, Min Wang, Zhengfei Yu, Hang Zhang, Jinshen Jiang, and Yun Zhou. Confidence-based and sample-reweighted test-time adaptation. *Knowledge-Based Systems*, 283:111164, 2024a.

Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. Test-time adaptation against multi-modal reliability bias. In *Proceedings of the International Conference on Learning Representations*, 2024b.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang Yoo. C-TPT: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *International Conference on Learning Representations*, volume 2024, pages 8636–8660, 2024.

Unified Confidence Adjustment for Robust Cross-Modal Retrieval under Test-Time Distribution Shifts (Supplementary Material)

Rui Zhou[1] **Yawen Hao**[1] **Hao Zuo**[1] **Xinhang Wan**[1] **Cheng Zhu**[*1] **Yun Zhou**[*1]

[1]National Key Laboratory of Information Systems Engineering, National University of Defense Technology, Changsha, China

[*]Co-corresponding authors. {zhucheng, zhouyun}@nudt.edu.cn

The structure of the Appendix is as follows.

- Appendix A defines the uniformity and cross-modal discrepancy used in the cross-modal matching task;
- Appendix B provides additional experimental results and analyses that are omitted from the main text due to space constraints;
- Appendix C includes the algorithm listing.
- Appendix D provides the missing proof in Section 3;

A INTRA-MODALITY UNIFORMITY AND CROSS-MODALITY DISCREPANCY

In cross-modal matching tasks, intra-modality uniformity and cross-modality discrepancy are two critical factors for achieving precise semantic alignment. We formally define them as follows.

Definition 4 (Uniformity)**.** *Given modality-specific encoder ϕ^m for modality $m \in \{Q, G\}$ and its normalized feature representations $\mathbf{z}_i^m = \frac{\phi^m(\mathbf{x}_i^m)}{\|\phi^m(\mathbf{x}_i^m)\|_2}$, intra-modality uniformity can be measured as:*

$$\mathcal{U}^m = \log \mathbb{E}_{i \neq j} \exp\left(-\left\|\mathbf{z}_i^m - \mathbf{z}_j^m\right\|_2^2\right). \tag{A1}$$

A lower $\mathcal{U}^m$ indicates that samples are more uniformly distributed on the hypersphere, reducing feature collapse and enhancing intra-modal discriminability.

Definition 5 (Cross-Modality Discrepancy)**.** *Let $\boldsymbol{\mu}^Q$ and $\boldsymbol{\mu}^G$ denote the feature centroids of the query and gallery modalities, respectively:*

$$\boldsymbol{\mu}^m = \frac{1}{N^m} \sum_{i=1}^{N^m} \mathbf{z}_i^m, \quad m \in \{Q, G\}. \tag{A2}$$

The cross-modality discrepancy can be defined as:

$$\mathcal{D}_{mod} = \left\|\boldsymbol{\mu}^Q - \boldsymbol{\mu}^G\right\|_2^2. \tag{A3}$$

A smaller $\mathcal{D}_{\text{mod}}$ implies better cross-modal alignment and reduced modality gap, which is essential for reliable retrieval performance.

Maintaining strong intra-modality uniformity while minimizing cross-modality discrepancy provides a principled foundation for robust cross-modal matching under distribution shifts. Therefore, these two properties can be jointly leveraged to estimate the discrepancy between the sample distribution and the model distribution, thereby enabling the identification of source-domain-like samples for reliable test-time adaptation.

Table A1: Ablation study on the source-domain-like similarity set (SI).

Method	B2Flickr	B2COCO	B2Fashion	B2Nocaps(ID)	B2Nocaps(ND)	B2Nocaps(OD)	Avg.
UCA (w/ SI)	**83.3**	**54.9**	10.2	**86.1**	**77.4**	**75.7**	**64.6**
UCA (w/o SI)	83.0	52.9	**10.5**	85.9	76.2	74.8	63.9

B EXPERIMENTS

B.1 TRADE-OFF ANALYSIS

We analyze the impact of the four loss weights on performance over the Flickr dataset under three settings: zero-shot transfer, image corruption, and text corruption. The detailed visualizations are provided in the appendix. All four weights are varied within the range $(0, 1]$ with a step size of 0.1. We group the parameters as $(\lambda_{\text{uca}}, \lambda_{\text{sem}})$ and $(\lambda_{\text{uni}}, \lambda_{\text{emg}})$, and visualize the results using 3D heatmaps.

Zero-shot setting. Fig. A1(a) analyzes the hyperparameter sensitivity of UCA under zero-shot transfer. In the top of Fig. A1(a), we fix λ_{uni} and λ_{emg} while varying λ_{uca} and λ_{sem}. The R@1 score ranges from 88.7% to 90.4%. The best performance is achieved at $\lambda_{\text{uca}} = 1$ and $\lambda_{\text{sem}} = 0.8$, whereas the worst result occurs at $\lambda_{\text{uca}} = 0.1$ and $\lambda_{\text{sem}} = 1$. Overall, larger λ_{uca} generally leads to better performance. In the bottom of Fig. A1(a), we fix λ_{uca} and λ_{sem} while varying λ_{uni} and λ_{emg}. The R@1 score ranges from 88.7% to 90.7%. The best performance is achieved at $\lambda_{\text{uni}} = 1$ and $\lambda_{\text{emg}} = 0.4$ or 0.5, while the worst result appears at $\lambda_{\text{uni}} = 0.1$ and $\lambda_{\text{emg}} = 0.9$.

