

Cross-Granularity Ranking Disagreement for Uncertainty-Aware Image-Text Retrieval

Wei Chen

470041595@qq.com

Fuzhou Technology and Business University

Yuhang Chen

Fuzhou Technology and Business University

Research Article

Keywords: image-text retrieval, cross-modal retrieval, uncertainty estimation

Posted Date: August 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10415604/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Cross-Granularity Ranking Disagreement for Uncertainty-Aware Image–Text Retrieval

Wei Chen^{1*} and Yuhang Chen²

¹²School of Artificial Intelligence, Fuzhou Technology and Business
University Geling, Yongtai, Fuzhou, 350715, Fujian, China.

*Corresponding author(s). E-mail(s): 470041595@qq.com;
Contributing authors: 13045973204@163.com;

Abstract

Image-text retrieval systems are commonly optimized for ranking accuracy but provide little indication of whether a returned list is trustworthy. Existing uncertainty-aware approaches typically represent each modality with a probability distribution, so their uncertainty is learned indirectly in the embedding space and may not reflect the error risk of the final ranking. We introduce Cross-granularity Ranking Uncertainty (CRU), a lightweight framework that estimates retrieval uncertainty directly from the candidate-ranking space. CRU constructs complementary global-semantic and local-detail matching distributions and decomposes query uncertainty into within-branch ranking ambiguity and cross-branch ranking disagreement. The resulting uncertainty controls a query-level dynamic fusion rule and is calibrated against retrieval correctness. Unlike sampling- or ensemble-based alternatives, CRU requires a single deterministic forward pass and supports efficient global retrieval followed by local re-ranking. Experiments on Flickr30K-1K, MS-COCO-1K, and MS-COCO-5K show rSum values of 520.4, 535.1, and 446.9, respectively. Relative to the strongest uncertainty baseline, CRU improves error-detection AUROC by 4.3–4.6 points and reduces AURC by 16.4–21.6%. These results demonstrate that ranking-space uncertainty improves retrieval accuracy, calibration, and selective prediction in a single deterministic pass.

Keywords: image–text retrieval, cross-modal retrieval, uncertainty estimation

1 Introduction

Image-text retrieval connects visual and linguistic content by ranking captions for an image query or images for a textual query. It underpins multimedia search, recommendation, electronic commerce, and digital-asset management. Large-scale contrastive pre-training has substantially improved transferable global representations [1, 2], while align-before-fuse and bootstrapped pre-training further strengthen cross-modal transfer [3, 4]. Fine-grained alignment methods exploit region-word or patch-token interactions to distinguish visually similar candidates [5–8]. Standard benchmarks therefore report increasingly strong Recall at K ($R@K$). Nevertheless, a deployed retrieval system normally returns a list even when the query is incomplete, several candidates are nearly indistinguishable, or a crucial attribute has been missed. A high ranking score alone does not reveal whether the result is reliable.

Consider the query “two men in red shirts are playing football.” A global matcher can recognize the dominant scene—people playing football—but may rank an image with three players or blue shirts highly. A local matcher can instead focus on number, color, and object-level correspondences. Agreement between their concentrated candidate distributions supplies evidence that the decision is stable. Conversely, disagreement between the two rankings reveals a conflict between scene semantics and local details; diffuse rankings in both branches reveal that neither view separates the candidates. These two failure modes are complementary and occur in the decision space that users actually observe.

Probabilistic cross-modal embeddings model one-to-many correspondence by predicting distributional representations [9, 10]. Their variance is useful for representing ambiguity, but it is optimized through an embedding objective and does not necessarily equal the probability that the top-ranked item is wrong. Monte Carlo dropout [11], deep ensembles [12], and prior networks [13] estimate different forms of predictive uncertainty, but typically require stochastic inference, multiple models, or uncertainty-specific training data. Large-scale studies further show that calibration can degrade under distribution shift [14]. Post-hoc confidence measures such as the top score, top-1/top-2 margin, or ranking entropy are inexpensive, yet observe only one ranking mechanism. They cannot expose a decision conflict in which both branches are individually confident but prefer different candidates.

We propose Cross-granularity Ranking Uncertainty (CRU), a deterministic framework that treats global and local matchers as complementary retrieval opinions. For each query, both branches define probability distributions over a shared candidate set. CRU measures (i) *ranking ambiguity*, the normalized entropy within each branch, and (ii) *granularity disagreement*, the Jensen–Shannon divergence between the two distributions. Their combination estimates query-level retrieval risk directly from the ranking space. Branch entropies further produce query-level fusion weights shared by all candidates, preserving a coherent and interpretable ordering. Finally, a lightweight correctness-aware calibrator maps the uncertainty score to the probability of a top-1 error. The method is trained end-to-end but needs neither stochastic sampling nor probabilistic embedding assumptions.

Our contributions are:

- a candidate-space formulation of retrieval uncertainty that exposes conflicts between global semantics and local details;
- a decomposable score combining ranking ambiguity and cross-granularity Jensen–Shannon disagreement, together with query-level confidence-aware fusion;
- correctness-aware calibration and confidence-weighted consistency for selective retrieval without extra inference passes; and
- an evaluation on three standard retrieval settings covering retrieval accuracy, error detection, calibration, selective retrieval, component ablation, and parameter sensitivity.

2 Related Work

2.1 Image–text retrieval

Early visual–semantic embedding methods map whole images and sentences into a shared space; VSE++ introduced hard-negative mining to strengthen this paradigm [15]. VSRN enriches global representations through graph-based semantic reasoning [16], while VSE ∞ learns pooling operators that adapt to the input feature structure [17]. Global dual encoders remain efficient because gallery embeddings can be indexed offline, and web-scale contrastive systems such as CLIP and ALIGN demonstrate how data scale improves transfer and retrieval [1, 2]. Their single-vector bottleneck can nevertheless suppress attributes and relations that determine fine-grained relevance.

Local-alignment methods explicitly compare visual regions or patches with words. SCAN infers latent region–word alignments through stacked cross-attention [5]; IMRAM refines these correspondences iteratively with recurrent attention memory [6]; and SGRAF reasons over local and global alignment evidence [18]. NAAF uses both matched and mismatched fragments [7], CHAN encodes hard cross-modal alignments efficiently [8], and confidence-aware matching suppresses unreliable region–word pairs [19]. Joint vision–language models process visual and linguistic tokens with expressive interaction, whereas ALBEF and BLIP combine unimodal alignment with cross-modal fusion [3, 4, 20]. These approaches mainly treat global and local modeling as mechanisms for improving accuracy. CRU instead uses their different inductive biases as retrieval opinions whose agreement is informative about reliability.

