Noise Self-Correction via Relation Propagation for Robust Cross-Modal Retrieval

Ruoxuan Li*
Nanjing University of Science and Technology
Nanjing, China
ruoxuan_li@njust.edu.cn

Xiangyu Wu*
Nanjing University of Science and Technology
Nanjing, China
wxy_yyjhl@njust.edu.cn

Yang Yang†
Nanjing University of Science and Technology
Nanjing, China
yyang@njust.edu.cn

Abstract

Cross-modal retrieval refers to identifying semantically relevant data across different modalities. However, annotation errors or inherent ambiguity can cause semantic inconsistency in sample pairs, degrading retrieval performance. Prior efforts either relied heavily on the quality of explicitly dividing clean and noisy subsets, or solely leveraged carefully selected single anchor information, neglecting relationships among diverse neighbors. In this paper, we propose a novel **G**raph-based **L**abel **P**ropagation (**GLP**) framework that learns pseudo-labels via label propagation on a sparse graph, enabling self-correction of noisy labels. Specifically, each modality's instances are treated as nodes, connected via k-nearest neighbor (kNN) search to form a sparse graph. Pseudo-label vectors are generated for all nodes within one modality to capture the matching degree of inter-modal nodes. Through iterative label propagation, the stabilized pseudo-labels implicitly exploit both intra- and inter-modal relationships to derive a reliable matching degree. A dynamic queue further enhances graph quality by updating high-quality nodes. Experiments on Flickr30K, MSCOCO, and CC120K show that our method outperforms state-of-the-art approaches, especially under high noise. Code is available at https://github.com/njustkmg/MM25-GLP.

CCS Concepts

• **Information systems** → **Multimedia and multimodal retrieval**; • **Computing methodologies** → *Machine learning*.

Keywords

Cross-Modal Retrieval, Noisy Correspondence Learning, Label Propagation

ACM Reference Format:
Ruoxuan Li, Xiangyu Wu, and Yang Yang. 2025. Noise Self-Correction via Relation Propagation for Robust Cross-Modal Retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, Dublin, Leinster, Ireland, 10 pages. https://doi.org/10.1145/3746027.3755585

*Ruoxuan Li and Xiangyu Wu are joint first authors.
†Corresponding author.

MM '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755585

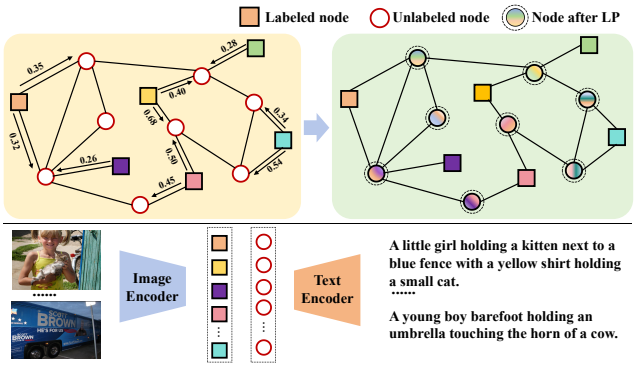

Figure 1: Illustration of GLP. Taking image-text retrieval as an example, each image and text feature is treated as a node in a sparse graph, with one modality's nodes serving as labeled nodes and the other modality's nodes as unlabeled. Through label propagation, the unlabeled nodes can acquire the matching degree with the labeled nodes.

1 Introduction

Cross-modal retrieval (CMR) [1, 2, 3, 4, 5, 6] aims to identify semantically relevant samples from one modality using a query from another. However, most existing studies overlook the potential data noise present in real-world scenarios, where the semantic relationships between sample pairs may not be perfectly matched. In reality, some widely used datasets, such as Conceptual Captions [7], are constructed through non-expert annotations or web crawling, which inevitably leads to mismatches or partial matches, known as noisy correspondences [8, 9, 10, 11]. Training CMR models on poorly matched datasets without discrimination can disrupt the alignment of modality representations in the feature space, ultimately degrading cross-modal retrieval performance [8, 12, 13, 14].

To mitigate the impact of noisy correspondence, a mainstream strategy focuses on enhancing the discriminability between clean and noisy pairs, while treating clean pairs as anchor points to correct noisy ones. Some attempts [11, 12, 15, 16] leverage the memory effect of neural networks, which prioritizes learning patterns from well-matched pairs early in training. Binary mixture models, such as Gaussian-Mixture-Model (GMM) or Beta-Mixture-Model (BMM), are employed to model the loss probability distribution, categorizing the datasets into clean and noisy subsets. These methods typically rely on a co-teaching framework to prevent noise accumulation. To avoid misclassification of ambiguous, noisy pairs as clean, recent works [10, 16, 9] design more nuanced data partition strategies. For instance, CTPR [10] and CREAM [16] integrate GMM

predictions to classify the training set into clean, vague, and noisy subsets. Similarly, UGNCL [9] divides the data into four subsets based on the uncertainty-aware metric: certain clean and noisy, and uncertain clean and noisy, subsequently reconstructing labels for the uncertain subsets.

Following the partitioning of clean and noisy subsets, these works identify anchor points–the most semantically aligned pairs within the clean subset, to correct labels in the noisy subset. NCR [8] and CTPR [10] compute soft labels by measuring the discrepancy between a target pair's similarity and batch-averaged similarity of other pairs. NPC [11] estimates sample noise levels by monitoring variations in cross-entropy loss when models trained on noisy data are evaluated against the corresponding clean anchor point.

Despite these advances, existing works always rely on intricate mechanisms to partition clean and noisy data, and even further subdivide into granular sets. Unfortunately, a critical limitation arises when hyperparameters such as confidence thresholds are misconfigured, leading to misclassification and exacerbating error accumulation. Furthermore, most methods rectify noisy labels using only a single clean anchor point, overlooking the broader neighborhood information embedded in the data manifold.

Inspired by **L**abel **P**ropagation (**LP**) theory in multi-modal classification, we regard cross-modal retrieval as a classification proxy, that is, to match the most similar item among candidate targets. Based on the assumption of "proximity similarity" of LP, we propose a novel label correction approach named **G**raph-based **L**abel **P**robagation (**GLP**), in an end-to-end manner. GLP treats each modality feature in a batch as a graph node and constructs intra-modality and inter-modality edges via k-nearest neighbors, leveraging local neighborhood relationships for robust cross-modal similarity estimation. Each node within the same modality is initialized with a one-hot pseudo-label vector, which is iteratively updated based on the probability transfer matrix of the graph until stabilized. Unlike existing works requiring explicit partitioning of clean and noisy subsets, GLP avoids this dependency by directly integrating both intra- and cross-modal neighborhood information into similarity computation. Conceptually, for a sample pair with inherent proximity, LP aggregates features from local neighbors to derive a robust matching degree estimate. Conversely, pairs lying outside the K-nearest neighbor radius are assigned a negligible matching degree due to the graph's sparsity, effectively truncating their influence. This selective propagation mechanism ensures cautious label refinement, mitigating the risk of overfitting to noisy correspondences. In addition, to further ensure the reliability of graph construction, we maintain a dynamically updated queue to store some higher-quality sample pairs. Benefiting from the nearly linear O(E) time complexity of GLP, the sparse graph we constructed offers significant advantages in training cost. To sum up, the main contributions of this work are outlined as follows:

- We propose **G**raph-based **L**abel **P**robagation (**GLP**) for robust cross-modal retrieval, a novel framework to rectify noisy labels in an end-to-end manner.
- **GLP** eliminates the need for explicit partitioning of clean and noisy subsets, fully leveraging both intra-modality and inter-modality neighborhood information to enable more reliable similarity estimation.
- The proposed **GLP** method is extensively evaluated on three benchmark datasets (Flickr30K, MSCOCO, and CC120K) and varying noisy levels, demonstrating its consistent advantages over current state-of-the-art (SOTA) methods.

