

Supplementary Material for Learning with Noisy Triplet Correspondence for Composed Image Retrieval

Shuxian Li^{1,*}, Changhao He^{1,*}, Xiting Liu², Joey Tianyi Zhou^{3,4}, Xi Peng^{1,5}, Peng Hu^{1,†}

¹College of Computer Science, Sichuan University, China. ²Georgia Institute of Technology, USA.

³Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore.

⁴Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore.

⁵National Key Laboratory of Fundamental Algorithms and Models for Engineering Simulation, China.

1. Introduction

In this supplementary material, we provide an extended experimental analysis to further validate the effectiveness of TME, alongside illustrative examples with noisy triplet correspondence (NTC) to highlight the prevalence of such instances even in well-annotated datasets. Specifically, we conduct further parametric analysis in Sec. 2.1. Then, We analyse the performance against overfitting and qualitative results for a detailed comparison in Sec. 2.2 and Sec. 2.3. Finally, we present NTC examples from CIRR [6] and FashionIQ [8] datasets in Sec. 3, emphasizing their ubiquitous nature.

2. Additional Experimental Analysis

2.1. Parametric Analysis

In this section, we conduct further parametric analysis of the GMM [1, 2] threshold and the prompt token length. In our implementation of TME, the default GMM threshold is set to 0.5, and the prompt token length is fixed at 32, consistent with the length of the query tokens [5]. The GMM threshold and the prompt tokens length are denoted as G and n_p , respectively. Sensitivity analyses of these parameters are summarized in Figure 1, from which we derive the following observations:

- For $G \in [0.1, 0.9]$, the performance is relatively insensitive to the threshold value. This is mainly because GMM often assigns probabilities p_i that are either significantly large or small for a given sample.
- A moderate value of G achieves the best performance. A very small G introduces too many noisy triplets into S^c , degrading performance. Conversely, an excessively high G excludes most triplets, resulting in underfitting. For example, When $G = 0.999$, no samples are retained in S^c , causing training failure.

- A prompt token length of $n_p = 32$ achieves optimal performance as it matches the length of the query tokens q and the image representations F^r encoded by E_I . This consistency benefits E_f -encoded representations z^{rm} and z^{pm} , which are derived from concatenations $[F^r, m]$ and $[p, m]$, respectively. A prompt token length of $n_p = 128$ exceeds memory limits on our GPU during experiments.

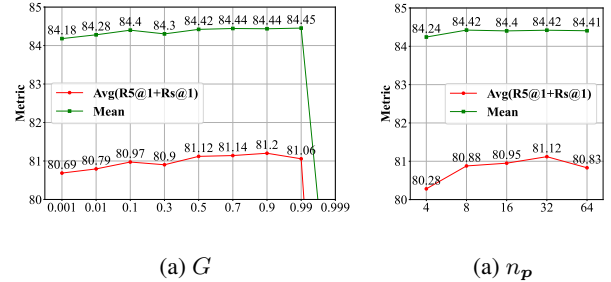


Figure 1. Variation of performance with different GMM thresholds G and lengths of prompt tokens n_p on the CIRR validation with a noise ratio of $\sigma = 0.2$.

2.2. Performance Against Overfitting

The performance of TME, SPRC [9], RCL [3], RDE [7], CaLa [4], and SSN [10] across epochs is presented in Figure 2, which highlights TME’s effectiveness, demonstrating its superiority over both general and robust methods. Specifically, TME achieves the highest 30-epoch accuracies and the highest number of the best accuracies across datasets of different noise levels. Ordinary methods such as SPRC and CaLa suffer significant performance degradation and overfitting, due to their lack of robustness. In contrast, through robust learning strategies, TME shows remarkable resistance to overfitting and maintains stable performance throughout training under any noise level. Compared to robust methods, TME exhibits greater robustness and better

*The first two authors contributed equally.

†Corresponding author: Peng Hu (penghu.ml@gmail.com).

