
Reducing Task Discrepancy of Text Encoders for Zero-Shot Composed Image Retrieval

Jaeseok Byun^{1*} Seokhyeon Jeong^{1*} Wonjae Kim² Sanghyuk Chun^{2†} Taesup Moon^{1,3†}

¹Department of ECE, Seoul National University ²NAVER AI Lab

³Department of ASRI/INMC/IPAI/AIIS, Seoul National University

Abstract

Composed Image Retrieval (CIR) aims to retrieve a target image based on a reference image and conditioning text, enabling controllable searches. Due to the expensive dataset construction cost for CIR triplets, a zero-shot (ZS) CIR setting has been actively studied to eliminate the need for human-collected triplet datasets. The mainstream of ZS-CIR employs an efficient projection module that projects a CLIP image embedding to the CLIP text token embedding space, while fixing the CLIP encoders. Using the projected image embedding, these methods generate image-text composed features by using the pre-trained text encoder. However, their CLIP image and text encoders suffer from the task discrepancy between the pre-training task (text \leftrightarrow image) and the target CIR task (image + text \leftrightarrow image). Conceptually, we need expensive triplet samples to reduce the discrepancy, but we use cheap *text* triplets instead and update the text encoder. To that end, we introduce the Reducing Task Discrepancy of text encoders for Composed Image Retrieval (**RTD**), a plug-and-play training scheme for the text encoder that enhances its capability using a novel target-anchored text contrastive learning. We also propose two additional techniques to improve the proposed learning scheme: a hard negatives-based refined batch sampling strategy and a sophisticated concatenation scheme. Integrating RTD into the state-of-the-art projection-based ZS-CIR methods significantly improves performance across various datasets and backbones, demonstrating its efficiency and generalizability.

1 Introduction

Composed Image Retrieval (CIR) is an emerging task aimed at retrieving a target image that closely resembles a reference image while reflecting the changes described in a conditioning text [1]. Using a query composed of image and text allows users to conduct more precise and flexible searches by specifying the desired modifications to the image through text. Supervised CIR methods [2–5] have been introduced to fuse the information from the bi-modal query, using labeled data in the form of triplets (I_r, T_c, I_t) , in which I_r is a reference image, T_c is a conditioning text, and I_t is a target image. However, unlike typical web-crawled image-text datasets [6, 7], acquiring sufficient triplets for training needs expensive manual human annotations. Hence, the existing CIR triplet datasets are typically small, limiting the zero-shot ability of supervised approaches trained on such datasets.

To overcome the dependency on small-scale human-verified triplet datasets, a new task, Zero-Shot Composed Image Retrieval (ZS-CIR), has been recently introduced. The first approach solves this novel task by utilizing the power of recent vision-language (VL) generative models. For example,

*Equal contribution

†Corresponding authors

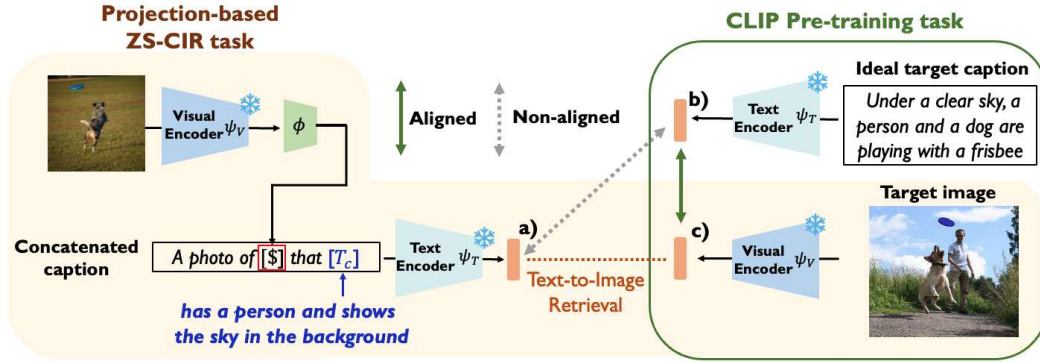


Figure 1: The task discrepancy of projection-based ZS-CIR methods [15–17] between the pre-training task (image-text alignment) and the ZS-CIR task (image-text composition).

a line of studies [8–11] uses text-to-image models like IP2P [12] to synthesize large-scale CIR triplets for supervised model training. Another example can be found in [13], which eliminates the need of training by using image captioning models and large-language models (LLM) during inference. While these approaches achieve decent performance, they are impractical due to their high computational and memory requirements for utilizing generative models. The second approach for removing the dependency on the triplet datasets, which has become the mainstream due to its simplicity, employs an integrable projection module on top of the pre-trained, frozen, and shared VL embedding space, such as CLIP [14]. Namely, a projection module ϕ to map a CLIP image embedding to the CLIP text embedding token space can be trained by solely using images [15, 16] or texts [17]. Then, as we illustrated in Figure 1, these methods first project the embedding of the query image to a special text token embedding [\$] using ϕ . Then, it is combined with the conditioning text $[T_c]$ to form a prompt “a photo of [\$] that $[T_c]$ ”. Finally, the combined prompt is used as a query for the text-to-image retrieval which utilizes the pre-trained VL embedding space.

The core assumption of the second approach, which often is referred to as projection-based ZS-CIR [15–17], is that the pre-trained text encoder should be robust enough to combine information from both the projected text token embedding and the conditioning text. However, we argue that this can cause significant *task discrepancy* for the pre-trained text encoder between the image-text alignment pre-training task and the ZS-CIR task. For example, in Figure 1, if we assume there is an *ideal caption* that accurately describes the target image, then the target image embedding (Fig. 1c) will align well with the embedding (Fig. 1b) of that caption due to the contrastive learning of the text and image encoders of CLIP. In contrast, in the ZS-CIR task, the text encoder instead receives a concatenated caption that combines the projected text token [\$] and the conditioning text. However, the text encoder typically is not trained to encode complex textual modifications—such as addition, negation, comparison, and spatial relationships—to the reference image, which are common in the conditioning text. As a result, there is no guarantee that the latent textual embedding of the concatenated caption (Fig. 1a) closely aligns with that of the target image embedding (Fig. 1c).

To that end, in this paper, we aim to reduce the task discrepancy of the text encoder only with cheap *text* triplets. The triplets (T_r, T_c, T_t) can be automatically generated without human labor [1, 18] and intensive resources [8–11], where T_r is the reference caption and T_t is a target caption. Using these triplets, we devise a *target-anchored text contrastive learning*, which trains the text encoder to update the embeddings of the concatenated caption T_{r+c} (formed by simple concatenation of reference caption T_r and conditioning text T_c) to align closely with the fixed embedding of the target caption T_t , which serves as an anchor point obtained from the frozen text encoder. We also propose two techniques to enhance the effectiveness of such language-only supervision further: a batch sampling strategy that incorporates hard negatives in each mini-batch and a refined concatenation scheme for T_r and T_c to reduce the training-inference task discrepancy. We note our approach can be seamlessly integrated with existing projection-based ZS-CIR methods [15–17] by replacing their text encoder with our updated text encoder while fixing other modules, *e.g.*, the image encoder and ϕ . Moreover, due to the benefits of language-only training, as highlighted by [17], our approach is not only efficient in the dataset generation process, but also in the training process.

Our experimental results demonstrate that our proposed method, dubbed as **RTD** (Reducing Task Discrepancy of text encoders for Composed Image Retrieval), substantially improves the ZS-CIR performance in diverse evaluation datasets (CIRR [1], CIRCO [16], FashionIQ [18], and GeneCIS [19]). Namely, when integrated into the existing projection-based ZS-CIR methods (SEARLE [16], Pic2Word [15], and LinCIR [17]), RTD consistently enhances performance across two different size backbones (ViT-B/32 and ViT-L/14), underscoring the generality of our approach. We even observe that the integrating RTD into LinCIR/SEARLE (ViT-B/32) has a more significant effect than switching to a larger backbone (ViT-B/32 \rightarrow ViT-L/14). Our systematic ablation analyses reveal that such performance enhancement primarily results from reducing the task discrepancy of the text encoder, rather than merely tuning the textual backbone network with additional data. Instead of updating all parameters of the text encoder, we also show that a more efficient approach, which selectively updates only a few layers of the text encoder, can be effective as well. Moreover, we verify that using template-based text triplets (without LLMs) is sufficient to achieve strong competitive ZS-CIR performance, highlighting the low data acquisition cost of our approach.

