SDA: Semantic Discrepancy Alignment for Text-conditioned Image Retrieval

Yuchen Yang[1,2], **Yu Wang**[2,3✉], **Yanfeng Wang**[2,3]
[1]University of Science and Technology of China [2]Shanghai AI Laboratory [3]Shanghai JiaoTong University

Abstract

In the realm of text-conditioned image retrieval, models utilize a query composed of a reference image and modification text to retrieve corresponding images. Despite its significance, this task is fraught with challenges, including small-scale datasets due to labeling costs and the complexity of attributes in modification texts. These challenges often result in models learning a generalized representation of the query, thereby missing the semantic correlations of image and text attributes. In this paper, we introduce a general boosting framework designed to address these issues by employing semantic discrepancy alignment. Our framework first leverages the ChatGPT to augment text data by modifying the original modification text's attributes. The augmented text is then combined with the original reference image to create an augmented composed query. Then we generate corresponding images using GPT-4 for the augmented composed query. We realize the cross-modal semantic discrepancy alignment by formulating distance consistency and neighbor consistency between the image and text domains. Through this novel approach, attribute in the text domain can be more effectively transferred to the image domain, enhancing retrieval performance. Extensive experiments on three prominent datasets validate the effectiveness of our approach, with state-of-the-art results on a majority of evaluation metrics compared to various baseline methods.

1 Introduction

Text-conditioned image retrieval makes the retrieval system more accurate and flexible by allowing the user to enter both a reference image and a text description. Recent years have witnessed some remarkable research efforts in the task of text-conditioned image retrieval (TCIR) (Vo et al., 2019) (Lee et al., 2021) (Wen et al., 2021) (Yang et al., 2021), with a focus on designing an appropriate composition module to learn joint visual-linguistic representations.

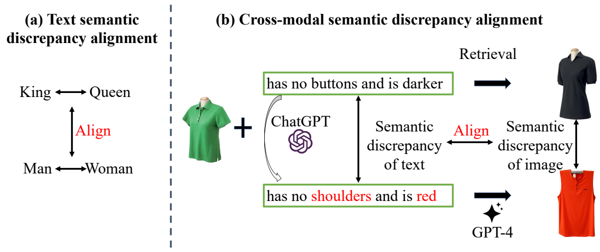


Figure 1: Illustration of our motivation. We use ChatGPT and GPT-4 to rewrite the modification text and generate corresponding images. Due to well-defined distribution of text domain, the semantic discrepancy between words are clearly distinguished. Inspired by this, we suppose to capture the alignment of semantic discrepancy across the image and text domain. This allows attributes such as color, layout, and style to be better understood by the model, which further facilitates a more discriminative joint image-text representation.

However, the existing methods face two challenges. First, the scale of the training set in text-conditioned image retrieval is typically very limited. This limitation arises because collecting data tuples for this task is much more labor-intensive than for other vision tasks, such as image classification and landmark recognition. Second, the modification text conveys complex semantics with attributes of high diversity. These challenges render existing methods sensitive to the dataset and prone to overfitting. A common problem is that the model learns "Image A + Text attribute C" and "Image B + Text attribute D" but still does not work on "Image A + Text attribute D" and "Image B + Text attribute C". This is because existing methods learn a general representation of the query and, therefore, overlook the semantic correlations between image and text attributes.

To address the aforementioned issues, we propose a general boosting framework for text-conditioned image retrieval by exploring semantic

✉: Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2024, pages 5250–5261
August 11-16, 2024 ©2024 Association for Computational Linguistics

discrepancy alignment. To illustrate our motivation, we provide an example. As shown in Fig. 1, the data tuple consists of a green T-shirt as the reference image with the modification text "has no buttons and is darker", the target image is a black T-shirt without buttons. Then we need to learn a compositor to synthesize visual features and textual features of the composed query to approach the feature of target image. If the modification text is partially changed, for instance, from "has no buttons and is darker" to "has no shoulders and is red", the existing compositor often not be able to output the correct composite feature for the unseen composed queries well. The reason is that the existing model direct matches the composed query to the target image rather than learning a projection of attributes across the image and text domain. On the other hand, thanks to recent progress in large language models (Kenton and Toutanova, 2019), semantically rich and accurate word embedding are provided and discrepancies in text semantics are clearly distinguished. A well-known example is that the discrepancy between the words "king" and "queen" is similar to that between "man" and "woman". Inspired by this, we seek to explore the consistency of semantic discrepancies across the image and text domain. We leverage powerful capabilities of ChatGPT and GPT-4 for data augmentation, which has proved successful in recent works (Xu et al., 2023) (Dai et al., 2023) (Yunxiang et al., 2023) (Xiong et al., 2023).

Based on above motivation, we propose a general boosting framework for semantic discrepancy alignment. To enable the model to understand diverse editing intentions and capture cross-domain consistency of attributes, we utilize the ChatGPT to generate qualified modification texts, such as "has no shoulders and is red" is generated by ChatGPT based on "has no buttons and is darker". Then, we combine these generated modification texts with the original reference image into new composed queries, which are named as augmented composed queries. Then we utilize the GPT-4 to generate corresponding images for the augmented composed queries. Since the original composed query and augmented composed queries only differs in the modification texts, we believe that the semantic discrepancy between them should be consistent across the composite domain (image domain) and text domain.

We formulate cross-domain semantic discrepancy alignment into two parts, namely neighbour consistency and distance consistency. Given the original composed query and augmented composed queries as input, the neighbour consistency means that the neighbour structure captured by similarity vectors in the composite domain and text domain should be similar. The distance consistency refers to the alignment of difference feature calculated by direct feature distance in the composite domain and text domain. The distance consistency ensures first-order alignment between the original composed query and augmented composed queries, while neighbour consistency aligns them in higher-order relationship.