Related studies in *Multimedia Systems* have addressed the heterogeneous gap from several complementary perspectives. Xie et al. modeled cross-modal semantic correlation through a probabilistic generation framework, while Jiang et al. learned discriminative shared semantic representations for heterogeneous modalities [21, 22]. More recent work captures modality consistency and diversity through public and private semantic subspaces [23], strengthens image–text correlations using multi-stage feature fusion [24], and improves retrieval under long-tailed training distributions with complementary experts [25]. These methods primarily improve representation alignment, semantic preservation, or data-balance robustness. In contrast, CRU uses complementary matching views to diagnose uncertainty in the resulting candidate ranking and calibrate its error risk.

2.2 Uncertainty in cross-modal representation learning

Monte Carlo dropout approximates Bayesian inference through repeated stochastic forward passes [11], while deep ensembles aggregate independently trained predictors [12]. Prior networks instead parameterize a distribution over predictive distributions to separate data and distributional uncertainty [13]. These approaches are general, but their repeated computation or additional uncertainty supervision is costly for large galleries, and uncertainty quality may deteriorate under dataset shift [14]. PCME represents images and texts as distributions to capture one-to-many correspondence [9]; PCME++ introduces a closed-form probabilistic distance and improves robustness to false negatives and noisy correspondences [10]. These methods estimate uncertainty in representation space. CRU asks whether the final ranking is ambiguous or contradicted by a complementary granularity, making no Gaussian assumption and requiring one deterministic pass.

2.3 Calibration and selective prediction

A confidence score is calibrated when it reflects empirical correctness frequency. Temperature scaling is a simple post-hoc calibration method [26], and maximum softmax probability provides a common baseline for detecting prediction errors [27]. Selective prediction permits abstention on uncertain inputs, trading coverage for lower risk [28]. Retrieval confidence is often approximated by maximum score, score margin, or entropy. Such measures are branch-internal: a sharp but incorrect ranking remains overconfident. CRU augments Shannon entropy [29] with cross-branch Jensen–Shannon divergence [30] and evaluates error-detection AUROC/AUPR, ECE, Brier score, and AURC.

3 Method

3.1 Overview and problem formulation

Let $\mathcal{D} = \{(I_i, C_i)\}_{i=1}^N$ denote paired images and captions. Given an image query, image-to-text (I2T) retrieval ranks a caption gallery; text-to-image (T2I) retrieval performs the reverse operation. We write q for a query, x for an opposite-modality candidate, $\mathcal{G}_q$ for its gallery, and $\mathcal{P}_q \subseteq \mathcal{G}_q$ for the set of annotated relevant candidates. The objective is not only to learn a similarity function that places $\mathcal{P}_q$ near the top, but also to estimate whether the resulting ranking should be trusted.

The proposed Cross-granularity Ranking Uncertainty (CRU) framework obtains two deterministic retrieval opinions from a shared pair of encoders. A global branch compares image- and sentence-level semantics, whereas a local branch aggregates patch–word correspondences. The two score lists are placed on a common candidate support and converted into categorical ranking distributions. We then quantify two distinct failure modes: *ambiguity*, when a branch spreads probability over many candidates, and *disagreement*, when global and local branches favor different candidates. These quantities determine query-adaptive fusion weights and are mapped to a calibrated top-1 error probability. The complete pipeline is symmetric for I2T and T2I; hence the derivation below applies to either direction unless stated otherwise.

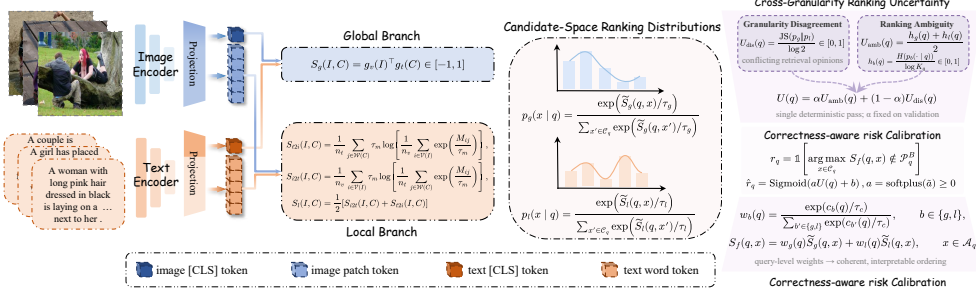


Fig. 1 Overall architecture of the proposed CRU framework.

The overall framework is illustrated in Fig. 1.

For every query, the model returns a fused similarity $S_f(q, x)$, an uncertainty value $U(q) \in [0, 1]$, and a calibrated risk $\hat{r}_q \in [0, 1]$. Here $U(q)$ is a ranking-derived uncertainty score, while $\hat{r}_q$ has the operational interpretation of the estimated probability that the top-ranked candidate is not in $\mathcal{P}_q$.

This formulation places uncertainty in the same decision space as the returned ranking rather than in an intermediate embedding. The diagnostic score $U(q)$ explains ranking instability, while $\hat{r}_q$ provides an actionable error probability.

3.2 Dual-granularity feature representation

The image encoder f_v is instantiated by a Vision Transformer (ViT) [31] or Swin Transformer [32], and the text encoder f_t by BERT [33]. They output one global token and a sequence of local tokens:

$$f_v(I) = \{v_{\text{cls}}, v_1, \dots, v_{L_v}\}, \quad v_i \in \mathbb{R}^{d_v}, \quad (1)$$

$$f_t(C) = \{t_{\text{cls}}, t_1, \dots, t_{L_t}\}, \quad t_j \in \mathbb{R}^{d_t}. \quad (2)$$

The visual class token and textual class token summarize the complete inputs, while the remaining tokens preserve patch- and word-level evidence. Because the information useful for global semantic matching is not necessarily optimal for local correspondence, we use separate projection heads:

$$g_v = \text{Norm}(W_v^g v_{\text{cls}}), \quad g_t = \text{Norm}(W_t^g t_{\text{cls}}), \quad (3)$$

$$\hat{v}_i = \text{Norm}(W_v^l v_i), \quad \hat{t}_j = \text{Norm}(W_t^l t_j), \quad (4)$$

where $g_v, g_t \in \mathbb{R}^{d_g}$, $\hat{v}_i, \hat{t}_j \in \mathbb{R}^{d_l}$, and

$$\text{Norm}(z) = \frac{z}{\max(\|z\|_2, \epsilon)}. \quad (5)$$

The small constant $\epsilon > 0$ prevents division by zero. We denote the valid visual and textual token index sets by $\mathcal{V}(I)$ and $\mathcal{W}(C)$, with cardinalities n_v and n_t . Padding tokens, textual special tokens, and the visual class token are excluded from these sets.

Explicit masking is important because otherwise sentence length and padding would alter both local similarity and its uncertainty.

