Text-oriented Image Query Representation for Zero-shot Composed Image Retrieval

Pavan Kartheek Rachabathuni
University of Florence - MICC
Florence
Italy
pavankartheek.rachabathuni@unifi.it

Andrea Ciamarra
CNIT Florence
Florence
Italy
andrea.ciamarra@unifi.it

Roberto Caldelli
CNIT, Florence and Universitas Mercatorum, Rome
Italy
roberto.caldelli@unifi.it

Marco Bertini
University of Florence - MICC
Florence
Italy
marco.bertini@unifi.it

2025 International Conference on Content-Based Multimedia Indexing (CBMI) | 979-8-3315-5500-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/CBMI66578.2025.11339329

***Abstract*—Zero-Shot Composed Image Retrieval (ZS-CIR) is the task of retrieving a target image based on a query that combines a reference image with a textual description specifying desired modifications in a zero-shot setting. Existing ZS-CIR models typically fuse visual and textual modalities into a single query representation, but often struggle to capture the fine-grained distinctions essential for accurate retrieval. In this paper, we present TEOZCIR, a transformer-based model that introduces a balanced semantic fusion module and an enhancement mechanism to more effectively integrate multimodal information. The model is built around two core components: the Text-Aware Query Combiner (TAQC) and the Query Enhancer Network (QENet). These components operate in tandem: TAQC dynamically adjusts the semantic contributions of the visual context based on the input text, generating a balanced query representation. This representation is then further refined by QENet, which enhances the fused features to better align with the target image. Throughout the entire process, the model maintains a lightweight architecture with significantly fewer trainable parameters compared to conventional training-based methods. Experiments carried out on three benchmark datasets CIRR, Fashion IQ, and CIRCO to demonstrate that TEOZCIR significantly improves ZS-CIR performance, setting a new benchmark for multimodal retrieval.**

***Index Terms*—Composed Image Retrieval, Zero-Shot Composed Image Retrieval, Fusion Strategies.**

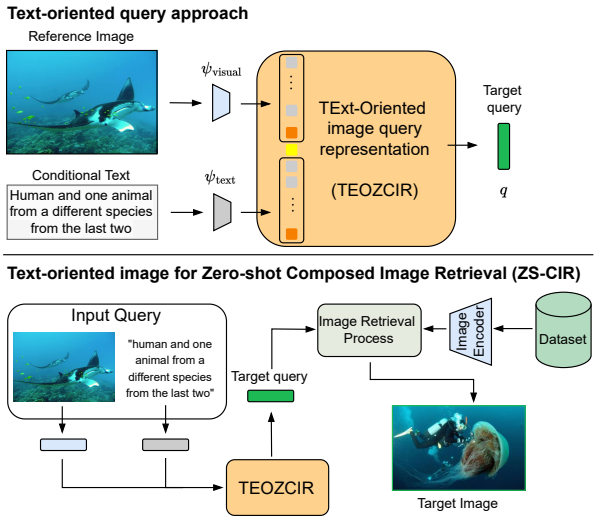

Fig. 1: Workflow Overview: (Top), TEOZCIR generates a target query by combining reference image features with conditional text. Bottom (inference), Image Retrieval Process: Matching the target query against the dataset to retrieve the target image in a zero-shot setting

I. Introduction

Content-Based Image Retrieval (CBIR) [1] is a core task in multimedia applications, with a wide range of use cases such as visual search [2]–[4], object and landmark localization [5]–[8], and person or vehicle re-identification [6], [9]. Traditional single modality approaches, like text-to-image [10]–[14] or image-to-image retrieval [15]–[19], typically align input queries with images. However, they often fail to capture nuanced user intent in image modification queries. To overcome this limitation, Composed Image Retrieval (CIR) [20]–[25] introduces a more expressive multimodal approach. CIR combines a reference image with textual instructions to form a single, composite query. This enables retrieval of images that match the reference image features incorporating specified textual modifications. For instance, a user might submit a photo and request changes like "add a red hat" or "change the background to a beach". By integrating visual and textual cues, CIR offers a more flexible and accurate search experience compared to traditional methods. Recent Vision-Language Models like CLIP [26] and BLIP [27] align text and image features in a shared space through large-scale contrastive pre-training. CIR leverages these learned representations to model the relationship between the reference image, the textual modification and the desired outcome. To achieve this objective, a carefully curated dataset is required with a large number of training triplets $(I^{Ref}, T^{Cond}, I^{Target})$, in which I^{Ref} is a reference image, T^{Cond} is a condition text and I^{Target} is a target image. Collecting or creating such triplets is very expensive and laborious. Constructing these datasets is a major bottleneck, not merely due to scale, but fundamentally because of the "fuzzy" nature of the CIR task. Since a given (I^{Ref}, T^{Cond}) input can correspond to multiple plausible target images reflecting diverse interpretations, establishing consistent ground truth for I^{Target} is inherently ambiguous, rendering dataset collection noticeably challenging and costly. To tackle these challenges, Zero-Shot Composed Image Retrieval (ZS-CIR) has emerged as a promising alternative.