Overall, we propose a general boosting framework for the text-conditioned image retrieval by exploring semantic discrepancy alignment, in which we leverage the capabilities of ChatGPT and GPT-4 to generate qualified augmented composed queries and corresponding images. The experiments on three popular datasets demonstrate the effectiveness of our framework, which achieves state-of-the-art performance on most evaluation metrics on the three datasets.

2 Related Work

2.1 Image Retrieval

Although traditional content-based image retrieval (Radenović et al., 2018) (Ng et al., 2020) (Revaud et al., 2019) (Gordo et al., 2017) (Teichmann et al., 2019) has developed rapidly and achieved good results, it still suffers from a fundamental difficulty, namely intention gap. The intention gap means that a single query image is difficult to accurately convey the search intention of users. To express the search intention of users more accurately, multi-modal queries have been explored, such as text and video. The task of cross-modal retrieval has attracted a wide range of attention. Cross-modal retrieval focuses on mapping different modalities into a common space to align heterogeneous modalities (Chen et al., 2020a) (Kuang et al., 2019) (Edwards et al., 2021) (Fei et al., 2021) (Li et al., 2023) (Wu et al., 2021) (Zhan et al., 2020b) (Han et al., 2023). However, the retrieval intention expressed by a single modality is still not enough to handle all scenarios.

To take the advantages of multiple modal queries, especially text and image, TIRG (Vo et al., 2019) first proposes the text-conditioned image retrieval task. In this setting, the input query is specified in the form of an image with a modification text that describes desired modifications to the reference image, which combines the advantages of rich image semantic information and text flexibility. Many researchers devote to learning the joint expression of vision-language. LBF (Hosseinzadeh and Wang, 2020) uses off-the-shelf region proposal network (Ren et al., 2015) to represent the input image as a set of local regions. Then it explores the bidirectional correlation between the words in the modification text and local areas in the image. VAL (Chen et al., 2020b) uses multi-scale techniques to deeply explore the composition of image and text semantics at both low and high levels. DCNet (Kim et al., 2021) leverages both the local and global features of the reference image for composition.

In contrast to previous work focusing on the design of compositors, we propose a general boosting framework for text-conditioned image retrieval. Through training with our framework, our model can better capture the semantic discrepancy of visual and textual information in the fashion domain, while learning a more discriminative joint image-text representation.

2.2 Visiolinguistic Representation Learning

Learning the joint representation of image and text forms the foundation of many multimedia tasks, such as VQA (Anderson et al., 2018) (Zhan et al., 2020a) (Antol et al., 2015) (Singh et al., 2019) and image captioning (Vinyals et al., 2015) (Guo et al., 2019a). To learn more robust joint image-text representations, data augmentation techniques have been widely explored in the cross-modal community. In cross-modal retrieval (Chen et al., 2020a) (Kuang et al., 2019), given a query image and a corresponding text, regular data augmentation method replaces some words in the original text as a negative sample of the original query image. The data augmentation technique currently utilized in the text-conditioned image retrieval task (Vo et al., 2019) (Kim et al., 2021) (Lee et al., 2021) is to make the reference image feature robust to transformations, which applies a random transformation to the reference image.

Our framework also utilizes a cross-domain data augmentation strategy. We use ChatGPT to reasonably replace attributes in the modification text without the need for a word-level substitution strategy. By entering a suitable prompt for ChatGPT, we can simply generate natural and appropriate modification texts.

3 Methodology

The text-conditioned image retrieval aims to return the relevant target images with an image and a modification text sentence as a composed query. Let (I_q, M, I_t) represents the reference image, the modification text and a candidate target image, respectively. Our goal is learning a joint representation $f(I_q, M)$ which is similar with the representation $f_{target}(I_t)$.

In the following, we start by discussing composed query augmentation in Sec. 3.1. Then we introduce our framework in Sec. 3.2 and the details of the semantic discrepancy alignment loss in Sec. 3.3. Finally, we elaborate the training and inference procedures in Sec. 3.4.

3.1 Composed Query Augmentation

Given an original training tuple (I_q, M, I_t), we use the ChatGPT to modify the attributes of the original modification texts to obtain a new modification text M_{edit}. Concretely, we first input the vocabulary to ChatGPT and then enter a prompt: "Replace the attribute of clothes in the following sentence." Then we combine M_{edit} with the original reference image I_q as an augmented composed query (I_q, M_{edit}). We repeat above operation three times and obtain three augmented composed queries (I_q, M^l_{edit}), $l \in \{1, 2, 3\}$ for each original tuple.

The reason we use ChatGPT to rewrite the original modification text is that our goal is to help the compositor learn the projection of attributes from the text domain to the image domain. We aim to replace attributes in a sentence while maintaining the syntax, which allows the semantic discrepancy between the new and original modification text to be reflected in the keywords. ChatGPT achieves exactly what we want and outputs reasonable generated modification texts. We provide some examples in our appendix.

Then we using GPT-4 to generate corresponding images I^l_{edit}, $l \in \{1, 2, 3\}$ for the augmented composed query. Due to GPT-4's powerful image generation and text comprehension capabilities, we can effectively augment the dataset.