Shared encoders keep computation efficient, while separate projection heads preserve the complementary evidence required for informative cross-granularity disagreement.

3.3 Complementary global and local matching

The branches are complementary retrieval views: global matching provides efficient semantic coverage, whereas local matching verifies fine-grained details. Their different error patterns make cross-granularity disagreement informative.

3.3.1 Global semantic branch

Owing to ℓ_2 normalization, the global score is cosine similarity:

$$S_g(I, C) = g_v(I)^\top g_t(C) \in [-1, 1]. \quad (6)$$

For a mini-batch containing B image–caption pairs, evaluating all cross-pair dot products produces $S_g^{i2t} \in \mathbb{R}^{B \times B}$, and $S_g^{t2i} = (S_g^{i2t})^\top$. At deployment, gallery-side global embeddings are computed once and indexed, which permits efficient approximate nearest-neighbor search over a large gallery.

The global branch captures scene-level intent and supplies a high-recall candidate neighborhood. Its indexability makes it the scalable first stage of the system.

3.3.2 Local detail branch

For an image–caption pair, the patch–word affinity matrix is

$$M_{ij} = \hat{v}_i^\top \hat{t}_j, \quad i \in \mathcal{V}(I), \quad j \in \mathcal{W}(C). \quad (7)$$

A hard maximum would retain only one correspondence for each token and yield unstable gradients. We instead use a temperature-controlled LogMeanExp operator in both directions:

$$S_{t2i}(I, C) = \frac{1}{n_t} \sum_{j \in \mathcal{W}(C)} \tau_m \log \left[\frac{1}{n_v} \sum_{i \in \mathcal{V}(I)} \exp \left(\frac{M_{ij}}{\tau_m} \right) \right], \quad (8)$$

$$S_{i2t}(I, C) = \frac{1}{n_v} \sum_{i \in \mathcal{V}(I)} \tau_m \log \left[\frac{1}{n_t} \sum_{j \in \mathcal{W}(C)} \exp \left(\frac{M_{ij}}{\tau_m} \right) \right], \quad (9)$$

$$S_l(I, C) = \frac{1}{2} [S_{i2t}(I, C) + S_{t2i}(I, C)]. \quad (10)$$

As $\tau_m \rightarrow 0^+$, each inner aggregation approaches a maximum; a larger τ_m distributes evidence over multiple alignments. The factors $1/n_v$ and $1/n_t$ inside the logarithms remove the additive $\tau_m \log n$ bias of an unnormalized LogSumExp, so candidates with

different valid token counts remain comparable. Averaging the two directions prevents either dense visual patches or a few salient words from unilaterally determining the local score.

Local comparison is more expensive than a global dot product. Therefore, at inference the global branch first retrieves a pool $\mathcal{G}_R(q) \subset \mathcal{G}_q$ of R candidates, and Eq. (10) is evaluated only inside this pool. This two-stage design retains indexability while providing fine-grained evidence for re-ranking and uncertainty estimation.

The local branch acts as a detail-sensitive verification view for attributes, quantities, and relations that may be suppressed by a global vector. Its conflict with the global branch therefore provides evidence of a granularity-dependent decision.

3.4 Candidate-space ranking distributions

Raw global and local scores have different scales and cannot be compared directly. Moreover, disagreement is meaningful only if both branches define probabilities over the same candidates. During training, the shared support $\mathcal{C}_q$ is the complete opposite-modality mini-batch. During inference, we first take the top K candidates from each branch within $\mathcal{G}_R(q)$ and form

$$\mathcal{C}_q = \text{TopK}_{x \in \mathcal{G}_R(q)} S_g(q, x) \cup \text{TopK}_{x \in \mathcal{G}_R(q)} S_l(q, x), \quad K_q = |\mathcal{C}_q|. \quad (11)$$

Thus $K \leq K_q \leq 2K$ when both lists contain K distinct elements. The union is essential: using only the global top- K list could conceal a locally preferred candidate, and vice versa. We use $K \geq 2$, ensuring $K_q > 1$ and making entropy normalization well-defined.

For branch $b \in \{g, l\}$, let the support mean and standard deviation be

$$\mu_b(q) = \frac{1}{K_q} \sum_{x \in \mathcal{C}_q} S_b(q, x), \quad (12)$$

$$\sigma_b(q) = \sqrt{\frac{1}{K_q} \sum_{x \in \mathcal{C}_q} (S_b(q, x) - \mu_b(q))^2}. \quad (13)$$

We standardize the branch scores using the same query-specific statistics:

$$\tilde{S}_b(q, x) = \frac{S_b(q, x) - \mu_b(q)}{\max(\sigma_b(q), \epsilon)}. \quad (14)$$

Equation (14) is computed from $\mathcal{C}_q$ but can be applied to every $x \in \mathcal{G}_R(q)$ when producing the final ranking. It removes affine scale differences without changing the ordering within a branch. The corresponding ranking distribution on the shared support is

$$p_b(x | q) = \frac{\exp(\tilde{S}_b(q, x)/\tau_b)}{\sum_{x' \in \mathcal{C}_q} \exp(\tilde{S}_b(q, x')/\tau_b)}, \quad x \in \mathcal{C}_q, \quad \tau_b > 0. \quad (15)$$

The temperature τ_b controls distribution sharpness and is selected on validation data. All softmax and LogMeanExp computations use the standard log-sum-exp stabilization in implementation.

This reformulation makes the two score lists comparable and focuses uncertainty on the competitive neighborhood around the returned results. The shared support ensures that entropy and divergence are computed over the same candidate alternatives.

3.5 Cross-granularity ranking uncertainty

CRU decomposes reliability into ambiguity within each ranking and disagreement between rankings. This distinguishes “neither branch separates the candidates” from “the branches confidently prefer different candidates.”

3.5.1 Within-branch ranking ambiguity

Let $H(p) = -\sum_{x \in \mathcal{C}_q} p(x) \log p(x)$ be Shannon entropy computed with natural logarithms [29]. The normalized entropy of branch b is

$$h_b(q) = \frac{H(p_b(\cdot | q))}{\log K_q} \in [0, 1]. \quad (16)$$

A value near zero indicates that the branch concentrates on one candidate, whereas a value near one indicates an almost uniform ranking. We average the two branch entropies:

$$U_{\text{amb}}(q) = \frac{h_g(q) + h_l(q)}{2}. \quad (17)$$

This term detects intrinsically difficult queries for which neither granularity separates the candidate set, but it cannot identify the case where both distributions are sharp and peak at different candidates.

High ambiguity indicates that several candidates receive similar evidence and that the top ordering is weakly supported.

