Rethinking Cross-Modal Anchor Alignment for Mitigating Error Accumulation

Supplementary Material

1. Hyperparameter δ Analysis

In this section, we present an analysis of the hyperparameter δ to validate our setting. The experiments are conducted on Flickr30K with a noise ratio of 0.4. As shown in Table 1, the retrieval metric RSum achieves the best result of 498.6 when $\delta = 0.5$. Both increasing and decreasing this value lead to degraded retrieval performance.

Table 1. Recall rates with different Threshold δ

δ	Image→Text			Text→Image			RSum
	R@1	R@5	R@10	R@1	R@5	R@10	
0.4	77.0	94.0	97.1	57.9	82.4	88.6	497.0
0.5	77.6	94.7	97.0	58.5	82.2	88.6	498.6
0.6	76.5	93.6	96.9	58.0	82.1	88.3	498.3

2. Computational Cost Analysis

In this section, we report the wall-clock time of GSL and several SOTA methods to evaluate the training efficiency of our approach. As shown in Table 2, the wall-clock time of NCR is approximately 3,670.46s and is used as the baseline. Compared with NCR, the training time of GSL, ESC, and SPS increases by 46%, 259.33%, and 6.3%, achieving RSum improvements of 10.8, 3.3, and 6.2, respectively. In contrast, BiCro and GSC reduce the training time by 17.08% and 4.31%, with RSum improvements of 8.0 and 5.8, respectively. Overall, the runtime of GSL remains within an acceptable range.

Table 2. Computational cost on Flickr30K with 0.2 noise

Method	Wall-Clock Time (S)	Relative Increase
NCR	3670.46	0.00%
BiCro	3043.67	-17.08%
GSC	3512.31	-4.31%
ESC	13191.06	259.33%
SPS	3901.82	6.30%
GSL	5378.99	46.55%

3. Noise Identification Accuracy Analysis

To evaluate the effectiveness of the Semantic-Constrained Triplet Loss Module in distinguishing clean and noisy samples, we conduct experiments on the COCO dataset with a noise ratio of 0.4. Specifically, we compare GSL with BiCro and SPS in terms of sample identification accuracy. The results are shown in Table 3. The results demonstrate that GSL achieves superior performance in noise identification, verifying the effectiveness of the Semantic-Constrained Triplet Loss Module in distinguishing clean and noisy samples.

Table 3. Noise identification results.

Method	Precision (%)	Recall(%)
BiCro	71.2	90.4
SPS	97.6	98.1
GSL	98.5	98.1

4. Training Pipeline

Our method is trained in a co-teaching manner[**?**]. The detailed training pipeline is shown in the Fig. 1 and illustrated in the Algorithm 1.

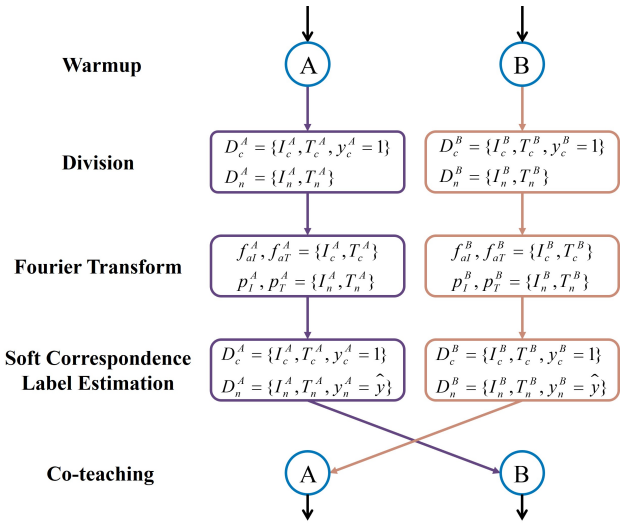

Figure 1. Training Pipeline.

5. Soft Correspondence Label Distribution

We visualize the distribution of corrected soft correspondence labels on the COCO dataset with a noise ratio of 0.4 to illustrate the advantage of the proposed label correction module, as shown in Fig. 2.

6. Retrieval Results

To further substantiate the effectiveness of GSL, we visualize retrieval results under both query modalities (image →

Algorithm 1 The training pipeline of our GSL method.