Image corruption setting. Fig. A1(b) analyzes the hyperparameter sensitivity of UCA under image corruption. In the top of Fig. A1(b), we fix λ_{uni} and λ_{emg} and vary λ_{uca} and λ_{sem}. The R@1 score ranges from 60.1% to 69.6%. The optimal performance is achieved at $\lambda_{\text{uca}} = 0.7$ and $\lambda_{\text{sem}} = 0.1$, whereas the worst results occur when $\lambda_{\text{uca}} = 0.1$ and λ_{sem} is relatively large (e.g., 0.7 or 0.8). Overall, performance is consistently better when $\lambda_{\text{uca}} > \lambda_{\text{sem}}$. In the bottom of Fig. A1(b), we fix λ_{uca} and λ_{sem} and vary λ_{uni} and λ_{emg}. The R@1 score ranges from 64.3% to 68.2%. In general, small λ_{uni} leads to significant performance degradation, whereas larger λ_{uni} consistently improves robustness.

Text corruption setting. Fig. A1(c) analyzes the hyperparameter sensitivity of UCA under text corruption. Overall, UCA demonstrates relatively stable behavior, with R@1 ranging from 69.6% to 70.3%, corresponding to less than a 1% variation. The overall trend remains consistent: larger λ_{uca} and smaller λ_{sem} tend to yield better performance.

Across zero-shot transfer, image corruption, and text corruption settings, UCA shows strong robustness to hyperparameter variations. In particular, placing more weight on confidence-aware regularization (larger λ_{uca}) while keeping the uniformity constraints balanced consistently yields strong performance. Moreover, under severe noise or corruptions, enlarging the trade-off of entropy minimization (λ_{sem}) can over-sharpen incorrect top-1 predictions, amplifying confirmation bias. In contrast, UCA encourages the model to recover a source-domain-like margin, effectively acting as a confidence regularizer and an anti-overfitting brake. These observations suggest that, for retrieval-time adaptation, the goal is not merely to increase confidence, but to preserve a source-consistent *relative* similarity structure among competing candidates.

B.2 ABLATION STUDIES

B.2.1 Loss Term

We conduct ablation experiments on both image-to-text and text-to-image retrieval over the B2COCO benchmark using BLIP to analyze the contribution of each loss term. Four evaluation metrics are reported in Table 3. **R@1** measures retrieval accuracy. **MD** computes the average distance between samples and their corresponding modality center, where larger values indicate better uniformity. **MG** quantifies the modality gap, with smaller values reflecting better cross-modal alignment. **CG** denotes the confidence gap between the best and second-best matched candidates; larger values imply stronger discriminability on B2COCO. Overall, the effects of different loss terms are consistent across text retrieval (TR) and image retrieval (IR). Removing UCA leads to a substantial performance drop, demonstrating the contribution of the proposed UCA.

Effectiveness of Uniformity. When removing $\mathcal{L}_{\text{uni}}$, MD slightly decreases in both directions (TR: $0.955 \rightarrow 0.951$, IR: $0.961 \rightarrow 0.956$), indicating that modality uniformity is compromised. Although the degradation appears relatively mild, it

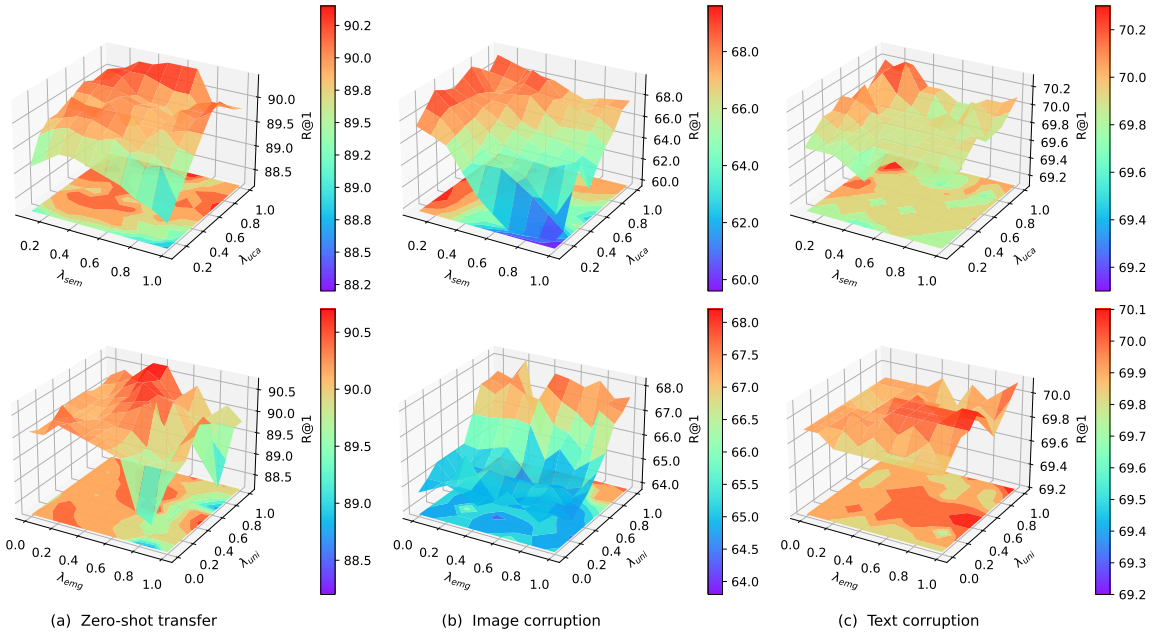

Figure A1: **Trade-off analysis on Flickr.** We report results under (a) zero-shot transfer, (b) image corruption, and (c) text corruption. In the first row, we fix λ_{uni} and λ_{emg} while varying λ_{uca} and λ_{sem}. In the second row, we fix λ_{uca} and λ_{sem} while varying λ_{uni} and λ_{emg}.

still results in an approximate 0.4% drop in retrieval accuracy.

