# Language-only Efficient Prompt Learning for Zero-shot Composed Image Retrieval

Seongwon Lee, Yong-Ju Lee*

Electronics and Telecommunications Research Institute (ETRI), South Korea
{sungonce,yongju}@etri.re.kr

**Abstract.** Composed image retrieval is the task of finding images that align with a given query image and target text modification. This task has been approached using zero-shot methodologies and language-only training paradigms to mitigate the high costs associated with collecting extensive image-text triplet datasets and the computational demands of training. However, existing approaches still under-utilize their potential in terms of adaptability and expressiveness due to their reliance on static, pre-defined prompts. To overcome this limitation, we introduce a novel approach called **L**anguage-**o**nly **Pro**mpt Learning (**LoPro**). LoPro advances the concept of language-only zero-shot learning in composed image retrieval by dynamically learning sentence prompts with text-only supervision. This enables LoPro to inherit the benefits of language-only training and significantly improve its expressiveness and findability.

## 1 Introduction

Image retrieval [3, 8, 9, 16], a fundamental task in computer vision, focuses on identifying and retrieving images from large databases that match a given query. An advanced variant of this task is composed image retrieval [2,4,7,10,14], where the goal is to find images that are not only visually similar to a query image but also align with specific textual modifications. This challenge is exacerbated by the high cost and effort required to collect and annotate large-scale image-text datasets necessary for training such models.

Recent advancements in the field have seen a shift towards zero-shot (ZS) learning methodologies, which aim to eliminate the dependency on vast, labeled datasets. Within this paradigm, language-only training approaches have emerged as a beacon of efficiency, proposing a model training strategy that leverages textual data alone, without the necessity for corresponding visual inputs. However, language-only zero-shot learning could not fully exploit its adaptability and expressiveness, because of their reliance on static, predefined prompts. These limitations hinder the potential of zero-shot learning, as the fixed, hand-crafted prompts can lead to sub-optimal retrieval performance.

To address this limitation, in this work, we propose **L**anguage-**o**nly **Pro**mpt Learning (**LoPro**). LoPro is designed to dynamically adapt and learn effective CIR prompts using text-only supervision, thereby enhancing the model's

* Corresponding Author.

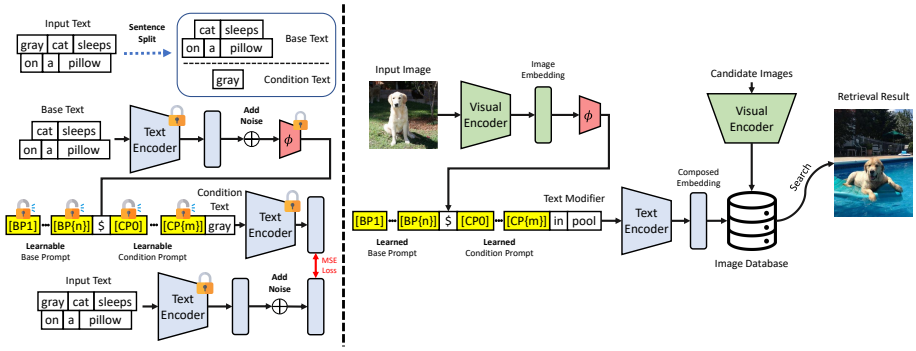


**Fig. 1: Overview of Language-only Prompt Learning (LoPro). Left:** Overview of training of LoPro. The learnable base prompts and learnable condition prompts are optimized with only single-sentence text data. **Right:** LoPro uses an elevated prompt in the inference stage to improve accuracy, which is different from previous methods.

expressiveness and its ability to interpret and respond to a diverse range of textual modifications in the context of composed image retrieval. By leveraging the inherent flexibility of language-only training, LoPro not only inherits the benefits of reduced dependency on extensive labeled datasets but also pushes the boundaries of what is achievable in composed image retrieval, particularly in a zero-shot learning framework.