Input: Given training data $\mathcal{D}$, matching models $\theta^A = \{f^A, g^A, S^A\}$ and $\theta^B = \{f^B, g^B, S^B\}$

Warmup the models (θ^A, θ^B) using L_w in Eq.(17)
for $i = 1$ to num_epochs **do**
  Compute the division triplet loss L_w using Eq.(17)
  Select anchor points in D and calculate the division regularization L_{intra} and L_{inter} using Eq.(13) and Eq.(15)
  Combine L_w L_{intra} and L_{inter} to acquire total division loss L_{SCT} using Eq.(16) and normalize the L_{SCT} as L_i
  $\mathcal{P}^A = \{p_i^A | p_i^A = p(k=0|l_i)\}_{i=1}^N \leftarrow BMM(\mathcal{D}, A)$
  $\mathcal{P}^B = \{p_i^B | p_i^B = p(k=0|l_i)\}_{i=1}^N \leftarrow BMM(\mathcal{D}, B)$
  for $k \in A, B$ **do**
    $\tilde{\mathcal{D}}_c^k = \{(I_i, T_i, y_i = 1) | p_i^k > \delta, \forall (I_i, T_i) \in \tilde{\mathcal{D}}\}$
    $\tilde{\mathcal{D}}_n = \{(I_i, T_i) | p_i^k > \delta, \forall (I_i, T_i) \in \tilde{\mathcal{D}}\}$
    for $k = num\_steps$ **do**
      Sample a mini-batch $\{\mathcal{B}_j^c = (I_c, T_c, y_c = 1), \mathcal{B}_j^n = (I_n, T_n)\}$ from $(\tilde{\mathcal{D}}_c^k, \tilde{\mathcal{D}}_n)$
      Select two anchor points (f_I^1, f_T^1) and (f_I^2, f_T^2) in the mini-batch
      Map the anchor points (f_I^1, f_T^1) and (f_I^2, f_T^2) to the frequency domain using Eq.(7)
      Refine the labels $y_i = \hat{y}_i$ of $\{\mathcal{B}_j^c, \mathcal{B}_j^n\}$ using Eq.(8)-Eq.(11)
      Train the network (θ^A, θ^B) on $\{\mathcal{B}_j^c, \mathcal{B}_j^n\}$ by optimizing L_{SCT} in Eq.(16)

Output: Matching models (θ^A, θ^B)

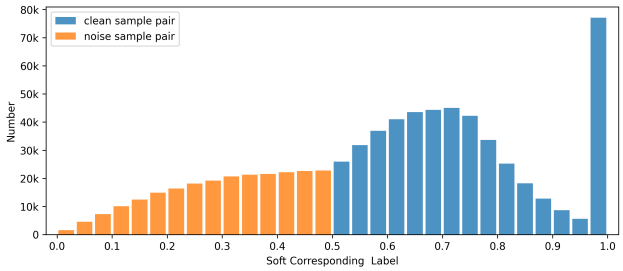

Figure 2. Soft correspondence label distribution.

text (Fig. 3) and text → image (Fig. 4)). For each query, we display the Top-3 retrieved items together with their similarity scores. In both directions, the model retrieves semantically correct content. These qualitative observations align with our quantitative results and further corroborate that GSL has stronger cross-modal alignment.

Top1: a person in a green jacket throws a stick into the water for a black and white dog to fetch whilst a brown and white dog looks on. **(sim=0.8530)**

Top2: a man is throwing a log into a waterway while two dogs watch. **(sim=0.7695)**

Top3: a man is throwing a big stick into the water so his two dogs and go fetch it. **(sim=0.7465)**

Top1: there are six men mountain biking in a forest terrain. **(sim=0.8059)**

Top2: six people riding bikes on a trail in the forest. **(sim=0.7497)**

Top3: six people ride mountain bikes through a jungle environment. **(sim=0.7218)**

Figure 3. Image-to-text matching results on Flickr30K.

(a) a black dog is retrieving a ball in water.

Sim=0.7884 Sim=0.4573 Sim=0.2339

(b) a young girl floats on her back in water and peers over her life jacket.

Sim=0.5597 Sim=0.3765 Sim=0.3585

Figure 4. Text-to-image matching results on Flickr30K.