Effectiveness of Modality Gap. Further removing $\mathcal{L}_{\text{emg}}$, MG increases significantly in both directions (TR: $0.591 \rightarrow 0.692$, IR: $0.632 \rightarrow 0.757$), suggesting that $\mathcal{L}_{\text{emg}}$ plays a critical role in narrowing the modality gap and stabilizing retrieval performance.

Effectiveness of Confidence Adjustment. When $\mathcal{L}_{\text{uca}}$ is further removed, performance degrades substantially. TR@1 decreases from 69.6% to 64.1%, and IR@1 drops from 48.9% to 46.9%, resulting in a 2.92% decline in average R@1. Meanwhile, CG decreases markedly, indicating a reduced confidence margin between the best and second-best matches and weakened discriminative ability.

B.2.2 SI-based source-like set

In addition, we further compare UCA with and without the SI-based source-like set. When the SI module is removed, the confidence calibration loss is reduced to directly enlarging the confidence gap between the top-1 and top-2 predictions, without using the source-likeness guidance. As shown in the following table, UCA consistently performs better than its variant without the SI set under severe distribution-shift settings, indicating the effectiveness of the SI-based source-like set.

B.3 FAILURE CASE ANALYSIS

To clarify the scope and limitations of our method, we further analyze two potential failure cases. First, we examine whether over-confident and under-confident samples may mutually cancel out their effects. Second, we investigate whether the source-similarity set can effectively approximate the similarity of the true candidate set.

First, we compare two confidence adjustment strategies:

- **Sample-wise confidence adjustment**, which computes the deviation of each individual sample from the source-domain-like Semantic Separation Margin and then averages these deviations.
- **Batch-wise confidence adjustment**, which performs calibration based on the batch-level mean of similarity gaps.

As indicated by Table A2, on Flickr-C, the model exhibits an overall under-confidence tendency. This suggests that, in

Table A2: The number of under-confident and over-confident samples before calibration, and the classification accuracy (%) after confidence calibration.

Metric	Noise	Blur	Weather	Digital	Avg.
Under-confident samples / batch	23.3	22.9	21.8	22.8	22.7
Over-confident samples / batch	5.0	5.3	5.9	5.2	5.4
Batch-wise confidence adjustment	72.3	71.9	86.8	77.0	77.0
Sample-wise confidence adjustment	70.4	70.8	86.5	75.9	75.9

Table A3: Failure case analysis of UCA under text perturbations.

Stage	Metric	Failure Scenarios		
		Flickr-IP	Flickr-Formal	COCO-IP
Confidence State Identification	Average under-confident samples per batch	29.2	27.6	32.9
	Average over-confident samples per batch	28.4	29.9	25.0
Batch-wise	$\left\| \frac{1}{B}\sum_{i=1}^{B}\hat{\eta}_i - \overline{\Delta\eta} \right\|_2 - \gamma$	0.0062	0.0074	0.0106
	R@1	81.9	81.6	56.8
Sample-wise	$\frac{1}{B}\sum_{i=1}^{B}\left\|\hat{\eta}_i - \overline{\Delta\eta}\right\|_2 - \gamma$	0.0401	0.0413	0.0522
	R@1	81.6	81.2	56.7

scenarios where the model performs well, there is no mutual cancellation between over-confident and under-confident samples. Moreover, batch-wise confidence adjustment consistently achieves better performance. In contrast, sample-wise confidence adjustment leads to an average performance drop of approximately 1.1%. A possible reason is that individual sample-level estimates are more susceptible to noise caused by hard negatives and mini-batch sampling fluctuations, making the per-sample margin relatively unstable.

However, the batch-mean-based strategy does not always yield effective results. By examining the 86 experimental settings reported in the submitted manuscript, we find that our method fails to achieve the best performance in three cases. We further conduct the above analysis on these failure cases, and the results are shown in Table A3. We observe that the over-confident and under-confident samples are almost evenly distributed in these cases, which helps explain why our method does not achieve the optimal performance on these tasks.

Further examining the possible reason, we hypothesize that when over-confident and under-confident samples are relatively balanced within a batch, the top-1/top-2 confidence gap estimated from the selected source-like set may become close to the average top-1/top-2 confidence gap of the current batch. This suggests that the selected source-like set may not provide a sufficiently distinct source-like reference. In this case, it becomes more difficult to identify the overall confidence tendency of the batch, thereby weakening the effectiveness of confidence adjustment.

This analysis also reveals an important applicability boundary of our method. UCA relies on the assumption that the selected source-like set can provide a meaningful reference margin for the current candidate set. When this assumption does not hold, especially under severe text perturbations where reliable source-like samples are difficult to identify, the effectiveness of confidence adjustment may become limited.

Nevertheless, the experimental results support three important observations.