2 Related Works

2.1 Noisy Correspondence Learning under CMR

Cross-Modal Retrieval (CMR) [17, 18] aims to search for relevant items across different modalities. Traditional image-text matching methods align cross-modal samples through similarity measurements, which can be categorized into two approaches: a) Coarse-grained metrics [19, 20, 21], focusing on global feature alignment, and b) Fine-grained metrics [1, 22, 23, 2, 24, 25], which emphasize evaluating semantic relationships between localized segments. Recent advances in vision-language models (VLMs) [26, 27], and pre-trained vision-language models (VLPs) [4] like CLIP [28] have demonstrated strong performance in cross-modal tasks [29, 30, 31]. Despite their zero-shot capabilities, VLPs remain sensitive to noisy training data in downstream tasks [11].

Noisy Correspondence Learning (NCL) refers to designing anti-noise methods that mitigate the negative effects caused by mismatched sample pairs within datasets. First introduced by [8], early works [11, 12, 15, 16, 32] primarily leverage the memory effect inherent in neural networks to identify clean pairs in the early training stage, subsequently estimating soft labels for noisy pairs via co-teaching architectures. Following these early efforts, subsequent work refines data partitioning strategies [10, 16, 9] to filter ambiguous mismatches or construct pseudo-correspondences [33, 34]. In addition, some studies [35, 36, 37] improve performance by applying constraints based on intrinsic properties observed in the data. Another approach involves constructing robust loss functions [13, 38, 39] to tackle the challenge of noisy correspondences (NC). Despite these efforts, existing NCL methods rely on complex mechanisms to partition clean and noisy data, where improper hyperparameter settings may lead to sample misclassification and error accumulation. Additionally, they overlook broader neighborhood information that could enhance label refinement. In contrast, we propose a lightweight framework that adaptively predicts pair-matching degrees while maintaining robustness to noise.

2.2 Label Propagation

Label propagation is a graph-based technique for label-efficient learning tasks [40, 41, 42]. The fundamental assumption behind LP is that labels change gradually across a graph, with neighboring nodes tending to have the same label. In transductive learning, where all test samples are accessible during inference, LP works by transferring labels from labeled nodes to the unlabeled ones. Recently, some works [30, 43, 44, 31] have applied label propagation to Vision-Language Models (VLMs) to enhance their performance in zero-shot or few-shot settings for downstream tasks. [30, 43, 44] are classification tasks where graphs are constructed on downstream test data, with text prompts from classes serving as labeled nodes, and labels are propagated to visual nodes. In a classification scenario, text nodes are similar to class prototypes, with fewer text nodes and more image nodes. In this work, LP is tailored for cross-modal retrieval. We perform a bidirectional inference process, and the

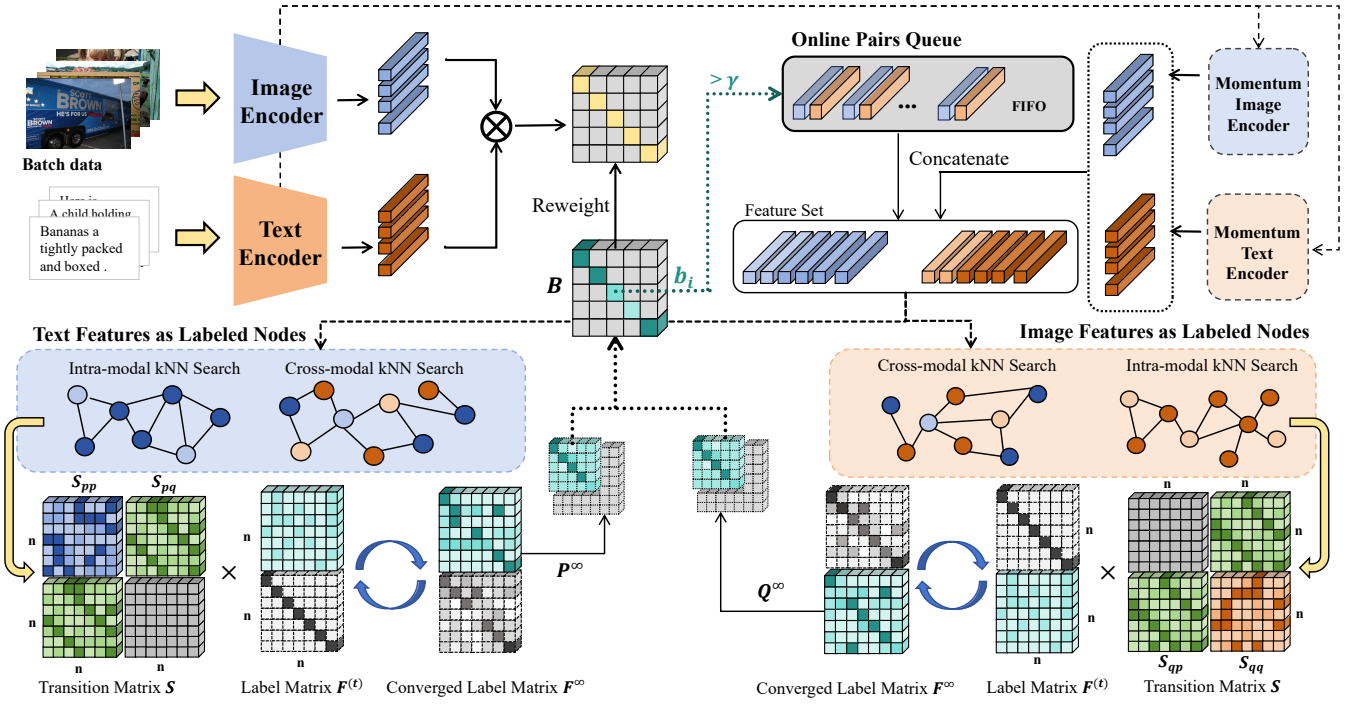

Figure 2: Framework of our proposed GLP. Given a batch of image-text pairs, features are extracted via momentum encoders, combined with historical features from a queue to form feature sets $\mathcal{P}$ and $\mathcal{Q}$. Intra-modal transition matrices S_{pp} and S_{qq} are constructed via k-NN search within each modality separately, while cross-modal transition matrices S_{pq} and S_{qp} are derived from k-NN search across modalities. A label matrix F is initialized and iteratively refined via LP until stabilized. The bidirectionally constrained iterative results are denoted as P^{∞} and Q^{∞}, respectively. Their fused output B is employed to reweight the original InfoNCE loss. High-confidence sample pairs are added to the queue following the First-In-First-Out (FIFO) policy.

number of labeled nodes in the graph is larger and more sparsely distributed.

3 Methodology

In this section, we present our GLP framework, which leverages the relationships between intra-modal and inter-modal neighboring nodes to optimize the matching degree of sample pairs. In Sec. 3.1, we define the cross-modal retrieval problem. In Sec. 3.2, we introduce the graph construction method, detailing the process of generating pseudo-label vectors and using label propagation to achieve a stable state for pseudo-labels, thereby enabling noise self-correction. In Sec. 3.3 we describe the integration of a queue mechanism to further enhance the stability of the graph.