3.5.2 Between-branch granularity disagreement

To capture conflicting retrieval opinions, we use the Jensen–Shannon divergence [30] and define the mixture distribution $m(x | q) = \frac{1}{2}[p_g(x | q) + p_l(x | q)]$. The Jensen–Shannon (JS) divergence is

$$\text{JS}(p_g \| p_l) = \frac{1}{2} \text{KL}(p_g \| m) + \frac{1}{2} \text{KL}(p_l \| m), \quad (18)$$

$$\text{KL}(p_b \| m) = \sum_{x \in \mathcal{C}_q} p_b(x | q) \log \frac{p_b(x | q)}{m(x | q)}. \quad (19)$$

Unlike an asymmetric KL divergence, JS is symmetric and finite because the softmax distributions have common positive support. With natural logarithms, $0 \leq$

$\text{JS}(p_g \| p_l) \leq \log 2$; hence we define

$$U_{\text{dis}}(q) = \frac{\text{JS}(p_g \| p_l)}{\log 2} \in [0, 1]. \quad (20)$$

The final uncertainty score is the convex combination

$$U(q) = \alpha U_{\text{amb}}(q) + (1 - \alpha) U_{\text{dis}}(q), \quad \alpha \in [0, 1]. \quad (21)$$

Because both components lie in $[0, 1]$, Eq. (21) also lies in $[0, 1]$. We select a single fixed α on the validation set rather than learning an unconstrained query-dependent coefficient, retaining a transparent decomposition between diffuse evidence and cross-granularity conflict.

Disagreement captures the important case in which both branches are sharp but peak at different candidates, which entropy alone would regard as confident. The coefficient α controls the balance between diffuse evidence and cross-granularity conflict.

3.6 Confidence-aware dynamic fusion

The complement of normalized entropy gives branch confidence $c_b(q) = 1 - h_b(q)$. We convert the two confidences into query-level weights:

$$w_b(q) = \frac{\exp(c_b(q)/\tau_c)}{\sum_{b' \in \{g, l\}} \exp(c_{b'}(q)/\tau_c)}, \quad b \in \{g, l\}, \quad (22)$$

$$S_f(q, x) = w_g(q) \tilde{S}_g(q, x) + w_l(q) \tilde{S}_l(q, x), \quad x \in \mathcal{A}_q, \quad (23)$$

Here $\mathcal{A}_q = \mathcal{C}_q$ during in-batch training and $\mathcal{A}_q = \mathcal{G}_R(q)$ during inference. The weights are non-negative and sum to one. A low τ_c favors the more concentrated branch, while a high τ_c approaches equal weighting. Importantly, $w_g(q)$ and $w_l(q)$ depend on the query but not on the candidate. Applying one weight pair to the entire pool yields a coherent ranking; candidate-dependent gating could change the scoring rule from item to item and make the resulting order difficult to interpret. Dynamic fusion and uncertainty serve different roles: fusion uses relative branch concentration to choose evidence, whereas $U(q)$ additionally preserves the disagreement signal that a fusion rule alone could hide.

For training and for the consistency objective below, the fused ranking distribution is

$$p_f(x | q) = \frac{\exp(S_f(q, x)/\tau_f)}{\sum_{x' \in \mathcal{C}_q} \exp(S_f(q, x')/\tau_f)}, \quad x \in \mathcal{C}_q. \quad (24)$$

Dynamic fusion gives greater influence to the branch that separates candidates more clearly. Disagreement is still retained for risk estimation, so fusion does not hide conflicts by merely averaging the two rankings.

3.7 Training objectives and correctness-aware calibration

The objective keeps both branches independently discriminative while optimizing an accurate and calibrated fused ranking. Separate supervision prevents either branch from becoming inactive.

3.7.1 Bidirectional retrieval losses

For branch $b \in \{g, l, f\}$ and a query set $\mathcal{Q}$ in a mini-batch, let $\mathcal{C}_q$ be its in-batch candidate set and $\mathcal{P}_q^B = \mathcal{P}_q \cap \mathcal{C}_q$ its in-batch positives. We use the multi-positive contrastive loss

$$\mathcal{L}_b^{q2x} = -\frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \log \frac{\sum_{x^+ \in \mathcal{P}_q^B} \exp(S_b(q, x^+)/\beta_b)}{\sum_{x \in \mathcal{C}_q} \exp(S_b(q, x)/\beta_b)}, \quad (25)$$

where $\beta_b > 0$ is the contrastive temperature. For the fused branch, S_b in Eq. (25) denotes S_f ; for the other branches it denotes the corresponding raw score. The symmetric objective is $\mathcal{L}_b = \frac{1}{2}(\mathcal{L}_b^{i2t} + \mathcal{L}_b^{t2i})$. The multi-positive numerator prevents other valid captions of the same image, when present in a batch, from being treated as negatives.

$\mathcal{L}_g$ preserves the indexable global representation, $\mathcal{L}_l$ trains fine-grained discrimination, and $\mathcal{L}_f$ directly optimizes their adaptive combination.

3.7.2 Correctness-aware risk calibration

The binary top-1 error target is

$$r_q = \mathbb{1} \left[\arg \max_{x \in \mathcal{C}_q} S_f(q, x) \notin \mathcal{P}_q^B \right]. \quad (26)$$

The discrete label is treated as a stop-gradient target. We map ranking uncertainty to predicted error probability using a monotone logistic calibrator:

$$\hat{r}_q = \text{Sigmoid}(aU(q) + b), \quad a = \text{softplus}(\tilde{a}) \geq 0. \quad (27)$$

Constraining the slope prevents the calibrator from reversing the intended ordering in which larger uncertainty implies no smaller predicted risk. Its binary cross-entropy loss is

$$\mathcal{L}_{\text{cal}} = -\frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} [r_q \log \hat{r}_q + (1 - r_q) \log(1 - \hat{r}_q)]. \quad (28)$$

For evaluation, the analogous error indicator is computed from the top-ranked item in the returned re-ranking pool $\mathcal{G}_R(q)$ and the full annotated set $\mathcal{P}_q$. The indicator is never needed to run the model: at inference, the calibrator produces $\hat{r}_q$ directly from $U(q)$ without access to relevance labels.

Calibration converts the relative uncertainty ordering into an actionable error probability that reflects the model’s empirical error rate. The monotone constraint preserves the ordering induced by $U(q)$.

3.7.3 Confidence-weighted cross-granularity consistency

For queries whose fused branch already places high probability on an annotated positive, global and local agreement is desirable. Conversely, forcing agreement on an unresolved query can make both branches confidently wrong and erase a useful uncertainty signal. We therefore use

$$\omega_q = \text{sg} \left[\sum_{x^+ \in \mathcal{P}_q^B} p_f(x^+ | q) \right], \quad (29)$$

$$\mathcal{L}_{\text{con}} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \omega_q \text{JS}(p_g(\cdot | q) \| p_l(\cdot | q)). \quad (30)$$

Here $0 \leq \omega_q \leq 1$, and $\text{sg}[\cdot]$ stops gradients through the reliability weight. The divergence itself remains differentiable, so agreement is strengthened mainly for already reliable training examples without allowing the model to reduce the loss merely by lowering ω_q .