- Batch-wise confidence adjustment consistently outperforms sample-wise confidence adjustment in both successful and failure scenarios.
- When the selected source-like set follows a distribution similar to that of the current batch, rather than representing a truly source-like distribution, our method may fail to achieve the optimal performance.
- Despite these limitations, the proposed method achieves the best performance on 96% of the 86 subtasks, demonstrating its effectiveness across a wide range of scenarios.

Table A4: Comparison of test-time adaptation performance using different candidate-set confidence proxies.

Method	Setting	Gauss.	Shot	Impul.	Speckle	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Avg.
UCA	$K=2$	**67.8**	**71.3**	**67.4**	**82.7**	**72.6**	**89.4**	**72.5**	**53.0**	**83.0**	**87.8**	**86.7**	**89.7**	**79.8**	**83.1**	**60.4**	**84.8**	**77.0**
Top-k margin mean	$K=3$	65.6	70.8	66.6	81.9	71.2	89.6	71.3	52.6	81.9	86.7	86.9	89.4	79.4	81.3	58.2	84.1	76.1
Top-k margin mean	$K=4$	64.4	67.3	66.3	81.3	71.9	88.4	70.0	51.0	82.2	87.2	86.2	89.0	79.1	80.7	56.7	82.9	75.3
Top-k margin mean	$K=5$	61.7	68.2	66.3	80.9	70.1	88.0	69.7	50.9	80.0	86.1	86.1	89.4	77.8	80.3	56.7	83.0	74.7
Top-k margin mean	$K=10$	61.8	66.4	65.0	79.9	69.8	87.7	67.4	50.9	79.8	86.1	86.6	89.7	78.1	79.5	56.7	82.3	74.2

Table A5: Comparison of confidence-state identification using different candidate-set confidence proxies.

Metric	K	Noise	Blur	Weather	Digital	Avg.
Under-confident	2	46.3	46.1	42.7	44.7	44.7
	3	48.2	48.0	46.2	46.8	46.8
	4	49.4	49.6	47.2	48.5	48.4
	5	50.0	50.2	48.1	50.0	49.6
	10	**50.4**	**50.8**	**49.0**	**51.0**	**50.6**
Over-confident	2	10.1	10.4	12.2	10.4	10.8
	3	8.8	8.8	8.9	9.0	8.9
	4	8.0	7.8	9.0	8.0	8.2
	5	7.4	7.0	8.6	7.4	7.6
	10	**7.0**	**6.2**	**8.2**	**6.4**	**7.0**

B.4 EXPERIMENT ON DIFFERENT CONFIDENCE PROXIES

We supplement an additional comparison experiment on different confidence proxies, including the gap between the top-1 and top-2 candidates and the top-K distribution.

To incorporate the top-K logits into our framework, we design a **top-K margin mean**, defined as

$$m(K) = \frac{1}{K-1} \sum_{i=1}^{K-1} \left(R^{(i)} - R^{(i+1)} \right),$$

where $R^{(i)}$ denotes the similarity score of the i-th ranked candidate. When replacing η with the top-K margin mean, the SI-based source-domain-like set is also constructed using the same top-K margin mean. When $K = 2$, $m(K)$ is equivalent to $\hat{\eta}$.

We conduct comparison experiments on Flickr-C, and the results are shown in Table A4. The results indicate that using the top-1/top-2 gap achieves the best overall performance. In contrast, incorporating more logits does not bring further improvement and even leads to a gradual performance drop as K increases.

As can be observed, compared with using only the top-1 and top-2 candidates, using more logits ($K > 2$) does not substantially change the overall observation that the model tends to be under-confident. Moreover, as K increases, the model is increasingly biased toward being identified as under-confident, as reflected by the gradual increase in under-confident samples and the decrease in over-confident samples.

This phenomenon may be attributed to the fact that lower-ranked candidates usually introduce additional noise and weakly relevant samples. In cross-modal retrieval, the top-1/top-2 gap directly captures the most discriminative boundary between the best-matched candidate and its strongest competitor. In contrast, including more lower-ranked logits dilutes this discriminative margin and makes the confidence proxy more sensitive to irrelevant candidates. Consequently, the top-K margin mean may underestimate the actual confidence state of the model, leading to more conservative confidence adjustment and inferior adaptation performance. These results support our choice of the top-1/top-2 gap as a simple yet effective confidence proxy.

B.5 STABILITY IN NON-STATIONARY ENVIRONMENTS

To further investigate this issue, we supplement two experiments to study how the model behaves under dynamic and continuous distribution shifts: 1) continuous shifts induced by different corruptions; 2) continuous distribution-shift induced by different datasets. First, we consider continuous shifts induced by corruptions, where the dataset remains fixed while the

Table A6: Performance comparison between Reset and Unreset settings under continuous cross-dataset distribution shifts.

Setting	Gauss.	Shot	Impul.	Speckle	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Avg.
Reset	67.8	71.2	67.4	82.7	72.3	**89.4**	**72.3**	52.7	**83.3**	**87.7**	**86.7**	**89.7**	**80.1**	**82.7**	60.5	**85.0**	77.0
Unreset	67.8	**75.9**	**73.8**	**85.3**	72.3	87.5	71.0	**62.9**	75.5	85.3	85.1	87.2	78.4	80.1	**65.1**	80.6	**77.1**

Table A7: Performance comparison under dynamic continuous distribution shifts across multiple datasets.

