Supplementary Material

Algorithm 1 The training pipeline of our L2RM.

```
Input: The training dataset $\mathbb{D}$ with PMPs, cross-modal retrieval model $(f_v, f_t, g)$, self-supervised learning cost function $f_c$,
       partial transport parameter $\rho$, Sinkhorn regularization parameter $\lambda$.
Warm up the model $(f_v, f_t, g)$ using $\mathcal{L}^{\text{InfoNCE}} + \mathcal{L}^{\text{RCE}}$
for $e = 1 : \mathit{num\_epochs}$ do
    // identifying mismatched pairs
    $\mathcal{W} = \{w_i\}_{i=1}^{N} \leftarrow \mathit{BetaMixtureModel}\,(\mathbb{D}, (f_v, f_t, g))$
    $\mathbb{D}_m = \{(V_i, T_i) \mid w_i \leq 0.5, \forall (V_i, T_i) \in \mathbb{D}\}$, $\mathbb{D}_{\widetilde{m}} = \{(V_i, T_i) \mid w_i > 0.5, \forall (V_i, T_i) \in \mathbb{D}\}$
    for $n = 1 : \mathit{num\_steps}$ do
        // update the learnable cost function
        Reconstruct the visual-text pairs $\mathbb{D}'$
        Sample a batched samples and get the corresponding matching matrix $(\boldsymbol{V}, \boldsymbol{T}, \boldsymbol{\pi}^{\text{sup}})$
        Train the cost function $f_c$ on $(\boldsymbol{V}, \boldsymbol{T}, \boldsymbol{\pi}^{\text{sup}})$ by minimizing $\mathcal{L}_{OT}$
        // rematching mismatched pairs
        Sample a batched samples $\mathbb{B}_{\widetilde{m}} = \{(V_i, T_i)\}_{i=1}^{N_b}$ from the mismatched subset $\mathbb{D}_{\widetilde{m}}$
        Compute the refined alignment $\tilde{\boldsymbol{\pi}}$ in the batch by optimizing the partial OT problem
        // update the cross-modal retrieval model
        Sample a batched samples $\mathbb{B}_m = \{(V_i, T_i)\}_{i=1}^{N_b}$ from the matched subset $\mathbb{D}_m$
        Train the retrieval model $(f_v, f_t, g)$ on $(\mathbb{B}_m, \mathbb{B}_{\widetilde{m}})$ by minimizing $\mathcal{L}^{\text{Final}}$
Output: Retrieval model $(f_v, f_t, g)$.
```

A. Limitations

Our work still has certain limitations, including (1) This work only explores the PMP problem among visual and textual modalities. Further research is needed to confirm the applicability of L2RM in other cross-modal domains against PMPs, *e.g.*, re-identification [9], video temporal learning [8], and graph matching [7]. (2) The effectiveness of our rematched method is limited by the batch size. When using smaller batch sizes, the likelihood of observing semantic relevant pairs will decrease. One possible improvement is to maintain a queue to compare more data. We also provide experimental analysis (see D.2 for details) to show the impact of batch size.

B. Fast Solver for Refined Alignment

In this section, we detail the fast approximation for computing the refined alignment. We will first introduce how to transform the original partial OT problem into a standard OT problem. Then, we will describe the solution by adopting the efficient Sinkhorn-Knopp algorithm.

Transform partial OT to OT-like problem. Recall that our partial OT problem seeks only ρ-unit mass of $\boldsymbol{p} = \sum_{i=1}^{m} p_i \delta(x_i)$ and $\boldsymbol{q} = \sum_{j=1}^{n} q_j \delta(y_j)$ is matched. To solve the exact partial OT problem, Chapel *et al.* [1] propose an ingenious method that transforms the original partial OT problem into an OT-like problem. Specifically, consider two virtual samples x_{m+1} and y_{n+1} are added to the original variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Intuitively, to ensure ρ-unit mass is transported between $\{x_i\}_{i=1}^{m}$ and $\{y_j\}_{j=1}^{n}$, we should constrain the transport mass from $\{x_i\}_{i=1}^{m}$ to y_{n+1} to $\|\boldsymbol{p}\|_1 - \rho$ and the transport mass from $\{y_j\}_{j=1}^{n}$ to x_{m+1} to $\|\boldsymbol{q}\|_1 - \rho$. Thus, the original partial OT problem from $\boldsymbol{X} = \{x_i\}_{i=1}^{m}$ to $\boldsymbol{Y} = \{y_j\}_{j=1}^{n}$ can be transformed into a standard OT problem from $\hat{\boldsymbol{X}} = \{x_i\}_{i=1}^{m+1}$ to $\hat{\boldsymbol{Y}} = \{y_j\}_{j=1}^{n+1}$, where the corresponding probability measures are extended to $\hat{\boldsymbol{p}} = [\boldsymbol{p}^\top, \|\boldsymbol{q}\|_1 - \rho]^\top$ and $\hat{\boldsymbol{q}} = [\boldsymbol{q}^\top, \|\boldsymbol{p}\|_1 - \rho]^\top$, respectively. Following [1], the original cost matrix $\boldsymbol{C}$ is extended to $\hat{\boldsymbol{C}} \in \mathbb{R}^{m+1 \times n+1}$:

$$\hat{\boldsymbol{C}} = \begin{bmatrix} \boldsymbol{C} & \xi \mathbb{1}_n \\ \xi \mathbb{1}_m^\top & 2\xi + A \end{bmatrix}, \tag{1}$$

where $A > \max(\boldsymbol{C}_{ij})$ and $\xi > 0$. Note that our original partial OT problem restricts the transport among the false positive pairs by imposing a mask matrix, which is extended by:

$$\hat{\boldsymbol{M}} = \begin{bmatrix} \boldsymbol{M} & \mathbb{1}_n \\ \mathbb{1}_m^\top & 1 \end{bmatrix}. \tag{2}$$

Based on these, computing the optimal transport plan in partial OT boils down to solve the following problem:

$$\begin{gathered} \min_{\hat{\boldsymbol{\pi}} \in \Pi(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}; \hat{\boldsymbol{M}})} \langle \hat{\boldsymbol{M}} \odot \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{C}} \rangle_F \\ \text{s.t. } \Pi(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}; \hat{\boldsymbol{M}}) = \{\hat{\boldsymbol{\pi}} \in \mathbb{R}_+^{m+1 \times n+1} | (\hat{\boldsymbol{M}} \odot \hat{\boldsymbol{\pi}}) \mathbb{1}_n = \hat{\boldsymbol{p}}, \\ (\hat{\boldsymbol{M}} \odot \hat{\boldsymbol{\pi}})^\top \mathbb{1}_m = \hat{\boldsymbol{q}}\}. \end{gathered} \tag{3}$$

Algorithm 2 Solving Eq.(3) with Sinkhorn algorithm.

```
Input: Distribution $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{q}}$, cost matrix $\hat{\boldsymbol{C}}$, mask matrix
       $\hat{\boldsymbol{M}}$, partial transport mass $\rho$, Sinkhorn regularization parameter $\lambda$, max iterations $it_{max}$.
Initialize $\hat{\boldsymbol{K}} = \hat{\boldsymbol{M}} \odot e^{\frac{-\hat{\boldsymbol{C}}}{\lambda}}$, $\boldsymbol{b} \leftarrow \mathbb{1}_{n+1}$, $it \leftarrow 0$
// Run Sinkhorn iterations
while $it \leq it_{max}$ and $\boldsymbol{a}$, $\boldsymbol{b}$ not convergence do
    $\boldsymbol{a} \leftarrow \frac{\hat{\boldsymbol{p}}}{\hat{\boldsymbol{K}}\boldsymbol{b}}$ // element-wise division
    $\boldsymbol{b} \leftarrow \frac{\hat{\boldsymbol{q}}}{\hat{\boldsymbol{K}}^{\top}\boldsymbol{a}}$
// Get the approximate solution
$\hat{\boldsymbol{\pi}} = \mathrm{diag}(\boldsymbol{a})\hat{\boldsymbol{K}}\mathrm{diag}(\boldsymbol{b})$
Output: Refined alignment $\tilde{\boldsymbol{\pi}} = (\hat{\boldsymbol{M}} \odot \hat{\boldsymbol{\pi}})[1:m, 1:n]$.
```

Eq.(3) is a standard OT problem and our objective $\tilde{\boldsymbol{\pi}} = (\hat{\boldsymbol{M}} \odot \hat{\boldsymbol{\pi}})[1:m, 1:n]$.