2 Related Work

Projection-based ZS-CIR. Projection-based CIR methods, such as Pic2Word [15], SEARLE [16] and LinCIR [17], are built upon the frozen CLIP [14] model, which includes a visual encoder ψ_V and a text encoder ψ_T . These methods first project the latent image embedding $v = \psi_V(I)$, where I is an input image, to the token embedding space using a projection module ϕ . Assuming the projected token embedding as a special token [\$], these methods predict the composed embedding by encoding “a photo of [\$] that [T_c]”, where [T_c] is the given conditioning text. See Figure 1 for the details of the inference stage of projection-based CIR methods. Each projection-based CIR methods employ a different training scheme for ϕ . We will explain their differences in Section 4.1. Although they show promising ZS-CIR performances, these methods rely on the pre-trained CLIP visual and text encoders. However, in practice, the target CIR task is different from the pre-training task of CLIP. In this paper, we aim to reduce the task discrepancy between the CLIP pretext task and the CIR task using an efficient language-only training scheme.

Previous attempts to reduce the task discrepancy between the CLIP pretext task and CIR. Combiner [20] additionally updates the text encoder to minimize the gap between the target caption feature and the summation of the reference image feature and the instruction text feature. However, this approach needs a number of expensive CIR triplets (I_r, T_c, I_t) for training. Our approach uses text-only triplets (T_r, T_c, T_t), cheap and automatically generated. As another example, Chen and Lai [21] synthesizes a triplet of an original image, the corresponding caption, and the masked image, where treating the original image as the target image, the caption as the conditioning text, and the masked image as the reference image. This approach, however, still has a gap between conditioning text (e.g., “change the dog to a cat”) and image caption (e.g., “a dog is jumping to catch a frisbee”); furthermore, it needs the full fine-tuning of the CLIP model, resulting in changing the visual embeddings in the retrieval database. On the other hand, RTD directly uses the instruction texts for training and does not change the target visual encoder, which enables the reuse of pre-extracted CLIP visual embeddings. Lastly, CIReVL [13] reduces the task discrepancy by making a descriptive caption of the composed query using a large captioning model and LLM. Although CIReVL shows great performance without any training, this method needs inefficient and expensive inferences of BLIP [22, 23] and GPT [24]. Furthermore, it needs a well-tuned task-specific prompt by a skilled user. RTD is much more efficient than CIReVL and fully automated without direct human intervention.

3 Main Method

In this section, we describe our main method, RTD: we first present the intuition and details of the generation of text triplets, then present the learning framework using them.

3.1 Generation of text triplets

As shown in Figure 1, the projection-based ZS-CIR methods may suffer from significant task discrepancy between the CLIP pre-training task and the target CIR task. Instead of directly using

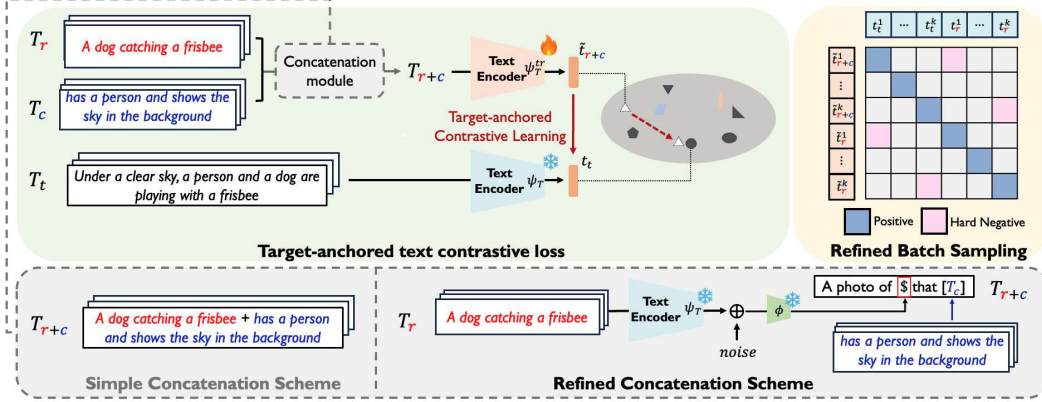


Figure 2: Overview of RTD.

expensive CIR triplets (I_r, T_c, I_t) , we aim to generate and employ text triplets (T_r, T_c, T_t) , which can be cheaply and automatically generated by LLM [8, 12] or rule-based templates [8], to resolve the task discrepancy. The LLM-based text triplet generation process uses a caption T_r as an input of the fine-tuned LLM, whose output predicts the corresponding conditioning text T_c and the target caption T_t . The template-based text triplet generation extracts keywords based on the pre-defined rule (e.g., noun) from the given caption T_r and randomly alters the keyword from the pre-extracted keyword sets to generate a target caption T_t . Then, the conditioning text T_c is automatically generated from the pre-defined templates (e.g., “change [original keyword] to [altered keyword]”). Our experiment shows that the fine-tuned LLM is not necessary for achieving strong performances; template-based triplets are sufficient. Note that unlike CompoDiff [8] and IP2P [12], which utilize the text generation step as a preliminary phase for subsequent text-to-image generation, we only require the text generation step. Detailed explanations and examples of text triplets are in the Appendix A.2.

3.2 Target-anchored text contrastive learning

Now, we explain our approach to update the text encoder for mitigating the task discrepancy solely with the generated text triplets (T_r, T_c, T_t) . We first assume that there exists a pre-trained projection module ϕ obtained by the projection-based ZS-CIR methods [15–17]. Recall that for a given reference image I_r and conditioning text T_c , the final composed feature is generated by passing the text prompt “a photo of $\phi(\psi_V(I_r))$ that T_c ” to the text encoder ψ_T , where ψ_V is the visual encoder and ϕ is the projection module (See Figure 1). We aim to update the text encoder ψ_T to reduce the discrepancy between the pretext task and ZS-CIR task using the text triplets while maintaining ψ_V and ϕ frozen.

[Target-anchored text contrastive loss] We apply contrastive learning using a paired caption (T_{r+c}, T_t) , where T_{r+c} denotes a concatenated caption of reference caption T_r and relative caption T_c . Namely, we let the representation of the concatenated caption closely approximate that of the target caption. However, solely updating the text encoder while fixing the image encoder can break the alignment between image and text encoders. To prevent the issue, we extract the text embedding of T_t using the frozen text encoder ψ_T , while the concatenated caption T_{r+c} is extracted from the learnable text encoder ψ_T^{tr} , initialized from ψ_T . Here, we assume that as the target caption T_t is a standard caption, a text embedding $\psi_T(T_t)$, is well-aligned with the frozen image embedding space. Following the assumption, we fix the target textual embedding to serve as an anchor point. This approach helps maintain the pre-trained alignment while learning new relationships. As shown in Section 4.4, anchoring the target textual embedding is essential for fine-tuning the text encoder with our objective.

Now, we define our target-anchored text contrastive loss \mathcal{L}_{TCL} using two text encoders: a frozen pre-trained text encoder ψ_T and a learnable text encoder ψ_T^{tr} which is initialized with ψ_T . Textual latent embeddings \tilde{t}_{r+c} and t_t are extracted from ψ_T^{tr} and ψ_T , respectively. Namely, $\tilde{t}_{r+c} = \psi_T^{tr}(E_w^{tr}(T_{r+c}))$ and $t_t = \psi_T(E_w(T_t))$, where E_w is a word embedding layer. We wish to tune ψ_T^{tr} to minimize the distance between the concatenated textual embedding \tilde{t}_{r+c} and the target tex-

tual embedding t_t while maximizing the distance from other textual embeddings within the batch. Therefore, we employ a symmetric InfoNCE loss [25, 26], which is defined as follows:

$$\mathcal{L}_{TCL} = \frac{1}{B} \sum_{k=1}^B -\log \frac{e^{(c(\tilde{i}_{r+c}^k, t_t^k)/\tau)}}{\sum_{j=1}^B e^{(c(\tilde{i}_{r+c}^k, t_t^j)/\tau)} + \sum_{j \neq k} e^{(c(t_t^k, \tilde{i}_{r+c}^j)/\tau)}} - \log \frac{e^{(c(t_t^k, \tilde{i}_{r+c}^k)/\tau)}}{\sum_{j=1}^B e^{(c(t_t^k, \tilde{i}_{r+c}^j)/\tau)} + \sum_{j \neq k} e^{(c(\tilde{i}_{r+c}^k, \tilde{i}_{r+c}^j)/\tau)}} \quad (1)$$

where $c(\cdot, \cdot)$ denotes the cosine similarity, B is the batch size, and τ is a temperature hyperparameter.