Setting	B2Flickr	B2COCO	B2Nocaps(ID)	B2Nocaps(ND)	B2Nocaps(OD)	B2Fashion	Avg.
Reset	83.3	54.8	85.6	77.4	75.7	**10.2**	**64.5**
Unreset	**83.6**	54.3	**85.9**	**79.0**	**75.9**	8.6	**64.5**

Table A8: **Sensitivity analysis w.r.t. κ.** Under Image Corruption, we evaluate Flickr with four noise-type perturbations using BLIP. Under Text Corruption, we evaluate Flickr with five character-level perturbations using BLIP. Under Zero-shot Transfer, we report results on three datasets using CLIP. The best and second-best results are highlighted in **bold** and underline, respectively. The colored rows indicate the hyperparameter setting adopted in our main experiments.

	Image Corr. (Noise)				Zero Shot Trans.			Text Corr. (Character-level)					Avg.
	Gauss.	Shot	Impul.	Speckle.	B2Flickr	B2COCO	B2Nocaps(ID)	OCR	CI	CR	CS	CD	
$\kappa = 0.1$	66.7	70.4	67.2	**83.4**	**83.7**	54.6	**85.9**	**57.9**	**23.1**	22.8	34.3	**26.3**	**56.3**
$\kappa = 0.5$	67.2	70.8	67.9	82.7	83.5	54.5	**85.9**	57.7	23.0	22.8	**34.4**	26.1	**56.3**
$\kappa = 1.$	67.3	**71.4**	66.7	82.9	83.3	**54.9**	**85.9**	57.7	22.7	23.0	34.1	26.1	56.2
$\kappa = 1.5$	**67.4**	**71.4**	67.7	83.0	83.0	54.7	85.4	57.8	**23.0**	**23.1**	**34.4**	25.8	**56.3**
$\kappa = 2.$	66.9	69.7	**68.0**	**83.4**	83.2	53.9	86.1	57.8	22.8	22.8	34.1	**26.3**	56.2

corruption type changes sequentially. As shown in Table R7, we adapt the base model from Noise to JPEG, covering 16 corruption types. In the table:

- Reset means that after adapting to each corruption type, both the model and optimizer are reset to the source-model state, matching the setting in the submitted manuscript.
- Unreset means that the model and optimizer are not reset after each corruption type, and adaptation continues on the next one, better reflecting dynamic continuous shifts.

As shown in Table A6, the Unreset setting achieves performance comparable to the Reset setting, indicating that UCA remains stable under dynamic and continuous distribution shifts. In particular, the two settings obtain identical performance on the first corruption type, while Unreset outperforms Reset on the 2nd to 4th corruption types. This observation suggests that the candidate-set similarity information accumulated from earlier corrupted data can remain beneficial for subsequent corruptions. After the Defoc. corruption, the performance exhibits some fluctuations, which may indeed be related to the accumulated effect of long-term adaptation. Nevertheless, the overall performance remains comparable to the Reset setting.

In addition, we further consider a more challenging continuous distribution-shift scenario, where the base model is sequentially adapted to multiple datasets, including Flickr, COCO, and Fashion. As shown in Table A7, the performance of the model does not degrade under this setting. This demonstrates that our method remains effective under dynamic continuous distribution shifts.

While the above experiments suggest that our method can maintain stable performance under dynamic continuous distribution shifts, we agree that this stability is not unconditional. When the selected source-like set does not provide a sufficiently distinct source-like reference, it becomes more difficult to clearly distinguish relatively reliable samples from unreliable ones, thereby weakening the effectiveness of the confidence adjustment strategy. Therefore, an adaptive strategy for source-like set construction is a promising direction for future work.

B.6 ADDITIONAL RESULTS

Table A8 shows the sensitivity analysis for κ. Table A10 shows comparison results under text natural corruptions with maximum severity level regarding the Recall@1 metric. Table A11 shows a comparison of time consumption among state-of-the-art TTA methods. Table A9 shows the sensitivity analysis for r_{SI} and P_{SI}.

Table A9: **Sensitivity analysis w.r.t. r_{SI} and P_{SI} using BLIP.** Under Image Corruption, we evaluate Flickr with gaussian perturbations. Under Text Corruption, we evaluate Flickr with OCR perturbations. Under Zero-shot Transfer, we report results on Flickr with Image-to-Text directions and Text-to-Image directions. The best and second-best results are highlighted in **bold** and underline, respectively. The colored columns indicate the hyperparameter setting adopted in our main experiments.

Settings	metric	r_{SI}				P_{SI}				
		0.1B	0.3B	0.5B	0.7B	1B	3B	5B	10B	30B
Image Corr.	TR@1	67.9	67.4	67.7	66.7	67.4	62	60.9	60.6	60.6
Text Corr.	IR@1	57.9	57.8	57.62	57.8	57.8	57.38	57.34	57.3	57.4
Zero-shot Trans.	TR@1	90.3	90.3	90.5	90.4	90.3	87.9	86.4	86.2	85.8
	IR@1	70.5	70.4	70.3	70.7	70.4	69.4	69.0	68.9	68.7

Table A10: **Comparison results under text natural corruptions** with maximum severity level ($s_{\text{char}} = 7, s_{\text{word}} = 2, s_{\text{sen}} = 1$) regarding the Recall@1 metric. All methods are evaluated with BLIP ViT-B/16 on Flickr30K and MSCOCO. The best and second-best results are highlighted in **bold** and underline, respectively.