This term reduces unnecessary branch conflicts on reliable examples while preserving informative disagreement on unresolved queries.

The complete training objective is

$$\mathcal{L} = \mathcal{L}_g + \lambda_l \mathcal{L}_l + \lambda_f \mathcal{L}_f + \lambda_c \mathcal{L}_{\text{cal}} + \lambda_d \mathcal{L}_{\text{con}}, \quad \lambda_l, \lambda_f, \lambda_c, \lambda_d \geq 0. \quad (31)$$

The global loss maintains an indexable representation, the local and fused losses improve fine-grained ranking, the calibration loss aligns uncertainty with empirical retrieval errors, and the consistency loss regularizes agreement only when supported by positive evidence.

3.8 Inference procedure and complexity

Gallery-side global embeddings and local tokens are cached before querying. For each query, CRU performs one encoder forward pass, retrieves R candidates using S_g , evaluates S_l only for these candidates, and then computes uncertainty, fusion, and calibrated risk.

Ignoring encoder cost, global exhaustive search requires $O(|\mathcal{G}_q|d_g)$ dot-product work and can be reduced with approximate indexing. Local re-ranking requires $O(Rn_v n_t d_l)$ elementary multiply-accumulate operations in the direct implementation, while uncertainty and fusion add only $O(R + K_q)$ scalar operations. Since $K_q \leq 2K \leq 2R$, the reliability computation is negligible relative to token matching. Unlike Monte Carlo dropout and ensembles, CRU neither repeats inference nor stores multiple independently trained encoders.

This separation keeps the existing global index unchanged and confines reliability estimation to the bounded re-ranking stage. The main accuracy-latency trade-off is therefore controlled by R and K without repeated stochastic inference.

4 Experiments

We evaluate CRU from four complementary perspectives: bidirectional retrieval accuracy, top-1 error detection and calibration, component contribution, and sensitivity to candidate size, local aggregation, and backbone capacity.

4.1 Datasets and protocols

Flickr30K. Flickr30K contains 31,783 images with five captions per image [34]. We use the standard Karpathy split of 29,000 training, 1,000 validation, and 1,000 test images. The 1K test protocol contains 1,000 image queries, 5,000 caption queries, and the corresponding galleries of 5,000 captions and 1,000 images.

MS-COCO. MS-COCO provides 123,287 images with five captions per image [35]. We use the standard split with 113,287 training, 5,000 validation, and 5,000 test images. The 1K protocol averages five disjoint folds, each with 1,000 images and 5,000 captions; the 5K protocol evaluates all 5,000 images against 25,000 captions. This distinction tests whether uncertainty remains useful as the gallery becomes more competitive.

Following the standard comparison protocol, we report Flickr30K-1K, MS-COCO-1K, and MS-COCO-5K. A query is labeled erroneous when none of its annotated relevant candidates is ranked first. Because caption annotations are not exhaustive, this operational label measures disagreement with the benchmark annotations rather than absolute semantic irrelevance; extended relevance judgments such as Crisscrossed Captions illustrate this limitation [36]. Calibration parameters and uncertainty hyperparameters are selected on validation data only. All CRU and reliability results are averaged over three independent seeds, while published retrieval baselines retain their reported values.

4.2 Evaluation metrics

Recall at K ($R@K$). $R@K$ is the percentage of queries for which at least one ground-truth match appears among the top K retrieved candidates. We report $R@1$, $R@5$, and $R@10$ for both image-to-text (i2t) and text-to-image (t2i) retrieval; higher values indicate better ranking accuracy. **rSum** is the sum of the six recall values from the two retrieval directions and provides a compact measure of overall retrieval quality.

AUROC and AUPR. Treating an incorrect top-1 retrieval as the positive class, AUROC measures how consistently uncertainty ranks erroneous queries above correct ones across all thresholds. AUPR summarizes the precision-recall trade-off for error detection and is particularly informative when correct and incorrect retrievals are imbalanced. Higher AUROC and AUPR indicate stronger error-separation ability.

ECE and Brier score. Expected calibration error (ECE) divides predictions into 15 equal-width confidence bins and measures the weighted gap between predicted error probability and empirical error frequency [26]. The Brier score is the mean squared difference between predicted error probability and the binary correctness target [37]. Lower values for both metrics indicate more reliable probability estimates.

AURC and latency. For selective retrieval, queries are retained from low to high uncertainty, and risk is the top-1 error rate among retained queries. The area under the

resulting risk–coverage curve (AURC) [28] summarizes performance over all coverage levels, with a lower value indicating that errors are rejected earlier. Latency reports the average per-query inference time in milliseconds and reflects the efficiency cost of candidate re-ranking and uncertainty estimation.

4.3 Implementation details

Unless stated otherwise, the visual encoder is ViT-B/16 and the text encoder is BERT-base. The global and local projection dimensions are $d_g = 512$ and $d_l = 256$. Images are resized to 224×224 , and captions are truncated to 32 wordpieces. We use AdamW with learning rates of 2×10^{-5} for the encoders and 2×10^{-4} for newly initialized heads, weight decay 0.02, batch size 128, a cosine schedule, and 5 warm-up epochs. Training lasts 30 epochs. Gradient norms are clipped at 1.0, and the checkpoint with the best validation rSum subject to the validation-AURC constraint is used for testing.

The numerical stability constant is $\epsilon = 10^{-8}$. We use $\tau_m = 0.07$ for local aggregation, $\tau_g = \tau_l = \tau_f = 0.07$ for ranking distributions, and $\beta_g = \beta_l = \beta_f = 0.07$ for contrastive training. At inference, global retrieval uses $R = 100$ and uncertainty uses $K = 32$ per branch before union. We set $\alpha = 0.5$, $\tau_c = 0.1$, and $(\lambda_l, \lambda_f, \lambda_c, \lambda_d) = (0.5, 1.0, 0.1, 0.05)$. Calibration begins after epoch 5. Hyperparameters are tuned on validation AURC while constraining rSum not to fall below fixed fusion. Latency is measured with cached gallery features, a query batch size of one, 100 warm-up queries, and 1,000 timed queries on one NVIDIA RTX 4090 GPU; data loading is excluded.