3.1 Problem Definition

Cross-modal retrieval [28, 4, 45] lies in mapping data from different modalities into an aligned shared space. Consider the image-text retrieval task as an example. Given a multi-modal dataset $\mathcal{D} = \{(I_i, T_i), y_i\}_{i=1}^{N}$ of size N, where (I_i, T_i) represents i^{th} image-text pair and y_i is the label ($y_i = 1$ for a match, $y_i = 0$ for a mismatch). The goal is to obtain a metric space where, for an input image query, text samples semantically aligned with it are closer, and misaligned ones are farther.

CLIP [28] is a pretrained vision-language model that performs a proxy task, specifically predicting the correct pairing between text and images. The CLIP model consists of an image encoder $f(\cdot)$ and a text encoder $g(\cdot)$. Following previous SOTA [11], we adopt the pretrained CLIP as the backbone for the subsequent construction of the cross-modal retrieval model.

InfoNCE [46] is the standard optimization objective in cross-modal retrieval tasks. The objective is to maximize the similarity between the target sample and the positive samples while minimizing the similarity between the target sample and the negative samples, thus learning feature vectors with representational capabilities. For the similarity matrix computed from a training batch of size n, InfoNCE considers the elements on the diagonal of the matrix as positive samples and the remaining $(n-1)$ samples as negative samples, which is defined as follows:

$$\mathcal{L}_{\text{InfoNCE}}(I,T) = -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{\exp(\text{Sim}(I_i,T_i)/\tau)}{\sum_{j=1}^{n}\exp(\text{Sim}(I_i,T_j)/\tau)}\right), \quad (1)$$

$$\mathcal{L}_{ce} = \mathcal{L}_{\text{InfoNCE}}(I,T) + \mathcal{L}_{\text{InfoNCE}}(T,I), \quad (2)$$

where $\text{Sim}(I_i, T_j)$ represents the similarity between the query sample I_i and the candidate sample T_j, which is typically measured

using cosine similarity:

$$\mathrm{Sim}(I_i, T_j) = \frac{f(I_i) \cdot g(T_j)}{\|f(I_i)\| \cdot \|g(T_j)\|}, \tag{3}$$

τ is the temperature parameter, which adjusts the scale of similarity scores. A smaller τ makes the contrastive learning process more sensitive.

3.2 Label Propagation on the Graph

3.2.1 Graph Construction

Given a training batch of size n, we define image and text instance subsets as $\mathcal{I} = \{I_1, I_2, ..., I_n\}$ and $\mathcal{T} = \{T_1, T_2, ..., T_n\}$. Both are encoded into feature representations $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n\}$ and $\mathcal{Q} = \{\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_n\}$ using uni-modal encoders $f(\cdot)$ and $g(\cdot)$ of CLIP, respectively. Each feature instance in $\mathcal{P}$ and $\mathcal{Q}$ is treated as a node in a graph, with the adjacency matrix $\mathbf{A} \in \mathbb{R}^{(n+n)\times(n+n)}$ encoding pairwise similarity scores. Inspired by the decomposition strategy in [47], we decompose $\mathbf{A}$ into intra-modal and inter-modal blocks as below:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{pp} & \mathbf{A}_{pq} \\ \mathbf{A}_{qp} & \mathbf{A}_{qq} \end{bmatrix}, \tag{4}$$

where $\mathbf{A}_{pp}, \mathbf{A}_{qq} \in \mathbb{R}^{n\times n}$ capture intra-modal (image-to-image, text-to-text) similarities, and $\mathbf{A}_{pq}, \mathbf{A}_{qp} \in \mathbb{R}^{n\times n}$ represent cross-modal relationships. Notably, the modality gap [48] inherent in vision-language models causes imbalanced similarity distributions between intra- and inter-modal features. To mitigate this, we avoid standard k-nearest neighbor (kNN) searches over the union set $\mathcal{U} = \mathcal{P} + \mathcal{Q}$, which disproportionately favors intra-modal connections. Instead, we perform separate kNN searches:

(1) Intra-modal: Within $\mathcal{P}$ and $\mathcal{Q}$ to construct $\mathbf{A}_{pp}$ and $\mathbf{A}_{qq}$.
(2) Cross-modal: Between $\mathcal{P}$ and $\mathcal{Q}$ to populate $\mathbf{A}_{pq}$ and $\mathbf{A}_{qp}$.

We denote by $\mathrm{kNN}_p(\cdot)$ the k-nearest neighbors searched *within* the image feature set $\mathcal{P}$, and by $\mathrm{kNN}_q(\cdot)$ those searched *within* the text feature set $\mathcal{Q}$. Formally, entries in $\mathbf{A}$ are defines as:

$$a_{ij} = \begin{cases} \mathbf{p}_i^\top \mathbf{p}_j, & \text{if } \mathbf{p}_j \in \mathrm{kNN}_p(\mathbf{p}_i) \text{ and } \mathbf{p}_i \in \mathrm{kNN}_p(\mathbf{p}_j) \text{ and } i \neq j, \\ \mathbf{q}_i^\top \mathbf{q}_j, & \text{if } \mathbf{q}_j \in \mathrm{kNN}_q(\mathbf{q}_i) \text{ and } \mathbf{q}_i \in \mathrm{kNN}_q(\mathbf{q}_j) \text{ and } i \neq j, \\ \mathbf{p}_i^\top \mathbf{q}_j, & \text{if } \mathbf{p}_i \in \mathrm{kNN}_p(\mathbf{q}_j) \text{ and } \mathbf{q}_j \in \mathrm{kNN}_p(\mathbf{p}_i), \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

As we only consider the affinities around each query node and its nearest neighbors, the adjacency matrix is sparse.

3.2.2 Label Propagation by Walking on the Graph

The idea of label propagation is to assign labels to unlabeled nodes in a graph by leveraging the information from labeled nodes. This is achieved by constructing a set of label vectors and iteratively propagating label information across the graph structure. To prepare for label propagation, we first convert the adjacency matrix $\mathbf{A}$ into a transition matrix through normalization. Given that $\mathbf{A}$ comprises blocks of similarity scores from distinct embedding spaces, we decompose $\mathbf{A}$ and normalize each block independently. The intra-modal block $\mathbf{A}_{pp}$ and $\mathbf{A}_{qq}$ are symmetrically normalized as follows:

$$\begin{cases} \mathrm{S}_{pp} = \mathbf{D}_{pp}^{-1/2} \mathbf{A}_{pp} \mathbf{D}_{pp}^{-1/2}, \\ \mathrm{S}_{qq} = \mathbf{D}_{qq}^{-1/2} \mathbf{A}_{qq} \mathbf{D}_{qq}^{-1/2}, \end{cases} \tag{6}$$

where $\mathbf{D}_{pp} = \mathrm{Diag}(\mathbf{A}_{pp}\mathbf{1}_n), \mathbf{D}_{qq} = \mathrm{Diag}(\mathbf{A}_{qq}\mathbf{1}_n)$ are the degree matrices, and $\mathbf{1}_n$ is an n-dimensional all-ones vector. Conversely, the cross-modal blocks $\mathbf{A}_{pq}$ and $\mathbf{A}_{qp}$ are l_1-normalized row-wise to produce S_{pq} and S_{qp}. The overall transition matrix S is then composed as:

$$\mathrm{S} = \begin{bmatrix} \mathrm{S}_{pp} & \mathrm{S}_{pq} \\ \mathrm{S}_{qp} & \mathrm{S}_{qq} \end{bmatrix}. \tag{7}$$