Solving OT with Sinkhorn algorithm. Exactly solving the OT problem with linear programming algorithms requires high computational overhead. To resolve Eq.(3) efficiently, we resort to the entropy-regularized OT problem by adding a entropic constraint $-\lambda H(\hat{\boldsymbol{M}} \odot \hat{\boldsymbol{\pi}})$, which enables the transport plan to be computed by the lightspeed Sinkhorn-Knopp algorithm [2]. Note that Gu *et al.* [4] show that the Sinkhorn's algorithm can be applied to solve the transport plan with mask operation. The detailed solution is presented in Algorithm. 2. We can see that the Sinkhorn's iteration only contains matrix multiplication and exponential operations, which can be computed efficiently.

C. Training Pipeline

In this section, we summarize our detailed training pipeline in Algorithm. 1. The code of L2RM is available at https://github.com/hhc1997/L2RM.

D. Additional Experiments

D.1. Implementation Details

Input preprocessing. Our experiments used the same input preprocessing as in the evaluation of NCR [6]. Specifically, all raw images are processed into the top 36 region proposals by the Faster-RCNN, where each is encoded as a 2048-dimensional feature.

Backbone architecture. L2RM is a general framework which could endow almost all existing cross-modal retrieval methods robust against PMPs. Same as previous robust methods [5, 6, 10, 11], we implement L2RM based on SGR, SAF, and SGRAF [3]. Specifically, the image regions and captions are projected into a common representation space by a full-connected network (*i.e.*,f_v) and a Bi-GRU model ((*i.e.*,f_t)), respectively. To calculate the cross-modal similarities, the similarity function g is based on the Similarity Graph Reasoning (SGR), Similarity Attention Filtration (SAF), or the combination of SGR and SAF.

Epochs	Flickr30K	MS-COCO	CC152K
warm up	5	10	10
training	35	20	40
total	40	30	50
update learning rate	15	10	20

Table 1. The epoch settings for training on three datasets.

Hyperparameters. We follow the same training setting as NCR where applicable. Specifically, the word embedding size is 300 and the common space size is 1024. The retrieval model is trained by a Adam optimizer (default settings) with a learning rate of 2×10^{-4} and a batch size of 128. The epoch setting for training is shown in Tab. 1. The learning rate will be decayed by 0.1 when the training achieves the update epoch. The margin α used in triplet loss is fixed as 0.2 for all experiments.

For hyperparameters specific to L2RM, we set the temperature parameter τ as 0.05. We train our learnable cost function using the Adam optimizer with the default settings and a learning rate of 2×10^{-6}. To solve the OT problem, we fix the partial transport mass $\rho = 0.1$ for all experiments. Note that for the experiments conducted on original datasets (0 MRate), we empirically find that disabling the positives masked strategy could achieve superior performance. In addition, we set the Sinkhorn regularization parameter λ as 0.01, 0.07, and 0.07 for Flickr30K, MS-COCO, and CC152K, respectively.

D.2. More Comparisons Results

Results under Synthesized PMPs. Tab. 2 shows the full comparison results on Flickr30K and MS-COCO under different mismatching rates. From the results, one could see that the existence of PMPs remarkably impair the performance of general cross-modal retrieval methods (*i.e.*, IMRAM, SAF, and SGR). With the mismatching rates increasing, their retrieval performance will degrade fast. Compared with the robust methods, we can find that our L2RM consistently outperforms them under different variants.

Results on well-annotated Datasets. The Flickr30K and MS-COCO are two well-annotated datasets (almost 0 MRate), thus we conduct comparison experiments on the original Flickr30K and MS-COCO to show L2RM's performance under well-matched pairs. The experimental results are reported in Tab. 3. From the results, one could observe that L2RM can boost the retrieval performance of