4.4 Comparison methods

Retrieval baselines cover global embedding, cross-attention, graph reasoning, hard alignment, patch slimming, asymmetric embedding, and region-uncertainty modeling: VSE++ [15], SCAN [5], SGR [18], CHAN [8], LAPS [38], AVSE [39], and GRM [40]. The matched-backbone numbers in Tables 1 and 2 follow the controlled ViT-Base-224 and BERT-base evaluation reported by GRM [40]. Reliability baselines include negative maximum score, the negative top-1/top-2 score margin, normalized ranking entropy, temperature-scaled confidence [26], and PCME++ embedding uncertainty [10]. Every reliability method uses the same candidate support, error definition, and validation split, so differences are attributable to the uncertainty score rather than to candidate construction.

5 Results and Analysis

5.1 Main retrieval results

Tables 1 and 2 report retrieval performance on Flickr30K and MS-COCO, respectively. CRU achieves rSum values of 520.4, 535.1, and 446.9 on Flickr30K-1K, MS-COCO-1K, and MS-COCO-5K. Compared with GRM, these results correspond to gains of 4.2, 2.6, and 3.9 points, respectively. The improvement is consistent across both retrieval directions: R@1 increases by 1.1/0.9 points on Flickr30K, 0.6/0.7 points on MS-COCO-1K, and 0.9/0.8 points on MS-COCO-5K for i2t/t2i. In fact, CRU improves all 18

Table 1 Bidirectional image-text retrieval on Flickr30K-1K under matched backbone settings.

Method	Flickr30K-1K						rSum
	Image to Text			Text to Image			
	R@1	R@5	R@10	R@1	R@5	R@10	
<i>ViT-Base-224 + BERT-base, 14×14 patches</i>							
VSE++	71.8	92.8	96.5	59.4	84.7	90.9	496.1
SCAN	69.5	90.9	95.6	56.4	83.1	90.0	485.6
SGR	69.7	90.8	95.2	59.1	84.1	89.9	488.7
CHAN	69.2	91.8	95.0	58.4	84.9	90.6	489.9
LAPS	74.0	93.4	97.4	62.5	87.3	92.7	507.3
AVSE	76.0	94.6	97.5	62.7	88.4	93.1	512.3
GRM	77.2	94.5	97.9	64.6	88.5	93.5	516.2
CRU	78.3	95.2	98.2	65.5	89.2	94.0	520.4

Table 2 Bidirectional image-text retrieval on MS-COCO-1K and MS-COCO-5K under the same setting as Table 1.

Method	MS-COCO-1K							MS-COCO-5K						
	Image to Text			Text to Image			rSum	Image to Text			Text to Image			rSum
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
<i>ViT-Base-224 + BERT-base, 14×14 patches</i>														
VSE++	75.0	94.6	98.0	62.7	89.4	94.9	514.6	52.4	80.3	88.8	40.6	70.4	81.1	413.4
SCAN	76.0	95.4	98.1	64.5	90.8	95.8	520.6	53.9	81.8	90.0	42.9	72.3	82.5	423.5
SGR	77.2	95.0	98.0	65.1	90.7	95.8	521.8	54.9	82.8	90.5	42.8	72.2	82.5	425.8
CHAN	77.1	95.1	98.1	65.0	91.0	96.0	522.2	56.3	83.2	90.1	43.0	72.6	82.8	428.0
LAPS	78.7	95.5	98.3	66.2	91.3	96.2	526.3	57.5	84.0	90.8	44.5	74.0	83.6	434.4
AVSE	79.8	95.6	98.3	67.0	91.5	96.3	528.5	58.8	84.3	91.0	45.1	74.3	83.9	437.4
GRM	80.8	96.7	98.9	67.6	91.8	96.7	532.5	60.1	86.7	93.2	44.4	74.5	84.1	443.0
CRU	81.4	97.0	99.1	68.3	92.3	97.0	535.1	61.0	87.2	93.6	45.2	75.3	84.6	446.9

reported recall values over GRM, showing that the rSum gain is not caused by one retrieval direction or one cutoff. Gains at R@5 and R@10 are smaller because the strongest baselines already approach saturation at these cutoffs, whereas R@1 remains sensitive to the ordering among near-duplicate candidates. The larger rSum gain on the full 5K gallery further suggests that confidence-aware fusion is especially useful when the candidate set contains more visually or linguistically similar distractors.

5.2 Uncertainty, calibration, and selective retrieval

Table 3 evaluates error detection, calibration, and selective retrieval. CRU obtains the best result for every metric on all three settings. Relative to PCME++, the strongest baseline, CRU improves AUROC by 4.5, 4.3, and 4.6 points and AUPR by 4.9, 4.7, and 4.3 points on Flickr30K-1K, MS-COCO-1K, and MS-COCO-5K, respectively. At

Table 3 Top-1 error detection, calibration, and selective retrieval on the three evaluation settings.

Setting	Method	AUROC $\uparrow$	AUPR $\uparrow$	ECE $\downarrow$	Brier $\downarrow$	AURC $\downarrow$
Flickr30K-1K	Maximum score	72.4	69.8	0.084	0.164	0.092
	Score margin	75.9	73.5	0.071	0.151	0.084
	Ranking entropy	78.6	76.1	0.063	0.145	0.078
	Temperature scaling	72.4	69.8	0.031	0.139	0.092
	PCME++	80.2	78.0	0.047	0.137	0.073
	CRU	84.7	82.9	0.024	0.121	0.061
MS-COCO-1K	Maximum score	73.1	70.6	0.087	0.169	0.098
	Score margin	77.0	74.8	0.073	0.154	0.087
	Ranking entropy	79.4	77.2	0.066	0.147	0.080
	Temperature scaling	73.1	70.6	0.034	0.143	0.098
	PCME++	81.3	79.1	0.049	0.139	0.074
	CRU	85.6	83.8	0.023	0.119	0.058
MS-COCO-5K	Maximum score	70.8	72.1	0.096	0.184	0.112
	Score margin	74.6	75.9	0.082	0.171	0.101
	Ranking entropy	77.8	79.0	0.074	0.162	0.093
	Temperature scaling	70.8	72.1	0.039	0.158	0.112
	PCME++	79.5	80.7	0.055	0.153	0.087
	CRU	84.1	85.0	0.028	0.132	0.069

the same time, ECE decreases from 0.047/0.049/0.055 to 0.024/0.023/0.028, corresponding to relative reductions of 48.9%, 53.1%, and 49.1%. Brier score is reduced by 11.7–14.4%, indicating that the improvement is not specific to a binning-based calibration measure. AURC decreases by 0.012, 0.016, and 0.018, or 16.4%, 21.6%, and 20.7% relative to PCME++, so retaining the least uncertain queries yields a consistently lower top-1 error rate. These gains indicate that cross-granularity disagreement captures error-related evidence that is not available from a single ranking distribution. Temperature scaling substantially reduces ECE but leaves AUROC, AUPR, and AURC unchanged from maximum score, confirming that a monotone post-hoc transformation calibrates probabilities but cannot improve the ordering of uncertain queries.