3.2 Framework

Figure 2 shows an overview of our framework. Existing text-conditioned image retrieval methods can be incorporated into our framework and achieve better performance. To simplify the presentation,

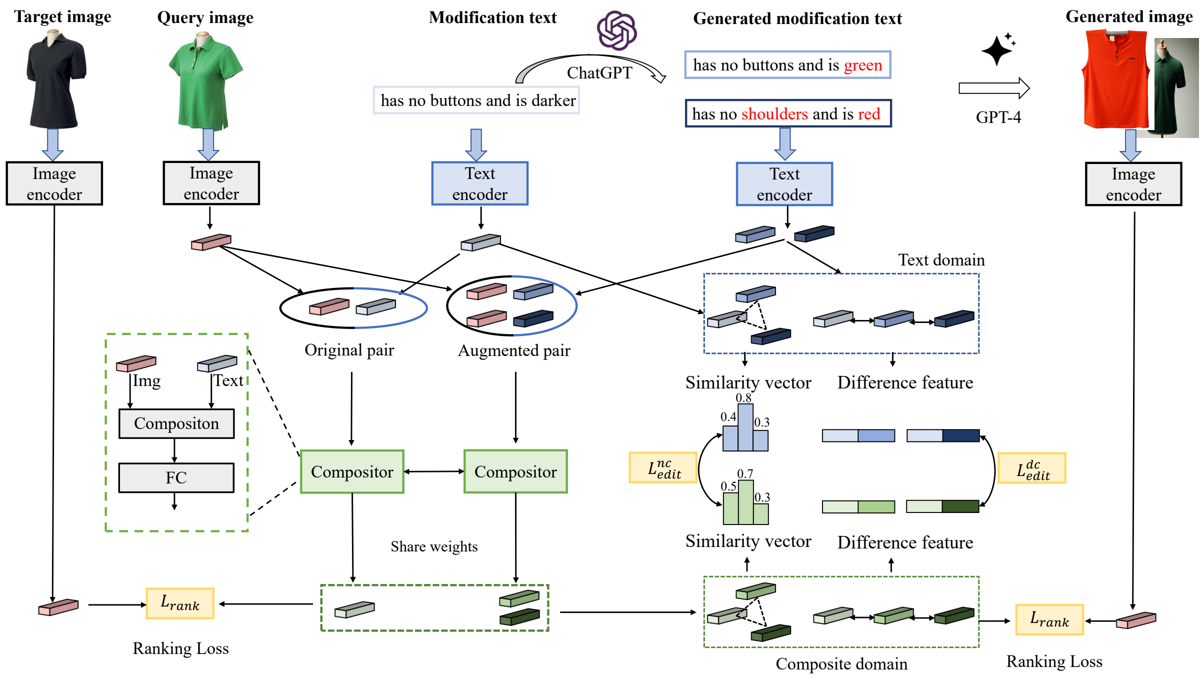


Figure 2: An overview of our framework. The original pair is consisted of the reference image and original modification text in the training set, and the augmented pairs are consisted of the reference image and generated modification texts. We believe that the semantic discrepancy should be consistent across the composite domain and the text domain. We aim to leverage the well-defined feature distribution in text domain to improve the compositor by exploring cross-domain consistency. We explore the cross-domain semantic discrepancy alignment by the distance consistency and neighbour consistency to ensure first-order and higher-order alignments.

we do not specify the network architecture used in the composed image retrieval method. Instead, we will focus on the key ingredient below.

Given an image encoder f_{img}, *e.g.* ResNet18 or ResNet50, the reference image feature V_q and target image feature V_t are formulated as $V_q = f_{img}(I_q)$ and $V_t = f_{img}(I_t)$. Similarly, given a text encoder f_{text}, *e.g,* LSTM (Hochreiter and Schmidhuber, 1997), the modification text feature T_{ori}, T_{edit} for M, M_{edit} can be obtaind by:

$$T_{ori} = f_{text}(M), \tag{1}$$

$$T_{edit}^{l} = f_{text}(M_{edit}^{l}), l \in \{1, 2, 3\}. \tag{2}$$

Then we feed the modification text feature T_{ori} and T_{edit} with the reference image feature V_q into the compositor to obtain the composite features:

$$\phi_{ori} = f_{comp}(V_q, T_{ori}), \tag{3}$$

$$\phi_{edit}^{l} = f_{comp}(V_q, T_{edit}^{l}), l \in \{1, 2, 3\}, \tag{4}$$

where f_{comp} is the compositor for image-text pair (f_{comp} takes image feature and text feature as input and outputs composite feature.) and can be flexibly instantiated by existing methods, such as TIRG (Vo et al., 2019), CoSMo (Lee et al., 2021), CLVC-Net (Wen et al., 2021)), *etc.* ϕ_{ori} and ϕ_{edit} are the composite features of the original composed query (I_q, M) and augmented composed queries (I_q, M_{edit}), respectively. These features are normalized before calculating the loss function.

3.3 Semantic Discrepancy Alignment Loss

Considering that the discrepancy between the original composed query (I_q, M) and the augmented composed query (I_q, M_{edit}) reflects the discrepancy of the modification text (M and M_{edit}), this semantic discrepancy should be consistent across the composite domain and text domain. We explore the semantic discrepancy alignment from two aspects, namely neighbour consistency and distance consistency.

Neighbour Consistency. To capture the structural knowledge in the composite domain and text domain separately, we calculate the KL divergence of the similarity vectors between the composed queries and modification texts. For ease of the following discussion, we introduce the variable ϕ_{edit}^{0}, which is equal to ϕ_{ori}. Similarly, we also introduce a other variable T_{edit}^{0}, which is equal to T_{ori}. To be specific, in the composite domain, given original composed query (I_q, M) and augmented composed queries (I_q, M_{edit}), the similarity vector is calculated as the dot product between each pair of them:

$$S^{p,q}_{comp} = \kappa(\phi^p_{edit}, \phi^q_{edit}), \;\; p,q \in \{0,1,2,3\}, \quad (5)$$

where κ is implemented as the dot product. Then, we concatenate all $S^{m,n}_{comp}$ and obtain S_{comp} as follows,

$$S_{comp} = [S^{0,1}_{comp}; S^{0,2}_{comp}; \cdots; S^{2,3}_{comp}]. \quad (6)$$

where ";" represents the concatenation.