Dataset	Method	Character-level					Word-level					Sentence-level					Avg.
		OCR	CI	CR	CS	CD	SR	RI	RS	RD	IP	Formal	Casual	Passive	Active	Backtrans	
Flickr	ViT-B/16	53.5	18.4	18.0	30.4	22.5	68.3	77.9	76.9	77.9	82.1	82.1	81.9	79.9	82.2	79.8	62.1
	Tent	55.4	18.6	18.2	31.1	23	69.6	78.8	77.7	77.9	82.2	81.9	81.8	79.6	82.0	79.9	62.5
	PL	56.6	19.5	13.6	31.9	23.7	69.7	79.1	77.9	77.8	**82.5**	82.3	82.0	80.5	82.4	79.9	62.6
	SHOT	56.8	14.4	10.7	31.2	21.4	69.7	79.0	77.9	77.9	82.3	82.2	82.0	80.6	**82.6**	80.0	61.9
	EATA	55.7	19.9	19.9	31.6	23.6	69.5	78.6	77.5	77.9	82.4	82.3	81.8	80.5	82.5	80.1	62.9
	SAR	53.5	20.1	19.1	32.1	23.8	68.3	77.9	76.9	77.7	82.1	82.1	81.9	79.9	82.2	79.8	62.5
	READ	55.8	19.7	20.6	32	23.5	69.3	78.6	77.6	77.8	82.4	82.2	81.8	80.5	**82.6**	80.1	63.0
	DeYO	53.5	18.4	18	30.4	22.5	68.3	77.9	76.9	77.9	82.1	82.1	81.9	79.9	82.2	79.8	62.1
	TCR	57.1	21.4	22.5	33.6	25.1	**69.8**	79.2	78.0	78.0	**82.5**	**82.4**	**82.1**	80.8	82.5	80.1	63.7
	Ours	**57.9**	**23.0**	**23.1**	**34.4**	**25.8**	**69.8**	**79.5**	**78.5**	**78.1**	82.4	82.3	**82.1**	**80.9**	**82.6**	**80.2**	**64.0**
COCO	ViT-B/16	31.4	11.3	9.4	18.9	11.4	43.6	51.5	50.3	50.6	56.8	56.6	56.2	54.9	56.8	54.2	40.9
	Tent	31.4	11.0	9.5	17.7	11.3	43.2	51.3	50.3	50.6	56.6	56.2	56.0	54.9	56.9	53.9	40.7
	PL	33.0	3.6	4.5	18.1	7.9	44.9	53.3	51.8	51.3	57.0	57.0	56.6	55.8	57.3	54.5	40.5
	SHOT	22.9	2.4	2.4	6.5	2.1	43.9	52.6	51.1	50.5	56.6	56.9	56.6	55.6	57.3	54.4	38.1
	EATA	33.1	11.9	10.5	18.4	12.0	44.9	53.0	51.6	50.3	56.2	56.8	56.8	56.0	56.8	54.3	41.5
	SAR	31.8	11.6	9.9	18.5	11.7	43.6	51.5	50.3	50.6	56.8	56.5	56.2	54.9	56.8	54.2	41.0
	READ	32.3	11.4	9.6	18.2	11.2	44.3	52.9	51.7	51.1	**57.6**	57.1	56.7	55.9	57.1	54.7	41.5
	DeYO	31.4	11.3	9.4	17.9	11.4	43.6	51.5	50.3	50.6	56.8	56.5	56.2	54.9	56.7	54.2	40.8
	TCR	34.1	13.7	11.8	**19.5**	**13.2**	45.1	53.4	51.8	51.2	57.0	57.1	56.8	55.8	57.3	54.7	42.2
	Ours	**34.7**	**14.2**	**12.2**	**19.5**	**13.2**	**45.7**	**54.4**	**52.4**	**51.5**	57.2	**57.4**	**57.2**	**56.2**	**57.8**	**55.0**	**42.6**

Table A11: **Comparison of time consumption among state-of-the-art TTA methods.** Experiments are conducted using CLIP ViT-B/16. For image-to-text adaptation, all methods adapt the LayerNorm parameters in the visual encoder, resulting in 0.040M trainable parameters and 17.582G FLOPs. For text-to-image adaptation, all methods adapt the LayerNorm parameters in the text encoder, resulting in 0.025M trainable parameters and 2.949G FLOPs. The total adaptation GPU runtime required to process an entire dataset is reported in seconds.

Method	B2Flickr		B2COCO		B2Fashion		B2Nocaps(ID)		B2Nocaps(ND)		B2Nocaps(OD)		Avg.
	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	
Tent	15s	38s	155s	14s	38s	25s	20s	66s	125s	18s	70s	41s	52s
SAR	20s	56s	273s	17s	40s	28s	27s	100s	187s	26s	109s	62s	79s
DeYO	20s	42s	267s	16s	39s	25s	24s	88s	167s	23s	100s	57s	72s
TCR	14s	36s	150s	13s	37s	24s	20s	66s	126s	19s	71s	42s	52s
UCA	15s	35s	167s	14s	38s	24s	20s	69s	126s	19s	73s	43s	54s

C ALGORITHM

Algorithm 1 summarizes the proposed UCA. To construct the source-domain-like set $\mathcal{D}_{SI}$, we collect, from each of the first ten mini-batches, the samples with the smallest SI values (ratio $r_{SI} = 30\%$) and insert them into $\mathcal{D}_{SI}$. WWe maintain

$\mathcal{D}_{SI}$ as a fixed-capacity queue of size $P_{SI} \times B$, where B denotes the batch size and P_{SI} is a hyperparameter with a default value of $P_{SI} = 1$. Once the queue reaches its capacity, newly selected samples are enqueued while the oldest samples are dequeued. To stabilize the estimation, we update $\mathcal{D}_{SI}$ for at most

$$N_{\text{update}} = \max\left(10, \frac{P_{SI}}{r_{SI}}\right)$$

iterations. Sensitivity analyses of P_{SI} and r_{SI} are reported in Table A9.

