Aligning Composed Query with Image via Discriminative Perception from Negative Correspondences

Yifan Wang[1], **Wuliang Huang**[2], **Chun Yuan**[1]*

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Institute of Computing Technology, Chinese Academy of Sciences
yifan-wa22@mails.tsinghua.edu.cn, huangwuliang19b@ict.ac.cn, yuanc@sz.tsinghua.edu.cn

Abstract

The task of composed image retrieval aims to match the multi-modal query composed of a reference image and a modification sentence with the target image. Most current approaches narrow the distances between the composed queries and targets by investigating matched correspondences in positive triplets. Nevertheless, they are inclined to exhibit heavy reliance on partial correlations. As the negative correspondences are underestimated, semantic clues that distinguish the target from mismatched candidates are obscured by incomplete associations. Moreover, the correlations between the modification textual features and the visual variations from the reference to candidates are imperative to further strengthen the semantic discriminations. In this paper, we propose *DIscriminative Perception from NEgative Correspondences* (DIPNEC) to address the aforementioned issues. To encourage awareness of the differences between matched and mismatched correspondences, DIPNEC introduces optimal transport with semantic preservation for reassignments on hard negative triplets. Besides, Difference Quantization Alignment (DQA) and Composed Word-level Alignment (CWA) jointly determine the matching scores between multi-modal queries and candidates. Specifically, DQA concentrates on the correlations of textual features with source-to-target visual differences, and CWA further emphasizes the differentiated semantics. DIPNEC has demonstrated competitive performances on the experimental results and ablation studies on widely-used datasets FashionIQ and CIRR.

Introduction

As the demand for flexible multi-modal retrieval rises in the era of massive data, composed image retrieval (CIR) (Vo et al. 2019) has recently been one research focus. Integrating with the need for image retrieval and cross-modal retrieval, CIR is framed as retrieving the matched images from the vision gallery with the hybrid-modal query of reference images and modification sentences. Hence, semantic correlations and comprehension on the query composed of different modalities are crucial to this matching task, and the progress could encourage the development in related domains, *e.g.*, visual reasoning (Gupta and Kembhavi 2023),

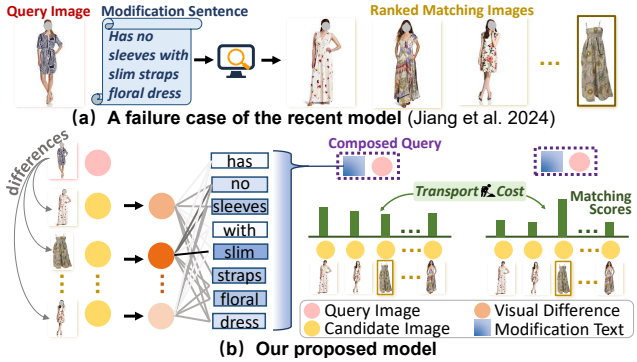

Figure 1: Illustration of our model compared with previous model. (a) A typical failure case that is misguided by partial correlations and overlooks semantics like *"slim"*. (b) DIPNEC enhances discriminative semantics by comparing words to visual differences and reassigns matching scores optimized by transport cost.

image captioning (Tu et al. 2024), and interactive dialogue system (Guo et al. 2023; Jia, Zhang, and Peng 2024).

As an emerging task to handle visual and linguistic understanding, CIR remains challenging for the textual modifiers exhibit partial associations with the references and target images. Semantic inconsistencies within these triplets lead to entangled alignment when matching the hybrid-modal queries with target images. Furthermore, inherited from multi-modal retrieval, CIR necessitates addressing distribution discrepancies across diverse modalities. Thus, the mainstream approaches focus on compositional learning (Wen, Gu, and Cheng 2021) to generate the composed query through reweighting modality contributions (Huang et al. 2024) or aggregating critical attributes (Zhao, Song, and Jin 2022), to enhance understanding on the query intentions. However, the lack of direct guidance from the targets when composing queries results in insufficient semantic learning. Recent models (Wen et al. 2023; Jiang et al. 2024) alleviate this issue by leveraging the alignment with target features.

In general, previous methods are primarily devoted to extracting correspondences from the matched triplets. However, the unaligned relationships for negative triplets have not been thoroughly investigated, which may trigger de-

*Corresponding Author

clines in generalization due to misguidance of partial correspondences. As illustrated in Figure 1(a), the retrieval model overlooked the inconsistency and became over-confident based on the biased alignments (*e.g.*, "no sleeves" and "floral dress"), which caused vulnerability to the distraction from the negative samples. Recently optimal transport has demonstrated promising performances in correcting the overconfidence in prediction distributions (Lu et al. 2023), yet directly exploiting this theory on CIR may introduce noise to the similarity distributions. Besides, textual modifiers, as the crucial components to clarify the visual differences between queries and targets in natural language, are usually underestimated as complements for the hybrid-modal query compositions and deserve more concentration. For instance, the existing model in Figure 1 is not sufficiently sensitive to the detailed semantics concerning "slim" in the textual features, resulting in oversight of the key information when composing query representations. Thus, we argue the perception of unaligned correspondences for negatives and textual discrimination are essential to improve the model robustness.

To this end, we propose a novel framework named *DIscriminative Perception from NEgative Correspondences* (DIPNEC) as shown in Figure 2. Given the encoded features, the matching score is synthesized by *difference quantization alignment* (DQA) and *composed word-level alignment* (CWA) to enhance the textual representations via discrimination learning from the negative samples (Figure 1(b)). To measure the consistency between the textual modifications and the visual changes, DQA is summarized from the semantic-aware association matrix reflecting the distances in the subspaces, and CWA further exploits the association matrix to estimate the significance of textual fragments. Furthermore, we leverage the *optimal transport with semantic preservation* to optimize the distributions for the aligned and unaligned relationships based on the transport costs. With the designed mask to avoid interferences on the established semantic correlations, the matching scores between the multi-modal queries and candidates adaptively adjust towards the optimized assignments, as demonstrated in FashionIQ (Wu et al. 2021) and CIRR (Liu et al. 2021). In summary, the main contributions are summarized as follows:

- We propose a novel discriminative perception approach to alleviate the over-confidence in the matching scores and enhance the distinctive features in textual modifiers.
- We design *difference quantization alignment* to highlight the correlations of textual modifications with visual variations and *composed word-level alignment* with emphasis on semantic distinguishments to comprehensively assess the matching scores.
- To perceive differences between aligned and unaligned correspondences, we optimize the transport cost on matched and mismatched samples with a semantic preservation mask to guide the similarity distributions.