[Refined batch sampling strategy for hard negatives] To further enhance the efficacy of updating the text encoder, we devise a simple yet effective batch sampling strategy that incorporates pairs of (T_{r+c}, T_t) and (T_r, T_r) within the same batch. Namely, the concatenated caption T_{r+c} and its corresponding reference caption T_r are sampled concurrently. Since all other captions in the batch are considered as negatives in Eq. (1), the concatenated caption T_{r+c} and reference caption T_r can serve as hard negatives for one another. Moreover, we believe including (T_r, T_r) pairs in the contrastive learning helps the learnable text encoder ψ_T^{tr} remain closely aligned with the pre-trained encoder ψ_T .

[Refined concatenation of reference and conditioning texts] A naive concatenation strategy also can suffer from training-inference task discrepancy because we actually use “a photo of [\$] that [T_c]” for inference. To tackle this issue, rather than simply concatenating the T_r and T_c , we also use the prompt “a photo of [\$] that [T_c]” for updating the text encoder, where [\$] is obtained by the reference caption T_r with the projection module ϕ . Instead of obtaining a pseudo-word token with latent image embedding v , we utilize a textual latent embedding from the reference caption T_r , *i.e.*, $\phi(t_r)$. However, [17] showed that naively replacing the image encoder with the text encoder for the input of ϕ will suffer from the modality gap [27], a phenomenon where text and image embeddings have a gap between them. Thus, to reduce the potential negative effect of the modality gap, following LinCIR [17], we inject noise sampled from $\text{Unif}(0, 1) \times \mathcal{N}(0, 1)$ into the textual token representation before it is processed by ϕ .

Figure 2 illustrates the overview of our RTD. We use CLIP backbone and pre-trained projection module ϕ produced by the existing projection-based ZS-CIR methods [15–17]. The text encoder is trained using the proposed loss function (Eq. (1)) while applying the refined batch sampling and concatenation scheme. During inference, the procedure mirrors that of existing ZS-CIR methods, except we utilize the updated text encoder ψ_T^{tr} instead of the frozen one ψ_T . Note that our method only updates the text encoder while the image encoder and the projection module are frozen.

Remark. We note that our entire process, including text triplet construction and the training step, is efficient due to the advantages of language-only training as highlighted in [17]. First, for the text triplet construction, we only require an efficient text triplet generation process without the resource-intensive text-to-image generation phase [8, 12]. If we choose the template-based text triplet generation process, it becomes even more efficient by eliminating the need for the LLM generation step. Moreover, our generated text triplets occupy just 100MB, whereas storing a similar quantity of images requires significantly more space (*e.g.*, around 400GB in the case of CC3M [6]). Second, the training complexity for the text encoder is substantially lower than that for the visual encoder due to the relatively short token lengths of texts (~ 12) compared to images (256). Consequently, the average inference time of the CLIP ViT-L/14 image encoder is $\times 3.5$ times slower than that of the text encoder. In Section 4.4, we demonstrate a more efficient implementation option by selectively updating only a few layers of the text encoder. Namely, we verify that the size of the learnable parameters of the text encoder can be reduced to the same size as the parameters of the projection module ϕ while achieving similar results.

3.3 Can RTD really reduce the task discrepancy of the text encoder?

In this subsection, we quantitatively verify whether RTD really can reduce the task discrepancy. We first conduct a toy experiment that measures the text-to-image (T2I) retrieval performance of the text encoder with the modification instruction. We retrieve the target images

Table 1: T2I retrieval performance of different text encoders on CIRCO validation dataset.

Query	Text encoder	mAP@5	mAP@10	mAP@25
T_t	Frozen	18.96	19.31	21.05
T_{r+c}	Frozen	10.12	10.71	12.34
T_{r+c}	RTD	15.12	15.80	17.77

I_t with the concatenated text query T_{r+c} or the ideal target caption T_t . If our text encoder successfully handles the discrepancy due to the concatenated caption, the text encoder updated by RTD will perform better than the frozen one. We use the CLIP ViT-L/14 and CIRCO [16] validation dataset for evaluation. Since the CIRCO dataset only has CIR triplets (I_r, T_c, I_t) , we use the BLIP [22] captioner to generate T_r and T_t corresponding to the I_r and I_t , respectively. Here, the simple concatenation scheme is applied for the text query T_{r+c} in all cases for a fair comparison. Table 1 shows that when the text encoder is frozen, the retrieval results using the concatenated caption T_{r+c} are significantly worse than those using the target caption T_t . It supports the claim that the frozen text encoder suffers from the negative effects of task discrepancy between the pretext task and the CIR task. In contrast, the text encoder updated by RTD shows a significant improvement over the frozen text encoder, showing that it successfully reduces the task discrepancy.

Moreover, we additionally measure the average cosine similarity between the composed textual features with the prompt “a photo of $\phi(\psi_V(I_r))$ that T_c ” (Figure 1a) and the target image features (Figure 1c). The similarity is measured by the LinCIR ViT-L/14 model [17] on the CIRCO validation split. When we use the frozen CLIP text encoder (ψ_T), the average similarity is 0.1. By changing the text encoder to our updated text encoder (ψ_T^{tr}), the similarity becomes 0.29 (+0.19). This result shows that RTD successfully aligns the composed query features using ϕ to the frozen CLIP image features.

4 Experiments

4.1 Experimental setup

Implementation details. We use the AdamW optimizer [28] with a weight decay of 0.01. The learning rate is set to 10^{-5} , with a batch size of 512. For a fair comparison, we select the text encoder model with the best zero-shot CIRR [1] dev R@1 score for evaluating RTD. We evaluate the CIR performances of the model in a zero-shot manner by evaluating it across four different benchmarks. We use the visual and textual encoders of the CLIP ViT-B/32 and ViT-L/14 [14] as our backbone. Unless otherwise noted, we use the LLM-based 2.5M text triplets provided by CompoDiff [8] for the training. We set the τ as 0.07 in Eq. (1) and scale the standard deviation of Gaussian distribution as 0.5 for the noise injection. More results on various noise distributions can be found in the Appendix. All experiments were conducted using four NVIDIA A100 GPUs with Python 3.8 and Pytorch [29].

Evaluation datasets and metrics. We compare ZS-CIR methods on five benchmark datasets: CIRR [1], CIRCO [16], FashionIQ [18], COCO [30], and GeneCIS [19]. Details of each dataset are in the Appendix A.1. For CIRR, FashionIQ, COCO, and GeneCIS, we have reported their recall scores at the top K retrieval results (R@K). Since the CIRCO dataset includes multiple positive images for each query, we use a ranking-based metric—mean Average Precision scores at the top K results (mAP@K)—which provides a more robust and reliable assessment [31, 32]. For the main results, we compare the results on the three categories (Shirt, Dress, Toptee) of the FashionIQ validation split, as well as the test sets of CIRR and CIRCO. GeneCIS and COCO object composition results and their detailed explanations can be found in the Appendix B.1.

Baselines. We evaluate the effect of our method when combined with three recent ZS-CIR methods: Pic2Word [15], SEARLE [16], and LinCIR [17]. All three methods share the same core concept shown in Figure 1, but use different training schemes. Pic2Word[15] optimizes contrastive loss between the image embedding and its projected text embedding of “a photo of [extract_itex]\\$” to obtain the projection module ϕ . Similarly, SEARLE [16] employs a two-stage approach, starting with an optimization-based textual inversion phase followed by a distillation phase for the projection module ϕ . LinCIR [17] introduces a language-only self-supervised task involving keyword token replacement by letting the original text embedding and the replaced text embedding whose keyword tokens are changed to the projected original text embedding by ϕ .

We train all these methods with the same backbone architectures (CLIP ViT-B/32 and ViT-L/14). We use the publicly available pre-trained model for SEARLE (ViT-B/32, ViT-L/14) and Pic2Word (ViT-L/14). Otherwise, we reproduce the results using the official implementation. When reproducing, we adhere to the same settings in the original papers. For example, we select the final last epoch model for the Pic2Word ViT-B/32 model and choose the model based on the best zero-shot CIRR [1] dev R@1 score for LinCIR. Although our method is not specifically designed for projection-based

Table 2: **FashionIQ validation results.** The results of RTD combined with Pic2Word [15], SEARLE [16], and LinCIR [17] across different CLIP backbones (ViT-B/32 and ViT-L/14) are shown. **Blue** denotes the performance gain achieved by our method.