5.3 Ablation studies

Table 4 separates the contribution of each component to accuracy and reliability. Adding local matching to the global branch increases rSum from 529.4 to 532.0, showing that token-level evidence recovers fine-grained correspondences missed by global embeddings. With dynamic fusion enabled, ambiguity and disagreement individually produce rSum values of 534.2 and 534.5 and AUROC values of 82.4 and 83.1. The slightly stronger disagreement-only result supports the central premise that conflicts between global and local rankings reveal errors that branch-internal concentration cannot detect. Their combination is nevertheless complementary: the full uncertainty formulation reaches 535.1 rSum and 85.6 AUROC while reducing AURC to 0.058.

Table 4 Component ablation on MS-COCO-1K. G/L: global/local; DF: dynamic fusion; Amb/Dis: ambiguity/disagreement; Cal: calibration.

G	L	DF	Amb	Dis	Cal	rSum $\uparrow$	AUROC $\uparrow$	ECE $\downarrow$	AURC $\downarrow$
✓						529.4	72.8	0.082	0.096
✓	✓					532.0	74.6	0.076	0.091
✓	✓	✓	✓			534.2	82.4	0.058	0.073
✓	✓	✓		✓		534.5	83.1	0.055	0.070
✓	✓	✓	✓	✓		535.1	85.6	0.061	0.058
✓	✓	✓	✓	✓	✓	535.1	85.6	0.023	0.058

Table 5 Sensitivity to candidate size, local aggregation, and backbone. Each block changes one factor from the default.

Factor	Setting	R@1 i2t	R@1 t2i	AUROC $\uparrow$	AURC $\downarrow$	Latency (ms) $\downarrow$
Candidate K	16	81.0	67.9	83.8	0.064	6.8
	32	81.4	68.3	85.6	0.058	8.9
	64	81.6	68.4	85.9	0.057	13.7
	128	81.7	68.5	86.0	0.057	22.8
Local aggregation	Max	80.8	67.5	83.9	0.064	8.1
	Mean	80.3	66.9	82.7	0.069	7.8
	LogSumExp	81.4	68.3	85.6	0.058	8.9
	Top- r mean	81.2	68.0	84.9	0.061	9.4
Backbone	ViT-B/16 + BERT	81.4	68.3	85.6	0.058	8.9
	Swin-B + BERT	84.1	71.4	86.3	0.054	9.6
	ViT-L/14 + BERT	85.7	73.2	87.1	0.049	15.8

Correctness-aware calibration reduces ECE from 0.061 to 0.023, a relative reduction of 62.3%, without changing rSum, AUROC, or AURC. This invariance is expected because the positive-slope logistic calibrator preserves both the fused ranking and the uncertainty ordering. Compared with the global-only variant, the complete model improves rSum by 5.7 and AUROC by 12.8 points, demonstrating that accuracy and reliability benefit from different but mutually supportive components.

5.4 Sensitivity analysis

Table 5 and Figure 2 analyze the effects of candidate size, local aggregation, and backbone capacity on MS-COCO-1K.

Increasing K from 16 to 32 improves AUROC from 83.8 to 85.6 and decreases AURC from 0.064 to 0.058, with only a 2.1-ms latency increase. Larger pools provide diminishing returns: $K = 128$ improves i2t R@1 by only 0.3 and AUROC by 0.4 over $K = 32$, but raises latency from 8.9 to 22.8 ms, making inference 2.6 times slower. We therefore use $K = 32$ as the default accuracy–efficiency balance. For local aggregation, LogSumExp performs best, exceeding max pooling by 0.6/0.8 R@1 points in i2t/t2i

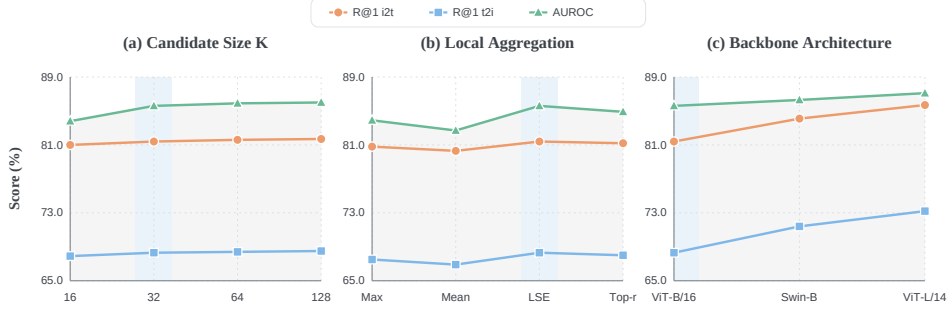


Fig. 2 Sensitivity trends for candidate size, local aggregation, and backbone architecture on MS-COCO-1K. Shaded columns indicate the default settings.

and 1.7 AUROC points. Its smooth approximation to maximum matching preserves salient token correspondences while allowing several compatible patches or words to contribute. Mean pooling is fastest but dilutes discriminative local evidence, whereas Top- r mean provides an intermediate trade-off. Finally, stronger visual backbones consistently improve all quality metrics. Moving from ViT-B/16 to Swin-B increases i2t/t2i R@1 by 2.7/3.1 points with only 0.7 ms additional latency, while ViT-L/14 reaches 85.7/73.2 R@1 and 87.1 AUROC at 15.8 ms. The monotonic AUROC improvement across all three backbones indicates that the ranking-disagreement formulation is not tied to a particular visual encoder, while the latency increase preserves the expected accuracy–efficiency trade-off.

6 Discussion

CRU estimates decision uncertainty from diffuse candidate evidence and conflicts between global and local rankings. This quantity is directly connected to retrieval errors, although it should not be interpreted as Bayesian epistemic uncertainty. Its value also depends on the candidate support, so comparisons across different gallery sizes or candidate-construction protocols require recalibration. The ablation results show that ambiguity and disagreement remain complementary after query-wise score standardization, indicating that the observed conflict is not merely caused by score-scale differences.

Several limitations remain. Flickr30K and MS-COCO do not enumerate every semantically relevant pair, and extended judgments reveal additional plausible matches [36]; consequently, calibration is defined relative to the available annotations. The method also assumes that global and local branches exhibit sufficiently different failure modes, and local token matching introduces additional cost. The two-stage design bounds this overhead by applying local matching only to a small candidate pool, while calibrated risk can support abstention or more expensive processing for difficult queries. Severe distribution shift may still degrade calibration, motivating future work on conformal guarantees and online recalibration.