Related Works

Composed Image Retrieval

The task of composed image retrieval combines the needs of unimodal image retrieval and text-based image matching to facilitate the interactive multi-modal retrieving process in real-life scenarios. With the sentences clarifying the modifications for the reference image, the queries are injected with cross-modal semantics that are coherent with the target images. Most existing models (Chen et al. 2024b; Chen, Zhou, and Peng 2024) strived for the composition of the visual and textual representations through multi-modal fusion (Zhang et al. 2021) to capture high-level interactions and powerful encoders to enhance the feature representations (Baldrati et al. 2022; Liu et al. 2024). For example, Huang *et al.* (Huang et al. 2024) introduced an editable modality de-equalizer to measure the importances of the modalities within the query. To narrow the distances between query and target features, models such as (Chen and Lai 2023; Zhang et al. 2024) resorted to data augmentation strategies to simulate the jittering of the target features in the limited domain and reconstructed the modifiers in the query respectively. Focused on the alignment between the composed query and the target, Delmas *et al.* (Delmas et al. 2022) exploited explicit and implicit matching, and TG-CIR (Wen et al. 2023) further improved it by distillation from targets. Despite progress, the semantic inconsistency between negative triplets are not fully explored, which hampers the model robustness. In this work, we introduce optimal transport with semantic preservation to percept the misalignment and enhance semantic discrimination by quantifying differences.

Optimal Transport

Optimal transport, also known as the earth mover distance, is defined as seeking the transport plan from the source to the target distributions with minimal costs. Flexibly measuring the distances while retaining the original structure information enables the extensive utilization of optimal transport in a variety of applications, *e.g.*, generation models, detection and unsupervised matching (Wang et al. 2024; Li et al. 2024). Kantorovich reformulated the original optimal transport mapping as optimizing a convex objective function and the Sinkhorn algorithm (Cuturi 2013) was utilized as an efficient solution with entropy regularization to minimize the weighted sum of cost and entropy. To extend the formulations of cost functions, GNOT (Asadulaev et al. 2024) addressed the continuous OT approach for image-to-image translation. To mitigate the distribution imbalances, UCLR (Li et al. 2024) combined K-means prototypes with optimal transport to encourage knowledge transferring across different domains, and SALAD (Izquierdo and Civera 2024) utilized optimal transport with dustbins to aggregate local features. In this work, DIPNEC exploits optimal transport with semantic mask concentrating on hard sample mining to re-assign distribution of matching scores in the composed image retrieval task.

Methodology

Problem Definitions

In the setting of composed image retrieval, the query is constituted of a reference image and a corresponding modification sentence, denoted as $\mathcal{I}^q$ and $\mathcal{T}^q$, respectively. We aim to retrieve the matched image $\mathcal{I}^t$ which fits the requirements of

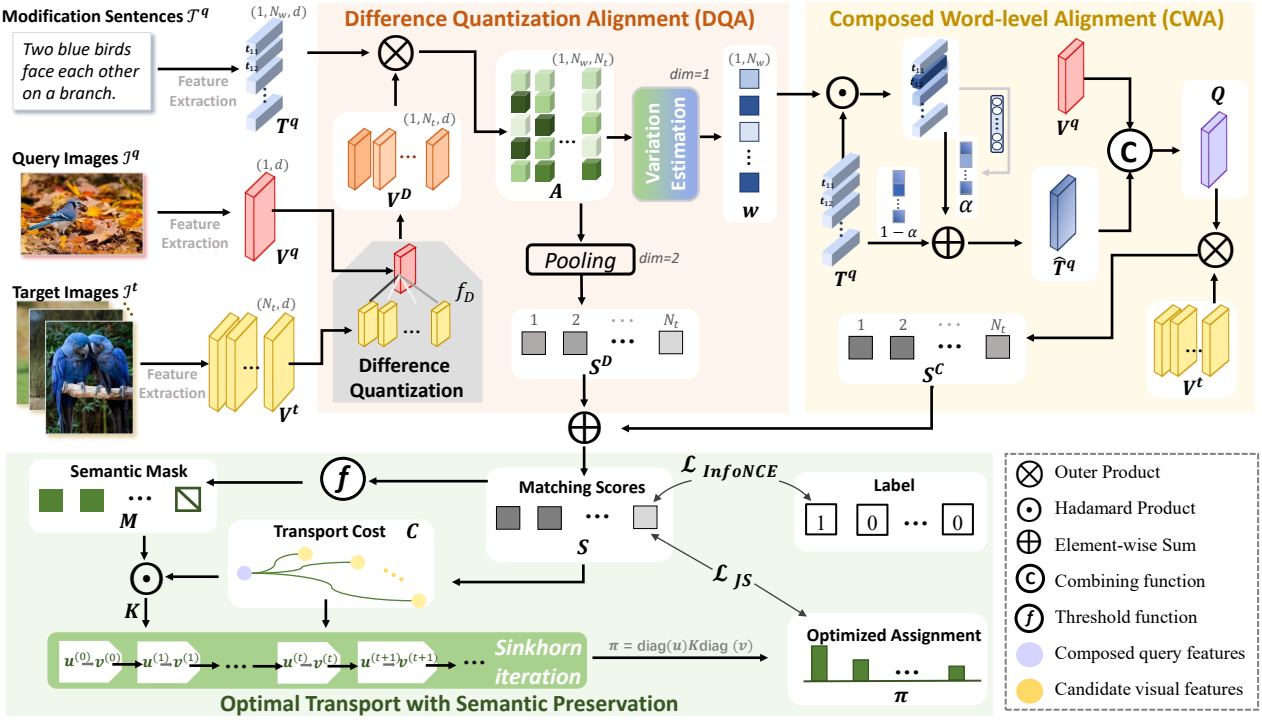

Figure 2: An overview of the proposed DIPNEC. Based on the encoded features, Difference Quantization Alignment and Composed Word-level Alignment together constitute the matching scores for the hybrid-modal query and candidate images. Afterwards, Optimal Transport with Semantic Preservation re-assigns the similarities especially on hard negatives.