		Shirt		Dress		Toptee		Average	
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
ViT-B/32	Pic2Word	13.40	28.46	8.48	20.77	13.31	29.68	11.73	26.30
	+RTD	23.06 (+9.66)	40.48 (+12.02)	20.33 (+11.85)	41.75 (+20.98)	24.12 (+10.81)	46.35 (+16.67)	22.5 (+10.77)	42.86 (+16.56)
	SEARLE	24.78	41.85	17.90	36.99	25.24	46.71	22.64	41.85
	+RTD	26.69 (+1.91)	44.31 (+2.46)	20.72 (+2.82)	43.13 (+6.14)	26.67 (+1.43)	48.75 (+2.04)	24.7 (+2.06)	45.4 (+3.55)
	LinCIR	18.55	34.64	15.67	33.86	20.19	40.08	18.14	36.20
	+RTD	23.65 (+5.10)	42.74 (+8.10)	19.98 (+4.31)	41.75 (+7.89)	24.73 (+4.54)	46.56 (+6.48)	22.79 (+4.65)	43.68 (+7.48)
ViT-L/14	Pic2Word	26.59	42.93	21.32	43.53	28.10	48.19	25.34	44.88
	+RTD	27.97 (+1.38)	46.96 (+4.03)	23.50 (+2.18)	46.65 (+3.12)	31.31 (+3.21)	53.09 (+4.90)	27.59 (+2.25)	48.90 (+4.02)
	SEARLE	26.94	45.34	19.58	40.80	28.45	49.77	24.99	45.30
	+RTD	32.63 (+5.69)	50.39 (+5.05)	23.2 (+3.62)	47.25 (+6.45)	32.18 (+3.73)	54.56 (+4.79)	29.34 (+4.35)	50.73 (+5.43)
	LinCIR	30.42	47.99	21.86	44.77	29.98	50.38	27.42	47.71
	+RTD	32.83 (+2.41)	50.44 (+2.45)	24.49 (+2.63)	48.24 (+3.47)	33.4 (+3.42)	54.56 (+4.18)	30.24 (+2.82)	51.08 (+3.37)

Table 3: **CIRR and CIRCO test results.** Details are the same as Table 2.

		R@1	CIRR		CIRCO			
			R@5	R@10	mAP@5	mAP@10	mAP@25	mAP@50
ViT-B/32	Pic2Word	13.64	37.45	52.22	2.85	3.24	3.89	4.31
	+RTD	23.59 (+9.95)	51.76 (+14.31)	65.16 (+12.94)	6.39 (+3.54)	6.66 (+3.42)	7.64 (+3.75)	8.16 (+3.85)
	SEARLE	23.71	53.3	66.84	8.90	9.42	10.64	11.34
	+RTD	26.29 (+2.58)	56.41 (+3.11)	69.74 (+2.90)	11.26 (+2.36)	12.11 (+2.69)	13.63 (+2.99)	14.37 (+3.03)
	LinCIR	18.87	45.66	58.43	6.25	6.74	7.62	8.10
	+RTD	24.82 (+5.95)	53.47 (+7.81)	66.87 (+8.44)	8.94 (+2.69)	9.35 (+2.61)	10.57 (+2.95)	11.21 (+3.11)
ViT-L/14	Pic2Word	24.22	51.49	64.05	8.27	9.10	10.09	10.75
	+RTD	27.86 (+3.64)	56.24 (+4.75)	68.48 (+4.43)	9.13 (+0.86)	9.63 (+0.53)	10.68 (+0.59)	11.27 (+0.52)
	SEARLE	24.89	52.31	65.69	11.62	12.72	14.33	15.13
	+RTD	26.63 (+1.74)	56.17 (+3.86)	68.96 (+3.27)	16.53 (+4.91)	17.89 (+5.17)	19.77 (+5.44)	20.68 (+5.55)
	LinCIR	23.76	52.89	66.46	13.00	14.11	15.81	16.68
	+RTD	26.63 (+2.87)	56.17 (+3.28)	68.96 (+2.50)	17.11 (+4.11)	18.11 (+4.00)	20.06 (+4.25)	21.01 (+4.33)

method, we do not compare our method with CIR methods that require massive external triplet datasets [8, 9, 11] or those utilizing large-scale captioners and LLMs during inference [13, 33] due to their inefficiency. Integrating RTD into the other CIR methods will be an interesting future work.

4.2 Main results

Table 2 summarizes the evaluation results on the FashionIQ dataset. In the table, we observe that the incorporation of our approach with ZS-CIR methods significantly improves the performance across all three existing ZS-CIR methods (SEARLE, Pic2Word, and LinCIR) and all backbones (ViT-B/32 and ViT-L/14). For example, regardless of the choice of ZS-CIR methods and backbones, the minimum performance gain for average R@10 and R@50 scores is greater than 2 and 3.5 points, respectively. Table 3 shows a similar trend on the CIRR and CIRCO datasets. Notably, in some metrics on the CIRR and CIRCO datasets, the performance improvements achieved through our method (ViT-B/32) even exceed those obtained by employing a larger backbone (ViT-L/14), which clearly demonstrates the effect of our method. Specifically, in the CIRR R@1 score, SEARLE + RTD (26.29) and LinCIR + RTD (24.82) using ViT-B/32 surpasses the original results of SEARLE (24.89) and LinCIR (23.76) using ViT-L/14. We verify that a similar tendency is observed in the GeneCIS and COCO object composition task datasets, as detailed in the Appendix.

We further evaluate the performance of RTD using the significantly larger backbone (ViT-G/14). In Table 4, by combining RTD and the state-of-the-art ZS-CIR method, LinCIR, we achieve the best ZS-CIR performances on all benchmarks. Details and the full results are provided in the Appendix. We also provide additional qualitative retrieval results in the Appendix.

Table 4: Results on LinCIR [17] with OpenCLIP ViT-G/14 backbone [34].

Method	CIRR		CIRCO		FashionIQ		Avg
	R@5	R@10	mAP@10	mAP@25	R@10	R@50	
LinCIR	64.51	76.12	21.93	24.12	44.53	65.53	49.46
+RTD	67.47 (+2.96)	78.31 (+2.19)	22.29 (+0.36)	24.46 (+0.34)	46.21 (+1.68)	67.26 (+1.73)	50.99 (+1.53)

Table 5: **Ablation study.** We measure the impact of TCL loss (Eq. (1)), refined batch sampling (RB), and refined concatenation scheme (RC) on the validation splits of three CIR datasets. We train RTD based on LinCIR ViT-L/14. The first row equals to the vanilla LinCIR without applying RTD.

Text pair	TCL			CIRR		CIRCO		FashionIQ		Avg
	Anchor	RB	RC	R@5	R@10	mAP@10	mAP@25	R@10	R@50	
-	-	✗	✗	54.29	67.76	12.67	14.45	27.42	47.71	37.38
(T_r, T_r)	✓	✗	✗	55.99	69.72	13.40	15.18	28.16	48.82	38.54
(T_{r+c}, T_r)	✓	✗	✗	58.19	71.54	14.36	16.03	26.93	47.94	39.17
(T_r, T_r)	✓	✓	✗	58.19	71.27	14.96	16.67	27.42	49.33	39.64
(T_{r+c}, T_r)	✗	✓	✗	54.34	66.97	12.23	13.64	25.02	45.31	36.25
(T_{r+c}, T_r)	✓	✓	✓	57.90	71.13	16.10	17.84	30.24	51.08	40.72

4.3 Ablation studies on the proposed method

Table 5 presents the effectiveness of the proposed components: target-anchored text contrastive loss (TCL), refined batch sampling (RB), and refined concatenation scheme (RC). All evaluation results are on the validation split of CIRR, CIRCO and FashionIQ. All model variants use ViT-L/14 and a projection module ϕ from LinCIR [17], making the results in row 1 indicative of the original performance of LinCIR. We first compare the impact of the text pairs fed into TCL loss. We compare our design choice (T_{r+c}, T_t) (from the generated text triplets) with (T_r, T_r) , which is the sole option for constructing a pair given a single conventional caption T_r . The results demonstrate that, on average, using generated triplets (3rd row) is more effective than using original conventional text pairs (2nd row), particularly in CIRR and CIRCO. In addition, RB (4th row) and RC (6th row) significantly enhance the overall performance, demonstrating the effectiveness of these components. Finally, we measure the impact of using the frozen text encoder for target caption T_t , denoted as “Anchor” in the table. Significant performance degradation is observed when the learnable text encoder is used for extracting the embedding of the target caption T_t (5th row) compared to the target-anchored case (4th row), supporting the importance of the anchoring design choice.