Algorithm 1: Unified Confidence Adjustment (UCA) for Test-Time Cross-Modal Retrieval

Input: Test set $\mathcal{D}_T = \{\mathbf{X}^Q, \mathbf{X}^G\}$; source model $\phi_{\Theta_s} = \{\phi^Q, \phi^G\}$; batch size B; temperature τ; loss weights $\lambda_{\text{uca}}, \lambda_{\text{sem}}, \lambda_{\text{uni}}, \lambda_{\text{emg}}$; constant ϵ; learning rate α; calibration intensity κ.

Output: Adapted parameters $\hat{\Theta}$ (update only normalization layers of the query encoder).

Initialize: $\hat{\Theta} \leftarrow \Theta_s$

foreach *incoming query batch* $\{\mathbf{x}_i^Q\}_{i=1}^B$ **do**

`// (1) Encode and retrieve nearest neighbors`

Compute normalized embeddings: $\mathbf{z}_i^Q = \phi^Q(\mathbf{x}_i^Q)/\|\phi^Q(\mathbf{x}_i^Q)\|_2$, and $\mathbf{z}_j^G = \phi^G(\mathbf{x}_j^G)/\|\phi^G(\mathbf{x}_j^G)\|_2$

for $i = 1$ **to** B **do**

Retrieve top-1 nearest neighbor $\mathbf{z}_i^{G'}$ in gallery for $\mathbf{z}_i^Q$

Obtain top-1/top-2 similarity scores $R_i^{(1)}, R_i^{(2)}$ from ranking $R(\mathbf{z}_i^Q, \mathbf{Z}^{G'}) = \text{sort}\left(\mathbf{z}_i^Q(\mathbf{Z}^{G'})^\top\right)$

`// (2) Source-likeness identification (SI) and build` $\mathcal{D}_{SI}$

Compute centroids $\mu^Q = \frac{1}{B}\sum_{i=1}^B \mathbf{z}_i^Q$ and $\mu^{G'} = \frac{1}{B}\sum_{i=1}^B \mathbf{z}_i^{G'}$

for $i = 1$ **to** B **do**

$SI_i \leftarrow 2\|\mathbf{z}_i^Q - \mathbf{z}_i^{G'}\|_2 - \left(\|\mathbf{z}_i^Q - \mu^Q\|_2 + \|\mathbf{z}_i^{G'} - \mu^{G'}\|_2\right)$

Select query-candidate pairs with the lowest 30% SI values to form $\mathcal{D}_{SI}$

`// (3) Estimate reference margins from` $\mathcal{D}_{SI}$

$\Delta_g \leftarrow \left\| \frac{1}{|\mathcal{D}_{SI}|}\sum_{(\mathbf{z}_j^Q, \mathbf{z}_j^{G'}) \in \mathcal{D}_{SI}} \mathbf{z}_j^Q - \frac{1}{|\mathcal{D}_{SI}|}\sum_{(\mathbf{z}_j^Q, \mathbf{z}_j^{G'}) \in \mathcal{D}_{SI}} \mathbf{z}_j^{G'} \right\|_2$

$\Delta_\eta \leftarrow \frac{1}{|D_{SI}|}\sum_{(\mathbf{z}_j^Q, \mathbf{z}_j^{G'}) \in D_{SI}} \left(R_{j(1)} - R_{j(2)}\right)$

`// (4) Compute losses on current batch`

$\bar{g} \leftarrow \left\| \frac{1}{B}\sum_{i=1}^B \mathbf{z}_i^Q - \frac{1}{B}\sum_{i=1}^B \mathbf{z}_i^{G'} \right\|_2$

$\mathcal{L}_{\text{uni}} \leftarrow \frac{1}{B}\sum_{i=1}^B \exp\left(-\frac{\|\mathbf{z}_i^Q - \overline{\mathbf{z}^Q}\|_2^2}{t}\right)$

$\mathcal{L}_{\text{emg}} \leftarrow (\bar{g} - \Delta_g)^2$

Compute query prediction distribution: $p \leftarrow \text{Softmax}\left(\mathbf{z}^Q(\mathbf{Z}^G)^\top/\tau\right)$

$\mathcal{L}_{\text{sem}} \leftarrow \frac{1}{B}\sum_{i=1}^B \min\left(\frac{\Delta_g}{\hat{g}_i + \epsilon}, 1\right)\mathcal{H}(\mathbf{p}_i), \quad \hat{g}_i = \left\|\mathbf{z}_i^Q - \mathbf{z}_i^{G'}\right\|_2$

$\hat{\eta}_i \leftarrow R_{i(1)} - R_{i(2)}$

$\gamma = \kappa\widehat{\sigma}_\eta/\sqrt{B}$

$\mathcal{L}_{\text{uca}} \leftarrow \max\left(0, \left|\frac{1}{B}\sum_{i=1}^B \hat{\eta}_i - \overline{\Delta\eta}\right| - \gamma\right)$

`// (5) Online update`

$\mathcal{L} \leftarrow \lambda_{\text{uca}}\mathcal{L}_{\text{uca}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{uni}}\mathcal{L}_{\text{uni}} + \lambda_{\text{emg}}\mathcal{L}_{\text{emg}}$

Update $\hat{\Theta} \leftarrow \hat{\Theta} - \alpha\nabla_{\hat{\Theta}}\mathcal{L}$

D THEORETICAL DERIVATION FOR THE CONFIDENCE STATE IDENTIFICATION

In this section, we provide the theoretical derivation for the confidence state identification mechanism (Definition 3) and the tolerance band γ. Specifically, we show that comparing the empirical similarity gap $\bar{\eta}$ with the reference margin $\overline{\Delta\eta}$ is mathematically equivalent to conducting a rigorous statistical hypothesis test on the local predictive dispersion.