the composed query. The matched triplets are formulated as $(\mathcal{I}_i^q, \mathcal{T}_i^q, \mathcal{I}_i^t)$ and mismatched ones as $(\mathcal{I}_i^q, \mathcal{T}_i^q, \mathcal{I}_j^t)_{j\neq i}$. N_q and N_t denote the number of queries and targets, respectively. To capture the intrinsic alignment, we design discriminative perception from the negative correspondences framework in Figure 2 (showing $N_q = 1$ for simplicity). For the reference images and target images, shared CLIP (Radford et al. 2021) visual encoders are exploited to obtain the global visual features, denoted as $V^q \in \mathbb{R}^d$ for $\mathcal{I}^q$ and $V^t \in \mathbb{R}^d$ for $\mathcal{I}^t$, respectively. To further explore the conclusive component of modifiers, we obtain the word-level features $T^q = \{t_k | k \in [1, N_w], t_i \in \mathbb{R}^d\}$ from the textual encoders for modification sentences $\mathcal{T}^q$, where N_w is the length of the sentence.

Different from previous methods directly using matching scores for ranking, the proposed DIPNEC emphasizes optimal transport with semantic preservation to refine similarity distributions and highlights the discriminative semantics, which will be illustrated in the following sections.

OT Assignment with Semantic Preservation

Preliminaries. Given two data distributions $\boldsymbol{p}$ and $\boldsymbol{q}$, the optimal transport theory is designed to seek the transport plan from $\boldsymbol{p}$ to $\boldsymbol{q}$ at the minimal cost. This process could be formulated as solving the coupling matrix $\boldsymbol{\pi}$ with the restriction that all the mass of $\boldsymbol{p}$ ought to be equal to the mass of $\boldsymbol{q}$. Based on the cost function $C \in \mathbb{R}_+^{m\times n}$, the optimization could be represented as:

$$\min_{\boldsymbol{\pi}\in\prod(\boldsymbol{p},\boldsymbol{q})} \langle \boldsymbol{\pi}, \boldsymbol{C} \rangle_F, \tag{1}$$

where $\prod(\boldsymbol{p},\boldsymbol{q}) = \{\boldsymbol{\pi} \in \mathbb{R}_+^{m\times n} | \boldsymbol{\pi}\mathbb{1}_n = \boldsymbol{p}, \boldsymbol{\pi}^\top \mathbb{1}_m = \boldsymbol{q}\}$. m and n are the lengths of distributions $\boldsymbol{p}$ and $\boldsymbol{q}$, respectively. $\mathbb{1}_D$ denotes all 1 vector in D dimensions. $\langle \cdot, \cdot \rangle_F$ refers to the Frobenius dot product.

The above objective function in Eq. 1 with restrictions is essentially a linear programming problem, which causes difficulty and high time complexity when solving high-dimensional data distributions. For effectively solving the transport problem, introducing an entropy regularization term on the objective function could achieve an approximate solution through solving in a smooth feasible region (Cuturi 2013), which transforms the objective function as:

$$\begin{aligned} &\min_{\boldsymbol{\pi}\in\prod(\boldsymbol{p},\boldsymbol{q})} \langle \boldsymbol{\pi}, \boldsymbol{C} \rangle_F - \epsilon H(\boldsymbol{\pi}), \\ &s.t. \boldsymbol{\pi} \in \mathbb{R}_+^{m\times n}, \boldsymbol{\pi}\mathbb{1}_n = \boldsymbol{p}, \boldsymbol{\pi}^\top \mathbb{1}_m = \boldsymbol{q}, \end{aligned} \tag{2}$$

where $H(\boldsymbol{\pi}) = -\sum_{ij} \boldsymbol{\pi}_{ij} \log \boldsymbol{\pi}_{ij}$ is the entropy regularization. Though constructing the Lagrangian function and computing the first-order condition, the optimized $\boldsymbol{\pi}^*$ could be computed as:

$$\boldsymbol{\pi}^* = diag(\boldsymbol{u}) K diag(\boldsymbol{v}), \tag{3}$$

where $K = e^{-C/\epsilon} \in \mathbb{R}_+^{m\times n}$. $\boldsymbol{u} \in \mathbb{R}_+^m$ and $\boldsymbol{v} \in \mathbb{R}_+^n$ are calculated through iterations:

$$\boldsymbol{u}^{(t+1)} = \frac{\boldsymbol{p}}{K\boldsymbol{v}^t}, \quad \boldsymbol{v}^{(t+1)} = \frac{\boldsymbol{q}}{K^\top \boldsymbol{u}^{(t+1)}}. \tag{4}$$

Semantic Preservation during Optimization. To suppress the misleading effects of noise and outliers in the primary matching score, optimal transport could be exploited to adaptively adjust the distances between the query and target based on the similarity distributions. Practically, given the reference images $\mathcal{I}^q$, modifiers $\mathcal{T}^q$ and target images $\mathcal{I}^t$, the transport cost is supposed to have a negative correlation with the similarity score for the matched triplet and impose penalty on the negative corrrespondences, defined as:

$$\boldsymbol{C}_{ij} = \begin{cases} 1 - S(\mathcal{I}_i^q, \mathcal{T}_i^q, \mathcal{I}_j^t), & j = i \\ S(\mathcal{I}_i^q, \mathcal{T}_i^q, \mathcal{I}_j^t), & j \neq i \end{cases} \tag{5}$$

where $S(\mathcal{I}_i^q, \mathcal{T}_i^q, \mathcal{I}_i^t) \in [0, 1]$ denotes the similarity score between the i-th query and j-th target and the specific calculation would be described in the next section.