4.4 More analyses on RTD

In this subsection, we show more analyses on RTD with the same setting to Table 5.

Table 6: **Impact of the text triplet generation method.** The details are the same as Table 5.

	CIRR		CIRCO		FashionIQ		Avg
	R@5	R@10	mAP@10	mAP@25	R@10	R@50	
Baseline (LinCIR)	54.29	67.76	12.67	14.45	27.42	47.71	37.38
+RTD (LLM-based)	57.90	71.13	16.10	17.84	30.24	51.08	40.72
+RTD (template-based)	56.71	70.34	15.01	16.98	30.37	51.94	40.23

[Impact of the text triplet generation method] There are two variants for the text triplet construction process: LLM-based generation and template-based generation. While our main experiments utilize LLM-based text triplets, we have verified that using rule-based template-based text triplets is sufficient; our template-based text triplets generation process is a fully text-only and automatic process without relying on human labor or LLMs. We describe the details of our template-based generation process in the Appendix A.2. Table 6 shows that the results of our template-based approach are comparable to those obtained with LLM-generated text triplets. We believe this finding implies the simplicity and efficiency of constructing text triplets, underscoring the practicality of RTD.

[Impact of the update rule for the text encoder] To verify that our improvements cannot be achievable solely by tuning the text encoder backbone without considering the task discrepancy, we additionally measure the results of previous methods (Pic2Word and LinCIR) when naively updating text encoders. Namely, after training ϕ while keeping all other networks frozen as in previous methods, we additionally update the text encoder using the original loss function, while fixing other modules including ϕ . We denote this update rule as “naïve tuning” in the Table 7. Unlike RTD, we observe that just naively updating the text encoder (“naïve tuning”) significantly degrades the performance

Table 7: **Impact of the update rule.** Two update rules are compared: (1) using the original objective from baseline and (2) using RTD. For a fair comparison, in both rules, ϕ is updated first and ψ_T is updated top on the frozen modules. We use the ViT-L/14 backbone, and numbers are measured on the CIRR dev split, CIRCO validation split, and Fashion IQ validation split.

	CIRR		CIRCO		FashionIQ		Avg
	R@5	R@10	mAP@10	mAP@25	R@10	R@50	
Baseline(Pic2Word)	51.40	64.43	8.77	10.12	25.34	44.88	32.15
+naïve tuning	19.21	27.51	1.29	1.61	4.4	11.15	10.86
+ RTD	56.64	69.77	8.83	9.81	27.59	48.90	36.92
Baseline(LinCIR)	54.29	67.76	12.67	14.45	27.42	47.71	36.86
+naïve tuning	52.67	66.78	11.40	12.99	26.34	45.92	35.52
+ RTD	57.90	71.13	16.10	17.84	30.24	51.08	40.72

Table 8: **More efficient variants.** “Learnable params (%)” denotes the percentage of learnable parameters relative to the entire set of parameters in the text encoder.

Training variants	Learnable params (%)	CIRR		CIRCO		FashionIQ		Avg
		R@5	R@10	mAP@10	mAP@25	R@10	R@50	
Baseline(LinCIR)	0%	54.29	67.76	12.67	14.45	27.42	47.71	37.38
+RTD (Full model)	100%	57.90	71.13	16.10	17.84	30.24	51.08	40.72
+RTD (Whole FCs)	45.8%	57.76	71.35	15.03	16.90	30.31	51.81	40.53
+RTD (Front 3 FCs)	11.5%	55.65	69.83	13.95	15.81	28.69	49.92	38.98
+RTD (Middle 3 FCs)	11.5%	56.69	70.03	14.66	16.58	28.55	49.84	39.39
+RTD (Last 3 FCs)	11.5%	56.84	69.74	14.81	16.70	29.16	50.43	39.61
+RTD (Interleave 3 FCs)	11.5%	57.21	70.65	15.20	17.13	28.91	50.17	39.88

of the baseline. The results indicate that merely updating the text backbone is not beneficial for ZS-CIR; instead, mitigating task discrepancy through RTD is necessary.

[More efficient variants] Table 8 presents the results of the more efficient implementations of our approach in terms of the number of updated parameters. Specifically, instead of updating the entire set of parameters of the text encoder, we explore updating only a few layers of the network when applying RTD. Our findings indicate that updating only the fully connected layers (denoted as “Whole FCs”) nearly matches the performance of the full model while using less than half the number of learnable parameters (40.72 vs. 40.53 average score). Additionally, we verify that updating only three fully connected layers, whose parameter size matches the projection module ϕ and constitutes 11.5% of the full model, is also sufficiently effective. We test various three-layer updating strategies: “First 3 FCs”: the first three layers (closest to the input), “middle 3 FCs”: the middle three layers, “Last 3 FCs”: the last three layers, and “Interleave 3 FCs”: an interleaved selection of three layers (first, middle, and last layers). Among these, we verify that the “Interleave 3 FCs” shows the best result, maintaining competitive performance with the full model (40.72 vs. 39.88 average score). We believe these findings suggest a promising direction for enhancing the training efficiency of our approach by selectively updating only specific layers of the text encoder.

5 Discussion and Limitations

As noted by [16, 17], the existing CIR benchmark datasets [1, 18, 19] are somewhat noisy due to the presence of false negatives, leading to unreliable evaluations. A similar issue is reported in the image-text cross-modal retrieval problem by [35, 36]. Specifically, although there is typically one ground truth positive for each query, there may be multiple ground truths within the database. Thus, Gu and Chun et al. [17] primarily use the CIRCO dataset as their main benchmark because CIRCO includes multiple positives and employs a more reliable ranking-based metric, mAP@K [31, 32]. Additionally, they focus on R@K with a larger K (e.g., 10) rather than R@1 in the other benchmarks. In our case, since the overall metrics improve regardless of their type (including mAP@K for CIRCO and R@K with a larger K for the other benchmarks), we believe the improvements from our method are sufficiently reliable despite the noisiness of the evaluation benchmarks.

6 Conclusion

Our research presents RTD, a novel plug-and-play training scheme designed to improve the capabilities of text encoders for ZS-CIR. By leveraging easily obtainable text triplets and implementing target-anchored text contrastive learning, RTD aligns projected and conditioning text embeddings with target embeddings, addressing the challenges posed by task discrepancies in ZS-CIR. Additionally, the integration of hard negatives-based refined batch sampling and a sophisticated concatenation scheme further enhances performance. Empirical evaluations demonstrate that RTD significantly boosts the performance of existing ZS-CIR methods across diverse datasets and model backbones, underscoring its effectiveness and versatility.

References

- [1] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*, 2021. 1, 2, 3, 6, 9, 12
- [2] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *CVPR*, 2022. 1
- [3] Ginger Delmas, Rafael S Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In *ICLR*, 2022.
- [4] Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *CVPR*, 2021.
- [5] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *CVPR*, 2019. 1
- [6] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypertexted, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 1, 5, 12
- [7] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS Dataset and Benchmark*, 2022. 1
- [8] Geonmo Gu, Sanghyuk Chun, HeeJae Jun, Yooheon Kang, Wonjae Kim, and Sangdoon Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*, 2023. 2, 4, 5, 6, 7, 12
- [9] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. CoVR: Learning composed video retrieval from web video captions. *arXiv preprint arXiv:2308.14746*, 2023. 7
- [10] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and quality assessment for composed image retrieval. In *AAAI*, 2024.
- [11] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhui Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*, 2024. 2, 7
- [12] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2, 4, 5
- [13] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. *arXiv preprint arXiv:2310.09291*, 2023. 2, 3, 7
- [14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3, 6
- [15] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *CVPR*, 2023. 2, 3, 4, 5, 6, 7, 12
- [16] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *ICCV*, 2023. 2, 3, 6, 7, 9, 12

