Cross-Modal Retrieval with Partially Mismatched Pairs

APPENDIX A
PROOF OF THEOREM 1

Inspired by [1], [2], we could give the following proof. First of all,

$$\begin{aligned} P(\mathbf{X}, \overline{\mathbf{Y}}) &= \frac{1}{C_{N-1}^{|\overline{\mathbf{Y}}|}} \sum_{y' \notin \overline{\mathbf{Y}}} P(\mathbf{X}, Y = y') \\ &= \frac{1}{C_{N-1}^{|\overline{\mathbf{Y}}|}} \left(P(\mathbf{X}) - \sum_{y' \notin \overline{\mathbf{Y}}} P(\mathbf{X}, Y = y') \right), \end{aligned}$$

where $X \in \{V, T\}$. Because the marginal distribution is equivalent for positive and negative labels, then we could obtain:

$$\begin{aligned} \sum_{y' \in \overline{\mathbf{Y}}} P(\overline{Y} = y' | \mathbf{X}) &= \sum_{y' \in \overline{\mathbf{Y}}} \frac{P(\mathbf{X}, \overline{Y} = y')}{P(\mathbf{X})} \\ &= \frac{1}{C_{N-1}^{|\overline{\mathbf{Y}}|}} \left(1 - \sum_{y' \in \overline{\mathbf{Y}}} P(Y = y' | \mathbf{X}) \right) \end{aligned}$$

To conduct $\sum_{\overline{\mathbf{Y}} \in \overline{\mathcal{Y}}_y}$ on both the left and the right sides of the above equation, and we could obtain:

$$\begin{aligned} \sum_{\overline{\mathbf{Y}} \in \overline{\mathcal{Y}}_y} \sum_{y' \in \overline{\mathbf{Y}}} P(\overline{Y} = y' | \mathbf{X}) &= \sum_{\overline{\mathbf{Y}} \in \overline{\mathcal{Y}}_y} \frac{1}{C_{N-1}^{|\overline{\mathbf{Y}}|}} \sum_{y' \notin \overline{\mathbf{Y}}} P(y' | \mathbf{X}) \\ &= \sum_{\overline{\mathbf{Y}} \in \overline{\mathcal{Y}}_y} \frac{1}{C_{N-1}^{|\overline{\mathbf{Y}}|}} \left(1 - \sum_{y' \in \overline{\mathbf{Y}}} P(y' | \mathbf{X}) \right) \\ &= \frac{C_{N-1}^{|\overline{\mathbf{Y}}|-1}}{C_{N-1}^{|\overline{\mathbf{Y}}|}} - \sum_{\overline{\mathbf{Y}} \in \overline{\mathcal{Y}}_y} \frac{1}{C_{N-1}^{|\overline{\mathbf{Y}}|}} \left(P(y|\mathbf{X}) + \sum_{\substack{y' \in \overline{\mathbf{Y}} \\ y' \neq y}} P(y' | \mathbf{X}) \right) \\ &= \frac{|\overline{\mathbf{Y}}|}{N - |\overline{\mathbf{Y}}|} - \sum_{\overline{\mathbf{Y}} \in \overline{\mathcal{Y}}_y} \frac{1}{C_{N-1}^{|\overline{\mathbf{Y}}|}} \left(P(y|\mathbf{X}) + \sum_{\substack{y' \in \overline{\mathbf{Y}} \\ y' \neq y}} P(y' | \mathbf{X}) \right) \\ &= \frac{|\overline{\mathbf{Y}}|}{N - |\overline{\mathbf{Y}}|} - \frac{1}{C_{N-1}^{|\overline{\mathbf{Y}}|}} \left(C_{N-1}^{|\overline{\mathbf{Y}}|-1} P(y|\mathbf{X}) + C_{N-2}^{|\overline{\mathbf{Y}}|-2} \sum_{y' \neq y} P(y' | \mathbf{X}) \right) \\ &= \frac{|\overline{\mathbf{Y}}|}{N - |\overline{\mathbf{Y}}|} - \frac{1}{C_{N-1}^{|\overline{\mathbf{Y}}|}} \left(C_{N-1}^{|\overline{\mathbf{Y}}|-1} P(y|\mathbf{X}) + C_{N-2}^{|\overline{\mathbf{Y}}|-2} (1 - P(y|\mathbf{X})) \right) \\ &= \frac{|\overline{\mathbf{Y}}|}{N - |\overline{\mathbf{Y}}|} - \frac{1}{C_{N-1}^{|\overline{\mathbf{Y}}|}} \left(C_{N-2}^{|\overline{\mathbf{Y}}|-2} + C_{N-2}^{|\overline{\mathbf{Y}}|-1} P(y|\mathbf{X}) \right) \\ &= \frac{|\overline{\mathbf{Y}}|}{N - |\overline{\mathbf{Y}}|} - \frac{|\overline{\mathbf{Y}}|(|\overline{\mathbf{Y}}| - 1)}{(N - |\overline{\mathbf{Y}}|)(N-1)} - \frac{|\overline{\mathbf{Y}}|}{N-1} P(y|\mathbf{X}) \\ &= \frac{|\overline{\mathbf{Y}}|}{N-1} - \frac{|\overline{\mathbf{Y}}|}{N-1} P(y|\mathbf{X}) \end{aligned}$$