D.1 STATISTICAL HYPOTHESIS TESTING FORMULATION

Let $\{\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_B\}$ be a sequence of independent and identically distributed (i.i.d.) random variables representing the sample-wise similarity gaps within a mini-batch of size B. Let $\mu = \mathbb{E}[\hat{\eta}_i]$ denote the true expected similarity margin of the current batch, and $\sigma^2 = \text{Var}(\hat{\eta}_i)$ denote its true variance.

We formulate the confidence state identification as a two-sided hypothesis test. Our goal is to determine whether the model's current predictive dispersion (represented by μ) deviates significantly from the target reference margin $\overline{\Delta\eta}$ obtained from reliable source-domain-like samples:

- **Null Hypothesis** ($\mathcal{H}_0$): $\mu = \overline{\Delta\eta}$. The expected similarity margin of the current test distribution remains consistent with the source-domain-like reference, indicating a well-calibrated state [Guo et al., 2017].
- **Alternative Hypothesis** ($\mathcal{H}_1$): $\mu \neq \overline{\Delta\eta}$. The model exhibits a systematic confidence deviation (either over-confidence or under-confidence) due to the test-time distribution shift.

D.2 TEST STATISTIC AND ASYMPTOTIC DISTRIBUTION

To test $\mathcal{H}_0$, we compute the sample mean $\bar{\eta} = \frac{1}{B}\sum_{i=1}^{B}\hat{\eta}_i$ and the sample standard deviation $\widehat{\sigma}_\eta$. According to the Central Limit Theorem (CLT) [Casella and Berger, 2024], for a sufficiently large batch size B, the standardized sample mean converges in distribution to a standard normal distribution:

$$\frac{\sqrt{B}(\bar{\eta} - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$$

Since the true population standard deviation σ is unknown, we substitute it with the consistent estimator $\widehat{\sigma}_\eta$. By Slutsky's Theorem [Van der Vaart, 2000], the test statistic T under the null hypothesis $\mathcal{H}_0$ follows an asymptotic standard normal distribution:

$$T = \frac{\bar{\eta} - \overline{\Delta\eta}}{\widehat{\sigma}_\eta/\sqrt{B}} \xrightarrow{d} \mathcal{N}(0, 1)$$

D.3 DERIVATION OF THE TOLERANCE BAND γ

Following standard hypothesis testing frameworks [Lehmann and Romano, 2005], given a significance level $\alpha \in (0, 1)$, we construct a $1 - \alpha$ confidence interval. Let κ be the critical value such that $\mathbb{P}(|Z| \leq \kappa) = 1 - \alpha$, where $Z \sim \mathcal{N}(0, 1)$ (i.e., $\kappa = Z_{1-\alpha/2}$). We fail to reject the null hypothesis $\mathcal{H}_0$ if the test statistic falls within the non-rejection region:

$$|T| \leq \kappa \implies \left| \frac{\bar{\eta} - \overline{\Delta\eta}}{\widehat{\sigma}_\eta/\sqrt{B}} \right| \leq \kappa$$

By rearranging the terms, we obtain the empirical bound for the well-calibrated state:

$$|\bar{\eta} - \overline{\Delta\eta}| \leq \kappa \frac{\widehat{\sigma}_\eta}{\sqrt{B}}$$

This perfectly recovers our definition of the tolerance band $\gamma = \kappa\widehat{\sigma}_\eta/\sqrt{B}$.

D.4 MAPPING TO CONFIDENCE STATES

Based on the testing criteria, the decision rules naturally map to the three confidence states defined in Eq. (10):

1. **Well-calibrated**: $|T| \leq \kappa \iff |\bar{\eta} - \overline{\Delta\eta}| \leq \gamma$. We do not have sufficient statistical evidence to reject $\mathcal{H}_0$, so no confidence adjustment is triggered.
2. **Over-confident**: $T > \kappa \iff \bar{\eta} > \overline{\Delta\eta} + \gamma$. $\mathcal{H}_0$ is rejected in the positive direction, indicating the model excessively separates candidates.

3. **Under-confident**: $T < -\kappa \iff \bar{\eta} < \overline{\Delta\eta} - \gamma$. $\mathcal{H}_0$ is rejected in the negative direction, indicating insufficient discriminability.

Remark. In our framework, κ acts as a unified tunable hyperparameter that explicitly controls the trade-off between confidence bounds and calibration intensity. We empirically set $\kappa = 1.5$ or 1. This highly sensitive threshold ensures that we preserve a fundamental degree of confidence while prioritizing immediate interventions upon even marginal deviations from a well-calibrated state, maximizing our overall calibration efforts. The empirical evidence presented in Table A8 further corroborates the superiority of this proactive setting.