Similarly, in the text domain, given original modification text M and generated modification text M_{edit}, the similarity vector is calculated as the dot product between each of them:

$$S^{p,q}_{text} = \kappa(T^p_{edit}, T^q_{edit}), \;\; p,q \in \{0,1,2,3\}. \quad (7)$$

Then, we concatenate all $S^{m,n}_{text}$ and obtain S_{text} as follows,

$$S_{text} = [S^{0,1}_{text}; S^{0,2}_{text}; \cdots; S^{2,3}_{text}]. \quad (8)$$

The similarity vectors represent higher-order relationship of original composed query and augmented composed queries. We expect this relationship should be consistent in both the composite domain and text domain. Hence, we compute the KL divergence of S_{comp} and S_{text} as the neighbour consistency loss:

$$\mathrm{L}^{nc}_{edit} = \sum_{i=1}^{K} KL(softmax(S_{comp_i}), softmax(S_{text_i})), \quad (9)$$

where K represents minibatch size, S_{comp_i} and S_{text_i} represent the composite domain and text domain similarity vectors for the i-th tuple in the batch, respectively.

Distance Consistency. Given the ϕ_{ori}, ϕ^l_{edit}, T_{ori} and T^l_{edit} as input, we first use two independent convolutional layers Θ_{comp} and Θ_{text} to calculate the domain difference features of composite domain and text domain:

$$\mathrm{D}^l_{comp} = \Theta_{comp}([\phi_{ori}; \phi^l_{edit}; \phi_{ori} - \phi^l_{edit}]), l \in \{1,2,3\}, \quad (10)$$

$$\mathrm{D}^l_{text} = \Theta_{text}([T_{ori}; T^l_{edit}; T_{ori} - T^l_{edit}]), l \in \{1,2,3\}, \quad (11)$$

where ";" represents the concatenation, D_{comp} and D_{text} are composite domain difference features and text domain difference features. Take Eq. (19) as an example, it consists of three components, the first two being the original and new composite features, and the third one is the direct subtraction, which is designed to reflect the difference of features in each channel dimension. Since the augmented composed queries (I_q, M_{edit}) and the original composed query (I_q, M) only differs in modification texts, D_{comp} should be similar to D_{text}. So we formulate the distance consistency loss for (I_q, M_{edit}) and (I_q, M) as follows:

$$\mathrm{L}^{dc}_{edit} = \frac{1}{3K} \sum_{l=1}^{3} \sum_{i=1}^{K} - \log \left\{ \frac{\exp(\kappa(D^l_{comp_i}, D^l_{text_i}))}{\sum_{j=1}^{K} \exp(\kappa(D^l_{comp_i}, D^l_{text_j}))} \right\}, \quad (12)$$

where K represents minibatch size, κ is an arbitrary similarity kernel function and is implemented as the dot product.

3.4 Training and Inference Procedures

Training. During the training stage, the whole framework is trained with the common ranking loss and our proposed semantic discrepancy alignment loss. Given a training minibatch B containing K triplets, each triplet consists of (I_{q_i}, M_i, I_{t_i}), which represents the i-th reference image, modification text and target image, respectively. We first obtain generated modification text for every triplet and obtain $3K$ augmented composed queries $(I_{q_i}, M^l_{edit_i}), l \in \{1,2,3\}$. Then we using GPT-4 to generate corresponding images I^l_{edit}, $l \in \{1,2,3\}$ for the augmented composed queries.

Then we calculate the semantic discrepancy alignment loss L^{nc}_{edit} and L^{dc}_{edit} as above mentioned.

As for the ranking loss, for ease of expression, we use V_{t_i} to represent the positive sample of the original composed query (I_{q_i}, M_i), which should be similar with ϕ_{ori_i}. Following TIRG, we consider the batch classification loss as ranking loss. The batch classification loss aims to reduce the distance between the query and positive sample meanwhile extending the distance between the query and negative sample. It aligns the pair (I_{q_i}, M_i) with the target image I_{t_i} through a batch-based classification, which assigns an independent label to each target image:

$$L_{rank} = \frac{1}{K} \sum_{i=1}^{K} - \log \left\{ \frac{\kappa(\phi_{ori_i}, V_{t_i})}{\sum_j^K \kappa(\phi_{ori_i}, V_{t_j})} \right\}. \quad (13)$$

Besides, the I^l_{edit}, $l \in \{1,2,3\}$ should be similar with ϕ^l_{edit}, $l \in \{1,2,3\}$:

$$L^{edit}_{rank} = \frac{1}{3K} \sum_{l=1}^{3} \sum_{i=1}^{K} - \log \left\{ \frac{\kappa(\phi^l_{edit}, I^l_{edit})}{\sum_j^K \kappa(\phi^l_{edit}, V_{t_j})} \right\}. \quad (14)$$

The overall loss to train the framework is formulated as follow:

$$L = L_{rank} + \alpha * L_{edit}^{nc} + \beta * L_{edit}^{dc} + \gamma * L_{rank}^{edit}, \quad (15)$$

where α, β, γ are learnable parameters and initialized with 1.

Since our main idea is to use the well-defined feature distribution in text domain to improve the compositor. We only update parameters of the compositor when training with the semantic discrepancy alignment loss. This means that the gradient returned from the semantic discrepancy alignment loss is truncated at the text encoder and image encoder to avoid confusion of information. In other words, we hope that the augmented composed queries are dedicated to boost the compositor to learn more diverse and robust image-text representations.

Inference. In the inference stage, we directly extract the composite feature ϕ_{ori} for each original composed query, and calculate its similarity with each database image to find the most similar target image.