MRate	Method	Flickr30K							MS-COCO						
		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	
0.2	IMRAM	59.1	85.4	91.9	44.5	71.4	79.4	431.7	69.9	93.6	97.4	55.9	84.4	89.6	490.8
	SAF	62.8	88.7	93.9	49.7	73.6	78.0	446.7	71.5	94.0	97.5	57.8	86.4	91.9	499.1
	SGR	55.9	81.5	88.9	40.2	66.8	75.3	408.6	25.7	58.8	75.1	23.5	58.9	75.1	317.1
	NCR	73.5	93.2	96.6	56.9	82.4	88.5	491.1	76.6	95.6	98.2	60.8	88.8	95.0	515.0
	BiCro	74.7	94.3	96.8	56.6	81.4	88.2	492.0	76.6	95.4	98.2	61.3	88.8	94.8	515.1
	DECL-SAF	73.4	92.0	96.4	53.6	79.7	86.4	481.5	74.4	95.3	98.2	59.8	88.3	94.8	510.8
	DECL-SGR	74.5	92.9	97.1	53.6	79.5	86.8	484.4	75.6	95.1	98.3	59.9	88.3	94.7	511.9
	DECL-SGRAF	77.5	93.8	97.0	56.1	81.8	88.5	494.7	77.5	95.9	98.4	61.7	89.3	95.4	518.2
	RCL-SAF	72.0	91.7	95.8	53.6	79.9	86.7	479.7	77.1	95.5	98.2	61.0	88.8	94.6	515.2
	RCL-SGR	74.2	91.8	96.9	55.6	81.2	87.5	487.2	77.0	95.5	98.1	61.3	88.8	94.8	515.5
	RCL-SGRAF	75.9	94.5	97.3	57.9	82.6	88.6	496.8	78.9	96.0	98.4	62.8	89.9	95.4	521.4
	L2RM-SAF	73.7	94.3	97.7	56.8	81.8	88.1	492.4	77.9	96.0	98.3	62.1	89.2	94.9	518.4
	L2RM-SGR	76.5	93.7	97.3	55.5	81.5	88.0	492.5	78.4	95.7	98.3	62.1	89.1	94.9	518.5
	L2RM-SGRAF	**77.9**	**95.2**	**97.8**	**59.8**	**83.6**	**89.5**	**503.8**	**80.2**	**96.3**	**98.5**	**64.2**	**90.1**	**95.4**	**524.7**
0.4	IMRAM	44.9	73.2	82.6	31.6	56.3	65.6	354.2	51.8	82.4	90.9	38.4	70.3	78.9	412.7
	SAF	7.4	19.6	26.7	4.4	12.0	17.0	87.1	13.5	43.8	48.2	16.0	39.0	50.8	211.3
	SGR	4.1	16.6	24.1	4.1	13.2	19.7	81.8	1.3	3.7	6.3	0.5	2.5	4.1	18.4
	NCR	68.1	89.6	94.8	51.4	78.4	84.8	467.1	74.7	94.6	98.0	59.6	88.1	94.7	509.7
	BiCro	70.7	92.0	95.5	51.9	77.7	85.4	473.2	75.2	95.3	98.1	60.0	87.8	94.3	510.7
	DECL-SAF	70.1	90.6	94.4	49.7	76.6	84.1	465.5	73.3	94.6	98.1	57.9	87.2	94.1	505.2
	DECL-SGR	69.0	90.2	94.8	50.7	76.3	84.1	465.1	73.6	94.6	97.9	57.8	86.9	93.9	504.7
	DECL-SGRAF	72.7	92.3	95.4	53.4	79.4	86.4	479.6	75.6	95.5	98.3	59.5	88.3	94.8	512.0
	RCL-SAF	68.8	89.8	95.0	51.0	76.7	84.8	466.1	74.8	94.8	97.8	59.0	87.1	93.9	507.4
	RCL-SGR	71.3	91.1	95.3	51.4	78.0	85.2	472.3	73.9	94.9	97.9	59.0	87.4	93.9	507.0
	RCL-SGRAF	72.7	92.7	96.1	54.8	80.0	87.1	483.4	77.0	95.5	98.3	61.2	88.5	94.8	515.3
	L2RM-SAF	72.1	92.1	96.1	52.7	78.8	85.9	477.7	74.4	94.7	98.3	59.2	87.9	94.4	508.9
	L2RM-SGR	73.1	92.4	96.3	52.3	79.4	86.3	479.8	75.2	94.8	98.1	59.4	87.8	94.1	509.4