Considering the two-modalities scenario, we treat nodes from one modality as labeled and nodes from the other as unlabeled. This setup enables tailoring label propagation to estimate cross-modal matching degrees by assigning labels to unlabeled nodes. Specifically, we adopt a bi-directional propagation strategy to fully exploit the mutual associations between modalities. For clarity of exposition, we designate nodes in subset $\mathcal{P}$ (image modality) as the labeled nodes and those in subset $\mathcal{Q}$ (text modality) as unlabeled nodes. For a given node $\mathbf{p}_i \in \mathcal{P}$, we construct a pseudo-label vector:

$$\mathbf{f}_i = [\mathbf{f}_{p,i}^\top, \mathbf{f}_{q,i}^\top]^\top, \tag{8}$$

where $\mathbf{f}_{p,i} \in \mathbb{R}^n$ and $\mathbf{f}_{q,i} \in \mathbb{R}^n$ represent similarity scores between $\mathbf{p}_i$ and nodes in subsets $\mathcal{P}$ and $\mathcal{Q}$, respectively. The initial vector $\mathbf{f}_{p,i}^{(0)}$ is a one-hot vector with a single non-zero element at index i, whereas $\mathbf{f}_{q,i}^{(0)}$ is initialized as the zero vector. The combination of all $\mathbf{f}_i$ for nodes in $\mathcal{P}$ forms the pseudo-label matrix $\mathrm{F} \in \mathbb{R}^{(2n)\times n}$. Given the transition matrix S, label propagation is an iterative process given by

$$\mathbf{f}_i^{(t+1)} = \alpha \mathrm{S} \mathbf{f}_i^{(t)} + (1-\alpha)\mathbf{f}_i^{(0)}, \tag{9}$$

where $\alpha \in (0, 1)$ is the propagation magnitude. Substituting Eq.7 and Eq.8 into Eq.9, the component-wise updates become:

$$\begin{cases} \mathbf{f}_{p,i}^{(t+1)} = \alpha \mathrm{S}_{pp} \mathbf{f}_{p,i}^{(t)} + \alpha \mathrm{S}_{pq} \mathbf{f}_{q,i}^{(t)} + (1-\alpha)\mathbf{f}_{p,i}^{(0)}, \\ \mathbf{f}_{q,i}^{(t+1)} = \alpha \mathrm{S}_{qp} \mathbf{f}_{p,i}^{(t)} + \alpha \mathrm{S}_{qq} \mathbf{f}_{q,i}^{(t)} + (1-\alpha)\mathbf{f}_{q,i}^{(0)}. \end{cases} \tag{10}$$

To prevent direct connections among labeled nodes, which is beneficial since each node corresponds to a different class, we set $\mathrm{S}_{pp} = 0$. Moreover, our primary interest lies in estimating the similarity between the labeled node $\mathbf{p}_i$ and all heterogeneous-modality nodes in $\mathcal{Q}$, i.e., the estimation of $\mathbf{f}_{q,i}$. Under these considerations, the update rule for $\mathbf{f}_{q,i}$ simplifies to:

$$\mathbf{f}_{q,i}^{(t+1)} = \alpha \mathrm{S}_{qq} \mathbf{f}_{q,i}^{(t)} + \alpha^2 \mathrm{S}_{qp} \mathrm{S}_{pq} \mathbf{f}_{q,i}^{(t-1)} + \alpha(1-\alpha)\mathbf{f}_{p,i}^{(0)}. \tag{11}$$

This iterative process admits a closed-form stationary solution:

$$\mathbf{f}_{q,i}^{\infty} = (I - \alpha \mathrm{S}_{qq} - \alpha^2 \mathrm{S}_{qp} \mathrm{S}_{pq})^{-1} \cdot \alpha(1-\alpha)\mathbf{f}_{p,i}^{(0)}. \tag{12}$$

Given that each $\mathbf{f}_{p,i}^{(0)}, i \in \{0, 1, ..., n\}$ is a one-hot vector, we can express the collective solution as a label matrix:

$$\mathbf{Q}^{\infty} = \alpha(1-\alpha)(I - \alpha \mathrm{S}_{qq} - \alpha^2 \mathrm{S}_{qp} \mathrm{S}_{pq})^{-1} \mathrm{S}_{qp}. \tag{13}$$

Similarly, we obtain

$$\mathbf{P}^{\infty} = \alpha(1-\alpha)(I - \alpha \mathrm{S}_{pp} - \alpha^2 \mathrm{S}_{pq} \mathrm{S}_{qp})^{-1} \mathrm{S}_{pq}. \tag{14}$$

The terms S_{qp} and S_{pq} serve as cross-modal similarity matrices responsible for transmitting information between different modalities. From a manifold learning perspective, the inverse matrix

$(I - \alpha \mathbf{S}_{qq} - \alpha^2 \mathbf{S}_{qp}\mathbf{S}_{pq})^{-1}$ and $(I - \alpha \mathbf{S}_{pp} - \alpha^2 \mathbf{S}_{pq}\mathbf{S}_{qp})^{-1}$ facilitate smoothing of predictions along the intrinsic data geometry within modality $\mathcal{Q}$ and $\mathcal{P}$, respectively, while incorporating second-order cross-modal paths to enhance the expressiveness of the propagation manifold. After column-wise normalization of $\mathbf{P}^\infty$ and $\mathbf{Q}^\infty$, the bi-directional propagation result is computed as:

$$\mathbf{B} = \lambda \mathbf{P}^\infty_{\text{norm}} + (1-\lambda)(\mathbf{Q}^\infty_{\text{norm}})^\top. \tag{15}$$

where the i-th diagonal element $\boldsymbol{b_i}$ of $\mathbf{B}$ denotes the matching degree of the i-th image-text pair in the current batch.

Leveraging these matching degrees, we re-weight the InfoNCE loss at the sample level, yielding the refined loss function:

$$\mathcal{L}_{\text{Re-InfoNCE}}(I,T) = -\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{b_i} \log\left(\frac{\exp(\text{Sim}(I_i,T_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{Sim}(I_i,T_j)/\tau)}\right). \tag{16}$$

3.3 Dynamic Queue for Robust Pseudo-label Estimation

In scenarios with severe noisy correspondence, the affinity matrix tends to become unreliable because of the inherent fragility in cross-modal connections. To enhance the robustness of label propagation, we incorporate two key components: (1) a queue mechanism storing high-quality historical sample pairs, and (2) a momentum-based model for stable feature representation.

Queue Mechanism. To improve the robustness of the transition matrix, we implement a fixed-capacity memory queue that archives high-consistency historical samples. For each batch, we compute pairwise matching degrees b_i and select reliable pairs exceeding a predefined threshold γ (where $b_i > \gamma$) for queue inclusion. The label propagation graph is constructed by augmenting current batch samples with historically preserved samples from the queue, thereby extending the context for pseudo-label generation. The queue follows a First-In-First-Out (FIFO) replacement policy with dynamic updating, ensuring only the most discriminative samples are retained. This mechanism guarantees that label propagation consistently operates on high-fidelity, information-rich exemplars.

Momentum Model. To ensure stable pseudo-label generation, we adopt a momentum-encoded model following the MoCo [49] paradigm. This design serves two critical purposes:

- Representation Space Stabilization: The momentum model maintains feature consistency across training iterations through exponential moving averaging (EMA), mitigating representation drift when utilizing historical samples from the queue.
- Smooth Transition Modeling: The gradually evolved feature space yields more reliable similarity measurements for graph construction in label propagation.