where $\overline{\mathcal{Y}}_y = \{\overline{\mathbf{Y}} | y \in \overline{\mathbf{Y}}, |\overline{\mathbf{Y}}| = c\}$, $\overline{\mathcal{Y}}_y = C_{N-1}^{|\overline{\mathbf{Y}}|-1}$, and c is the constant size of $\overline{\mathbf{Y}}$, i.e., the number of the selected negatives. Therefore, we could obtain

$$P(y|\mathbf{X}) = 1 - \frac{N-1}{|\overline{\mathbf{Y}}|} \sum_{\overline{\mathbf{Y}} \in \overline{\mathcal{Y}}_y} \sum_{y' \in \overline{\mathbf{Y}}} P(\overline{Y} = y' | \mathbf{X})$$

Finally, we could transform the classification risk as below:

$$\begin{aligned}
R(h; \mathcal{L}) &= \mathbb{E}_{(\mathbf{X},\mathbf{Y}) \sim \mathcal{D}} \mathcal{L}(h(\mathbf{X}), \mathbf{Y}) \\
&= \mathbb{E}_{\mathbf{X} \sim \mathcal{M}} \sum_{y \in \mathcal{Y}} P(y|\mathbf{X}) \mathcal{L}(h(\mathbf{X}), y) \\
&= \mathbb{E}_{\mathbf{X} \sim \mathcal{M}} \sum_{y \in \mathcal{Y}} \left(1 - \frac{N-1}{|\overline{\mathbf{Y}}|} \sum_{\overline{\mathbf{Y}} \in \overline{\mathcal{Y}}_y} \sum_{y' \in \overline{\mathbf{Y}}} P(\overline{Y} = y' | \mathbf{X}) \right) \mathcal{L}(h(\mathbf{X}), y) \\
&= \mathbb{E}_{\mathbf{X} \sim \mathcal{M}} \left(\sum_{y \in \mathcal{Y}} \mathcal{L}(h(\mathbf{X}), y) - \frac{N-1}{|\overline{\mathbf{Y}}|} \sum_{y \in \mathcal{Y}} \sum_{\overline{\mathbf{Y}} \in \overline{\mathcal{Y}}_y} \sum_{y' \in \overline{\mathbf{Y}}} P(\overline{Y} = y' | \mathbf{X}) \mathcal{L}(h(\mathbf{X}), y) \right) \\
&= \mathbb{E}_{\mathbf{X} \sim \mathcal{M}} \left(\sum_{y \in \mathcal{Y}} \mathcal{L}(h(\mathbf{X}), y) - \frac{N-1}{|\overline{\mathbf{Y}}|} \sum_{\overline{\mathbf{Y}} \in \overline{\mathcal{Y}}} \sum_{y' \in \overline{\mathbf{Y}}} P(\overline{Y} = y' | \mathbf{X}) \mathcal{L}(h(\mathbf{X}), y') \right) \\
&= \mathbb{E}_{(\mathbf{X},\overline{\mathbf{Y}}) \sim \overline{\mathcal{D}}} \left(\sum_{y \in \mathcal{Y}} \mathcal{L}(h(\mathbf{X}), y) - \frac{N-1}{|\overline{\mathbf{Y}}|} \sum_{y \in \overline{\mathbf{Y}}} \mathcal{L}(h(\mathbf{X}), y) \right) \\
&= \mathbb{E}_{(\mathbf{X},\overline{\mathbf{Y}}) \sim \overline{\mathcal{D}}} \left(\sum_{y \notin \overline{\mathbf{Y}}} \mathcal{L}(h(\mathbf{X}), y) - \frac{N - |\overline{\mathbf{Y}}| - 1}{|\overline{\mathbf{Y}}|} \sum_{y \in \overline{\mathbf{Y}}} \mathcal{L}(h(\mathbf{X}), y) \right) \\
&= \mathbb{E}_{(\mathbf{X},\overline{\mathbf{Y}}) \sim \overline{\mathcal{D}}} \overline{\mathcal{L}}(h(\mathbf{X}), \overline{\mathbf{Y}}) \\
&= \overline{R}(h; \overline{\mathcal{L}}),
\end{aligned}$$

Therefore, we could obtain the complementary/negative loss $\overline{\mathcal{L}}(h(\mathbf{X}), \overline{\mathbf{Y}}) = \sum_{y \notin \mathcal{Y}} \mathcal{L}(h(\mathbf{X}), y) - \frac{N - |\overline{\mathbf{Y}}| - 1}{|\overline{\mathbf{Y}}|} \sum_{y \in \overline{\mathbf{Y}}} \mathcal{L}(h(\mathbf{X}), y)$, which concludes the proof.