The above optimization could guide a more reliable assignments for similarity score. However, introducing entropy regularization in Eq. 2 may encourage average distributions to some extent. To enhance the awareness on the similarity of hard negative triplets that exhibit high matching scores with the query and preserve the learned semantic relationships simultaneously, we exploit an optimization mask to prompt the transport assignments concentrating on the difficultly distinguished samples. Hence, the optimization mask M is designed as:

$$M_{ij} = \begin{cases} 1, & S(\mathcal{I}_i^q, \mathcal{T}_i^q, \mathcal{I}_j^t) > \max\limits_K \{S(\mathcal{I}_i^q, \mathcal{T}_i^q, \mathcal{I}_j^t)\}_{j=1}^{N_t} \\ 0, & S(\mathcal{I}_i^q, \mathcal{T}_i^q, \mathcal{I}_j^t) \leq \max\limits_K \{S(\mathcal{I}_i^q, \mathcal{T}_i^q, \mathcal{I}_j^t)\}_{j=1}^{N_t}, \end{cases} \tag{6}$$

where $\max_K$ means K largest values.

Similarly, we follow the solution in Eq. 2 and 3 and yield the assignment with the optimization mask within the mini-batch B, as:

$$\boldsymbol{\pi}^* = diag(\boldsymbol{u}) K diag(\boldsymbol{v}), K = M \odot e^{-C/\epsilon} \in \mathbb{R}_+^{B \times B}. \tag{7}$$

After acquiring the optimal assignment $\boldsymbol{\pi}^*$ based on the semantic preservation mask M, the following training objective based on the Jensen-Shannon divergence (Nielsen 2019) is adopted to maximize the distribution alignment between the optimized assignment and the similarity distribution, as:

$$\mathcal{L}_{JS} = \frac{1}{2}(\mathrm{KL}(\boldsymbol{\pi} \| \hat{\boldsymbol{S}}) + \mathrm{KL}(\boldsymbol{S} \| \hat{\boldsymbol{S}})), \hat{\boldsymbol{S}} = \frac{1}{2}(\boldsymbol{\pi} + \boldsymbol{S}), \tag{8}$$

where $\boldsymbol{S}_{ij} = S(\mathcal{I}_i^q, \mathcal{T}_i^q, \mathcal{I}_j^t)$ for short.

Similarity Composition

In this section, we will elaborate on the similarity composition in two essential similarity components: 1) *Difference Quantization Alignment* to perceive the interactions between textual modifiers and visual differences from query to the candidates, and 2) *Composed Word-level Alignment* to capture correlations between targets and queries from salient local modifiers and reference images.

Difference Quantization Alignment. Textual modification information is crucially important to guide the transition from the reference image to the target, yet previous methods have not fully exploited the modifiers when merely combining them in query features. In this module, we regard the features of modifiers as textual descriptions of the visual transformations from the reference images to the target images. In other words, the distance between the textual representation T_i^q and the visual difference feature comparing V_i^q and V_i^t should be compressed. As for features of negative pairs, textual feature T_i^q and difference feature comparing V_i^q and V_j^t are separated in the subspace. Given the encoded visual features, the visual difference feature V_{ij}^D for the i-th query image and j-th candidate is computed through:

$$V_{ij}^D = f_D(V_i^t - V_j^q), \tag{9}$$

where $V^D \in \mathbb{R}^{N_q \times N_t \times d}$. $f_D(\cdot)$ is implemented as an MLP network to transform the visual features into a common space. Hence, we further obtain the semantic-aware association matrix between visual difference features and word fragment $T_i^q = \{t_{ik}^q\}_{k=1}^{N_w}$ on the word level as:

$$\boldsymbol{A}_{ijk} = V_{ij}^D \otimes g_t(t_{ik}^q), \tag{10}$$

where $g_t(\cdot)$ exploits similar design to $f_d(\cdot)$ to encode the word-level textual features. Note that $\boldsymbol{A} \in \mathbb{R}^{N_q \times N_t \times N_w}$ evaluates the proximity of each word fragment in each query to the target, which could be interpreted from two perspectives. Through mean-pooling strategy on the lengths of the sentence N_w, it summarizes the local textual clues to generate similarities from the i-th query to the j-th target based on the correlations between visual changes and textual modifications, which is formulated as:

$$\boldsymbol{S}_{ij}^D = \mathrm{softmax}(\sum_k^{N_w} (\boldsymbol{A}_{ijk} / N_w)). \tag{11}$$

From another perspective, the responses of word tokens with discriminative semantics vary significantly to the target and negative samples. Hence, variability estimation on the dimension of candidate numbers N_t enables evaluation on the sensitivity of each word to different candidate images as $w_{ik} = \mathrm{softmax}(\max_j \boldsymbol{A}_{ijk} - \min_j \boldsymbol{A}_{ijk})$, to be utilized as weights of salient words.

Composed Word-level Alignment. To incorporate the differentiated semantics in the text modality, this branch further explores the fusion between the updated query text features and reference image features for effective alignment across queries and targets. Based on the weight w_{ik} derived from the above section for the k-th word in the i-th query, the enhanced modification textual features is conducted as:

$$\begin{aligned} \hat{T}_i^q &= \alpha_i \cdot \sum_k^{N_w} t_{ik}^q + (1 - \alpha_i) \cdot \sum_k^{N_w} (w_{ik} \odot t_{ik}^q), \\ \alpha_i &= \phi_t(\sum\nolimits_k^{N_w} t_{ik}^q), \end{aligned} \tag{12}$$

where ϕ_t is designed as an MLP network followed by a $Sigmoid(\cdot)$ function.

With the refined textual representations concentrating on the salient word tokens, the active components of the modifiers that have a clear reference to the target are strengthened.