The parameters are updated via:

$$\theta^{(t)}_{\text{mom}} = \beta \theta^{(t-1)}_{\text{mom}} + (1-\beta)\theta^{(t)}, \tag{17}$$

where $\theta^{(t)}_{\text{mom}}$ denotes the momentum-encoded parameters at step t. This EMA operation produces noise-resistant features with reduced temporal variance, facilitating more optimized affinity matrix construction compared to direct usage of raw features. The algorithmic workflow is presented in Algorithm 1.

Algorithm 1 Pipeline of learning with our GLP method.

Input: Multi-modal dataset $\mathcal{D} = \{(I_i, T_i), y_i\}_{i=1}^N$
Initialize parameters for model M and its momentum version M_{mom}, an empty queue $\mathcal{J}$.
for each epoch $t = 1, 2, \ldots, T$ **do**
  for each minibatch B from $\mathcal{D}$ **do**
    Compute instance features using M_{mom}: $f(\cdot), g(\cdot)$
    Combine batch features and queue features to get subsets $\mathcal{P}$ and $\mathcal{Q}$
    Construct transition matrix $\mathbf{S}$ by Eq.5,6,7
    Get stabilized pseudo-label matrix $\mathbf{P}^\infty$ and $\mathbf{Q}^\infty$ by Eq.14,13
    Combine bi-directional LP result $\mathbf{B}$ by Eq.15
    Select high-fidelity pairs and add them to queue $\mathcal{J}$
    Compute instance features using M: $f'(\cdot), g'(\cdot)$
    Train M by optimizing the sample-wise reweighted loss using Eq.16
    Update M_{mom} by Eq.17
  end for
end for
Output: Refined model M and its momentum version M_{mom}

4 Experiments

4.1 Datasets and Performance Measurements

Datasets. Following the experimental settings and dataset splits in NPC [11], we evaluate our method on three widely-used image-text matching benchmarks, i.e., Flickr30K [50], MSCOCO [51], and Conceptual Captions [7], where Flickr30K and MSCOCO contain the synthetic noise and Conceptual Captions contains the real-world noise. The details in these datasets are delineated as follows:

- **Flickr30K** contains 31,783 images with 5 captions each. We assign 1,000 image-text pairs for validation, 1,000 image-text pairs for testing, and the rest for training.
- **MSCOCO** includes 123,287 images with 5 captions each. We assign 25,000 image-text pairs for validation, 5,000 image-text pairs for testing, and the rest for training. MSCOCO can be either evaluated using the whole 5,000 test set or an average of 5-fold 1,000 test sets [52].
- **Conceptual Captions** is a large-scale dataset automatically harvested from the Internet; therefore, about 3%-20% image-text pairs in the dataset are mismatched or weakly-matched [7]. Following NPC [11], we use a subset named CC120K in our experiments, with splits of 93,656 training, 1,000 validation, and 1,000 test image-text pairs.

Evaluation Protocol. We evaluate the retrieval performance with the recall rate at K (R@K) metric. In a nutshell, R@K measures the proportion of relevant items retrieved within the top K items closest to the query. In our experiments, we take image and text as queries, respectively, and report R@1, R@5, R@10 results and their sum RSUM for a comprehensive evaluation.

4.2 Implementation Details

As a general framework, GLP can be applied to many existing cross-modal matching models. Same as NPC [11], we implement GLP

Table 1: Retrieval Performance Comparison on Flickr30K and MSCOCO datasets on varying noisy levels.

Noise Ratio	Method	Flickr30K						MSCOCO 1K					
		Image → Text			Text → Image			Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
0%	NCR	77.3	94.0	97.5	59.6	84.4	89.9	78.3	95.8	98.5	63.3	90.4	95.8
	DECL	79.8	94.9	97.4	59.5	83.9	89.5	79.1	96.3	98.7	63.3	90.1	95.6
	BiCro	81.7	95.3	98.4	61.6	85.6	90.8	79.1	96.4	98.6	63.8	90.4	96.0
	PC^2	78.7	94.8	97.0	60.0	84.4	89.8	79.1	96.5	98.8	64.0	90.3	95.6
	CLIP	86.2	97.6	99.2	72.9	92.3	96.0	79.9	95.1	98.1	65.0	90.3	98.1
	NPC	87.9	**98.1**	**99.4**	75.0	**93.7**	**97.2**	**82.2**	**96.5**	**98.7**	**68.3**	**92.0**	**98.7**
	GLP	**88.9**	**98.1**	**99.4**	**75.1**	93.4	96.7	81.2	96.0	98.5	67.2	91.4	98.5
20%	NCR	73.5	93.2	96.6	56.9	82.4	88.5	77.7	95.5	98.2	62.5	89.3	95.3
	DECL	77.5	93.8	97.0	56.1	81.8	88.5	77.5	95.9	98.4	61.7	89.3	95.4
	BiCro	78.1	94.4	97.5	60.4	84.4	89.9	78.8	96.1	**98.6**	63.7	90.3	95.7
	L2RM	77.9	95.2	97.8	59.8	83.6	89.5	80.2	96.3	98.5	64.2	90.1	95.4
	PC^2	78.7	94.9	96.9	59.8	83.9	89.6	77.8	95.7	98.4	62.8	89.7	95.3
	CLIP	82.3	95.5	98.3	66.0	88.5	93.5	75.0	93.1	97.2	58.7	86.1	97.2
	NPC	87.3	97.5	98.8	72.9	92.1	95.8	79.9	95.9	98.4	66.3	90.8	98.4
	GLP	**88.1**	**98.2**	**99.4**	**74.1**	**93.3**	**96.3**	**81.0**	**95.9**	98.4	**66.4**	**90.9**	**98.4**
40%	NCR	68.1	89.6	94.8	51.4	78.4	84.8	74.7	94.6	98.0	59.6	88.1	94.7
	DECL	72.7	92.3	95.4	53.4	79.4	86.4	75.6	95.5	98.3	59.5	88.3	94.8
	BiCro	74.6	92.7	96.2	55.5	81.1	87.4	77	**95.9**	98.3	61.8	89.2	94.9
	L2RM	75.8	93.2	96.9	56.3	81.0	87.3	77.5	95.8	98.4	62.0	89.1	94.9
	PC^2	75.8	93.5	96.9	57.5	81.9	88.2	77.4	95.8	98.4	62.1	89.4	95.1
	CLIP	76.2	93.3	96.5	59.4	85.0	90.9	70.7	91.7	96.2	54.7	83.4	96.2
	NPC	85.6	97.5	98.4	71.3	91.3	95.3	79.4	95.1	98.3	65.0	90.1	98.3
	GLP	**87.9**	**98.1**	**99.4**	**73.8**	**92.9**	**96.5**	**80.2**	95.8	**98.5**	**66.6**	**90.8**	**98.5**
60%	NCR	13.9	37.7	50.5	11.0	30.1	41.4	0.1	0.3	0.4	0.1	0.5	1.1
	DECL	65.2	88.4	94.0	46.8	74.0	82.2	73.0	94.2	97.9	57.0	86.6	93.8
	BiCro	67.6	90.8	94.4	51.2	77.6	84.7	73.9	94.4	97.8	58.3	87.2	93.9
	L2RM	70.0	90.8	95.4	51.3	76.4	83.7	75.4	94.7	97.9	59.2	87.4	93.8
	PC^2	70.8	90.3	94.4	53.1	79.0	85.9	74.2	94.4	97.8	58.9	87.5	93.8
	CLIP	66.3	87.3	93.0	52.1	78.8	87.4	67.0	88.8	95.0	49.7	79.6	75.0
	NPC	83.0	95.9	98.6	68.1	89.6	94.2	78.2	94.4	97.7	63.1	89.0	97.7
	GLP	**88.2**	**98.0**	**99.2**	**72.2**	**92.2**	**96.2**	**80.5**	**95.9**	**98.5**	**66.5**	**90.6**	**98.6**
80%	NCR	1.5	6.2	9.9	0.3	1.0	2.1	0.1	0.3	0.4	0.1	0.5	1.0
	DECL	53.4	78.8	86.9	37.6	63.8	73.9	64.8	90.5	96.0	49.7	81.7	90.3
	BiCro	2.3	9.2	17.2	2.6	10.2	16.8	62.2	88.6	94.6	47.4	79.2	88.5
	L2RM	55.7	80.8	87.8	39.4	65.4	74.9	69.0	91.9	96.4	52.6	82.4	90.3
	CLIP	65.9	88.5	93.6	48.7	76.0	85.1	67.5	90.2	95.3	49.7	78.8	95.3
	NPC	79.3	93.1	97.1	59.6	84.6	90.8	73.9	92.8	96.6	59.1	86.4	96.6
	GLP	**84.9**	**97.3**	**99.3**	**69.9**	**90.7**	**95.0**	**79.0**	**95.0**	**98.1**	**64.6**	**89.7**	**98.1**