- [17] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, , Yooheon Kang, and Sangdoon Yun. Language-only efficient training of zero-shot composed image retrieval. In *CVPR*, 2024. 2, 3, 4, 5, 6, 7, 8, 9, 12, 14, 15
- [18] 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 *CVPR*, 2021. 2, 3, 6, 9, 12
- [19] Sagar Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *CVPR*, 2023. 3, 6, 9, 12, 13
- [20] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *CVPR*, 2022. 3
- [21] Junyang Chen and Hanjiang Lai. Pretrain like you inference: Masked tuning improves zero-shot composed image retrieval. *arXiv preprint arXiv:2311.07622*, 2023. 3
- [22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3, 6
- [23] 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 *ICML*, 2023. 3
- [24] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020. 3
- [25] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 5
- [26] Niv Cohen, Rinon Gal, Eli A Meir, Gal Chechik, and Yuval Atzmon. “this is my unicorn, fluffy”: Personalizing frozen vision-language representations. In *ECCV*, 2022. 5
- [27] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *NeurIPS*, 2022. 5
- [28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6
- [29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019. 6
- [30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6, 12
- [31] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *ECCV*, 2020. 6, 9, 12
- [32] Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, and Seong Joon Oh. Eccv caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for ms-coco. In *ECCV*, 2022. 6, 9, 12
- [33] Shitong Sun, Fanghua Ye, and Shaogang Gong. Training-free zero-shot composed image retrieval with local concept reranking. *arXiv preprint arXiv:2312.08924*, 2023. 7
- [34] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, july 2021. URL <https://doi.org/10.5281/zenodo.5143773>. If you use this software, please cite it as below. 7, 14, 15
- [35] Sanghyuk Chun. Improved probabilistic image-text representations. *arXiv preprint arXiv:2305.18171*, 2023. 9
- [36] Jaeseok Byun, Dohoon Kim, and Taesup Moon. Converting and smoothing false negatives for vision-language pre-training. *arXiv preprint arXiv:2312.06112*, 2023. 9
- [37] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018. 12
- [38] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *CVPR*, 2021. 12
- [39] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 12

A Additional Implementation Details

A.1 CIR Datasets

FashionIQ [18] is a dataset that contains fashion-related images from three main categories: Shirts, Dresses, and Tootie. It has a total of 30,134 triplets, which were created from 77,684 images. As the ground truth labels are not publicly available, we utilize the results from the validation set for our analysis and comparison. CIR [1] encompasses a wider range of domains and contains images with more complex descriptions compared to FashionIQ. It contains 36,554 triplets extracted from 21,552 images, which are sourced from the well-known NLR2 natural language inference dataset [37]. As pointed out in previous works [15, 17, 16], CIR and FashionIQ suffer from a significant number of false negatives, which can potentially lead to inaccurate retrieval evaluations [16, 15]. To address this issue, CIRCO [16], based on COCO images [30], is recently introduced by providing multiple positive images for each query. This approach enables a more reliable and robust mAP metric [31, 32], which is essential for accurate evaluation of retrieval performance.

We additionally provide results on two more benchmark datasets, GeneCIS [19] and COCO Object Composition introduced by [15], in Appendix B.1. GeneCIS [19] is also constructed based on COCO images and the Visual Attributes in the Wild dataset [38]. GeneCIS introduces four task variations: (1) focus on an attribute, (2) change an attribute (3) focus on an object and (4) change an object. These tasks explore different aspects of image retrieval and manipulation. For the COCO Object Composition task, we utilize 5000 images from the COCO validation dataset to evaluate object composition. Our objective is to retrieve an image that contains an object specified by a query image, along with scenes or objects described using text. The composed query is constructed by combining "a photo of [\$], [obj₁], [obj₂] ... and [obj_n]" where [obj_i] are text descriptions.

A.2 Training text triplets

In Figure A.1, A.2, examples of both LLM-based and template-based triplets are presented. Both methods generate natural and coherent text.

[Detailed explanation on template-based triplets] Here, we describe the detailed procedure for generating template-based text triplets. We mainly follow the procedure of Compodiff [8], but a text-to-image generation step is not involved. For template-based triplets, we use captions from the CC3M dataset [6] as reference captions T_r . Firstly, given reference captions, important keywords like nouns are extracted with a part-of-speech (POS) tagger via the Spacy library. Then, the keyword is filtered by frequency filtering with hard thresholding to focus only on frequently occurring keywords. Specifically, we only use keywords that appear more than 100. After applying keyword frequency filtering, the remaining keyword list is used to create caption triplets (T_r, T_c, T_t) . To generate the triplets, a keyword from the given T_r is selected, and alternative keywords are extracted based on text similarity scores ranging from 0.5 to 0.7, using the SBERT all-MiniLM-L6-v2 [39]. The target caption T_t is then constructed by substituting the original keyword with a similar alternative. The conditioning text T_c is generated using randomly selected pre-defined templates, as detailed in Table A.1. Here, most of the templates are similar to that of Compodiff [8].

Since the quality of the generated triplets with the above procedure may not be optimal, we employ an additional filtering process. Compodiff [8] employs an additional filtering process that uses cosine similarities between generated images and texts, calculated by CLIP encoders. However, as we do not have images for captions, we filter the inappropriate texts using only textual information inspired by LinCIR [17]. Namely, we calculate the similarity between the CLIP text embedding of T_t and the CLIP text embedding of "a photo of [\$]" where [\$] is obtained by T_t projected by ϕ from LinCIR (ViT-L/14). Following LinCIR noise ($\text{Unif}(0, 1) \times \mathcal{N}(0, 1)$) is injected before passing through ϕ . After calculating the above similarity, texts whose similarities are less than the threshold (0.75) are removed. The same process is also applied to the reference caption T_r and the intersection of filtering processes for T_t and T_r is used for the final dataset whose size becomes 1.3M.

```

{
  "source_caption": "what do you do with automobile model for $60 k",
  "target_caption": "what do you do with model for $60 k",
  "relative_caption": "without automobile"
},
{
  "source_caption": "a collage of my latest artwork includes oil pastel and acrylic paintings",
  "target_caption": "a collage of my latest artwork includes water pastel and acrylic paintings",
  "relative_caption": "alter oil to match water"
},
{
  "source_caption": "baseball player hits a home run against sports team",
  "target_caption": "baseball customer hits a home run against sports team",
  "relative_caption": "player is removed and customer takes its place"
},
{
  "source_caption": "another wall at my home",
  "target_caption": "another bedroom at my home",
  "relative_caption": "bedroom is added in place of wall"
},
{
  "source_caption": "tennis player faces a tough schedule if she is to advance",
  "target_caption": "tennis player faces a tough routine if she is to advance",
  "relative_caption": "change schedule to routine"
},
}

```

Figure A.1: Example of template-based triplet datasets

```

{
  "source_caption": "Christopher Nolan got advice from Steven Spielberg before making",
  "target_caption": "Steven Spielberg got advice from Walter Mitty before making",
  "relative_caption": "get advice from Walter Mitty"
},
{
  "source_caption": "by Koh Chip Whye - Black & White Buildings & Architecture",
  "target_caption": "by Koh Chip Whye - Colorful Buildings & Architecture",
  "relative_caption": "make the buildings more colorful"
},
{
  "source_caption": "The Most Hyperrealistic Images Of Beautiful Bathing Women With Their Heads Underwater",
  "target_caption": "The Most Hyperrealistic Images Of Beautiful Bathing Women With Their Heads Underwater and Octopus Arms",
  "relative_caption": "make the women have octopus arms"
},
{
  "source_caption": "Mountains above clouds - p312m1472749 by Mikael Svensson",
  "target_caption": "Mountains on Mars - p312m1472749 by Mikael Svensson",
  "relative_caption": "Put the mountains on Mars"
},
{
  "source_caption": "Le bouquiniste Paris -60x60",
  "target_caption": "The New York City book store -60x60",
  "relative_caption": "Instead of Paris, make it New York."
},
}

```

Figure A.2: Example of LLM-based triplet datasets

B Additional experiments

B.1 Results on GeneCIS [19] and COCO object composition

We observe that incorporating our approach with ZS-CIR methods leads to marginal but consistent performance improvements on GeneCIS as shown in Table B.1. The relatively smaller performance difference compared to other datasets can be attributed to the discrepancy between the format of the conditioning text of GeneCIS and the ZS-CIR methods training methodology. Namely, GeneCIS only uses the fixed four text instructions “change attribute”, “focus attribute”, “change object” and “focus object”, which is different from the usual text instruction we expected (*e.g.*, “change the dog to a cat”).