Appendix B
Proof of Robustness

Following [3], [4], we could give the following proof.

Lemma 1. *In an instance-level retrieval problem, $\mathcal{L}_{\mathrm{mae}}$ is noise tolerant against uniform PMPs, if mismatching noise $\eta < \frac{N-1}{N}$.*

Proof. Let $\mathcal{L}_{\mathrm{mae}}(p, Y) = \sum_{p \in \overline{\mathcal{P}}_Y} p$, for uniform mismatching noise, the noise risk can be defined as:

$$\begin{aligned}
R^{\eta}(h) = \mathbb{E}_{\mathbf{X},\hat{Y}} \mathcal{L}_{\mathrm{mae}} \left(h(\mathbf{X}), \hat{Y} \right) &= \mathbb{E}_{\mathbf{X}} \mathbb{E}_{Y|\mathbf{X}} \mathbb{E}_{\hat{Y}|\mathbf{X},Y} \mathcal{L}_{\mathrm{mae}} \left(h(\mathbf{X}), \hat{Y} \right) \\
&= \mathbb{E}_{\mathbf{X}} \mathbb{E}_{Y|\mathbf{X}} \left((1-\eta) \mathcal{L}_{\mathrm{mae}} \left(h(\mathbf{X}), Y \right)) + \frac{\eta}{N-1} \sum_{K \neq Y} \mathcal{L}_{\mathrm{mae}} \left(h(\mathbf{X}), K \right)) \right) \\
&= (1-\eta) R(h) + \frac{\eta}{N-1} \left(\mathbb{E}_{\mathbf{X},Y} \sum_{K=1}^{N} \mathcal{L}_{\mathrm{mae}} \left(h(\mathbf{X}), K \right)) - R(h) \right) \\
&= R(h) \left(1 - \frac{\eta N}{N-1} \right) + \frac{\eta}{N-1} C,
\end{aligned} \tag{1}$$

where the last equality holds due to $\mathbb{E}_{\mathbf{X},Y} \sum_{K=1}^{N} \mathcal{L}_{\mathrm{mae}} \left(h(\mathbf{X}), K \right)) = C = |\overline{\mathbf{Y}}|$. Therefore,

$$R^{\eta}(h^*) - R^{\eta}(h) = \left(1 - \frac{\eta N}{N-1} \right) \left(R(h^*) - R(h) \right) \leqslant 0, \tag{2}$$

because $\eta < \frac{N-1}{N}$ and h^* is a global minimizer of $R(h)$. This proves h^* is also the global minimizer of risk $R^{\eta}(h)$, that is, $\mathcal{L}_{\mathrm{mae}}$ is noise tolerant to symmetric label noise. □

APPENDIX C
IMPACT ANALYSIS FOR THE NUMBER OF NEGATIVES

In this section, we conducted experiments to investigate the influence of the number of negative pairs, i.e., the retrieval performance under different ratios of negatives in a mini-batch. The experimental results are shown in Table 9. From the table, one could find that with more negative samples the underfitting problem is alleviated, which verified the effectiveness of our motivation.

TABLE 9: Comparison with different number of negative samples.

Ratio	Image-to-Text			Text-to-Image			rSum
	R@1	R@5	R@10	R@1	R@5	R@10	
0.1	0.1	0.5	1.3	0.1	0.4	1.0	3.4
0.5	0.2	3.0	4.8	0.5	2.4	4.8	15.7
0.6	1.5	4.8	7.2	0.8	3.7	6.6	24.6
0.7	50.8	79.0	86.8	34.4	60.4	70.5	381.9
0.8	59.9	83.4	89.3	40.8	66.3	75.5	415.2
0.9	61.0	85.0	**91.5**	42.3	68.3	76.7	424.8
0.99	62.5	84.9	91.3	42.9	**68.9**	**77.9**	428.4
0.999	**62.6**	**85.2**	91.3	**43.6**	68.8	77.3	**428.8**

APPENDIX D
COMPARISON RESULTS ON MS-COCO 5K

In this section, we conduct experiments on MS-COCO 5K. The experimental results are shown in Table 11. From the table, one could find that our method also achieves the best performance under different mismatching noises on MS-COCO 5K. From the experimental results, one could see that our method could remarkably improve the robustness of models against PMPs.