7 Conclusion

We presented CRU, a cross-granularity framework for uncertainty-aware image–text retrieval. Global semantics and local details act as complementary ranking opinions; uncertainty decomposes into within-branch ambiguity and between-branch disagreement; and branch confidence controls query-level fusion. A correctness-aware calibrator converts this score into empirical top-1 error probability, enabling selective retrieval in one deterministic pass. Experiments on Flickr30K-1K, MS-COCO-1K, and MS-COCO-5K yield rSum values of 520.4, 535.1, and 446.9. CRU also improves AUROC by 4.3–4.6 points over PCME++ while reducing ECE to 0.023–0.028 and AURC to 0.058–0.069. The framework therefore provides a practical route from accurate ranking to retrieval systems that can also indicate when their outputs should not be trusted.

Declarations

- **Funding:** The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.
- **Conflict of interest:** The authors declare that they have no conflict of interest.
- **Ethics approval and consent to participate:** Not applicable. This study does not involve human participants, human data, or animals.
- **Consent for publication:** All authors have read and approved the final manuscript and consent to its publication.
- **Data availability:** The datasets analyzed in this study are publicly available. Flickr30K is available at <https://shannon.cs.illinois.edu/DenotationGraph/>, and MS-COCO is available at <https://cocodataset.org/>.
- **Materials availability:** Not applicable.
- **Code availability:** The code supporting the findings of this study is available from the corresponding author upon reasonable request.
- **Author contributions:** Both authors contributed to the study conception and design, methodology development, experimental analysis, and manuscript preparation. Both authors read and approved the final manuscript.

References

- [1] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763 (2021)
- [2] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y.-H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 4904–4916 (2021)

- [3] Li, J., Selvaraju, R.R., Gotmare, A.D., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: *Advances in Neural Information Processing Systems*, vol. 34, pp. 26182–26196 (2021)
- [4] Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: *Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 162, pp. 12888–12900 (2022)
- [5] Lee, K.-H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: *European Conference on Computer Vision*, pp. 201–216 (2018)
- [6] Chen, H., Ding, G., Liu, X., Lin, Z., Liu, J., Han, J.: IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12655–12663 (2020)
- [7] Zhang, K., Mao, Z., Wang, Q., Zhang, Y.: Negative-aware attention framework for image-text matching. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15661–15670 (2022)
- [8] Pan, Z., Wu, F., Zhang, B.: Fine-grained image-text matching by cross-modal hard aligning network. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19275–19284 (2023)
- [9] Chun, S., Oh, S.J., Rezende, R.S., Kalantidis, Y., Larlus, D.: Probabilistic embeddings for cross-modal retrieval. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8415–8424 (2021)
- [10] Chun, S.: Improved probabilistic image-text representations. In: *International Conference on Learning Representations* (2024)
- [11] Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: *Proceedings of the 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 48, pp. 1050–1059 (2016)
- [12] Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: *Advances in Neural Information Processing Systems*, vol. 30 (2017)
- [13] Malinin, A., Gales, M.: Predictive uncertainty estimation via prior networks. In: *Advances in Neural Information Processing Systems*, vol. 31 (2018)
- [14] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J.V., Lakshminarayanan, B., Snoek, J.: Can you trust your model’s uncertainty?

- evaluating predictive uncertainty under dataset shift. In: *Advances in Neural Information Processing Systems*, vol. 32 (2019)
- [15] Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: Improving visual-semantic embeddings with hard negatives. In: *British Machine Vision Conference* (2018)
 - [16] Li, K., Zhang, Y., Li, K., Li, Y., Fu, Y.: Visual semantic reasoning for image-text matching. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4654–4662 (2019)
 - [17] Chen, J., Hu, H., Wu, H., Jiang, Y., Wang, C.: Learning the best pooling strategy for visual semantic embedding. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15789–15798 (2021)
 - [18] Diao, H., Zhang, Y., Ma, L., Lu, H.: Similarity reasoning and filtration for image-text matching. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 1218–1226 (2021)
 - [19] Zhang, H., Mao, Z., Zhang, K., Zhang, Y.: Show your faith: Cross-modal confidence-aware network for image-text matching. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 3262–3270 (2022). <https://doi.org/10.1609/aaai.v36i3.20235>
 - [20] Kim, W., Son, B., Kim, I.: ViLT: Vision-and-language transformer without convolution or region supervision. In: *Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 139, pp. 5583–5594 (2021)
 - [21] Xie, L., Pan, P., Lu, Y.: Analyzing semantic correlation for cross-modal retrieval. *Multimedia Systems* **21**(6), 525–539 (2015) <https://doi.org/10.1007/s00530-014-0397-6>
 - [22] Jiang, A., Li, H., Li, Y., Wang, M.: Learning discriminative representations for semantical crossmodal retrieval. *Multimedia Systems* **24**, 111–121 (2018) <https://doi.org/10.1007/s00530-016-0532-7>
 - [23] Li, Y., Tang, X., Lu, J., Huang, Y.: Dual graph-structured semantics multi-subspace learning for cross-modal retrieval. *Multimedia Systems* **30**, 294 (2024) <https://doi.org/10.1007/s00530-024-01471-0>
 - [24] Wang, H., Zhao, K., Zhao, D.: A triple fusion model for cross-modal deep hashing retrieval. *Multimedia Systems* **29**, 347–359 (2023) <https://doi.org/10.1007/s00530-022-01005-6>
 - [25] Liu, P., Liu, X.: Complementary expert balanced learning for long-tail cross-modal retrieval. *Multimedia Systems* **30**, 113 (2024) <https://doi.org/10.1007/s00530-024-01317-9>

- [26] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330 (2017)
- [27] Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017)
- [28] Geifman, Y., El-Yaniv, R.: SelectiveNet: A deep neural network with an integrated reject option. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 2151–2159 (2019)
- [29] Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27**(3–4), 379–423 (1948) <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- [30] Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory **37**(1), 145–151 (1991) <https://doi.org/10.1109/18.61115>
- [31] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
- [32] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical vision transformer using shifted windows. In: IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
- [33] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186 (2019)
- [34] Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)
- [35] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision, pp. 740–755 (2014)
- [36] Parekh, Z., Baldridge, J., Cer, D., Waters, A., Yang, Y.: Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO. In: Conference of the European Chapter of the Association for

Computational Linguistics, pp. 2855–2870 (2021)

- [37] Brier, G.W.: Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**(1), 1–3 (1950) [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)
- [38] Fu, Z., Zhang, L., Xia, H., Mao, Z.: Linguistic-aware patch slimming framework for fine-grained cross-modal alignment. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26307–26316 (2024)
- [39] Liu, Y., Liu, M., Huang, S., Lv, J.: Asymmetric visual semantic embedding framework for efficient vision-language alignment. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 5676–5684 (2025)
- [40] Liu, J., Zhou, H., Liu, Y., Chen, B., Jiang, Y.: Cross modal fine-grained alignment via granularity-aware and region-uncertain modeling. In: *Proceedings of the AAAI Conference on Artificial Intelligence* (2026). In press