	L2RM-SGRAF	**75.8**	**93.2**	**96.9**	**56.3**	**81.0**	**87.3**	**490.5**	**77.5**	**95.8**	**98.4**	**62.0**	**89.1**	**94.9**	**517.7**
0.6	IMRAM	16.4	38.2	50.9	7.5	19.2	25.3	157.5	18.2	51.6	68.0	17.9	43.6	54.6	253.9
	SAF	0.1	1.5	2.8	0.4	1.2	2.3	8.3	0.1	0.5	0.7	0.8	3.5	6.3	11.9
	SGR	1.5	6.6	9.6	0.3	2.3	4.2	24.5	0.1	0.6	1.0	0.1	0.5	1.1	3.4
	NCR	13.9	37.7	50.5	11.0	30.1	41.4	184.6	0.1	0.3	0.4	0.1	0.5	1.0	2.4
	BiCro	64.1	87.1	92.7	47.2	74.0	82.3	447.4	73.2	93.9	97.6	57.5	86.3	93.4	501.9
	DECL-SAF	56.6	82.5	89.7	40.4	66.6	76.6	412.4	68.6	92.9	97.4	54.1	84.9	92.7	490.6
	DECL-SGR	64.5	85.8	92.6	44.0	71.6	80.6	439.1	69.7	93.4	97.5	54.5	85.2	92.6	492.9
	DECL-SGRAF	65.2	88.4	94.0	46.8	74.0	82.2	450.6	73.0	94.2	97.9	57.0	86.6	93.8	502.5
	RCL-SAF	63.9	84.8	91.7	43.0	71.2	79.4	434.0	70.1	93.1	96.8	54.5	84.4	91.9	490.8
	RCL-SGR	62.3	86.3	92.9	45.1	71.3	80.2	438.1	71.4	93.2	97.1	55.4	84.7	92.3	494.1
	RCL-SGRAF	67.7	89.1	93.6	48.0	74.9	83.3	456.6	74.0	94.3	97.5	57.6	86.4	93.5	503.3
	L2RM-SAF	66.1	88.8	93.8	47.8	74.2	82.2	452.9	71.2	93.4	97.5	56.5	85.9	93.0	497.5
	L2RM-SGR	65.1	87.8	93.6	47.0	73.5	81.5	448.5	72.7	93.9	97.5	56.9	86.2	93.3	500.5
	L2RM-SGRAF	**70.0**	**90.8**	**95.4**	**51.3**	**76.4**	**83.7**	**467.6**	**75.4**	**94.7**	**97.9**	**59.2**	**87.4**	**93.8**	**508.4**
0.8	IMRAM	3.1	9.7	5.2	0.3	0.9	1.9	21.1	1.3	5.0	8.3	0.2	0.6	1.3	16.7
	SAF	0.0	0.8	1.2	0.1	0.5	1.1	3.7	0.2	0.8	1.4	0.1	0.5	1.0	4.0
	SGR	0.2	0.3	0.5	0.1	0.6	1.0	2.7	0.2	0.6	1.0	0.1	0.5	1.0	3.4
	NCR	1.5	6.2	9.9	0.3	1.0	2.1	21.0	0.1	0.3	0.4	0.1	0.5	1.0	2.4
	BiCro	2.3	9.2	17.2	2.6	10.2	16.8	58.3	62.2	88.6	94.6	47.4	79.2	88.5	460.5
	DECL-SAF	46.9	73.7	83.0	32.1	59.0	69.4	364.1	59.3	87.9	94.8	46.3	79.1	88.9	456.3
	DECL-SGR	44.4	72.6	82.0	33.9	59.5	69.0	361.4	60.0	88.7	94.5	45.9	78.8	88.3	456.2
	DECL-SGRAF	53.4	78.8	86.9	37.6	63.8	73.9	394.4	64.8	90.5	96.0	49.7	81.7	90.3	473.0
	RCL-SAF	45.0	72.8	80.8	30.7	56.5	67.3	353.1	62.9	89.3	94.9	47.1	77.9	87.4	459.5
	RCL-SGR	47.1	70.5	79.4	30.3	56.1	66.3	349.7	63.2	89.3	95.2	47.6	78.7	88.0	462.0
	RCL-SGRAF	51.7	75.8	84.4	34.5	61.2	70.7	378.3	67.4	90.8	96.0	50.6	81.0	90.1	475.9
	L2RM-SAF	50.8	77.9	85.5	35.6	62.6	72.7	385.1	64.7	90.8	95.8	50.0	80.9	89.4	471.6
	L2RM-SGR	50.5	77.2	83.9	34.2	61.1	71.6	378.5	65.2	90.3	96.1	49.8	81.0	88.2	470.6
	L2RM-SGRAF	**55.7**	**80.8**	**87.8**	**39.4**	**65.4**	**74.9**	**404.0**	**69.0**	**91.9**	**96.4**	**52.6**	**82.4**	**90.3**	**482.6**