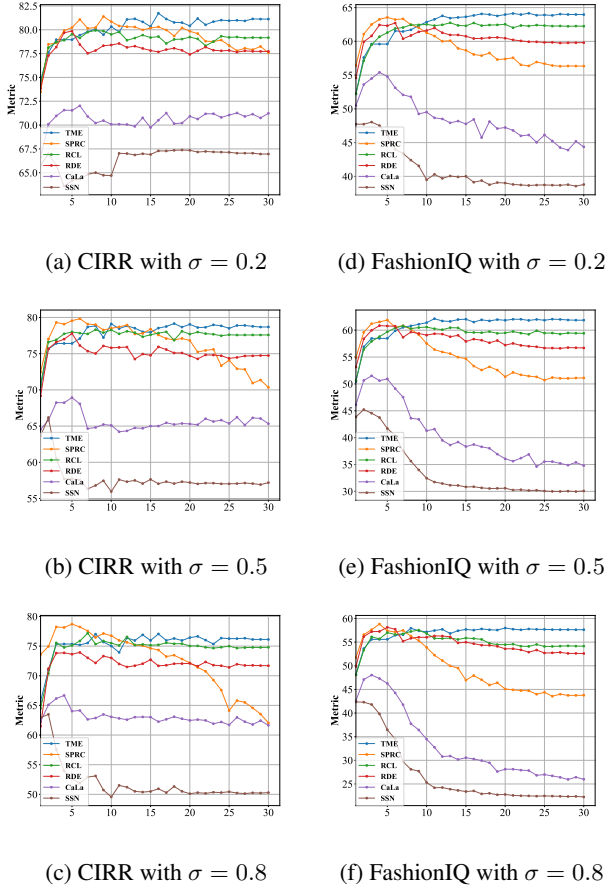


Figure 2. Performance w.r.t. epoch on CIRR validation dataset and FashionIQ validation dataset with various noise ratios σ .

performance as well. Existing robust methods rigidly apply noisy dual correspondence solutions to the CIR task, limiting their ability to exploit intrinsic relationships. Conversely, TME uses visual variation modeling to reconstruct associations and effectively learn from noisy data, enhancing performance. This robustness achieved by TME is crucial for real-world scenarios without clean validation sets, ensuring reliable training without early stopping.

2.3. Qualitative Results

Retrieving results of SPRC and TME at the last checkpoint when $\sigma = 0.5$ is shown in Figure 4. These results demonstrate that TME more effectively handles some queries and pushes irrelevant images away from low-rank results under a high noise ratio. A comparison of similarity matrices of multimodal query z^{rm} and target image representation z^t between SPRC and TME is presented in Figure 3, revealing that TME produces a sharper diagonal with better background contrast, showing a better alignment of the positive pairs and greater dilution of noisy training data. This indi-

cates TME’s superior robustness and ability to learn effectively under a high noise ratio.

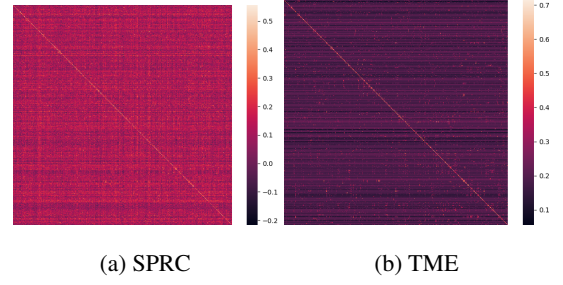


Figure 3. Visualization of the similarity matrix of multimodal query and target image on CIRR validation dataset with a noise ratio of $\sigma = 0.5$.

3. Examples with Noisy Triplet Correspondence

We present examples with noisy triplet correspondence (NTC) from the original CIRR and FashionIQ datasets **without** synthetic noise. Specifically, from 100 randomly selected images from CIRR, we identify **17** clearly NTC examples and display 12 noisy triplets in Figure 5. Similarly, from 100 randomly selected images from FashionIQ, we find **19** NTC examples and show 12 noisy triplets in Figure 6. In FashionIQ, modifications are often insufficient or inaccurate, leading to many false negatives and false positives. In CIRR, annotators tend to describe only the target image’s characteristics, resulting in modification-target matched pairs, i.e., partially matched triplets. Consequently, FashionIQ contains several completely mismatched triplets, and CIRR has a few. Despite both datasets being well-annotated, we observe that partially matched triplets with partially inaccurate modifications or modifications matching only the target images are common, highlighting the need for methods capable of learning with noisy triplet correspondence for composed image retrieval.

References

- [1] Eric Arazo, Diego Ortego, Paul Albert, Noel O’Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pages 312–321. PMLR, 2019. 1
- [2] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017. 1
- [3] Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. Cross-modal retrieval with partially mismatched pairs.

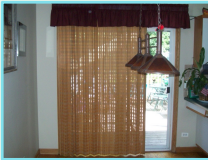






















































Query		Retrieval Results				
1.	<p>Target the corner of room in dim light mode rather making sunshine light into scene</p> 	    	SPRC			
		    	TME			
2.	<p>Target on igloo shapes houses instead of tall buildings</p> 	    	SPRC			
		    	TME			
3.	<p>is black wit thin straps Has straps and is black</p> 	    	SPRC			
		    	TME			
4.	<p>is a mutlicolored abstract pattern button down shirt with collar is much more patterned and colourful</p> 	    	SPRC			
		    	TME			
5.	<p>is white colored and has sleeves is white in color and has long sleeves</p> 	    	SPRC			
		    	TME			