based on CLIP [28]. The baselines of our method include CLIP with ViT-B/32 and NPC [11]. All the methods are trained on a single RTX 3090 GPU optimized by AdamW optimizer coupled with a cosine annealing rate scheduler. We start training GLP with the learning rate of 5e-7 with a weight decay of 0.2. In all experiments, we train the model for 5 epochs with a mini-batch size of 256, and the hyperparameter β is set to 0.99.

4.3 Comparison with Advanced Methods

Quantitative Comparison. To illustrate the effectiveness, we compare GLP with various approaches, including noise-robust learning methods based on SGRAF [22] such as NCR [8], DECL [38], BiCro [12], L2RM [34] and PC^2 [33], as well as CLIP-based method including CLIP with fine-tuning [28] and NPC [11]. The results are shown in Table1. Our method, GLP, significantly outperforms all methods across all noise levels. Notably, on Flickr30K with an 80% noise ratio, GLP shows a substantial improvement over NPC, with a significant R@1 performance gap. Specifically, GLP achieves an average improvement of 2.7% in R@1 score, higher than NPC in image-to-text (i2t) matching, and an average improvement of 3.2% higher R@1 score in text-to-image (t2i) matching. Moreover, as the noise ratio increases, the performance gap between GLP and clip-based methods becomes even more pronounced. For example, on the MSCOCO 1K set, when the noise ratio increases from 20% to 80%, the R@1 performance gap between GLP and NPC widens from 0.8% to 5.6% on i2t, and from 1.1% to 5.1% on t2i. This phenomenon is powerful to prove the effectiveness of GLP on robust learning.

Stability Comparison. In addition to the effectiveness of the methods, we further analyze the stability advantage of GLP under varying noise levels. Fig 3 shows the average R@1 performance

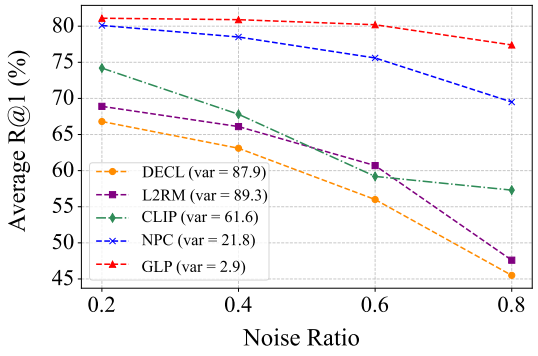

Figure 3: Average R@1 with Variance Comparison.

variations of different methods under different noise ratios. For fairness, we selected two SGRAF-based methods, DECL and L2RM, which perform optimally at high noise levels, as well as the NPC method based on CLIP. It can be observed that GLP outperforms all other methods across all noise ratios. Meanwhile, as the noise ratio increases, the performance decline of GLP is significantly smaller than that of other methods. Furthermore, we calculated the variance of each method at different noise ratios to quantify the stability of the methods. GLP exhibits remarkable stability, with a variance of only 2.9%, significantly outperforming all other methods. Compared to the baseline CLIP, GLP shows a 58.7% reduction in variance, and compared to NPC, a 18.9% reduction. The decrease in variance indicates that GLP significantly improves the stability of performance.

Table 2: Retrieval Performance Comparison on CC120K.

Method	Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	67.9	88.9	93.1	67.6	88.0	92.8
NPC	71.2	90.9	94.7	71.8	**90.8**	94.2
PAU	70.7	**91.0**	94.7	69.6	90.0	94.2
GLP	**72.0**	90.9	**94.9**	**72.2**	90.7	**94.3**

Results on Real-world Noise To substantiate the comprehensive performance assessment, we also report quantitative results on the CC120K dataset, which contains real-world noisy correspondences (NCs) and better reflects industrial scenarios. According to the results shown in Table 2, GLP demonstrates competitive performance in both Image-to-Text and Text-to-Image retrieval tasks. Specifically, GLP achieves an R@1 of 72% for Image-to-Text, which outperforms CLIP (67.9%) and is comparable to NPC (71.2%). In the Text-to-Image task, GLP achieves an R@1 of 72.2%, slightly outperforming NPC (71.8%) and surpassing CLIP (67.6%). Overall, GLP surpasses CLIP in both retrieval tasks and provides similar or better performance compared to NPC, establishing it as a strong method for cross-modal retrieval tasks on the CC120K dataset.

4.4 Comparison with CLIP-based Methods

To make a fairer performance comparison, we also compared GLP with more CLIP-based methods. These methods [3] [53] [54] focus on studying the ambiguous relationships between cross-modal data pairs, thus implicitly addressing the NC problem. We present results on different noise levels using the MSCOCO 1K and MSCOCO 5K sets. In Table 3, under different noise levels, GLP maintains a strong performance, particularly with 20% and 50% noise. For 20% noise, GLP achieves 73.7% in 1K R@1 and 55.2% in 5K R@1, significantly outperforming methods like VSE and PCME++. Even with 50% noise, GLP still outperforms most methods, demonstrating its robustness in handling noisy data. Specifically, GLP achieves 71.3% in 1K R@1, which is much higher than VSE (38.5%) and PCME (65.8%).