Table 2. Image-text retrieval performance under different mismatching rates (MRate) on Flickr30K and MS-COCO.

existing methods, *i.e*., SAF, SGR, and SGRAF, even though it is proposed to improve robustness. On the one hand, the dataset cannot be absolutely well-matched; it still contains a few mismatched pairs. On the other hand, our rematching

Method	Flickr30K							MS-COCO						
	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	
IMRAM	68.8	91.6	96.0	53.0	79.0	87.1	475.5	74.0	95.6	98.4	60.6	88.9	94.6	512.1
SAF	73.7	93.3	96.3	56.1	81.5	88.0	488.9	76.1	95.4	98.3	61.8	89.4	95.3	516.3
SGR	75.2	93.3	96.6	56.2	81.0	86.5	488.8	78.0	95.8	98.2	61.4	89.3	95.4	518.1
SGRAF	77.8	94.1	97.4	58.5	83.0	88.8	499.6	79.6	96.2	98.5	63.2	90.7	96.1	524.3
NCR	77.3	94.0	97.5	59.6	84.4	89.9	502.7	78.7	95.8	98.5	63.3	90.4	95.8	522.5
BiCro	79.5	94.2	97.4	59.4	83.6	89.8	503.9	78.4	95.6	98.5	62.6	89.7	95.7	520.5
DECL-SGRAF	78.9	94.7	97.4	59.3	84.1	89.8	504.2	79.3	96.5	98.7	63.3	90.6	95.0	523.4
RCL-SAF	76.7	93.7	97.3	56.2	82.6	88.8	495.3	78.5	96.1	98.6	62.7	90.0	95.4	521.3
RCL-SGR	77.5	94.7	97.4	58.8	83.3	88.9	500.6	78.2	96.2	98.4	62.9	90.0	95.7	521.4
RCL-SGRAF	**79.9**	**96.1**	97.8	**61.1**	**85.4**	**90.3**	**510.6**	80.4	96.4	98.7	64.3	90.8	96.0	526.6
L2RM-SAF	77.1	93.2	96.7	57.5	82.4	87.8	494.7	78.2	95.7	98.6	63.4	89.6	95.1	520.6
L2RM-SGR	79.1	94.1	97.7	58.1	83.6	88.9	501.5	79.0	96.4	98.3	63.7	90.2	95.8	523.4
L2RM-SGRAF	79.6	95.9	**98.4**	60.7	84.8	89.0	508.4	**80.5**	**96.6**	**98.9**	**65.7**	**90.8**	**96.1**	**528.6**

Table 3. Image-text retrieval performance on original Flickr30K and MS-COCO datasets.

strategy augments more positive pairs to a certain extent by comparing unpaired samples, which could enhance the generalization of the model.

Results on MS-COCO 5K Datasets. Tab. 4 shows the quantitative results on MS-COCO with full 5K test images. From the results, we could observe that co-trained models offer bigger gains when the test data becomes complex.

MRate	Method	Image-to-Text			Text-to-Image			rSum
		R@1	R@5	R@10	R@1	R@5	R@10	
0.2	L2RM-SAF	56.6	83.3	90.9	40.1	69.5	80.0	420.4
	L2RM-SGR	56.6	83.4	90.6	40.6	69.5	80.0	420.7
	L2RM-SGRAF	59.6	85.1	92.0	42.5	71.5	81.3	432.0
0.4	L2RM-SAF	53.1	81.6	89.8	38.4	67.5	78.2	408.6
	L2RM-SGR	53.5	81.0	89.5	38.0	66.9	77.7	406.6
	L2RM-SGRAF	57.1	83.4	91.0	40.8	69.4	79.7	421.4
0.6	L2RM-SAF	51.0	78.4	86.8	34.9	63.1	74.7	388.9
	L2RM-SGR	50.2	79.0	87.8	34.5	63.0	74.6	389.1
	L2RM-SGRAF	53.5	81.0	88.9	37.3	65.7	76.7	403.1
0.8	L2RM-SAF	40.7	71.2	80.9	28.2	55.8	68.0	344.8
	L2RM-SGR	42.6	71.5	81.7	28.8	55.7	67.3	347.6
	L2RM-SGRAF	45.7	74.4	83.9	30.9	58.5	69.8	363.2

Table 4. Performance under different MRates on MS-COCO 5K.