TABLE 10: Image-text matching with different mismatching rates (MRates) on MS-COCO 5K.

Method	MRate	Image-to-Text			Text-to-Image			rSum	MRate	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	
SCAN [5]	0.2	37.3	69.1	81.0	23.4	53.3	67.0	360.8	0.4	20.9	47.1	60.1	9.7	27.6	38.9	329.9
PolyLoss [6]		44.6	74.9	84.9	27.8	57.8	70.8	317.2		14.5	41.6	55.9	6.8	22.0	33.1	173.9
VSRN [7]		38.0	67.4	78.5	28.7	57.5	68.8	338.9		11.5	31.5	44.9	5.7	19.3	29.6	142.5
GSMN [8]		39.6	72.2	83.2	29.5	59.5	72.0	356.0		5.4	18.2	26.7	3.6	14.7	24.0	92.6
IMRAM [9]		45.3	76.2	86.4	34.7	63.3	74.4	380.3		27.4	56.8	70.6	19.4	44.7	57.3	276.2
SAF [10]		48.4	78.0	87.7	35.9	65.2	76.8	392.0		3.5	14.1	20.7	6.5	18.0	26.1	88.9
SGR [10]		8.7	26.8	40.1	8.9	27.0	40.1	151.6		0.3	1.4	2.2	0.2	0.7	1.3	6.1
SGRAF [10]		48.4	78.0	87.7	35.9	65.2	76.8	392.0		2.4	10.9	17.3	5.3	15.1	22.7	73.7
NCR* [11]		50.7	80.1	88.3	36.4	65.8	77.0	398.3		50.0	77.7	87.0	35.3	64.4	76.0	390.4
NCR [11]		54.9	82.6	90.5	39.0	68.4	79.2	414.6		53.5	80.5	88.9	37.9	67.2	78.2	406.2
RCL-VSRN		47.6	76.8	86.7	34.3	65.0	76.8	387.2		44.1	73.6	84.0	31.2	60.6	72.7	366.2
RCL-GSMN		55.9	83.0	90.2	39.3	67.6	78.4	414.4		52.1	80.3	88.9	37.3	65.2	75.7	399.5
RCL-IMRAM		52.0	80.9	89.6	37.6	66.4	77.0	403.5		51.7	79.9	88.4	36.3	64.0	74.4	394.7
RCL-SAF		55.1	82.8	90.7	39.6	68.5	79.3	416.0		52.2	80.8	89.0	37.7	66.2	77.0	402.9
RCL-SGR		54.9	83.4	90.8	39.7	68.9	79.4	417.1		53.2	81.1	89.6	37.7	66.5	77.3	405.4
RCL-SGRAF		**58.4**	**84.9**	**91.4**	**41.6**	**70.4**	**80.8**	**427.5**		**56.2**	**83.3**	**90.6**	**39.8**	**68.4**	**79.0**	**417.3**
SCAN [5]	0.6	11.3	31.5	45.0	0.4	1.0	1.5	90.7	0.8	3.1	11.4	18.1	0.0	0.1	0.2	2.9
PolyLoss [6]		4.1	14.3	23.0	0.1	0.6	1.0	43.1		0.3	1.1	1.8	0.0	0.1	0.2	3.5
VSRN [7]		3.2	12.2	20.0	1.2	4.9	8.4	49.9		0.4	1.4	2.7	0.2	0.7	1.2	6.6
GSMN [8]		1.3	4.7	7.6	0.9	3.4	5.5	23.4		0.4	1.4	2.7	0.4	1.6	2.9	9.4
IMRAM [9]		4.2	18.0	32.0	6.5	20.7	30.8	112.2		0.5	1.4	2.4	0.0	0.2	0.3	4.8
SAF [10]		0.0	0.2	0.2	0.2	0.9	1.6	3.1		0.0	0.2	0.3	0.0	0.1	0.2	0.8
SGR [10]		0.0	0.2	0.3	0.0	0.1	0.2	0.8		0.0	0.1	0.2	0.0	0.1	0.2	0.6
SGRAF [10]		0.0	0.2	0.3	0.0	0.3	0.6	1.4		0.0	0.1	0.2	0.0	0.1	0.2	0.6
NCR* [11]		0.0	0.1	0.1	0.0	0.1	0.2	0.5		0.0	0.1	0.1	0.0	0.1	0.2	0.5
NCR [11]		0.0	0.1	0.1	0.0	0.1	0.2	0.5		0.0	0.1	0.1	0.0	0.1	0.2	0.5
RCL-VSRN		37.0	67.0	78.3	25.3	53.1	66.3	327.0		26.8	53.5	66.1	15.9	39.1	52.1	253.5
RCL-GSMN		47.5	76.3	84.8	33.6	61.5	72.5	376.2		37.1	66.2	77.6	25.7	51.3	63.4	321.3
RCL-IMRAM		48.5	77.9	86.6	31.6	60.9	72.8	378.3		37.1	66.5	76.9	26.1	50.6	61.9	319.1
RCL-SAF		48.3	77.2	86.3	33.5	61.8	73.1	380.2		39.4	69.1	80.0	26.8	53.4	65.5	334.2
RCL-SGR		50.1	78.0	86.1	34.3	62.4	73.7	384.6		40.8	69.7	80.6	27.7	54.4	66.0	339.2
RCL-SGRAF		**53.4**	**79.9**	**88.4**	**36.5**	**64.9**	**76.0**	**399.1**		**44.5**	**73.2**	**82.7**	**30.4**	**57.9**	**69.1**	**357.8**