4 Experiments

To verify the effectiveness of our framework, we conduct experiments on three benchmarks including FashionIQ (Guo et al., 2019b), Shoes (Berg et al., 2010) and Fashion200k (Han et al., 2017). In this section, we will introduce the implementation details, the experimental results and ablation studies in Sec. 4.1, Sec. 4.2 and Sec. 4.3, respectively.

4.1 Implementation Details

We propose a general boosting framework for existing text-conditioned image retrieval methods such as CoSMo (Lee et al., 2021) and CLVC-Net (Wen et al., 2021) by exploring semantic discrepancy alignment. We conduct the experiments in Pytorch (Paszke et al., 2019). The image encoder is ResNet-18 (He et al., 2016) for Fashion200k dataset and ResNet-50 (He et al., 2016) for FashionIQ and Shoes datasets. We adopt the output from layer 4 of the backbone networks as image feature. The text encoder is composed of an embedding layer and an LSTM (Hochreiter and Schmidhuber, 1997), followed by a single linear layer. The output of the embedding layer is a 512-dimensional vector, and the hidden size of LSTM is 1024. In the semantic discrepancy alignment loss, we implement Θ_{comp} and Θ_{text} as two 1×1 convolutional layers with output size 512. In the training stage, we use a rectified Adam (Liu et al., 2019) optimizer with a base learning rate of 0.0004, which decays once after 20 epochs by a factor of 10 and the batch size K is set to 32. We repeat each experiment five times and report the mean and deviation of results. For ChatGPT, we utilize the GPT-3.5 model (*i.e.*, textdavinci-003).

4.2 Experimental Results

FashionIQ Dataset. FashionIQ is a natural language-based interactive fashion product retrieval dataset. It contains 77,684 images, covering three categories: Dress, Toptee and Shirt. There are 18,000 image pairs in the 46,609 training images. Each pair is accompanied with around two natural language sentences as modification text. Compared to other datasets, the modification text in FashionIQ is more natural and complicated with an average length of 10.69 words.

Table 1 shows our results on FashionIQ. To verify the effectiveness of our proposed method, we conduct experiment following the setting of CoSMo, CLVC-Net and CLIP4Cir. When incorporating with CLVC-Net, our proposed method obviously outperforms the referred method in all the metrics on all categories, and with a 4.77% and 4.54% performance improvement in terms of the AvgRecall@10 and AvgRecall@50 metric, respectively. When incorporating with CLIP4Cir, our proposed method obtain a 1.67% and 2.95% performance improvement in terms of the AvgRecall@10 and AvgRecall@50 metric, respectively. Notably, our method improves the compositor while CLIP4Cir has relatively few parameters in compositor and relies more on the capabilities of the CLIP model itself. Hence, the improvement of our method on the CLIP4Cir is not as significant as that on CoSMo and CLVC-Net.

Besides, our method based on BERT have gained an improvement of 1.27%/1.20% in Recall@10 compared to based on LSTM. Thus we believe that a better feature distribution in the text domain can improve the performance of our method.

Remarkably, FashionVLP (Goenka et al., 2022) utilizes side information, including landmark detection and object detection models for the fashion datasets. Despite not utilizing any side information in our method, we still outperforms them in terms of performance.

Shoes Dataset. The Shoes dataset is originally proposed for attribute discovery. It consists of

Method	Dress		Toptee		Shirt		Avg	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Image Only	2.92	10.10	4.53	11.63	5.34	14.62	4.26	12.12
Text Only	8.67	25.08	9.68	28.25	8.30	25.02	8.88	26.11
Concat	9.06	27.27	10.45	29.83	9.66	28.06	9.72	28.33
TIRG (Vo et al., 2019)	14.87	34.66	19.08	39.62	18.26	37.89	17.40	37.39
VAL (Chen et al., 2020b)	21.12	42.19	25.64	49.49	21.03	43.44	22.60	45.04
MAAF (Dodds et al., 2020)	23.80	48.60	27.90	53.60	21.30	44.20	24.30	48.80
RTIC (Shin et al., 2021)	27.37	52.95	27.33	53.60	22.03	45.29	25.58	50.61
RTIC-GCN (Shin et al., 2021)	27.71	53.50	29.63	56.30	22.72	44.16	26.69	51.32
ComposeAE (Anwaar et al., 2021)	11.99	31.38	11.01	27.48	11.04	26.49	11.34	28.45
TRACE (Jandial et al., 2020)	26.13	52.10	31.16	59.05	26.20	50.93	27.83	54.02
CIRR (Liu et al., 2021)	17.45	40.41	21.64	45.38	17.53	38.81	18.87	41.53
DCNet (Kim et al., 2021)	28.95	56.07	30.44	58.29	23.95	47.30	27.78	53.89
ARTEMIS (Delmas et al., 2022)	27.16	52.40	29.20	54.83	21.78	43.64	26.05	50.29
FashionVLP* (Goenka et al., 2022)	32.42	60.29	38.51	**68.79**	31.89	58.44	34.27	62.51
AACL (Tian et al., 2023)	24.82	48.85	30.88	56.85	29.89	55.85	28.53	53.85
CoSMo (Lee et al., 2021)	25.64	50.30	29.21	57.46	24.90	49.18	26.58	52.31
CoSMo + Ours	27.85 ± 0.16	53.92 ± 0.20	33.74 ± 0.25	63.39 ± 0.31	28.87 ± 0.13	54.71 ± 0.32	30.15	57.34
CLVC-Net (Wen et al., 2021)	29.85	56.47	33.50	64.00	28.75	54.76	30.70	58.41
CLVC-Net + Ours	**33.47** ± 0.19	**60.56** ± 0.28	39.14± 0.26	68.50 ± 0.30	33.82 ± 0.27	59.80 ± 0.31	35.47	**62.95**
CLVC-Net + Ours w/ BERT	**34.91** ± 0.25	**61.40** ± 0.24	**40.07**± 0.20	**69.33** ± 0.27	35.02 ± 0.29	**61.28** ± 0.20	**36.67**	**64.00**
CLIP4Cir (Baldrati et al., 2022)	31.63	56.67	38.19	62.42	**36.36**	58.00	35.39	59.03
CLIP4Cir + Ours	32.85 ± 0.20	59.04 ± 0.37	**40.41** ± 0.28	66.39 ± 0.41	**37.93** ± 0.24	**60.52** ± 0.36	**37.06**	61.98