Methods	Dress		Shirt		Toptee		Average		
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	Mean
VAL (Chen, Gong, and Bazzani 2020)	21.12	42.19	21.03	43.44	25.64	49.49	22.60	45.04	33.82
CIRPLANT (Liu et al. 2021)	14.38	34.66	13.64	33.56	16.44	38.34	14.82	35.52	25.17
CoSMo (Lee, Kim, and Han 2021)	21.39	44.45	16.90	37.49	21.32	46.02	19.87	42.65	31.26
DCNet (Kim et al. 2021)	28.95	56.07	23.95	47.30	30.44	58.29	27.78	53.89	40.84
CLVC-Net (Wen et al. 2021)	29.85	56.47	28.75	54.76	33.50	64.00	30.70	58.41	44.56
ARTEMIS (Delmas et al. 2022)	27.16	52.40	21.78	43.64	29.20	54.83	26.05	50.29	38.17
FashionVLP (Goenka et al. 2022)	26.77	53.20	22.67	46.22	28.51	57.47	25.98	52.30	39.14
CLIP4Cir (Baldrati et al. 2022)	31.63	56.67	36.36	58.00	38.19	62.42	35.39	59.03	47.21
CRN (Yang et al. 2023a)	30.20	57.15	29.17	55.03	33.70	63.91	31.02	58.70	44.86
DWC (Huang et al. 2024)	33.61	58.80	37.09	62.46	40.80	68.38	37.17	63.21	50.19
MGUR (Chen et al. 2024a)	32.61	61.34	33.23	62.55	41.40	72.51	35.75	65.47	50.61
SPN (Feng, Zhang, and Nie 2024)	38.82	62.92	45.83	66.44	48.80	71.29	44.48	66.88	55.68
FAME-ViL (Han et al. 2023)	42.19	67.38	47.64	68.79	50.69	73.07	46.84	69.75	58.29
TG-CIR (Wen et al. 2023)	45.22	69.66	52.60	72.52	56.14	77.10	51.32	73.09	62.21
CaLa (Jiang et al. 2024)	42.38	66.08	46.76	68.16	50.93	73.42	46.69	69.22	58.05
DIPNEC (Ours)	**46.90**	**71.29**	**56.92**	**77.77**	**58.18**	**80.88**	**54.00**	**76.64**	**65.32**

Table 1: Experiments results on FashionIQ. Best results are marked in bold.

Hence, we further integrate enhanced modification features $\hat{T}_i^q$ with the reference image features V_i^q to yield the query features $Q_i = \psi(V_i^q, \hat{T}_i^q)$ through the combining function $\psi(\cdot)$. Afterwards, the relevance between the query fusion Q_i and the target image feature V_j^t is modeled as:

$$\boldsymbol{S}_{ij}^C = \frac{Q_i^\top \cdot V_j^t}{\|Q_i\|\|V_j^t\|}, \tag{13}$$

where $\|\cdot\|$ means the L2 normalization. Finally, both the correlation between quantified visual difference and the text modifier $\boldsymbol{S}_{ij}^D$ and the relevance from the query to the target $\boldsymbol{S}_{ij}^C$ determine the ultimate matching score as follows:

$$\boldsymbol{S}_{ij} = \boldsymbol{S}_{ij}^C + \boldsymbol{S}_{ij}^D. \tag{14}$$

Objective Functions

Following models (Baldrati et al. 2022; Han et al. 2023), we adopt InfoNCE loss on the matching score $\boldsymbol{S}_{ij}$ to narrow the distances between positive samples and meanwhile separate the query and the negative samples, as:

$$\mathcal{L}_{InfoNCE} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\tau \boldsymbol{S}_{ij})}{\sum_{j=1}^{B} \exp(\tau \boldsymbol{S}_{ij})}, \tag{15}$$

where τ is the scaling parameter and B is the batch size. With the assignment by optimal transport to supervise the similarity ranking as mentioned above, the whole training objective is defined as follows with the weight γ:

$$\mathcal{L}_{all} = \mathcal{L}_{InfoNCE} + \gamma \mathcal{L}_{JS}. \tag{16}$$

Experiments

Experimental Setup

Datasets. We conduct experiments of DIPNEC on the commonly-used datasets for composed image retrieval, *i.e.*, FashionIQ (Wu et al. 2021) and CIRR (Liu et al. 2021). **FashionIQ** (Wu et al. 2021) contains the triplets of reference images, modification sentences, and candidate images from the fashion platforms, and it covers categories including dresses, toptees, and shirts. With 77,684 fashion images total, the training, validation, and test sets are split by the proportion of 3:1:1. **CIRR** (Liu et al. 2021) reconstructed from NLVR^2 dataset (Suhr et al. 2019) concerns about complex objects in real-world scenarios. It is comprised of 36,554 query-to-target triplets in total with 21,552 images, and split by 8:1:1 for training, validating, and testing, respectively. Besides, CIRR provides subset validation and test where each subset is composed of 6 pictures that are highly similar to evaluate the model robustness.

Evaluation Metrics. Recall rate at K (R@K) is a widely-used evaluation metric in retrieval tasks, defined as the proportion of matched target images ranked in the top-K results from the model given the composed query in this task. We follow previous works (Huang et al. 2024; Yang et al. 2023b) to compare R@10 and R@50 on dresses, toptees, and shirts and mean recall rates in FashionIQ. For CIRR, ranking results in the subset setting as $\text{Recall}_{\text{subset}}@K$ are also presented with R@1, R@5, R@10, and R@50 metrics.

Implementation Details. We implemented the visual and textual encoders as $\text{CLIP}_{\text{ViT}-\text{L}/14}$ (Radford et al. 2021) and BERT structures respectively. The combining function followed Combiner (Baldrati et al. 2022) to compose the query image and modification text features. We set K as 20% of the batch size in Eq. 6, $\epsilon = 0.1$ in Eq. 7, $\tau = 100$ in Eq. 15, and $\gamma = 1.0$ in Eq. 16. More analysis of parameter sensitivities and ablation studies could be referred to following sections. The encoders were first fine-tuned for 10 epochs with an initial learning rate of 2×10^{-6}. Then the proposed DIPNEC was trained for 100 epochs with the Adam optimizer at the learning rate of 2×10^{-5} with the encoder parameters

Methods	Recall@K				Recall$_{\text{subset}}$@K			CM
	K=1	K=5	K=10	K=50	K=1	K=2	K=3	
TIRG (Vo et al. 2019)	11.04	35.08	51.27	83.29	23.82	45.65	64.55	29.45
MAAF (Dodds et al. 2020)	10.31	33.03	48.30	80.06	21.05	41.91	61.60	27.04
CIRPLANT (Liu et al. 2021)	19.55	52.55	68.39	92.38	39.20	63.03	79.49	45.88
ARTEMIS (Delmas et al. 2022)	16.96	46.10	61.31	87.73	39.99	62.20	75.67	43.05
CLIP4Cir (Baldrati et al. 2022)	33.59	65.35	77.35	95.21	62.39	81.81	92.02	63.87
CompoDiff (Gu et al. 2024)	22.35	54.36	73.41	91.77	35.84	56.11	76.60	45.10
TG-CIR (Wen et al. 2023)	45.25	78.29	87.16	97.30	72.84	89.25	95.13	75.57
BLIP4CIR (Liu et al. 2024)	40.17	71.81	83.18	95.69	72.34	88.70	95.23	72.07
SSN (Yang et al. 2023b)	43.91	77.25	86.48	97.45	71.76	88.63	95.54	74.51
SPN (Feng, Zhang, and Nie 2024)	45.33	78.07	87.61	**98.17**	73.93	89.28	95.61	76.00
DIPNEC (Ours)	**47.24**	**80.20**	**89.07**	97.87	**73.97**	**89.74**	**95.72**	**77.09**

Table 2: Experiments results on CIRR. Best results are marked in bold. "CM" means (R@5+Recall$_{\text{subset}}$@1)/2.

frozen. All experiments were implemented in Pytorch on a single NVIDIA GeForce RTX 3090 Ti GPU.