## 2 LoPro: Language-only Prompt Learning

LinCIR, the pioneering language-only zero-shot composed image retrieval approach, has significantly reduced the need for expensive datasets and opened new avenues for efficient and scalable learning. However, despite their promising results, the full potential of language-only training paradigms remains untapped. One major limitation is the reliance on hand-crafted, pre-defined prompts (*e.g.* "a photo of", "that"), which restricts the model's ability to adapt to varied and complex query modifications. In this section, we introduce a novel approach, Language-only Prompt Learning (LoPro), to overcome these limitations.

### 2.1 Text Splitting Projection (TSP)

We introduce a novel text-only self-supervision, named Text Splitting Projection (TSP) for language-only training. The key to this approach is splitting the input text into base text and condition text to construct a text-only triplet. This allows the base text and input text to act as stand-ins for the base image and target image, respectively, enabling effective text-only training. In detail, the first step involves dissecting the input text $x$ into two distinct parts: the base text $x_b$ and the condition text $x_c$. This is achieved by extracting keywords from the input text and then randomly selecting a subset of these keywords to serve as the conditional text $x_c$. The remaining text is designated as the base text $x_b$. In this context, we consider the consecutive nouns and adjectives that appear in the sentence as the keywords. For example, in the "gray cat sleeps on a pillow"

sentence, the keywords will be "gray", "cat", and "pillow". Each keyword is individually and probabilistically included in the condition text. For example, if the keyword gray passed the probability condition, "cat sleeps on a pillow" would be the base text, and "gray" would be the condition text.

### 2.2 Language-Only Prompt Learning

After splitting, we extract the base text embedding $z_b$ and project it to projected base text embedding $e_b$ with the projection module $\phi$. Then, The base text passes through a text encoder and is then projected into tokens that represent its essential information. Then, we get a composed caption by concatenating the base prompt $[BP1]$ to $[BP\{n\}]$, the projected base text embedding $e_b$, the condition prompt $[CP1]$ to $[CP\{m\}]$, the condition text token embedding $e_c$ of a given text $x_c$. Finally, using the composed caption, we can extract a composed latent feature $\hat{z}$ and minimize the MSE loss between the original latent feature $\hat{z}$ of a given input text $x$. Here, we only train the learnable prompts, base prompts $[BP1]$ to $[BP\{n\}]$ and condition prompt $[CP1]$ to $[CP\{m\}]$, while keeping the text encoder, visual encoder and projection module $\phi$ frozen. Additionally, following LinCIR, we adapt noise addition to mitigate the modality gap between vision and language.

## 3 Experiments

### 3.1 Implementation Details

We use official CLIP ViT-L [12] as visual and text encoders. LinCIR does not provide an official checkpoint, so we reproduce the LinCIR with the official code and use it as our projection module $\phi$. As a result, in this paper, we denote the reproduced LinCIR as LinCIR$^\dagger$, and LoPro is based on LinCIR$^\dagger$. We set $n = 3$ and $m = 1$, which is computationally the same prompt length as Pic2Word or LinCIR, which utilize fixed base prompt *"a photo of"* and condition prompt *"that"*. We use a learning rate of 1e-4 and weight decay of 1e-2 with the AdamW [11] optimizer. The total batch size is 512, and we apply dropout with a probability of 0.5. We use CompoDiff [5] captions for prompt training. For a fair comparison, we select the best model based on the CIRR [7] dev R@1 score, following the other early papers. We also adopt the early stopping strategy following LinCIR. Keywords are selected by the POS tagger of the spacy library. Each keyword in the input text has a probability of 0.2 of being included in the condition text. The condition text can also be blank. If the condition text contains more than one keyword, it is followed by an "and".

### 3.2 Experimental Results

Following LinCIR, we adopt CIRCO [1] as the main benchmark and FashionIQ [15] as the sub-benchmark. All models and comparisons are based on ViT-L backbone.

