

Supplementary Materials:

Anonymous Authors

1 HYPER-PARAMETER CHOICE

For both datasets, we set c_0 to 10000. For FashionIQ, we set c_1 to 20000. For CIRR, we set c_1 to 15000. For the CIR task we manually tuned $\tau \in \{0.01, 0.02, 0.03, 0.05\}$ and $learning_rate \in \{5e-6, 6e-6\}$, the best parameters are reported in Table 1. For each model, we use the largest batch size that V100 can run, that is, 256 for TG-CIR, 256 for CLIP4CIR, 128 for BLIP4CIR, and 32 for SPRC. For the first stage of ZS-CIR, we set $batch_size$ to 48, $learning_rate$ to $2e-6$, and τ to 0.01 in the first stage. For the second stage of ZS-CIR, we set $batch_size$ to 128, $learning_rate$ to $2e-6$, and τ to 0.02.

Model	Dataset	$learning_rate$	τ
TG-CIR	FashionIQ	$\{2e-6, 5e-6, 6e-6, 1e-5, 2e-5\}$	$\{0.01, 0.02, 0.03, 0.05\}$
TG-CIR	CIRR	$\{2e-6, 5e-6, 6e-6, 1e-5, 2e-5\}$	$\{0.01, 0.02, 0.03, 0.05\}$
CLIP4CIR	FashionIQ	$\{2e-6, 5e-6, 6e-6, 1e-5, 2e-5\}$	$\{0.01, 0.02, 0.03, 0.05\}$
CLIP4CIR	CIRR	$\{2e-6, 5e-6, 6e-6, 1e-5, 2e-5\}$	$\{0.01, 0.02, 0.03, 0.05\}$
BLIP4CIR	FashionIQ	$\{2e-6, 5e-6, 6e-6, 1e-5, 2e-5\}$	$\{0.01, 0.02, 0.03, 0.05\}$
BLIP4CIR	CIRR	$\{2e-6, 5e-6, 6e-6, 1e-5, 2e-5\}$	$\{0.01, 0.02, 0.03, 0.05\}$
SPRC	FashionIQ	$\{2e-6, 5e-6, 6e-6, 1e-5, 2e-5\}$	$\{0.01, 0.02, 0.03, 0.05\}$
SPRC	CIRR	$\{2e-6, 5e-6, 6e-6, 1e-5, 2e-5\}$	$\{0.01, 0.02, 0.03, 0.05\}$

Table 1: Hyper-parameters used for the second stage of all datasets and models in main results.

2 BASELINE DETAILS

TIRG [7] is the first CIR model proposed that uses gating and residual connection to retain important semantics of reference images.

CIRPLANT [6] learns modified representations of the reference image conditioned on the modified text and aligns the modified representation with the target image.

DCNET [4] proposes a correction network where the difference representation between the reference image and the target image is aligned with the sentence representation.

CoSMo [5] introduces the content-style modulation that independently modulates the content and style of a reference image according to the modified text.

MAAF [3] employs the dot-product attention to fuse the reference image representation and the modified text representation.

ARTEMIS [2] uses an explicit matching module to assess the fitness between the sentence and the target image and an implicit similarity module to match the queries with possible target images.

ComqueryFormer [8] establishes a unified Transformer structure to encode the cross-modal inputs and a global-local alignment module to reduce the distance between the query and the target image.

AMC [10] develops an adaptive multi-expert network that activates multiple experts with different levels of image-text interaction.

PL4CIR [9] leverages an extra fashion image-text dataset to enhance the encoding ability of fashion-style examples.

CLIP4CIR2 [1] fine-tune both the image encoder and the text encoder in the first stage. And fine-tune the combiner in the second

Table 2: Comparison with More Baselines.

Model	FashionIQ		CIRR		
	R@10	R@50	R@1	R@5	R _{subset} @1
Combiner [1]	43.82	66.98	45.35	78.43	72.95
+SPN [1]	44.10	67.10	45.64	78.45	74.10

stage. As shown in Table 2, SPN also improves the performance of Combiner in both FashionIQ and CIRR datasets. Due to space constraints, we do not put this result in the paper.

3 DISCUSSION ON DIFFERENT COMBINATIONS OF NEGATIVE EXAMPLE TYPES

We explore all possible combinations of four types of negative examples and find that using negative examples obtained by replacing the target image produces good results and is straightforward. There is no improvement in performance; in fact, it even leads to a decrease in performance and incurs additional computational overhead.

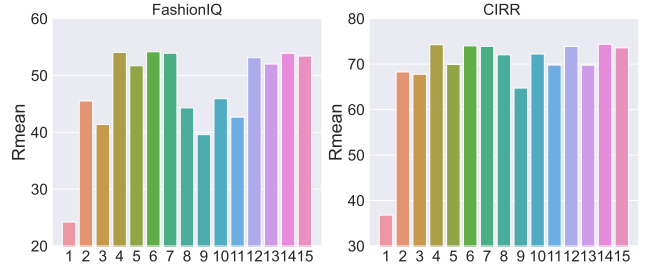


Figure 1: Performance of Utilizing Different Types of Negative Examples in Contrastive Learning. Negative example type numbers range from 1 to 15, with each number having four binary digits. The four binary digits from 1 to 4 respectively represent replacing the reference image, the modified text, the target image, and the query pair, representing four types of negative examples. For example, 10 (binary 1010) represents using negative examples by replacing query pairs and replacing modified text. We find that example type 4, the negative examples obtained by replacing target images, works best.