* denotes the results of one single model for NCR.

APPENDIX E
COMPARISONS WITH STATE OF THE ARTS ON THE SETTING OF NCR

In this section, we conduct some experiments under the mismatched pairs generated by NCR [11] on MS-COCO 1K and Flick30K as shown in Table 11. From the experimental results, one could find that our method performs remarkably better than all baselines in the setting used in [11]. Even in the absence of PMPs, our method still achieves the best performance, which indicates that the proposed loss function could improve the performance of cross-modal models under arbitrary noise rates. Besides, one could see that the baselines and our method could achieve much better results compared with Table 2, which indicates that the setting of NCR [11] is easier than our setting. Even under a noise rate of 0.2, our method has very little performance drop, *e.g.*, only 0.6 drop in terms of score sum.

TABLE 11: Image-text matching with different mismatching rates (MRate) on MS-COCO 1K and Flick30K. Notably, the mismatched pairs are given by NCR [11].

Noise	Methods	MS-COCO							Flickr30K						
		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%	SCAN [5]	69.2	93.6	97.6	56.0	86.5	93.5	496.4	67.4	90.3	95.8	48.6	77.7	85.2	465.0
	VSRN [7]	76.2	94.8	98.2	62.8	89.7	95.1	516.8	71.3	90.6	96.0	54.7	81.8	88.2	482.6
	IMRAM [9]	76.7	95.6	98.5	61.7	89.1	95.0	516.6	74.1	93.0	96.6	53.9	79.4	87.2	484.2
	SAF [10]	76.1	95.4	98.3	61.8	89.4	95.3	516.3	73.7	93.3	96.3	56.1	81.5	88.0	488.9
	SGR [10]	78.0	95.8	98.2	61.4	89.3	95.4	518.1	75.2	93.3	96.6	56.2	81.0	86.5	488.8
	SGRAF [10]	79.6	96.2	98.5	63.2	90.7	**96.1**	524.3	77.8	94.1	97.4	58.5	83.0	88.8	499.6
	NCR* [11]	75.9	95.4	98.0	61.1	89.2	95.1	514.7	72.7	91.8	95.8	55.7	82.3	88.3	486.6
	NCR [11]	78.7	95.8	98.5	63.3	90.4	95.8	522.5	77.3	94.0	97.5	59.6	84.4	89.9	502.7
	RCL-SAF	78.5	96.1	98.6	62.7	90.0	95.4	521.3	76.7	93.7	97.3	56.2	82.6	88.8	495.3
	RCL-SGR	78.2	96.2	98.4	62.9	90.0	95.7	521.4	77.5	94.7	97.4	58.8	83.3	88.9	500.6
	RCL-SGRAF	**80.4**	**96.4**	**98.7**	**64.3**	**90.8**	96.0	**526.6**	**79.9**	**96.1**	**97.8**	**61.1**	**85.4**	**90.3**	**510.6**
20%	SCAN [5]	66.2	91.0	96.4	45.0	80.2	89.3	468.1	59.1	83.4	90.4	36.6	67.0	77.5	414.0
	VSRN [7]	25.1	59.0	74.8	17.6	49.0	64.1	289.6	58.1	82.6	89.3	40.7	68.7	78.2	417.6
	IMRAM [9]	68.6	92.8	97.6	55.7	85.0	91.0	490.7	63.0	86.0	91.3	41.4	71.2	80.5	433.4
	SAF [10]	67.3	92.5	96.6	53.4	84.5	92.4	486.7	51.0	79.3	88.0	38.3	66.5	76.2	399.3
	SGR [10]	67.8	91.7	96.2	52.9	83.5	90.1	482.2	62.8	86.2	92.2	44.4	72.3	80.4	438.3
	SGRAF [10]	75.4	95.2	97.9	60.1	88.5	94.8	511.9	72.8	90.8	95.4	56.4	82.1	88.6	486.1
	NCR* [11]	76.7	95.2	97.8	60.8	88.6	94.9	514.0	69.6	92.7	96.0	54.2	80.8	87.4	480.7