Table 3: Retrieval performance under different noise levels.

noise	method	1K R@1	5K R@1	1K RSUM
20%	VSE	72.0	51.4	520.2
	PCME++	70.8	49.5	522.4
	PAU	71.4	51.7	521.5
	CLIP	66.8	47.2	507.2
	NPC	73.1	53.8	529.8
	GLP	**73.7**	**55.2**	**530.9**
50%	VSE	38.5	18.4	390.5
	PCME++	65.7	44.0	503.9
	PAU	69.3	45.3	513.4
	CLIP	60.9	41.4	486.0
	NPC	71.3	51.9	523.4
	GLP	**72.9**	**54.0**	**528.4**

4.5 Further Analysis

4.5.1 Ablation Study

To investigate the contribution of each component in our framework, we conduct ablation studies under two noise settings (40% and 60%) on the Flickr30K dataset. The results are summarized in Table 4. Specifically, we examine the effect of three key components: matching degree re-weight (b_i), online queue (Que), and momentum updating (Mom). We observe that removing any individual component leads to a performance drop across both image-to-text and text-to-image retrieval tasks. Under the 40% noise setting, removing the queue or momentum update leads to a noticeable decrease in performance, suggesting their essential roles in stabilizing training and enhancing robustness. Notably, when all components are removed, the performance drops significantly (e.g., Image-to-Text R@1 drops from 88.6 to 76.2, and Text-to-Image R@1 from 73.5 to 59.4), indicating that each module contributes positively. When the noise level increases to 60%, the performance gap becomes more pronounced. The full model achieves the best performance with image-to-text R@1 of 88.2 and text-to-image R@1 of 72.2. These results validate the necessity of each component and demonstrate the effectiveness and robustness of our proposed design in noisy correspondence scenarios.

Table 4: Ablation studies. The best results are marked in bold.

Noise	Components			Image → Text			Text → Image		
	b_i	Que	Mom	R@1	R@5	R@10	R@1	R@5	R@10
40%	✓	✓	✓	**88.6**	**98.1**	**99.4**	**73.5**	**92.9**	**96.5**
	✓	✓		86.0	97.8	99.3	72.4	91.9	96.0
	✓		✓	85.6	98.0	99.3	72.6	92.5	96.0
	✓			87.4	97.7	99.0	72.6	92.3	95.9
				76.2	93.3	96.5	59.4	85.0	90.9
60%	✓	✓	✓	**88.2**	**98.0**	**99.2**	**72.2**	**92.2**	**96.2**
	✓	✓		87.2	97.3	99.0	70.3	91.3	95.4
	✓		✓	86.4	97.1	99.2	70.5	91.1	95.4
	✓			86.2	96.7	99.1	69.9	91.1	95.0
				66.3	87.3	93.0	52.1	78.8	87.4

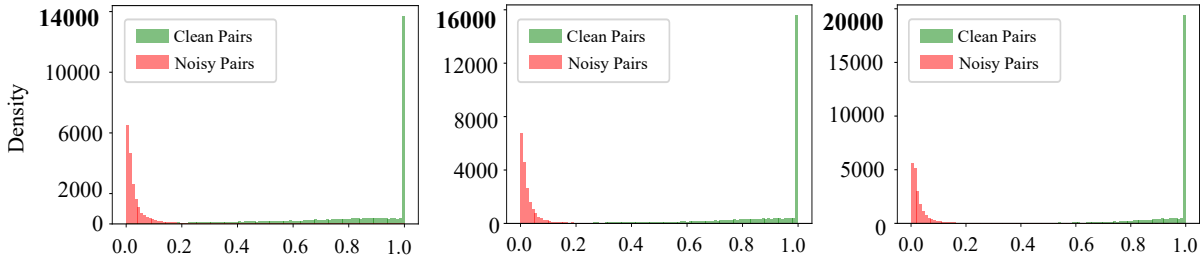

Figure 4: We visualize the distribution of matching degrees for clean and noisy pairs at different training stages of our GLP, which is conducted on Flickr30K under 40% NCs.

1. White dog with brown ears standing near water. (bi=1)
2. Car racing on a dirt road. (bi=0.001)
3. A white dog shakes on the edge of a beach with an orange ball. (bi=1)
4. Attendees for a presentation or lecture sit in blue chairs. (bi=0.022)
5. White dog playing with a red ball on the shore near the water. (bi=0.987)

1. Someone in a blue shirt and hat is standing on stair and leaning against a window. (bi=0.998)
2. Two people use a plumb line on the ground while a woman in a gray sweater looks on. (bi=0.037)
3. A man in a blue shirt is standing on a ladder cleaning a window. (bi=1)
4. The black dog is running on the grass. (bi=0.012)
5. A blond-haired boy sits on a red chair next to a black guitar and looks at a book. (bi=0.055)

Figure 5: Examples from Flickr30K dataset. Each image has 5 annotated captions. The GT captions are marked in green, while the NC captions are highlighted in red. The corresponding average refined weight (b_i) for each pair is displayed alongside.

4.5.2 Hyperparameter Analysis

GLP involves three key hyperparameters: α, knn_i and knn_c, with their respective impacts shown in Fig 6. Parameter α controls the strength of label propagation, balancing the influence of the propagated labels and the original labels during each update. Within the range of $[0, 1]$, GLP consistently delivers stable performance, with the optimal outcome achieved when $\alpha = 0.9$. This setting was maintained across all subsequent experiments. knn_i and knn_c refer to the number of k-nearest neighbors searched within the same modality and across different modalities, respectively. Optimal performance is achieved when using 2 intra-modal and 15 cross-modal nearest neighbors. Moreover, the retrieval performance is not highly sensitive to the specific choices of knn_i and knn_c, indicating the robustness of GLP to hyperparameter settings. Furthermore, we

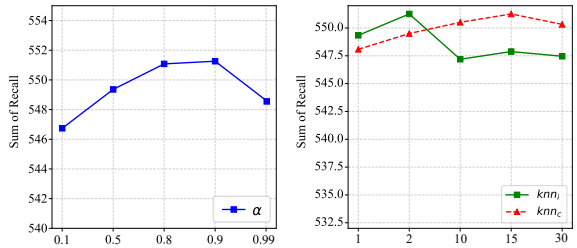

Figure 6: Performance under different hyper-parameters of GLP on Flickr30K under 40% NCs.

investigated the impact of queue capacity on performance when incorporating features from the queue into the graph construction process. With a batch size of 256, we compared the results for different queue lengths, specifically 10, 100, and 500. The experimental results indicate that a larger queue capacity leads to a greater improvement in performance. This is consistent with our hypothesis that utilizing high-quality matching pairs for graph construction contributes to the stability of the graph.

Table 5: Comparison of experimental results with different queue capacities in graph construction.

Q_len	Image → Text			Text → Image			Avg R@1
	R@1	R@5	R@10	R@1	R@5	R@10	
10	86.10	97.80	99.30	72.42	91.91	96.01	79.26
100	**87.90**	98.10	99.40	73.84	92.92	**96.46**	**80.87**
500	87.5	**98.5**	**99.5**	**73.86**	**92.98**	96.36	80.68

4.6 Visualize Analysis

To better demonstrate the scientific validity of GLP, we carry out experiments under 40% noise on Flickr30K to visually investigate the evolution of matching degree in the training process. As shown in Fig 4, as the training progressed, we observe an improvement in the model's ability to distinguish clean samples, which demonstrates the effectiveness of our method. Furthermore, we present representative examples from the Flickr30K in Fig 5, which illustrates the average refined weight (b_i) for each sample pair across five training epochs. The results demonstrate a remarkable discrepancy between the same image with Ground-Truth (GT) annotations and Noisy-Correspondence (NC) annotations, confirming that the GLP can effectively discriminate between clean and noisy pairs.

5 Conclusion

This paper focuses on noisy correspondence in cross-modal retrieval, which introduces mismatched pairs and degrades performance. To address this problem, we present GLP, a novel framework that utilizes both intra-modal neighbors and cross-modal neighbors to estimate an image-text pair's matching degree, thus reweighting the InfoNCE loss. By tailoring label propagation to the cross-modal retrieval task, our GLP is a concise framework without using co-teaching. Meanwhile, we conduct experiments on three widely used datasets to verify the effectiveness of our method in both synthetic and real-world noise correspondences.