Table A.1: The full 50 keyword converting templates

"replace \${source} with \${target}"	"substitute \${target} for \${source}"
"apply \${target}"	"\${source} is removed and \${target} takes its place"
"convert \${source} to \${target}"	"modify \${source} to become \${target}"
"replace \${source} with \${target}"	"customize \${source} to become \${target}"
"update \${source} to \${target}"	"change \${source} to match \${target}"
"substitute \${target} for \${source}"	"\${target} is introduced after \${source} is removed"
"alter \${source} to match \${target}"	"\${target} is added in place of \${source}"
"upgrade \${source} to \${target}"	"\${target} is introduced as the new option after"
"amend \${source} to fit \${target}"	"\${source} is removed and \${target} is added"
"opt for \${target}"	"\${source} is removed and \${target} is introduced"
"\${source} is removed"	"\${target} is added as a replacement for \${source}"
"add \${target}"	"\${target} is the new option available"
"if it is \${target}"	"\${target} is added after \${source} is removed"
"\${target} is the updated option"	"\${target} is introduced after \${source} is retired"
"\${target} is the updated choice"	"tweak \${source} to become \${target}"
"\${source} is replaced with \${target}"	"has no \${source}"
"change \${source} to \${target}"	"alter \${source} to \${target}"
"swap \${source} for \${target}"	"redesign \${source} as \${target}"
"turn \${source} into \${target}"	"adapt \${source} to fit \${target}"
"choose \${target} instead of \${source}"	"\${target} is the new choice"
"\${target} is the new selection"	"exchange \${source} with \${target}"
"transform \${source} into \${target}"	"show no \${source}"
"no \${source}"	"remove \${source}"
"delete \${source}"	"not a \${source}"
"with no \${source}"	"without \${source}"

In the experiment on COCO object composition, we observe a significant performance improvement, similar to the results obtained on other datasets in Table B.2. This finding reaffirms that our approach, when combined with ZS-CIR methods, consistently achieves strong performance, demonstrating its generalizability.

Table B.1: GeneCIS results

		R@1	Average R@2	R@3
ViT-B	Pic2Word	11.13	21.08	31.05
	+RTD	12.03 (+0.90)	21.61 (+0.53)	31.09 (+0.04)
	SEARLE	12.19	22.56	32.03
	+ours	12.82 (+0.63)	22.97 (+0.41)	32.44 (+0.41)
	LinCIR	12.23	21.29	30.80
	+ours	12.83 (+0.60)	22.83 (+1.54)	32.22 (+1.42)
ViT-L	Pic2Word	11.18	21.45	30.55
	+ours	11.92 (+0.74)	22.32 (+0.87)	31.33 (+0.78)
	SEARLE	12.30	22.08	31.29
	+ours	12.40 (+0.10)	22.82 (+0.74)	32.37 (+1.08)
	LinCIR	12.45	22.66	32.06
	+ours	13.18 (+0.73)	23.12 (+0.46)	32.77 (+0.71)

B.2 Results on larger backbone (ViT-G)

We further evaluate the performance of RTD using the significantly larger backbone (OpenCLIP ViT-G/14 [34]). As described in Section 4.3, we use the projection module ϕ from LinCIR [17]. Since the pre-trained projection module ϕ for LinCIR [17] (ViT-G/14) is not publicly available, we reproduce it and integrate RTD with it. We emphasize that similar to our previous results, RTD again achieves remarkable gains across all datasets. Here, we set the learning rate as 10^{-6} .

B.3 Ablations on noise injection

We conduct an ablation study of the different noise types employed for the ‘‘refined concatenation scheme’’ shown in Figure 2. We compare three different noise types, uniform distribution, Gaussian distribution, and LinCIR-ish noise ($\text{Unif}(0, 1) \times \mathcal{N}(0, 1)$). We also examine the scale of LinCIR-

Table B.2: COCO object composition results

		COCO		
		R@1	R@5	R@10
ViT-B	Pic2Word	6.88	13.6	17.52
	+ours	7.62 (+0.74)	20.23 (+6.63)	28.69 (+11.17)
	SEARLE	9.52	21.45	29.38
	+ours	11.01 (+1.49)	24.34 (+2.89)	32.84 (+3.46)
	LinCIR	7.15	18.38	27.3
+ours	9.59 (+2.44)	21.66 (+3.28)	30.66 (+3.36)	
ViT-L	Pic2Word	10.26	23.67	32.53
	+ours	10.26 (+0.00)	24.66 (+0.99)	33.56 (+1.03)
	SEARLE	12.07	26.13	35.17
	+ours	14.38 (+2.31)	29.74 (+3.61)	38.09 (+2.92)
	LinCIR	11.37	24.53	33.85
+ours	14.6 (+3.23)	29.84 (+5.31)	38.87 (+5.02)	

Table B.3: FashionIQ results on larger OpenCLIP ViT-G/14 backbone [34].

Method	Shirt		Dress		Toptee		Average	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
LinCIR (reported in [17])	46.76	65.11	38.08	60.88	50.48	71.09	45.11	65.69
LinCIR (reproduced)	46.61	64.72	38.18	60.54	49.26	70.83	44.68	65.36
+RTD	47.20 (+0.59)	66.24 (+1.52)	39.86 (+1.68)	63.01 (+2.47)	51.56 (+2.30)	72.51 (+1.68)	46.21 (+1.54)	67.26 (+1.90)

Table B.4: CIRR and CIRCO results on larger OpenCLIP ViT-G/14 backbone [34].

ViT-G	R@1	CIRR		CIRCO			
		R@5	R@10	mAP@5	mAP@10	mAP@25	mAP@50
LinCIR (reported in [17])	35.25	64.72	76.05	19.81	21.01	23.03	24.18
LinCIR (reproduced)	34.94	64.51	76.12	20.63	21.93	24.12	25.20
+RTD	36.31 (+1.37)	67.47 (+2.96)	78.31 (+2.19)	21.08 (+0.45)	22.29 (+0.36)	24.46 (+0.34)	25.44 (+0.24)

ish noise from 0.1, 0.5, and 1. We report the test scores for CIRR and CIRCO, as well as the FashionIQ validation scores for Pic2Word, SEARLE, and LinCIR in Table D.1 and Table D.2. In the tables, we observe that all noise distributions show decent performance and LinCIR-like noises show slightly better performances than uniform distribution and normal distribution. We also observe that the different scale choice for the LinCIR-like noise somewhat affects the overall performances. In the main experiments, we choose 0.5 for the noise scale, following the observed performance improvements.

C Qualitative example on CIRCO

We qualitatively illustrate the results of incorporating RTD into LinCIR on the CIRCO dataset in Figure D.1. The visual examples provide an intuitive demonstration of how the integration of RTD enhances the performance of LinCIR, effectively capturing the semantic meaning of the modification descriptions while preserving the relevant visual information from the reference image.

D Societal Impacts

Although our paper demonstrates promising outcomes in the ZS-CIR task, further examination of the data and the model is essential prior to practical deployment. Since our method focuses mainly on optimization for accuracy, unwanted social implications can occur. For example, real-world images from databases and user-generated text may inadvertently cause harmful cases.