	NCR [11]	77.7	95.5	98.2	62.5	89.3	95.3	518.5	75.0	93.9	**97.5**	58.3	83.0	89.0	496.7
	RCL-SAF	**77.9**	**95.6**	**98.5**	62.5	89.3	95.1	**518.9**	**75.2**	93.0	96.4	55.9	81.6	88.0	490.1
	RCL-SGR	**78.3**	**95.8**	**98.5**	**62.5**	**89.6**	95.1	**519.8**	75.0	93.2	96.6	57.5	81.7	88.1	492.1
	RCL-SGRAF	**79.4**	**96.3**	**98.8**	**63.8**	**90.3**	**95.5**	**524.1**	**77.5**	**94.6**	97.0	**59.5**	**83.9**	**89.8**	**502.3**
50%	SCAN [5]	40.8	73.5	84.9	5.4	15.1	21.0	240.7	27.7	57.6	68.8	16.2	39.3	49.8	259.4
	VSRN [7]	23.5	54.7	69.3	16.0	47.8	65.9	277.2	14.3	37.6	50.0	12.1	30.0	39.4	183.4
	IMRAM [9]	21.3	60.2	75.9	22.3	52.8	64.3	296.8	9.1	26.6	38.2	2.7	8.4	12.7	97.7
	SAF [10]	30.4	67.8	82.3	33.5	69.0	82.8	365.8	30.3	63.6	75.4	27.9	53.7	65.1	316.0
	SGR [10]	60.6	87.4	93.6	46.0	74.2	79.0	440.8	36.9	68.1	80.2	29.3	56.2	67.0	337.7
	SGRAF [10]	71.7	94.1	97.7	57.0	86.6	93.7	500.8	69.8	90.3	94.8	50.1	77.5	85.2	467.7
	NCR* [11]	72.7	94.1	97.7	57.5	87.1	93.8	502.9	67.7	91.4	95.2	50.4	77.3	84.7	466.7
	NCR [11]	74.6	94.6	97.8	59.1	87.8	94.5	508.4	72.9	**93.0**	**96.3**	54.3	79.8	86.5	482.8
	RCL-SAF	**76.0**	**94.7**	**98.1**	**60.0**	87.6	94.0	**510.4**	69.6	90.8	94.3	51.9	77.0	85.1	468.7
	RCL-SGR	**75.7**	**94.6**	**97.9**	**60.1**	87.7	94.2	**510.2**	72.2	90.5	94.4	52.2	77.9	85.5	472.7
	RCL-SGRAF	**77.6**	**95.3**	**98.2**	**61.8**	**88.7**	**94.8**	**516.4**	**75.8**	92.0	96.0	**55.4**	**80.7**	**87.2**	**487.1**

* denotes the results of one single model for NCR.

Appendix F
Examples of PMPs

In this section, we provide some examples to illustrate the practicality of PMPs in real applications. In the experiments, we have studied a real-world dataset (i.e., Conceptual Captions), which is collected from the Internet, and it is inevitable to introduce a lot of mismatched pairs as shown in Fig. 9. From Fig. 9, one could see that the captions are fully irrelevant to the corresponding images. Although rigorous filtering and post-processing steps are conducted to reduce the PMPs, there still exists $3\% \sim 20\%$ PMPs in the dataset [12]. Therefore, the PMP problem is common in real applications. Although we did not intentionally generate partially relevant pairs, the random PMPs still contain many partially relevant pairs as you stated, some examples of which are shown in Fig. 10. Note that, although our experiments mainly focus on evaluating the effectiveness of our method on random PMPs for convenience, our method can not only handle fully irrelevant pairs but also all kinds of PMPs even with clean data, which has been demonstrated in our experiments. Therefore, this work studied an applicable and challenging problem, and presented a novel and effective method for different PMPs.