4 DISCUSSION ON MODIFIED TEXT GENERATION

As discussed in the main text, we have three ways of generating modified text from captions of image pairs. We find that using LLM to rewrite modified text simply is better than in-context learning. However, as shown in Fig.2, both methods of using LLM are worse than using the prompt template directly.

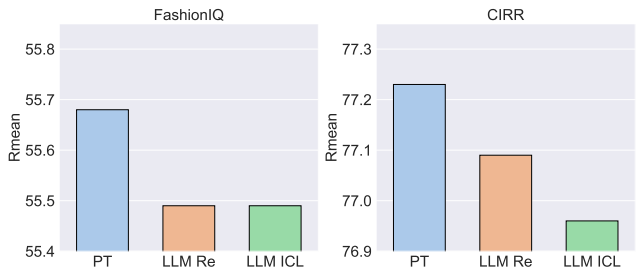


Figure 2: Performance on Different Ways of Modified Text Generation.

4.2 ICL Prompt for FashionIQ

You are a researcher tasked with rewriting the source sentence to mimic the target sentence while trying to keep the original meaning. Please ensure that your responses are close to the style of the target sentences in the examples. Remember to only output the new sentence without other additional words. Output the answer in one string

source: The dress is a sleeveless, black, fitted, and stylish dress
target: is solid black with no sleeves

source: Red, flowy, short, sequined, and elegant.
target: is red and flowy

source: Obama Mama shirt, black color.
target: has the words Obama Mama on front

source: Striped, black and white, sleeveless, fitted, and stylish.
target: has sleeveless black and white stripes

source: Colorful striped top with a v-neck.
target: Has stripes.

source: {simple modified text}
target:

4.1 Rewrite Prompt

Rewrite the sentence to maintain the original meaning while reducing grammatical errors and increasing the variety of expressions. Remember to only output the new sentence without other additional words.

sentence: {simple modified text}
new sentence:

4.3 ICL Prompt for CIRR

You are a researcher tasked with rewriting the source sentence to mimic the target sentence while trying to keep the original meaning. Please ensure that your responses are close to the style of the target sentences in the examples. Remember to only output the new sentence without other additional words. Output the answer in one string

source: A large, brown dog with a black nose is sitting on the grass, looking up instead of A cute baby panda is being held by a person in a zoo

target: Dog in grass instead of a panda.

source: A street with several blue buildings, including churches, and a park with trees and bushes instead of A large, old stone church with a tower and a wall, surrounded by a grassy field and a dirt road

target: instead of an old fortress with a rampart, an Orthodox church with a courtyard.

source: A colorful parrot is standing on a perch in a cage instead of Two parrots are sitting on a branch, sharing a piece of fruit

target: Remove one of the parrots.

source: Two colorful parrots are kissing on a branch instead of A colorful parrot is perched on a tree branch, looking at the camera.

target: two birds, facing each other.

source: A monkey is standing on a grassy field, looking at the camera instead of A group of monkeys is sitting on the ground, with some of them touching each other.

target: I want the pic to show just one monkey.

source: {simple modified text}

target:

[4] Jongseok Kim, Youngjae Yu, Hoesong Kim, and Gunhee Kim. 2021. Dual Compositional Learning in Interactive Image Retrieval. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 1771–1779. <https://doi.org/10.1609/AAAI.V35I2.16271>

[5] Seungmin Lee, Dongwan Kim, and Bohyung Han. 2021. CoSMo: Content-Style Modulation for Image Retrieval With Text Feedback. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 802–812. <https://doi.org/10.1109/CVPR46437.2021.00086>

[6] Zheyuan Liu, Cristian Rodriguez Opazo, Damien Teney, and Stephen Gould. 2021. Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2105–2114. <https://doi.org/10.1109/ICCV48922.2021.00213>

[7] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing Text and Image for Image Retrieval - an Empirical Odyssey. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 6439–6448. <https://doi.org/10.1109/CVPR.2019.00660>

[8] Yahui Xu, Yi Bin, Jiwei Wei, Yang Yang, Guoqing Wang, and Heng Tao Shen. 2023. Multi-Modal Transformer With Global-Local Alignment for Composed Query Image Retrieval. *IEEE Trans. Multim.* 25 (2023), 8346–8357. <https://doi.org/10.1109/TMM.2023.3235495>

[9] Yida Zhao, Yuqing Song, and Qin Jin. 2022. Progressive Learning for Image Retrieval with Hybrid-Modality Queries. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 1012–1021. <https://doi.org/10.1145/3477495.3532047>