Table 1: Retrieval performance on the FashionIQ official validation set under VAL evaluation protocols. * denotes the use of additional side information (*e.g.* landmark detection) during training. "w/ BERT" denotes using BERT (Kenton and Toutanova, 2019) as text encoder. The "Avg" column refers to the average results on three categories. Overall $1^{st}/2^{nd}$ in **black/blue**

10,000 training queries and 4,658 validation examples. The modification texts on this dataset are also artificially annotated and has a format similar to FashionIQ. According to Table 2, we conduct our experiments following the setting of CoSMo and CLVC-Net, and the Recall@10 is improved by about 2.06%/1.88% when incorporating these two methods into our framework. Our method with CLVC-Net outperforms ARTEMIS (Delmas et al., 2022) by 3.16% on the Recall@10 and 1.86% on the Recall@50.

Method	Shoes		
	R@1	R@10	R@50
TIRG (Vo et al., 2019)	7.89	26.53	51.05
VAL (Chen et al., 2020b)	16.49	49.12	73.53
RTIC (Shin et al., 2021)	–	43.66	72.11
RTIC-GCN (Shin et al., 2021)	–	43.38	72.09
ComposeAE (Anwaar et al., 2021)	3.46	20.84	52.58
TRACE (Jandial et al., 2020)	18.11	52.41	75.42
DCNet (Kim et al., 2021)	–	53.82	79.33
ARTEMIS (Delmas et al., 2022)	**18.72**	53.11	79.31
FashionVLP* (Goenka et al., 2022)	–	49.08	77.32
CoSMo (Lee et al., 2021)	16.72	48.36	75.64
CoSMo + Ours	17.81 ± 0.31	50.42 ± 0.29	78.55 ± 0.39
CLVC-Net (Wen et al., 2021)	17.64	**54.39**	**79.47**
CLVC-Net + Ours	**18.43** ± 0.25	**56.27** ± 0.29	**81.17** ± 0.32

Table 2: Retrieval performance on the Shoes dataset. * denotes the use of additional side information during training. Overall $1^{st}/2^{nd}$ in **black/blue**

Fashion200K Dataset. Fashion200K is a diverse dataset consisting of about 200K clothes images of various styles. Each image is equipped with some tags describing attributes. The modification text is automatically generated on this dataset.

Method	Fashion200k		
	R@1	R@10	R@50
TIRG (Vo et al., 2019)	14.1	42.5	63.8
JGAN (Zhang et al., 2020)	17.3	45.2	65.7
LBF (Hosseinzadeh and Wang, 2020)	17.8	48.4	68.5
VAL (Chen et al., 2020b)	21.2	49.0	68.8
MAAF (Dodds et al., 2020)	18.9	–	–
ComposeAE (Anwaar et al., 2021)	22.8	**55.3**	**73.4**
DCNet (Kim et al., 2021)	–	46.9	67.6
FashionVLP* (Goenka et al., 2022)	–	49.9	70.5
AACL (Tian et al., 2023)	19.64	52.3	71.0
CoSMo (Lee et al., 2021)	23.3	50.4	69.3
CoSMo + Ours	**24.6** ± 0.19	51.5 ± 0.23	70.2 ± 0.39
CLVC-Net (Wen et al., 2021)	22.6	53.0	72.2
CLVC-Net + Ours	**24.7** ± 0.33	**54.8** ± 0.22	**73.0** ± 0.39

Table 3: Retrieval performance on the Fashion200K dataset. * denotes the use of additional side information during training. Overall $1^{st}/2^{nd}$ in **black/blue**

We use the training split of around 172k images for training and the testset of 33,480 test queries for evaluation. As shown in Table 3, we have gained an improvement of 1.3% in Recall@1 compared to CoSMo and 2.1% in Recall@1 compared to CLVC-Net.

4.3 Ablation Studies

In this subsection, we conduct ablation studies to analyze the influence of semantic discrepancy alignment loss, the text generation strategies and the number of the augmented composed queries L of our proposed method. Particularly, we conduct experiments on FashionIQ based on CLVC-Net and use the same evaluation metric as before. We have included more ablation studies in our supplementary file.

distance consistency	neighbour consistency	L_{edit}	AvgR@10	AvgR@50
×	×	×	30.70	58.41
×	✓	×	30.65	58.97
×	✓	✓	31.98	60.32
✓	×	×	31.44	60.15
✓	×	✓	34.70	61.53
✓	✓	✓	**35.47**	**62.95**

Table 4: Ablation study on semantic discrepancy alignment loss of our proposed method.

Effect of Semantic Discrepancy Alignment Loss. In our method, we calculate two semantic discrepancy alignment loss $L_{edit}^{nc,dc}$ by exploring neighbour consistency and distance consistency. We make an ablation experiment to study on their impact.