Impact of Batch Size. To study the influence of different batch sizes for our method, we conducted the ablation study on Flickr30K with 0.6 MRate. Note that our method can flexibly adapt to different batch sizes by adjusting the transport mass ρ, and we set ρ to 0.05, 0.1, and 0.2 for the batch size 64, 128, and 256, respectively. From Tab. 5, one could observe that our method still achieves superior results with a small batch size, *i.e.*, 64, and even surpasses the second-best baseline RCL-SGRAF (in terms of the rSum metric) using a 128 batch size. We could also see that our L2RM can gain from a larger batch size, *i.e.*, 256, while some methods may suffer a performance drop.

Batch	Method	Image-to-Text			Text-to-Image			rSum
		R@1	R@5	R@10	R@1	R@5	R@10	
64	L2RM-SAF	63.5	86.4	93.2	45.8	73.0	81.4	443.3
	L2RM-SGR	62.9	87.4	92.7	46.1	72.8	81.3	443.2
	RCL-SGRAF	66.9	88.3	94.1	48.3	**75.3**	82.5	455.4
	L2RM-SGRAF	**67.2**	**89.4**	**94.2**	**49.2**	**75.3**	**83.4**	**458.7**
128	L2RM-SAF	66.1	88.8	93.8	47.8	74.2	82.2	452.9
	L2RM-SGR	65.1	87.8	93.6	47.0	73.5	81.5	448.5
	RCL-SGRAF	67.7	89.1	93.6	48.0	74.9	83.3	456.6
	L2RM-SGRAF	**70.0**	**90.8**	**95.4**	**51.3**	**76.4**	**83.7**	**467.6**
256	L2RM-SAF	66.7	89.0	93.5	48.0	74.2	82.1	453.5
	L2RM-SGR	66.0	88.5	94.2	48.2	73.9	82.2	453.0
	RCL-SGRAF	66.4	88.9	94.0	47.0	73.3	81.3	450.9
	L2RM-SGRAF	**69.7**	**91.4**	**95.6**	**51.6**	**77.1**	**83.6**	**469.0**

Table 5. Performance with different batch sizes on Flickr30K.

D.3. Analysis on Refined Alignment

Our refined alignment is derived from a partial OT problem, which only allows ρ unit mass to be transported. We further analyze how the transported and untransported data can benefit robust cross-modal retrieval. In Fig. 1, we plot the distribution of averaged refined alignments (image to caption) for both transported and untransported data drawn from each batch of the MS-COCO training set. The normalized distribution is ranked in descending order of probability. The upper subplot shows that the probability of transport data tends to concentrate on one dominant target. It is in line with our expectations that L2RM captures the semantic similarity among some unpaired samples. Interestingly, for those untransported data, the down subplot shows that the distribution of averaged refined alignments approximates a uniform distribution. Such refined alignments are formally equivalent to the label smoothing strategy, wherein