Figure 4. Qualitative results on CIRRR and FashionIQ validation datasets. We show the results of both SPRC and TME for a clear comparison. Images in the green box are the target images corresponding to reference images. Note that the FashionIQ dataset contains apparent false negatives. Most of TME’s retrieval results align well with the query, even if they are not the specified target images.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(8):9595–9610, 2023. 1

- [4] Xintong Jiang, Yaxiong Wang, Mengjian Li, Yujiao Wu, Bingwen Hu, and Xueming Qian. Cala: Complementary association learning for augmenting composed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Informa-*

tion Retrieval, pages 2177–2187, 2024. 1

- [5] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1
- [6] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and

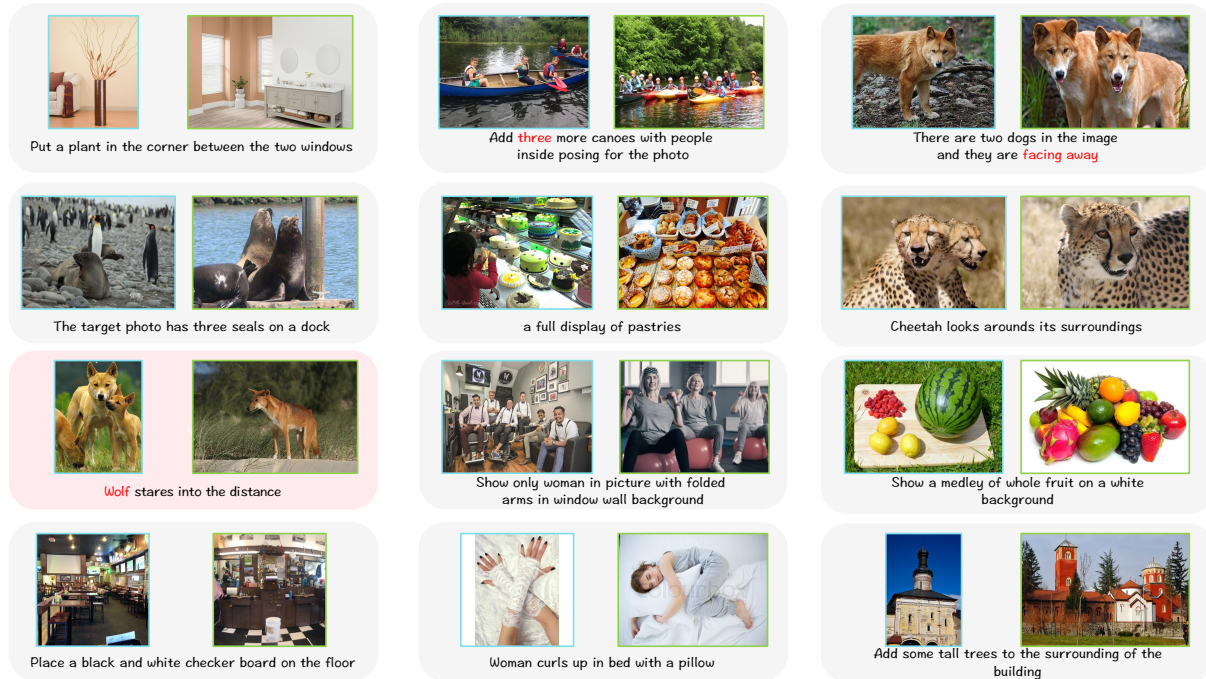


Figure 5. Examples with noisy triplet correspondence in CIRR. Reference images and corresponding target images are marked with blue and green boxes, respectively. Inaccurate modifications are marked in red. The triplet with a red background is a completely mismatched triplet, whose modification describes only the target image but incorrectly identifies a dingo as a wolf. The rest triplets are partially matched triplets.

Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134, 2021. [1](#)

- [7] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27197–27206, 2024. [1](#)
- [8] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317, 2021. [1](#)
- [9] Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, Chun-Mei Feng, et al. Sentence-level prompts benefit composed image retrieval. In *The Twelfth International Conference on Learning Representations*, 2024. [1](#)
- [10] Xingyu Yang, Daqing Liu, Heng Zhang, Yong Luo, Chaoyue Wang, and Jing Zhang. Decompose semantic shifts for composed image retrieval. *arXiv preprint arXiv:2309.09531*, 2023. [1](#)



Figure 6. Examples with noisy triplet correspondence in FashionIQ. Reference images and corresponding target images are marked with blue and green boxes, respectively. Inaccurate modifications are marked in red. Triplets with a red background are completely mismatched triplets whose modifications are completely inaccurate. The rest triplets are partially matched triplets.