Quantitative Experimental Results

We report the quantitative comparison of the proposed DIPNEC with the state-of-the-art approaches on FashionIQ and CIRR datasets in Table 1 and Table 2, respectively.

Results on FashionIQ. The proposed DIPNEC has demonstrated competitive performance with an improvement of 3.11% on the global metric comparing TG-CIR. The recent method CaLa (Jiang et al. 2024) constructed twin-attention-based visual compositors and hinge-based attention by complicated model designs including Q-former (Li et al. 2023) and additional transformers. Our DIPNEC still outperforms CaLa by a large margin. Specifically, we can observe an evident growth in the shirt category, as a vast array of candidate images of shirts with similar semantics may lead to confusion in previous models. The improvement is credited to the difference quantization module to augment the discriminative semantics and optimal transport assignment with semantic preservation to guide the distributions.

Results on CIRR. Apart from the fashion datasets, the overall performance of our DIPNEC on CIRR dataset with complicated textual modifications and semantically rich images also manifests the superiority of the proposed architecture. Through R@50 of SPN (Feng, Zhang, and Nie 2024) is slightly higher, SPN adopted LLaVA-v1 (Liu et al. 2023) to generate captions with additional LLM models. Note the gain on the evaluation metric R@1 of our proposed DIPNEC is comparatively highlighted. It verifies that the cooperation of difference quantization alignment and composed word-level alignment could effectively enhance the semantic correlations between queries and targets by perceiving the distinguished semantics across multiple candidate images.

Ablation Studies

Analysis on Optimal Transport. To explore the specific design of optimal transport, we have implemented the ablation experiment on FashionIQ in Table 3 and come to the

Models	Dress	Shirt	Toptee	Mean
Ours w/o OT	44.42	55.05	56.81	52.09
Ours w/o Mask	44.52	56.32	56.85	52.57
Ours w Mask on HN	45.11	55.49	56.55	52.39
Ours w KL	46.35	56.62	58.18	53.72
Ours w symKL	46.15	56.97	**58.23**	53.79
Ours	**46.90**	**56.92**	58.18	**54.00**

Table 3: Analysis of optimal transport on R@10 metric.

the conclusions as follows: 1) Optimal transport with emphasis on the hard negative candidate samples is essential to the re-assignment with semantic preservation and the model robustness. An obvious performance drop could be observed in the "Ours w/o OT", which substantiates the effectiveness of optimal transport. Furthermore, through comparing "Ours w/o Mask" with our DIPNEC, the improvement after introducing the mask in Eq. 6 verifies that concentrating on the transport of the hard negative instances enables effective refinement on the matching score, and simultaneously maintains the learned semantics would not be disturbed during optimization. However, "Ours w Mask on HN" (using the opposite mask as Eq. 6) that filters the hard negative triplets achieves unsatisfactory results, which implies that the OT assignment on the easy negatives might be futile and bring noises. 2) Jensen-Shannon (JS) divergence outperforms the loss functions with the listed distribution measurements generally. The supervision of the refined optimal transport assignment is indispensable to guide the distance adjustment between the queries and the targets. JS divergence in our DIPNEC narrowly leads the KL and symmetrical KL divergence for its symmetrical design and computation stability.

Ablation Study of Model Designs. To investigate the effectiveness of the core components in the proposed DIPNEC, we have conducted ablative experiments in Table 4. "Ours w/o $\hat{T}^q$" refers to $Q_i = \phi(V_i^q, T^q)$ when generating query compositions, and "Ours w/o $\boldsymbol{S}^D$" means directly using $\boldsymbol{S}_{ij}^C$ as the final similarity scores. DQA and CWA

Models	FashionIQ		CIRR	
	R@10	R@50	R@5	R_{subset}@1
Ours w/o DQA	52.96	74.60	79.12	71.66
Ours w/o CWA	52.08	74.62	78.45	70.39
Ours w/o $\hat{T}^q$	53.65	75.92	79.91	72.59
Ours w/o $\boldsymbol{S}^D$	53.35	75.45	79.89	72.71
Ours	**54.00**	**76.64**	**80.20**	**73.97**

Table 4: Ablation experiments on model designs.

both are conductive to improve the overall performance, and it could be further promoted after the combination, as the growth in recall rates comparing our model with first two rows shows. The apparent effects of integrating CWA stem in part from enhanced word-level modifier features based on weights from DQA, which is also verified by the model disabling the weights (*i.e.*, "Ours w/o $\hat{T}^q$") for updating sentence features. From the comparison between the last two rows, it can be inferred that the correlation matrix $\boldsymbol{S}^D$ injects the strengthened alignment between the quantified visual difference and the sentence clues to the matching scores.

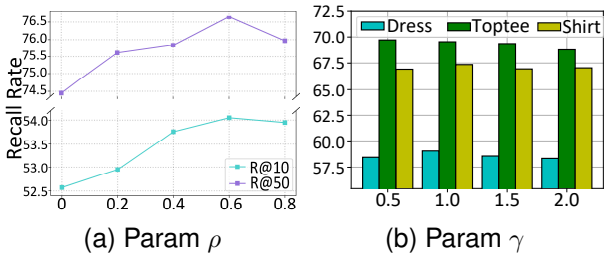

Figure 3: Parameter sensitivity analysis of ρ and γ.