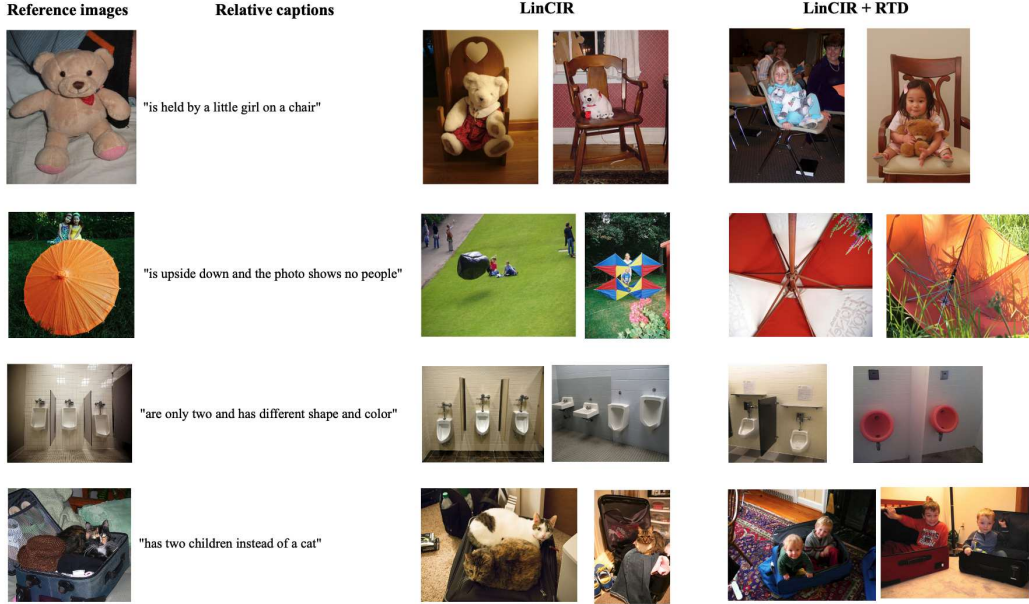


Figure D.1: Qualitative Results on CIRCO dataset

Table D.1: Noise type variation on CIRR/CIRCO dataset

	Noise type	Scale	CIRR			CIRCO					
			R@1	R@5	R@10	mAP@5	mAP@10	mAP@25	mAP@50		
ViT-B/32	Pic2Word	-	13.64	37.45	52.22	2.85	3.24	3.89	4.31		
	+ours	Unif(-1,1)	1	23.23	50.55	64.28	4.29	4.57	5.19	5.57	
		$\mathcal{N}(0,1)$	1	21.18	47.78	61.47	4.09	4.26	4.83	5.17	
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.1	23.52	51.13	64.53	5.13	5.46	6.17	6.62	
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.5	23.01	51.18	64.84	4.29	4.57	5.19	5.57	
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	1	23.59	51.76	65.16	6.39	6.66	7.64	8.16	
	SEARLE	-	23.71	53.3	66.84	8.9	9.42	10.64	11.34		
	+ours	Unif(-1,1)	1	26.07	55.98	69.18	10.87	11.55	12.97	13.65	
		$\mathcal{N}(0,1)$	1	26.41	56.68	69.47	10.91	11.53	12.88	13.6	
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.1	26.02	55.47	68.15	10.43	11.07	12.37	13.07	
$\mathcal{N}(0,1) \times \text{Unif}(0,1)$		0.5	26.29	56.41	69.74	11.26	12.11	13.63	14.37		
$\mathcal{N}(0,1) \times \text{Unif}(0,1)$		1	26.43	56.58	69.76	11.42	12.04	13.38	14.1		
ViT-L/14	LinCIR	-	18.87	45.66	58.43	6.25	6.74	7.62	8.1		
	+ours	Unif(-1,1)	1	24.39	52.77	66.39	6.81	7.27	8.28	8.84	
		$\mathcal{N}(0,1)$	1	24.63	53.52	66.63	7.6	7.97	8.92	9.49	
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.1	24.58	53.3	66.65	9.6	10.11	11.47	12.15	
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.5	24.82	53.47	66.87	8.94	9.35	10.57	11.21	
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	1	25.4	54.58	67.69	8.17	8.53	9.72	10.35	
	Pic2Word	-	-	24.22	51.49	64.05	8.27	9.1	10.09	10.75	
		+ours	Unif(-1,1)	1	28.24	55.95	68.77	8.14	8.81	9.83	10.37
			$\mathcal{N}(0,1)$	1	27.06	53.95	66.43	7.08	7.66	8.57	9.07
			$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.1	28.24	57.35	68.65	10.04	10.63	11.71	12.31
$\mathcal{N}(0,1) \times \text{Unif}(0,1)$			0.5	27.86	56.24	68.48	9.13	9.63	10.68	11.27	
$\mathcal{N}(0,1) \times \text{Unif}(0,1)$			1	27.71	55.68	68.02	8.14	8.78	9.84	10.35	
SEARLE		-	24.89	52.31	65.69	11.62	12.72	14.33	15.13		
+ours		Unif(-1,1)	1	26.96	56.99	69.52	15.82	16.78	18.54	19.39	
		$\mathcal{N}(0,1)$	1	27.66	57.54	69.57	15.24	15.93	17.65	18.44	
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.1	26.31	55.88	69.4	16.05	17.26	19.12	20.01	
	$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.5	27.04	56.82	69.95	16.53	17.89	19.77	20.68		
	$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	1	27.93	57.76	70.19	17.35	18.66	20.52	23.44		
Pic2Word	LinCIR	-	23.76	52.89	66.46	13	14.11	15.81	16.68		
	+ours	Unif(-1,1)	1	26.58	56.31	68.94	17.23	18.2	20.11	21.03	
		$\mathcal{N}(0,1)$	1	26.75	55.64	68.48	16.45	17.57	19.37	20.3	
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.1	26.7	56.22	69.08	17.24	18.27	20.24	21.19	
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.5	26.63	56.17	68.96	17.11	18.11	20.06	21.01	
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	1	26.99	56.1	69.01	17.33	18.3	20.21	21.13	

Table D.2: Noise type variation on FashionIQ dataset

	Noise type	Scale	Shirt		Dress		Toptee		Average		
			R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	
ViT-B/32	Pic2Word	-	13.4	28.46	8.48	20.77	13.31	29.68	11.73	26.3	
	+ours	Unif(-1,1)	1	21.84	37.63	18.49	39.61	23.0	43.91	21.11	40.38
		$\mathcal{N}(0,1)$	1	20.36	37.54	16.16	38.18	21.67	42.48	19.4	39.4
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.1	22.23	39.35	19.98	41.7	23.81	45.23	22.01	42.09
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.5	24.53	43.82	20.33	41.55	26.01	48.75	23.62	44.7
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	1	23.06	40.48	20.33	41.75	24.12	46.35	22.5	42.86
	SEARLE	-	24.78	41.85	17.90	36.99	25.24	46.71	22.64	41.85	
	+ours	Unif(-1,1)	1	23.75	42.25	20.18	40.36	25.14	46.46	23.02	43.02
		$\mathcal{N}(0,1)$	1	24.14	42.25	20.23	40.16	24.17	46.35	22.85	42.92
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.1	25.12	44.85	20.92	41.40	26.57	47.63	24.20	44.62
$\mathcal{N}(0,1) \times \text{Unif}(0,1)$		0.5	26.69	44.31	20.72	43.13	26.67	48.75	24.70	45.40	
$\mathcal{N}(0,1) \times \text{Unif}(0,1)$		1	25.07	44.01	20.43	41.00	26.11	47.12	23.87	44.04	
LinCIR	-	18.55	34.64	15.67	33.86	20.19	40.08	18.14	36.20		
+ours	Unif(-1,1)	1	21.79	39.35	18.89	40.21	23.66	45.33	21.45	41.63	
	$\mathcal{N}(0,1)$	1	22.37	38.67	19.53	40.11	23.71	44.37	21.87	41.05	
	$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.1	23.95	44.11	19.83	41.99	26.62	47.58	23.47	44.56	
	$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.5	23.65	42.74	19.98	41.75	24.73	46.56	22.79	43.68	
	$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	1	22.82	41.12	19.78	41.70	25.09	47.07	22.56	43.29	
ViT-L/14	Pic2Word	-	26.59	42.93	21.32	43.53	28.10	48.19	25.34	44.88	
	+ours	Unif(-1,1)	1	27.87	45.93	23.90	46.80	31.21	52.22	27.66	48.32
		$\mathcal{N}(0,1)$	1	26.94	44.95	23.45	45.56	30.34	51.45	26.91	47.32
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.1	28.26	47.64	24.05	47.20	31.21	53.70	27.84	49.51
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.5	27.97	46.96	23.50	46.65	31.31	53.09	27.59	48.90
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	1	28.41	46.91	24.10	46.21	31.11	52.27	27.87	48.46
	SEARLE	-	26.94	45.34	19.58	40.80	28.45	49.77	24.99	45.30	
	+ours	Unif(-1,1)	1	30.13	46.57	22.16	46.90	28.76	50.74	27.02	48.07
		$\mathcal{N}(0,1)$	1	26.99	43.23	21.17	44.82	27.54	49.06	25.23	45.70
		$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.1	32.63	50.39	23.20	47.25	32.18	54.56	29.34	50.73
$\mathcal{N}(0,1) \times \text{Unif}(0,1)$		0.5	31.80	49.31	23.20	47.30	31.41	54.00	28.80	50.20	
$\mathcal{N}(0,1) \times \text{Unif}(0,1)$		1	30.03	47.06	22.41	47.05	30.39	52.42	27.61	48.84	
LinCIR	-	30.42	47.99	21.86	44.77	29.98	50.38	27.42	47.71		
+ours	Unif(-1,1)	1	31.94	50.10	24.44	48.19	33.04	54.26	29.81	50.85	
	$\mathcal{N}(0,1)$	1	31.70	49.41	23.90	48.19	33.23	53.54	29.27	50.38	
	$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.1	32.92	50.64	24.49	48.74	33.50	55.02	30.31	51.47	
	$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	0.5	32.83	50.44	24.49	48.24	33.40	54.56	30.24	51.08	
	$\mathcal{N}(0,1) \times \text{Unif}(0,1)$	1	32.43	50.54	24.64	48.79	33.25	54.77	30.11	51.36	