*Results on Evaluation Benchmarks (Tab. 1a, Tab. 1b).* Our proposed LoPro outperforms the existing state-of-the-art ZS-CIR methods (Pic2Word [13], SEARLE [1], LinCIR [6]), showing its promising possibilities.

| | mAP@5 | mAP@10 | mAP@25 | mAP@50 |
|---|---|---|---|---|
| Pic2Word | 8.72 | 9.51 | 10.64 | 11.29 |
| SEARLE | 11.68 | 12.73 | 14.33 | 15.12 |
| LinCIR | 12.59 | 13.58 | 15.00 | 15.85 |
| LinCIR† | 12.42 | 13.48 | 14.98 | 15.87 |
| LoPro (Ours) | **13.25** | **14.28** | **15.99** | **16.84** |

**(a) CIRCO evaluation.**

| | Shirt | | Dress | | Toptee | | Average | |
|---|---|---|---|---|---|---|---|---|
| | R@10 | R@50 | R@10 | R@50 | R@10 | R@50 | R@10 | R@50 |
| Pic2Word | 26.20 | 43.60 | 20.00 | 40.20 | 27.90 | 47.40 | 24.70 | 43.70 |
| SEARLE | 26.89 | 45.58 | 20.48 | 43.13 | 29.32 | 49.97 | 25.56 | 46.23 |
| LinCIR | 29.10 | 46.81 | 20.92 | 42.44 | 28.81 | 50.18 | 26.28 | 46.49 |
| LinCIR† | 29.69 | 46.96 | 20.72 | 42.98 | 28.96 | 49.62 | 26.46 | 46.52 |
| LoPro (Ours) | **31.75** | **49.21** | **22.21** | **44.87** | **30.55** | **51.91** | **28.17** | **48.67** |

**(b) FashionIQ evaluation.**

**Table 1: Evaluation on Composed Image Retrieval Benchmark.**

| | Training Time (h) | | | Inference Time (s) | CIRCO | FashionIQ | Training GPUs |
|---|---|---|---|---|---|---|---|
| | Stage 1 | Stage 2 | Total | | mAP@25 | R@50 | |
| Pic2Word | 3.0 | - | 3.0 | 0.02 | 10.64 | 43.70 | A100 x 8 |
| SEARLE | 1.7 | 2.5 | 4.2 | 0.02 | 14.33 | 46.23 | A100 x 8 |
| LinCIR | 0.5 | - | **0.5** | 0.02 | 15.00 | 46.49 | A100 x 8 |
| LoPro (Ours) | 0.5 | 0.1 | 0.6 | 0.02 | **15.99** | **48.67** | A100 x 8 |
| CompoDiff | 82 | 41 | 123 | 0.12 | 15.83 | 48.64 | A100 x 128 |

**Table 2: Comparison with Zero-Shot Composed Image Retrieval methods.**

| | Base Prompt | Condition Prompt | CIRCO | | | | FashionIQ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | mAP@5 | mAP@10 | mAP@25 | mAP@50 | Shirt R@10 | Shirt R@50 | Dress R@10 | Dress R@50 | Toptee R@10 | Toptee R@50 |
| LinCIR† | a photo of | that | 12.42 | 13.48 | 14.98 | 15.87 | 29.69 | 46.96 | 20.72 | 42.98 | 28.96 | 49.62 |
| | | which | 11.90 | 12.88 | 14.35 | 15.21 | 29.74 | 47.20 | 20.58 | 41.94 | 28.81 | 49.11 |
| | observe | that | 11.97 | 13.00 | 14.41 | 15.19 | 30.18 | 48.87 | **23.10** | **45.31** | 30.24 | **52.28** |
| | | which | 12.73 | 13.60 | 15.13 | 15.90 | 30.37 | 48.87 | 21.96 | 44.32 | 29.53 | 50.89 |
| | retrieve | that | 11.44 | 12.30 | 13.46 | 14.2 | 29.24 | 46.37 | 22.26 | 43.98 | 29.02 | 50.13 |
| | | which | 10.85 | 11.57 | 12.70 | 13.42 | 28.56 | 46.32 | 21.17 | 43.58 | 28.51 | 48.39 |
| | search for | that | 12.21 | 13.18 | 14.57 | 15.34 | 29.49 | 47.15 | 21.91 | 43.98 | 29.47 | 49.72 |
| | | which | 11.96 | 12.70 | 14.15 | 14.89 | 29.34 | 47.20 | 21.86 | 44.27 | 29.37 | 49.92 |
| LoPro (Ours) | Learnable | Learnable | **13.25** | **14.28** | **15.99** | **16.84** | **31.75** | **49.21** | 22.21 | 44.87 | **30.55** | 51.91 |