Further Analysis

Parameter Analysis. To further analyze the parameter sensitivity, we have also reported results on different ratios of masked samples in each mini-batch ($\rho = \frac{K}{\text{batch size}}$) in Figure 3(a). Within certain limits, as the masked samples increase, the recall metric has a continuous gain accordingly, owing to the more emphasis on the OT assignments imposed on the match scores on hard negative candidates. However, note that the mask ratio exceeding around 60% may lead to performance impairment due to the complicated restrictions of optimization and the growing complexity in the masked sparse matrix. Additionally, we present the analysis of the hyper-parameter γ in Figure 3(b). It is encouraged to set γ within the range of 0.5 to 1.5 to strike a balance of the optimal transport loss $\mathcal{L}_{JS}$ with the ranking loss.

Visualization on the OT assignment. The visualization comparison of matching scores in Figure 4 shows that the gap between the match scores for the positive triplets (in the diagonal) and negative triplets is enlarged, which further

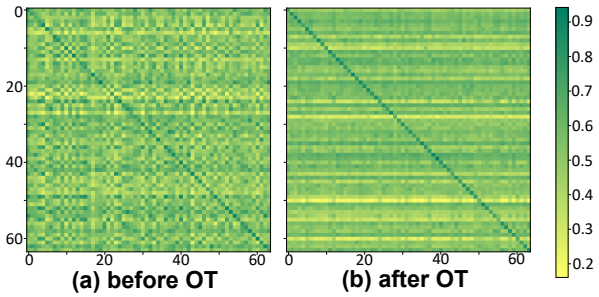

Figure 4: Comparison of matching scores before and after optimal transport assignment with semantic preservation.

verifies the robustness of the optimization on the transport cost with semantic preservation.

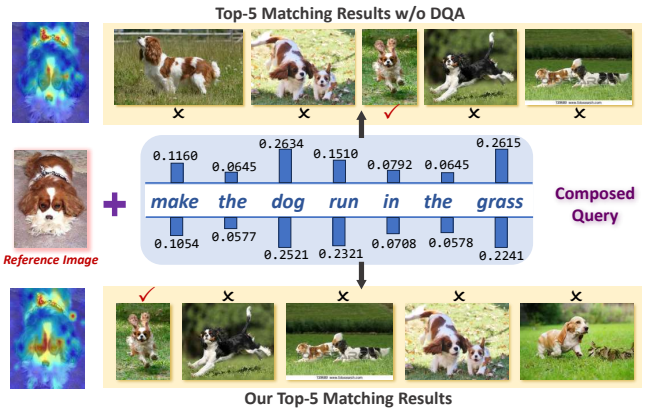

Figure 5: Visualization results on CIRR.

Qualitative Results. Figure 5 displays the top matching results from the proposed DIPNEC with the weights on each word of the query sentences. As the figure shows, the word "run" containing the essence of the modification requirement is assigned with higher weight than models without DQA, which promotes semantic discrimination and captures query intentions precisely in the final top matching results.

Conclusion

In this paper, we proposed a novel DIPNEC framework for discriminative perception from negative correspondences in composed-query image retrieval. To augment the semantic distinguishment across samples, Difference Quantization Alignment was introduced to assess the relevances between visual differences and textual modifiers, and Composed Word-level Alignment exploited enhanced textual features to evaluate the distances between queries and candidates. Furthermore, we deployed an optimization process on transport costs with a semantic preservation mask to guide the assignment of matching scores. Extensive experiments demonstrated DIPNEC could effectively boost retrieval performance. In the future, we would explore optimal transport with flexible boundaries in multi-modal learning tasks.

Acknowledgments

This work was supported by the National Key R&D Program of China (2022YFB4701400/4701402), SSTIC Grant (KJZD20230923115106012, KJZD20230923114916032, GJHZ20240218113604008), and Beijing Key Lab of Networked Multimedia.

References

Asadulaev, A.; Korotin, A.; Egiazarian, V.; Mokrov, P.; and Burnaev, E. 2024. Neural Optimal Transport with General Cost Functionals. In *ICLR*. OpenReview.net.

Baldrati, A.; Bertini, M.; Uricchio, T.; and Bimbo, A. D. 2022. Conditioned and composed image retrieval combining and partially fine-tuning CLIP-based features. In *CVPR Workshops*, 4955–4964. IEEE.

Chen, J.; and Lai, H. 2023. Ranking-aware Uncertainty for Text-guided Image Retrieval. *CoRR*, abs/2308.08131.

Chen, Y.; Gong, S.; and Bazzani, L. 2020. Image Search With Text Feedback by Visiolinguistic Attention Learning. In *CVPR*, 2998–3008. IEEE.

Chen, Y.; Zheng, Z.; Ji, W.; Qu, L.; and Chua, T. 2024a. Composed Image Retrieval with Text Feedback via Multi-grained Uncertainty Regularization. In *ICLR*. OpenReview.net.

Chen, Y.; Zhong, H.; He, X.; Peng, Y.; Zhou, J.; and Cheng, L. 2024b. FashionERN: Enhance-and-Refine Network for Composed Fashion Image Retrieval. In *AAAI*, 1228–1236. AAAI Press.

Chen, Y.; Zhou, J.; and Peng, Y. 2024. SPIRIT: Style-guided Patch Interaction for Fashion Image Retrieval with Text Feedback. *ACM Trans. Multim. Comput. Commun. Appl.*, 20(6): 167:1–167:17.

Cuturi, M. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *NeurIPS*, 2292–2300.

Delmas, G.; de Rezende, R. S.; Csurka, G.; and Larlus, D. 2022. ARTEMIS: Attention-based Retrieval with Text-Explicit Matching and Implicit Similarity. In *ICLR*. OpenReview.net.

Dodds, E.; Culpepper, J.; Herdade, S.; Zhang, Y.; and Boakye, K. 2020. Modality-Agnostic Attention Fusion for visual search with text feedback. *CoRR*, abs/2007.00145.

Feng, Z.; Zhang, R.; and Nie, Z. 2024. Improving Composed Image Retrieval via Contrastive Learning with Scaling Positives and Negatives. In *ACM MM*, 1–10. ACM.

Goenka, S.; Zheng, Z.; Jaiswal, A.; Chada, R.; Wu, Y.; Hedau, V.; and Natarajan, P. 2022. FashionVLP: Vision Language Transformer for Fashion Retrieval with Feedback. In *CVPR*, 14085–14095. IEEE.