Our experimental results are presented in Table 4. We observe that $L_{edit}^{nc,dc}$ both obviously improve the performance. This reflects that for a model with strong generalisability, it is necessary to understand not only the substitution of semantics but also the lack of semantics. Besides, we observe that without the distance consistency loss or the neighbour consistency loss, the performance both degrades. It reveals that both distance consistency and neighbour consistency delivers improvements to our framework. .

Study of Text Generation Strategies. To investigate the impact of text generation strategies, we make a comparison with a number of automated word replacement strategies. we also try both random substitution of arbitrary words and random substitution of semantically similar words, referred to as "Arbitrary" and "Semantically similar", respectively. Specifically, semantically similar words are defined as any of the Top 10 words with the highest similarity between word embedding. We make an ablation experiment based on CLVC-Net to study the impact of text generation strategies.

Our experimental results are presented in Table 5. We observe that the strategy of random substitution of arbitrary words leads to a performance degradation of 0.87% in AvgR@10 compared to the baseline model (CLVC-Net). And ChatGPT generation clearly outperforms the automatic word substitution strategies. This illustrates that it is important to make the generated modification text conform to common sense.

Effect of The Augmented Composed Query Number L. In Sec. 3.1, we construct multiple augmented composed queries to study cross-modal semantic discrepancies. Here we make an ablation study based on CLVC-Net to investigate the effect of the number of the augmented composed queries L. As shown in Table 6, we did experiments to study the effect of L from 0 to 5 (0 represents the baseline), and the best result was achieved when L was 3. We believe that an appropriate L can better reveal the semantic discrepancies of the composed queries in different modalities. However, a too large L will make learning more difficult. As a result, we set L to 3 to strike a balance.

Method	AvgR@10	AvgR@50
Baseline	30.70	58.41
Arbitrary	29.83	58.02
Semantically similar	31.67	60.08
ChatGPT	**35.47**	**62.95**

Table 5: Ablation study on text generation strategies.

Numbers	AvgR@10	AvgR@50
0	30.70	58.41
1	31.75	59.33
2	33.39	61.01
3	**35.47**	**62.95**
4	34.55	62.28
5	32.77	60.96

Table 6: Ablation study on the augmented composed query number L.

5 Conclusions

We propose a general boosting framework for the text-conditioned image retrieval task by exploring semantic discrepancy alignment. By leveraging the strong capability of ChatGPT and GPT-4, we generate suitable modification texts to construct augmented composed queries with corresponding images. In our framework, we capture the semantic discrepancy alignment by introducing two novel losses: neighbour consistency and distance consistency. We leverage the well-defined feature distribution in text domain to improve the ability of the compositor and further ensure the first-order and higher-order alignment between composite domain and text domain. Through extensive experiments, we demonstrate that our proposed method achieves a new state-of-the-art performance on most datasets.

6 Limitations

Since our method involves data augmentation using ChatGPT and GPT-4, this will incur additional overhead. It will also limit the generalisation of our approach to larger application scenarios.

Acknowledgements

This work is supported by National Key R&D Program of China (No. 2022ZD0162101), National Natural Science Foundation of China (No. 62106140) and STCSM (No. 21511101100, No. 22DZ2229005).

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. 2021. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1140–1149.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21466–21474.

Tamara L Berg, Alexander C Berg, and Jonathan Shih. 2010. Automatic attribute discovery and characterization from noisy web data. In *Proceedings of the European Conference on Computer Vision*, pages 663–676.

Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020a. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10638–10647.

Yanbei Chen, Shaogang Gong, and Loris Bazzani. 2020b. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3001–3011.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. 2023. Chataug: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. 2022. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. *arXiv preprint arXiv:2203.08101*.

Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. 2020. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607.

Hongliang Fei, Tan Yu, and Ping Li. 2021. Cross-lingual cross-modal pretraining for multimodal retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3644–3650.

Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. 2022. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14105–14115.

Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. 2017. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, pages 237–254.

Longteng Guo, Jing Liu, Jinhui Tang, Jiangwei Li, Wei Luo, and Hanqing Lu. 2019a. Aligning linguistic words and visual semantic units for image captioning. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 765–773.

Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven Rennie, and Rogerio Feris. 2019b. Fashion iq: A new dataset towards retrieving images by natural language feedback. *arXiv preprint arXiv:1905.12794*.

Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. 2017. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1463–1471.

Yunpeng Han, Lisai Zhang, Qingcai Chen, Zhijian Chen, Zhonghua Li, Jianxin Yang, and Zhao Cao. 2023. Fashionsap: Symbols and attributes prompt for fine-grained fashion vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15028–15038.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Mehrdad Hosseinzadeh and Yang Wang. 2020. Composed query image retrieval using locally bounded features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3596–3605.

Surgan Jandial, Ayush Chopra, Pinkesh Badjatiya, Pranit Chawla, Mausoom Sarkar, and Balaji Krishnamurthy. 2020. Trace: Transform aggregate and compose visiolinguistic representations for image search with text feedback. *arXiv preprint arXiv:2009.01485*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. 2021. Dual compositional learning in interactive image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1771–1779.

Zhanghui Kuang, Yiming Gao, Guanbin Li, Ping Luo, Yimin Chen, Liang Lin, and Wayne Zhang. 2019. Fashion retrieval via graph reasoning networks on a similarity pyramid. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3066–3075.

Seungmin Lee, Dongwan Kim, and Bohyung Han. 2021. Cosmo: Content-sstyle modulation for image retrieval with text feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 802–812.

Zejun Li, Zhihao Fan, Jingjing Chen, Qi Zhang, Xuan-Jing Huang, and Zhongyu Wei. 2023. Unifying cross-lingual and cross-modal modeling towards weakly supervised multilingual vision-language pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5939–5958.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*.

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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 2125–2134.