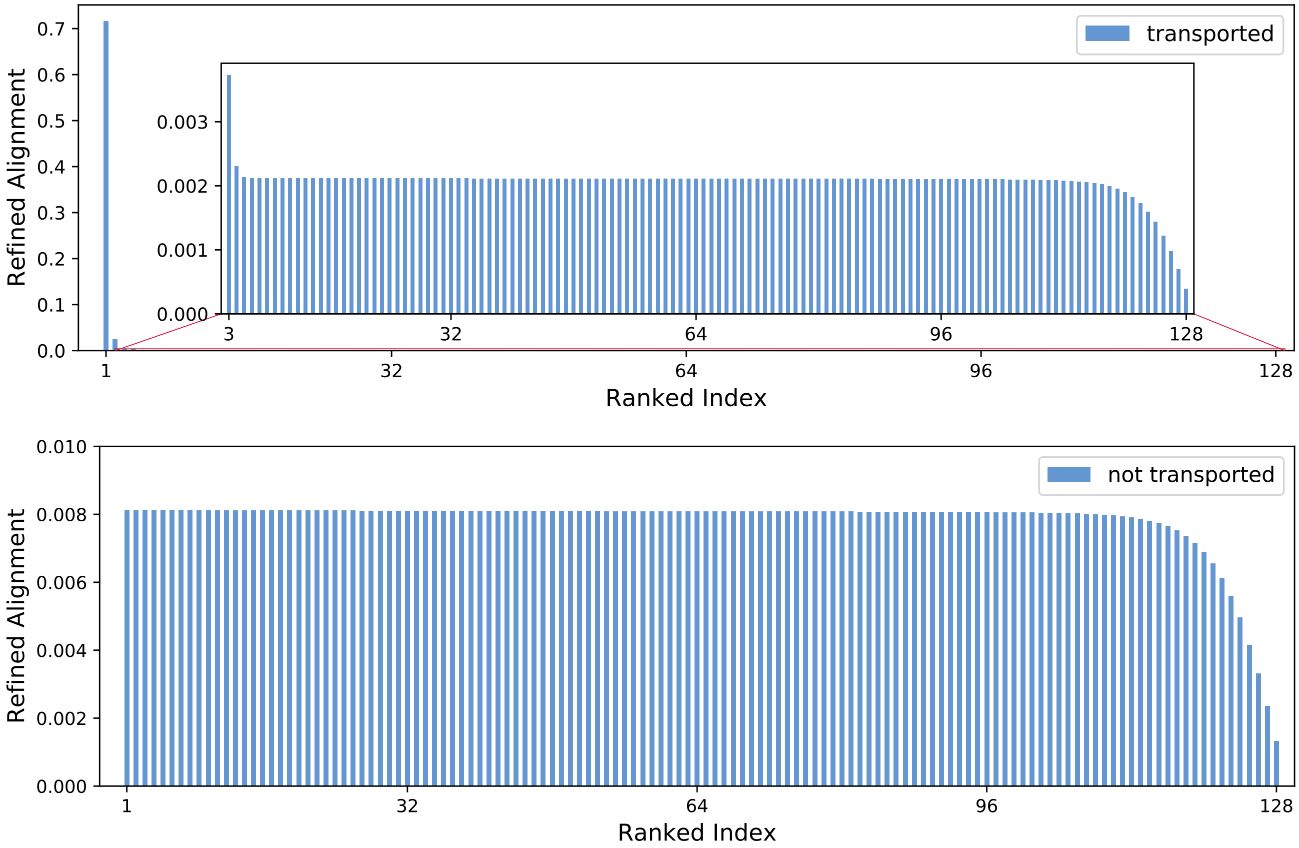

Figure 1. The averaged ranked distribution of normalized refined alignments (image to caption) about transported (upper subplot) and untransported (down subplot) data on MS-COCO under 0.4 PMPs.

the original one-hot targets are mixed with uniform target vectors, *i.e.*,

$$\boldsymbol{y}_i^{LS} = (1-\gamma)\boldsymbol{y}_i + \frac{\gamma}{N_b - 1}(\mathbb{1}_{N_b} - \boldsymbol{y}_i), \qquad (4)$$

where γ is a smoothing parameter. As the original targets provide incorrect supervision for those mismatched pairs, increasing the value of γ as much as possible can alleviate the impact of the wrong matching relation. Our refined alignments accord with this rule, which reveals that the untransported data can also improve the robustness against mismatched pairs.

References

[1] Laetitia Chapel, Mokhtar Z Alaya, and Gilles Gasso. Partial optimal tranport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33:2903–2913, 2020. 1

[2] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 2

[3] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1218–1226, 2021. 2

[4] Xiang Gu, Yucheng Yang, Wei Zeng, Jian Sun, and Zongben Xu. Keypoint-guided optimal transport with applications in heterogeneous domain adaptation. *Advances in Neural Information Processing Systems*, 35:14972–14985, 2022. 2

[5] Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[6] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419, 2021. 2

[7] Yijie Lin, Mouxing Yang, Jun Yu, Peng Hu, Changqing Zhang, and Xi Peng. Graph matching with bi-level noisy correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 23362–23371, 2023. 1

[8] Yijie Lin, Jie Zhang, Zhenyu Huang, Jia Liu, Zujie Wen, and

Xi Peng. Multi-granularity correspondence learning from long-term noisy videos. *arXiv preprint arXiv:2401.16702*, 2024. 1

[9] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. *arXiv preprint arXiv:2308.09911*, 2023. 1

[10] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4948–4956, 2022. 2

[11] Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. 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*, pages 19883–19892, 2023. 2