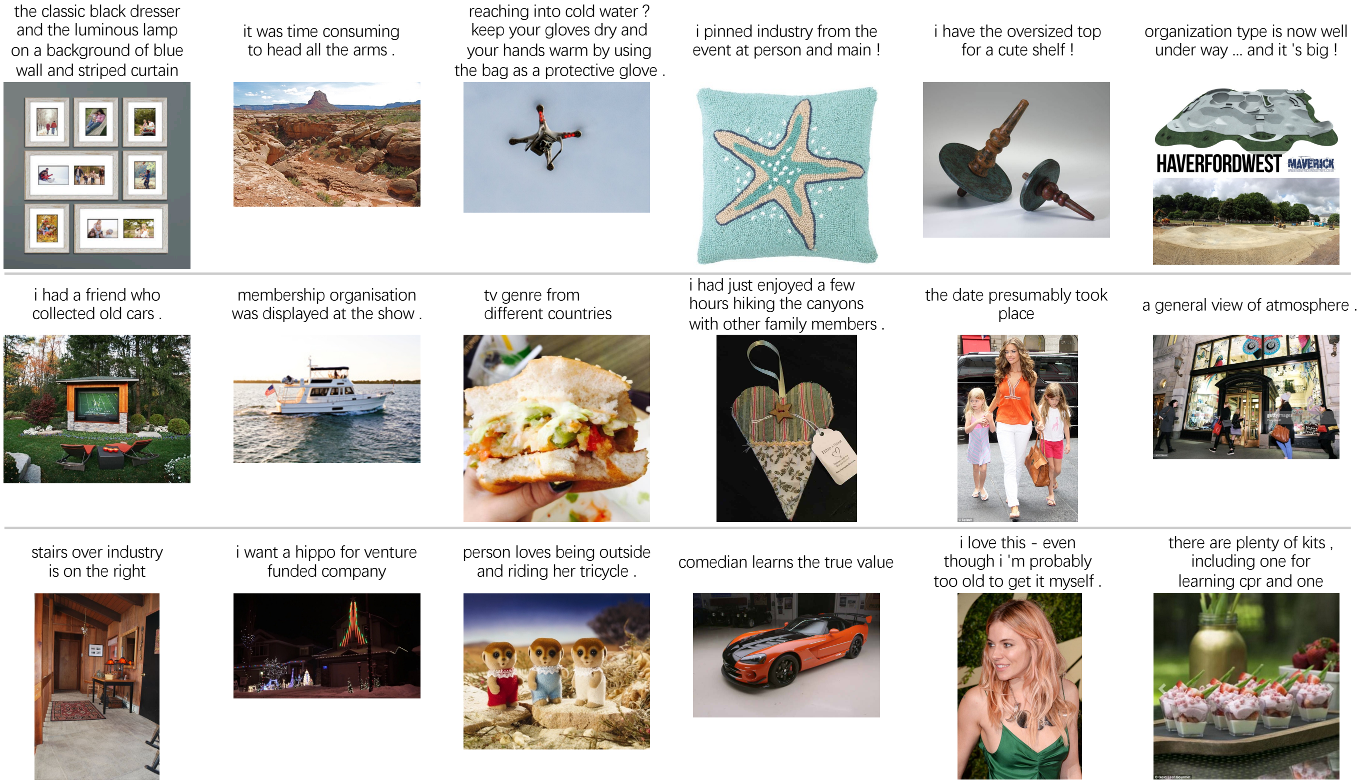
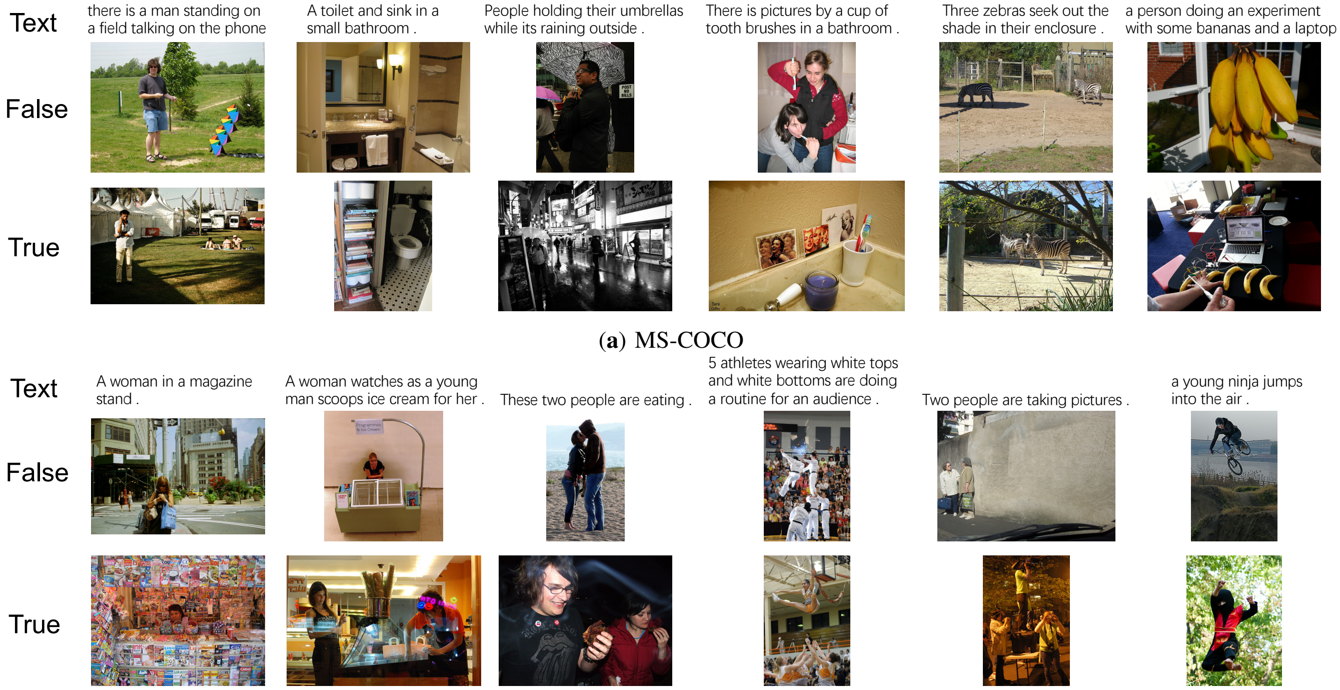