Gu, G.; Chun, S.; Kim, W.; Jun, H.; Kang, Y.; and Yun, S. 2024. CompoDiff: Versatile Composed Image Retrieval With Latent Diffusion. *Trans. Mach. Learn. Res.*, 2024.

Guo, J.; Shuang, K.; Zhang, K.; Liu, Y.; Li, J.; and Wang, Z. 2023. Learning to Imagine: Distillation-Based Interactive Context Exploitation for Dialogue State Tracking. In *AAAI*, 12845–12853. AAAI Press.

Gupta, T.; and Kembhavi, A. 2023. Visual Programming: Compositional visual reasoning without training. In *CVPR*, 14953–14962. IEEE.

Han, X.; Zhu, X.; Yu, L.; Zhang, L.; Song, Y.; and Xiang, T. 2023. CVPR. 2669–2680. IEEE.

Huang, F.; Zhang, L.; Fu, X.; and Song, S. 2024. Dynamic Weighted Combiner for Mixed-Modal Image Retrieval. In *AAAI*, 2303–2311. AAAI Press.

Izquierdo, S.; and Civera, J. 2024. Optimal Transport Aggregation for Visual Place Recognition. In *CVPR*, 1–11. IEEE.

Jia, X.; Zhang, R.; and Peng, M. 2024. Multi-domain gate and interactive dual attention for multi-domain dialogue state tracking. *Knowl. Based Syst.*, 286: 111383.

Jiang, X.; Wang, Y.; Li, M.; Wu, Y.; Hu, B.; and Qian, X. 2024. CaLa: Complementary Association Learning for Augmenting Comoposed Image Retrieval. In *ACM SIGIR*, 2177–2187. ACM.

Kim, J.; Yu, Y.; Kim, H.; and Kim, G. 2021. Dual Compositional Learning in Interactive Image Retrieval. In *AAAI*, 1771–1779. AAAI Press.

Lee, S.; Kim, D.; and Han, B. 2021. CoSMo: Content-Style Modulation for Image Retrieval With Text Feedback. In *CVPR*, 802–812. IEEE.

Li, B.; Shi, Y.; Yu, Q.; and Wang, J. 2024. Unsupervised Cross-Domain Image Retrieval via Prototypical Optimal Transport. In *AAAI*, 3009–3017. AAAI Press.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, 19730–19742. PMLR.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Liu, Z.; Opazo, C. R.; Teney, D.; and Gould, S. 2021. Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models. In *ICCV*, 2105–2114. IEEE.

Liu, Z.; Sun, W.; Hong, Y.; Teney, D.; and Gould, S. 2024. Bi-directional Training for Composed Image Retrieval via Text Prompt Learning. In *WACV*, 5741–5750. IEEE.

Lu, Y.; Qin, Y.; Zhai, R.; Shen, A.; Chen, K.; Wang, Z.; Kolouri, S.; Stepputtis, S.; Campbell, J.; and Sycara, K. P. 2023. Characterizing Out-of-Distribution Error via Optimal Transport. In *NeurIPS*.

Nielsen, F. 2019. On the Jensen-Shannon Symmetrization of Distances Relying on Abstract Means. *Entropy*, 21(5): 485.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, 8748–8763. PMLR.

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2019. A Corpus for Reasoning about Natural Language

Grounded in Photographs. In *ACL*, 6418–6428. Association for Computational Linguistics.

Tu, Y.; Li, L.; Su, L.; Zha, Z.; and Huang, Q. 2024. SMART: Syntax-Calibrated Multi-Aspect Relation Transformer for Change Captioning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(7): 4926–4943.

Vo, N.; Jiang, L.; Sun, C.; Murphy, K.; Li, L.; Fei-Fei, L.; and Hays, J. 2019. Composing Text and Image for Image Retrieval - an Empirical Odyssey. In *CVPR*, 6439–6448. IEEE.

Wang, Y.; Wang, F.; Dong, J.; and Luo, H. 2024. CL2CM: Improving Cross-Lingual Cross-Modal Retrieval via Cross-Lingual Knowledge Transfer. In *AAAI*, 5651–5659. AAAI Press.

Wen, H.; Song, X.; Yang, X.; Zhan, Y.; and Nie, L. 2021. Comprehensive Linguistic-Visual Composition Network for Image Retrieval. In *SIGIR*, 1369–1378. ACM.

Wen, H.; Zhang, X.; Song, X.; Wei, Y.; and Nie, L. 2023. Target-Guided Composed Image Retrieval. In *ACM MM*, 915–923. ACM.

Wen, K.; Gu, X.; and Cheng, Q. 2021. Learning Dual Semantic Relations With Graph Attention for Image-Text Matching. *IEEE Trans. Circuits Syst. Video Technol.*, 31(7): 2866–2879.

Wu, H.; Gao, Y.; Guo, X.; Al-Halah, Z.; Rennie, S.; Grauman, K.; and Feris, R. 2021. Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. In *CVPR*, 11307–11317. IEEE.

Yang, Q.; Ye, M.; Cai, Z.; Su, K.; and Du, B. 2023a. Composed Image Retrieval via Cross Relation Network With Hierarchical Aggregation Transformer. *IEEE Transactions on Image Processing*, 32: 4543–4554.

Yang, X.; Liu, D.; Zhang, H.; Luo, Y.; Wang, C.; and Zhang, J. 2023b. Decompose Semantic Shifts for Composed Image Retrieval. *CoRR*, abs/2309.09531.

Zhang, G.; Li, S.; Wei, S.; Ge, S.; Cai, N.; and Zhao, Y. 2024. Multimodal Composition Example Mining for Composed Query Image Retrieval. *IEEE Trans. Image Process.*, 33: 1149–1161.

Zhang, G.; Wei, S.; Pang, H.; and Zhao, Y. 2021. Heterogeneous Feature Fusion and Cross-modal Alignment for Composed Image Retrieval. In *ACM MM*, 5353–5362. Association for Computing Machinery.

Zhao, Y.; Song, Y.; and Jin, Q. 2022. Progressive Learning for Image Retrieval with Hybrid-Modality Queries. In *SIGIR*, 1012–1021. ACM.