Acknowledgments

This work is supported by the National Key RD Program of China (2022YFF0712100), NSFC (62276131), Natural Science Foundation of Jiangsu Province of China under Grant (BK20240081), the Fundamental Research Funds for the Central Universities (No.30922010317, No.30923011007)

References

[1] Yuhao Cheng et al. "Cross-modal graph matching network for image-text retrieval". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2022), pp. 1–23.

[2] Kuang-Huei Lee et al. "Stacked cross attention for image-text matching". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 201–216.

[3] Sanghyuk Chun et al. "Probabilistic embeddings for cross-modal retrieval". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 8415–8424.

[4] Haoyu Lu et al. "Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval". In: *Proceedings of the IEEE/CVF conference on computer Vision and pattern recognition*. 2022, pp. 15692–15701.

[5] Fengqiang Wan et al. "Covlr: Coordinating cross-modal consistency and intra-modal relations for vision-language retrieval". In: *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2024, pp. 1–6.

[6] Yang Yang et al. "Rethinking Label-Wise Cross-Modal Retrieval from A Semantic Sharing Perspective." In: *IJCAI*. 2021, pp. 3300–3306.

[7] Piyush Sharma et al. "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 2556–2565.

[8] Zhenyu Huang et al. "Learning with noisy correspondence for cross-modal matching". In: *Advances in Neural Information Processing Systems* (2021), pp. 29406–29419.

[9] Quanxing Zha et al. "Ugncl: Uncertainty-guided noisy correspondence learning for efficient cross-modal matching". In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2024, pp. 852–861.

[10] Zerun Feng et al. "Learning from noisy correspondence with tri-partition for cross-modal matching". In: *IEEE Transactions on Multimedia* (2023), pp. 3884–3896.

[11] Xu Zhang, Hao Li, and Mang Ye. "Negative Pre-aware for Noisy Cross-Modal Matching". In: *Proceedings of the AAAI Conference on Artificial Intelligence* (2024), pp. 7341–7349.

[12] Shuo Yang et al. "Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 19883–19892.

[13] Peng Hu et al. "Cross-modal retrieval with partially mismatched pairs". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.8 (2023), pp. 9595–9610.

[14] Zhongtian Fu et al. "Noise-aware image captioning with progressively exploring mismatched words". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. 11. 2024, pp. 12091–12099.

[15] Haochen Han et al. "Noisy correspondence learning with meta similarity correction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 7517–7526.

[16] Xinran Ma et al. "Cross-modal retrieval with noisy correspondence via consistency refining and mining". In: *IEEE Transactions on Image Processing* (2024), pp. 2587–2598.

[17] Yang Yang et al. "Rebalanced Vision-Language Retrieval Considering Structure-Aware Distillation". In: *IEEE Transactions on Image Processing* (2024).

[18] Yang Yang et al. "Alignment efficient image-sentence retrieval considering transferable cross-modal representation learning". In: *Frontiers of Computer Science* 18.1 (2024), p. 181335.

[19] Fartash Faghri et al. "Vse++: Improving visual-semantic embeddings with hard negatives". In: *arXiv preprint arXiv:1707.05612* (2017).

[20] Kunpeng Li et al. "Visual semantic reasoning for image-text matching". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 4654–4662.

[21] Shengsheng Qian et al. "Dual adversarial graph neural networks for multi-label cross-modal retrieval". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021, pp. 2440–2448.

[22] Haiwen Diao et al. "Similarity reasoning and filtration for image-text matching". In: *Proceedings of the AAAI conference on artificial intelligence*. 2021, pp. 1218–1226.

[23] Yi He et al. "Cross-graph attention enhanced multi-modal correlation learning for fine-grained image-text retrieval". In: *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 2021, pp. 1865–1869.

[24] Zhengxin Pan, Fangyu Wu, and Bailing Zhang. "Fine-grained image-text matching by cross-modal hard aligning network". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 19275–19284.

[25] Kun Zhang et al. "Negative-aware attention framework for image-text matching". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 15661–15670.

[26] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 4171–4186.

[27] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[28] Alec Radford et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PmLR. 2021, pp. 8748–8763.

[29] Ding Jiang and Mang Ye. "Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 2787–2797.

[30] Yannis Kalantidis, Giorgos Tolias, et al. "Label propagation for zero-shot classification with vision-language models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 23209–23218.

[31] Vladan Stojnić et al. "LPOSS: Label Propagation Over Patches and Pixels for Open-vocabulary Semantic Segmentation". In: *arXiv preprint arXiv:2503.19777* (2025).

[32] Mouxing Yang et al. "Learning with twin noisy labels for visible-infrared person re-identification". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 14308–14317.

[33] Yue Duan et al. "PC2: Pseudo-Classification Based Pseudo-Captioning for Noisy Correspondence Learning in Cross-Modal Retrieval". In: *Proceedings of the 32nd ACM International Conference on Multimedia*. 2024, pp. 9397–9406.

[34] Haochen Han et al. "Learning to rematch mismatched pairs for robust cross-modal retrieval". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 26679–26688.

[35] Yuchen Yang et al. "Robust noisy correspondence learning with equivariant similarity consistency". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 17700–17709.

[36] Zihua Zhao et al. "Mitigating noisy correspondence by geometrical structure consistency learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 27381–27390.

[37] Quanxing Zha et al. "ReCon: Enhancing True Correspondence Discrimination through Relation Consistency for Robust Noisy Correspondence Learning". In: *arXiv preprint arXiv:2502.19962* (2025).

[38] Yang Qin et al. "Deep evidential learning with noisy correspondence for cross-modal retrieval". In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 4948–4956.

[39] Yang Qin et al. "Cross-modal active complementary learning with self-refining correspondence". In: *Advances in Neural Information Processing Systems* (2023), pp. 24829–24840.

[40] Ahmet Iscen et al. "Label propagation for deep semi-supervised learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 5070–5079.

[41] Xun Xu and Gim Hee Lee. "Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 13706–13715.

[42] Hao Zhu and Piotr Koniusz. "Transductive few-shot learning with prototype-based label propagation by iterative graph refinement". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 23996–24006.

[43] Xuefeng Hu et al. "Reclip: Refine contrastive language image pre-training with source free domain adaptation". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 2994–3003.

[44] Yushu Li et al. "Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model". In: *arXiv preprint arXiv:2412.18303* (2024).

[45] Haoxuan You et al. "Cobit: A contrastive bi-directional image-text generation model". In: *arXiv preprint arXiv:2303.13455* (2023).

[46] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).

[47] Ahmet Iscen et al. "Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2077–2086.

[48] Victor Weixin Liang et al. "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 17612–17625.

[49] Kaiming He et al. "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738.
[50] Peter Young et al. "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions". In: *Transactions of the association for computational linguistics* 2 (2014), pp. 67–78.
[51] Tsung-Yi Lin et al. "Microsoft coco: Common objects in context". In: *European conference on computer vision*. 2014, pp. 740–755.
[52] Andrej Karpathy and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3128–3137.
[53] Sanghyuk Chun. "Improved Probabilistic Image-Text Representations". In: *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. 2024.
[54] Hao Li et al. "Prototype-based aleatoric uncertainty quantification for cross-modal retrieval". In: *Advances in Neural Information Processing Systems* (2023), pp. 24564–24585.