the classic black dresser and the luminous lamp on a background of blue wall and striped curtain
it was time consuming to head all the arms .
reaching into cold water ? keep your gloves dry and your hands warm by using the bag as a protective glove .
i pinned industry from the event at person and main !
i have the oversized top for a cute shelf !
organization type is now well under way ... and it 's big !
HAVERFORDWEST MAVERICK
i had a friend who collected old cars .
membership organisation was displayed at the show .
tv genre from different countries
i had just enjoyed a few hours hiking the canyons with other family members .
the date presumably took place
a general view of atmosphere .
stairs over industry is on the right
i want a hippo for venture funded company
person loves being outside and riding her tricycle .
comedian learns the true value
i love this - even though i 'm probably too old to get it myself .
there are plenty of kits , including one for learning cpr and one

Fig. 9: Mismatched examples from Conceptual Captions [12]. From the mismatched image-text pairs, one could see that incorrect captions are fully irrelevant to the corresponding images.


Text
False
True
there is a man standing on a field talking on the phone
A toilet and sink in a small bathroom .
People holding their umbrellas while its raining outside .
There is pictures by a cup of tooth brushes in a bathroom .
Three zebras seek out the shade in their enclosure .
a person doing an experiment with some bananas and a laptop
Text
False
True
A woman in a magazine stand .
A woman watches as a young man scoops ice cream for her .
These two people are eating .
5 athletes wearing white tops and white bottoms are doing a routine for an audience .
Two people are taking pictures .
a young ninja jumps into the air .

(**a**) MS-COCO

(**b**) Flick30K

Fig. 10: Some partially irrelevant pairs in our random PMPs. From ((a)) MS-COCO [13] and ((b)) Flick30K [14], one could see that some captions are partially relevant to the corresponding images. Namely, some texts in the caption are likely to be relevant, e.g., "there is a man standing on a field" is relevant to the mismatched/false image in the first group.

REFERENCES

[1] T. Ishida, G. Niu, W. Hu, and M. Sugiyama, "Learning from complementary labels," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5644–5654.

[2] L. Feng, T. Kaneko, B. Han, G. Niu, B. An, and M. Sugiyama, "Learning with multiple complementary labels," in *International Conference on Machine Learning (ICML)*. PMLR, 2020, pp. 3072–3081.

[3] A. Ghosh, H. Kumar, and P. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 31, no. 1, 2017.

[4] X. Ma, H. Huang, Y. Wang, S. Romano, S. Erfani, and J. Bailey, "Normalized loss functions for deep learning with noisy labels," in *International Conference on Machine Learning (ICML)*, 2020.

[5] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 201–216.

[6] J. Wei, X. Xu, Y. Yang, Y. Ji, Z. Wang, and H. T. Shen, "Universal weighting metric learning for cross-modal matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 005–13 014.

[7] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4654–4662.

[8] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, "Graph structured network for image-text matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 921–10 930.

[9] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, "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 (CVPR)*, 2020, pp. 12 655–12 663.

[10] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," Technical Report, Tech. Rep., 2021.

[11] Z. Huang, G. Niu, X. Liu, W. Ding, X. Xiao, X. Peng *et al.*, "Learning with noisy correspondence for cross-modal matching," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[12] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "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.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 740–755.

[14] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.