**Table 3: Comparison with various pre-defined prompts.**

*Comparison with state-of-the-art Zero-Shot Composed Image Retrieval methods. (Tab. 2).* In comparison with state-of-the-art Zero-Shot Composed Image Retrieval methods, as shown in Table 2, our proposed method is evaluated in terms of training times and performance metrics against existing approaches for composed image retrieval tasks. The table underscores our method's reduced training time and superior performance even over the supervised approach CompoDiff, illustrating the effectiveness of our finely tuned prompts and the benefits derived from language-only training.

*Comparison with various pre-defined prompts (Tab. 3).* In comparison to various pre-defined prompts, as detailed in Table 3, it is evident that the learnable prompt approach surpasses the traditionally used hand-crafted prompts in terms of both performance and stability, achieving significantly better results.

## 4 Conclusion

In this work, we introduced a new approach called Language-only Prompt Learning (LoPro) that can improve the expressiveness and searchability of composed image retrieval by using language-only training with zero-shot learning. This method enables us to learn prompts using text-only supervision, eliminating the need for extensive image-text datasets and showing superior adaptability and performance compared to existing fixed-prompt methods.

## 5 Acknowledgements

This study was supported by the following grant: the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration)

## References

1. Baldrati, A., Agnolucci, L., Bertini, M., Del Bimbo, A.: Zero-shot composed image retrieval with textual inversion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15338–15347 (2023) 3
2. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: Effective conditioned and composed image retrieval combining clip-based features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21466–21474 (2022) 1
3. Cao, B., Araujo, A., Sim, J.: Unifying deep local and global features for image search. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 726–743. Springer (2020) 1
4. Chen, Y., Gong, S., Bazzani, L.: Image search with text feedback by visiolinguistic attention learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3001–3011 (2020) 1
5. Gu, G., Chun, S., Kim, W., Jun, H., Kang, Y., Yun, S.: Compodiff: Versatile composed image retrieval with latent diffusion. arXiv preprint arXiv:2303.11916 (2023) 3
6. Gu, G., Chun, S., Kim, W., Kang, Y., Yun, S.: Language-only efficient training of zero-shot composed image retrieval. arXiv preprint arXiv:2312.01998 (2023) 3
7. Hosseinzadeh, M., Wang, Y.: Composed query image retrieval using locally bounded features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3596–3605 (2020) 1, 3
8. Lee, S., Lee, S., Seong, H., Kim, E.: Revisiting self-similarity: Structural embedding for image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23412–23421 (2023) 1
9. Lee, S., Seong, H., Lee, S., Kim, E.: Correlation verification for image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5374–5384 (June 2022) 1
10. Lee, S., Kim, D., Han, B.: Cosmo: Content-style modulation for image retrieval with text feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 802–812 (2021) 1
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 3
12. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 3
13. Saito, K., Sohn, K., Zhang, X., Li, C.L., Lee, C.Y., Saenko, K., Pfister, T.: Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19305–19314 (2023) 3

14. Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.J., Fei-Fei, L., Hays, J.: Composing text and image for image retrieval-an empirical odyssey. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6439–6448 (2019) 1
15. Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., Feris, R.: 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. pp. 11307–11317 (2021) 3
16. Yang, M., He, D., Fan, M., Shi, B., Xue, X., Li, F., Ding, E., Huang, J.: Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11772–11781 (2021) 1