[10] Hongguang Zhu, Yunchao Wei, Yao Zhao, Chunjie Zhang, and Shujuan Huang. 2023. AMC: Adaptive Multi-expert Collaborative Network for Text-guided Image Retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 6, Article 188 (may 2023), 22 pages. <https://doi.org/10.1145/3584703>

5 CASE STUDY

REFERENCES

[1] Alberto Baldradi, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2024. Composed Image Retrieval using Contrastive Learning and Task-oriented CLIP-based Features. *ACM Trans. Multim. Comput. Commun. Appl.* 20, 3 (2024), 62:1–62:24. <https://doi.org/10.1145/3617597>

[2] Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. 2022. ARTEMIS: Attention-based Retrieval with Text-Explicit Matching and Implicit Similarity. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. <https://openreview.net/forum?id=CVfLvQq9gLo>

[3] Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. 2020. Modality-Agnostic Attention Fusion for visual search with text feedback. *CoRR*

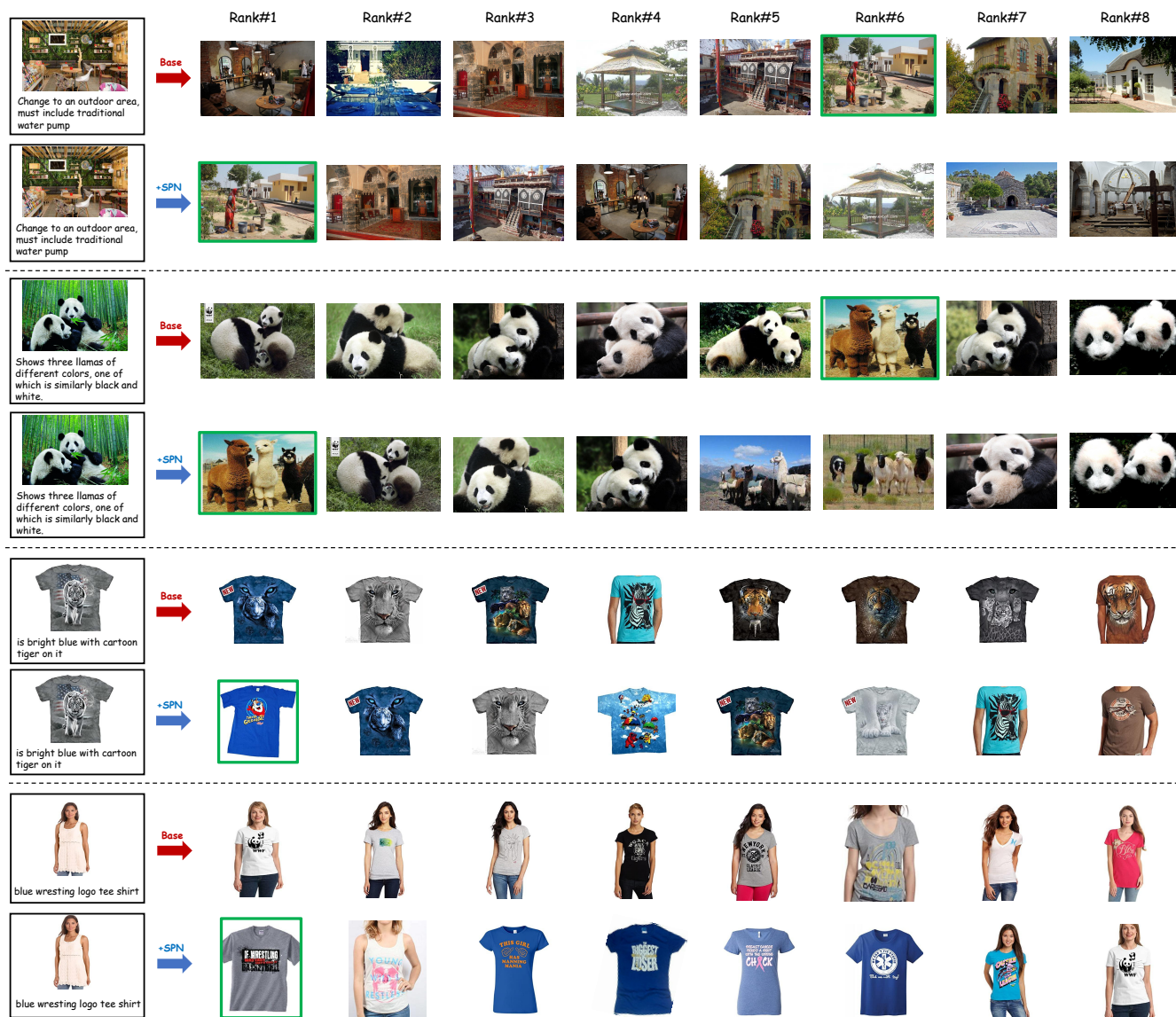


Figure 3: Comparison of retrieval results between the CLIP4CIR model w/o and w SPN.