Tony Ng, Vassileios Balntas, Yurun Tian, and Krystian Mikolajczyk. 2020. SOLAR: Second-order loss and attention for image retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 253–270.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Filip Radenović, Giorgos Tolias, and Ondřej Chum. 2018. Fine-tuning cnn image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems*, pages 91–99.

Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. 2019. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5107–5116.

Minchul Shin, Yoonjae Cho, Byungsoo Ko, and Geonmo Gu. 2021. Rtic: Residual learning for text and image composition using graph convolutional network. *arXiv preprint arXiv:2104.03015*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Marvin Teichmann, Andre Araujo, Menglong Zhu, and Jack Sim. 2019. Detect-to-retrieve: Efficient regional aggregation for image search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5109–5118.

Yuxin Tian, Shawn Newsam, and Kofi Boakye. 2023. Fashion image retrieval with text feedback by additive attention compositional learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1011–1021.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6439–6448.

Haokun Wen, Xuemeng Song, Xin Yang, Yibing Zhan, and Liqiang Nie. 2021. Comprehensive linguistic-visual composition network for image retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1369–1378.

Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu. 2021. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4730–4738.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

Yuchen Yang, Min Wang, Wengang Zhou, and Houqiang Li. 2021. Cross-modal joint prediction and alignment for composed query image retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3303–3311.

Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

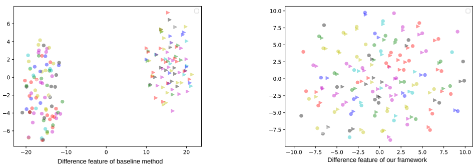
Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. 2020a. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2345–2354.

Yu-Wei Zhan, Xin Luo, Yongxin Wang, and Xin-Shun Xu. 2020b. supervised hierarchical deep hashing for cross-modal retrieval. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3386–3394.

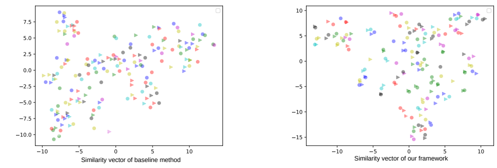
Feifei Zhang, Mingliang Xu, Qirong Mao, and Chang-sheng Xu. 2020. Joint attribute manipulation and modality alignment learning for composing text and image to image retrieval. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3367–3376.

A Appendix

Visualization of Similarity Vectors and Difference features. To analyze the distribution of similarity vectors and difference features obtained from the original composed query and augmented composed queries, we use t-SNE (Van der Maaten and Hinton, 2008) to display visualization results on the testset of FashionIQ. As shown in Fig. 3, we sample 100 original composed queries on the test set and generate corresponding augmented composed queries for each pair as we discussed in the main paper. We then compute the similarity vectors and difference features as Eq. (10)(18)(19)(20) in the composite domain (represented by circles) and text domain (represented by triangles) and use t-SNE to visualize them in a two-dimensional space (the same colour indicates the corresponding pair).



(a) Visualization results of difference features on FashionIQ



(b) Visualization results of similarity vectors on FashionIQ

Figure 3: Visualization results of difference features and similarity vectors on FashionIQ. In both (a) and (b), the left figure represents the baseline method while the right figure represents “CLVC-Net + Ours”. Similarity vectors and difference features are calculated as Eq. (10)(18)(19)(20).

For an in-depth analysis of this figure, since our motivation is that semantic discrepancies should be consistent across the two domains, we argue that the similarity vectors and difference features should exhibit a similar distribution in both domains. We observe that the vast majority of the similarity vectors and difference features in our framework are pairwise matching, while this pairwise matching relationship is not maintained in the baseline method (we adopt the CLVC-Net as baseline method), which means that our framework aligns the semantic discrepancy across the two domains. These results validate the effectiveness of our cross-domain semantic discrepancy alignment optimization objective.

Effect of Stop Gradient Training Strategy. We only update parameters of the compositor when

Original modification text	Generated modification text
Has no buttons and is darker.	Is brighter and has buttons. Is lighter and has no pockets. Has no shoulders and is red.
The shirt is gray with a floral design and is lighter in color.	The shirt is black with a polka dot design and are darker in color. The shirt is blue with a geometric design and is brighter in color. The shirt is green with a striped design and is darker in color.
Is darker and more sporty.	Is more elegant and less dark. Is more colorful and casual. Is lighter and more formal.
Is less casual and white checkered button up shirt.	Is a relaxed and black striped polo shirt. Is a more formal and black striped button-up shirt. Is a brighter and floral print short-sleeved shirt.
Has short sleeves and has a peasant neckline.	Has a v-neckline and has long sleeves. Has no sleeves and has a turtleneck neckline. Has a scoop neckline and has long sleeves.
Are more solid in black with orange and-beige trim.	Are more cozy in beige with black and orange trim. Are more classic in beige with black and orange trim. Are more lightweight in gray with blue and yellow trim.

Table 7: Examples of generated modification texts of ChatGPT.

training with the semantic discrepancy alignment loss. This means that the gradient returned from the semantic discrepancy alignment loss is truncated at the image encoder and text encoder to avoid confusion of information. We make an ablation study on effect of this stop gradient training strategy. As shown in Table 8, we observe that without this stop gradient training strategy, the overall performance degrades 4.76% on R@10.

Method	AvgR@10	AvgR@50
Ours w/o stop	26.59	52.46
Ours	**30.15**	**57.34**

Table 8: Ablation study on effect of stop gradient training strategy (based on CoSMo).

Examples of Generated Modification Texts of ChatGPT. As shown in Table 7, ChatGPT generate smooth and reasonable modification texts based on the original modification text while making random changes to the attributes. This allows us to better transfer the semantics of the text domain to the image domain in order to learn a more discriminative joint image-text representation.