Rather than relying on curated triplets, ZS-CIR utilizes large-scale image-text pairs for training, enabling it to generalize more effectively to unseen data. In this work, we present TEOZCIR, a lightweight multimodal text-guided late-fusion approach designed to address the Zero-Shot CIR task. The fusion process dynamically adjusts the importance of textual and visual contexts based on a learnable parameter, which is induced between the two input modalities. It accounts for the semantic modifications of the visual content while attending the information from the input (conditional) text. Additionally, the fusion mechanism amplifies both the edited features and the conditional text, ensuring the intermediate representation to align closely with the target image. To balance the contributions from both text and visual modalities, the intermediate representation is further refined to obtain the final target query. The overall workflow of our approach is shown in Fig. 1. The main contributions can be summarized as follows:

- We introduce **TEOZCIR**, a novel multimodal **TE**xt-**O**riented image query approach for **Z**ero-shot **C**omposed **I**mage **R**etrieval task, to retrieve a target image based on the context, coming from the reference image and the conditional text, with few trainable parameters;
- we design the Text-Aware Query Combiner (TAQC) and the Query Enhancer Network (QENet) modules, to adjust the semantic contributions from both modalities and enhance the final target query, with a simple training, involving fewer trainable parameters compared to existing approaches;
- we conduct experiments on three popular datasets in zero-shot CIR task. We demonstrate that, without relying on LLMs, our proposed method outperforms existing fusion-based approaches and shows good generalization capabilities w.r.t LLM-based methods.

II. Related Works

Recent advancements in Composed Image Retrieval (CIR) have introduced state-of-the-art models that fuse both image and text modalities to form a composed query that tackles the retrieval task effectively. Methods addressing CIR can be divided into two categories: supervised learning-based approaches, where the model is trained and evaluated on the same benchmark dataset, and methods for Zero-Shot Composed Image Retrieval (ZS-CIR), where the model is trained on one dataset and evaluated on another. Existing CIR solutions are broadly classified into Late Fusion and Early Fusion approaches. Early Fusion methods integrate visual and textual features at the input or embedding level, allowing the model to learn joint representations before making predictions. For instance, SPRC [28] model tackles the CIR task in an early fusion strategy and employs a lightweight querying transformer to generate text prompts that blend the context of the reference image with the modification text. Late Fusion methods, instead, merge the two modalities after the features are extracted from the input text and image. CLIP4CIR [29] utilizes CLIP as the backbone and an MLP as the image-text combiner to integrate the reference image and the conditional text into a unified representation, which is then matched with the target image. BLIP4CIR+Bi [30] introduces a novel mapping between the target image and the reference image using ⟨target image, reversed modification text⟩ pairs. All these methods mentioned above are supervised learning approaches which require curated triplets for training.

In zero-shot settings, many existing approaches address the CIR task in late fusion mechanism. PALAVRA [31] employs a two-stage process, first by applying textual inversion with a mapping function and then optimizing the pseudo-word token. Pic2Word [13] is a training-dependent method that leverages a textual inversion network optimized by contrastive loss to capture the pseudo-word token. Similarly, SEARLE [10] generates pseudo-word tokens through textual inversion and distills them into a dedicated network. SEARLE-OTI [10] represents a variant that operates without the distillation network. Additionally, LinCIR [32] projects text embeddings into the token space for retrieval. These methods overcome the reliance on curated triplets. However, the models suffer from a limitation as they primarily focus on fine-tuning the text inversion mechanism that projects the image into the text latent space, while overlooking the alignment between the projected image and the target image. MagicLens [33] adopts an attention pooling mechanism for the ZS-CIR task, showing better performance. Its strength lies in handling their own dataset which comprises samples with rich semantic relations, and also employing stronger backbones. TransAgg [34] is another transformer-based model, which is trained on their proposed dataset (named Laion_combined) that fuses multimodal input to address the ZS-CIR task. However, these models [32]–[34] often fail to capture and fuse intricate fine-grained distinctions crucial for accurate retrieval. Instead, our late fusion model is designed by using a tailored weighting mechanism that operates on multimodal inputs, leveraging a learnable token to guide the retrieval process. This mechanism injects textual information from the conditional text into the visual content, enabling more context-aware retrieval.

Some emerging techniques [11], [12] are based on a non-training manner. Among them, CIReVL [12] is a training-free approach that employs a generative vision language model alongside an LLM to recompose captions based on textual modifications. Furthermore, SEIZE [11] (inspired by LDRE [14]) is a training-free approach that utilizes a visual-language model (VLM) and a large language model (LLM) to generate initial captions and their edited variants. By comparing the similarity between the aggregated edited captions and the target image, the method facilitates retrieval. Despite the strong performance of LLM and VLM models, there are limitations, such as the induction of hallucinations in the retrieval pipeline and the use of static embeddings due to the training-free approach. In contrast, we propose a novel late fusion text oriented mechanism, that captures and fuses the semantic fine-grained features with the improved interpretation from incorporated textual aspects guiding the visual edited content of the reference image for target image retrieval.

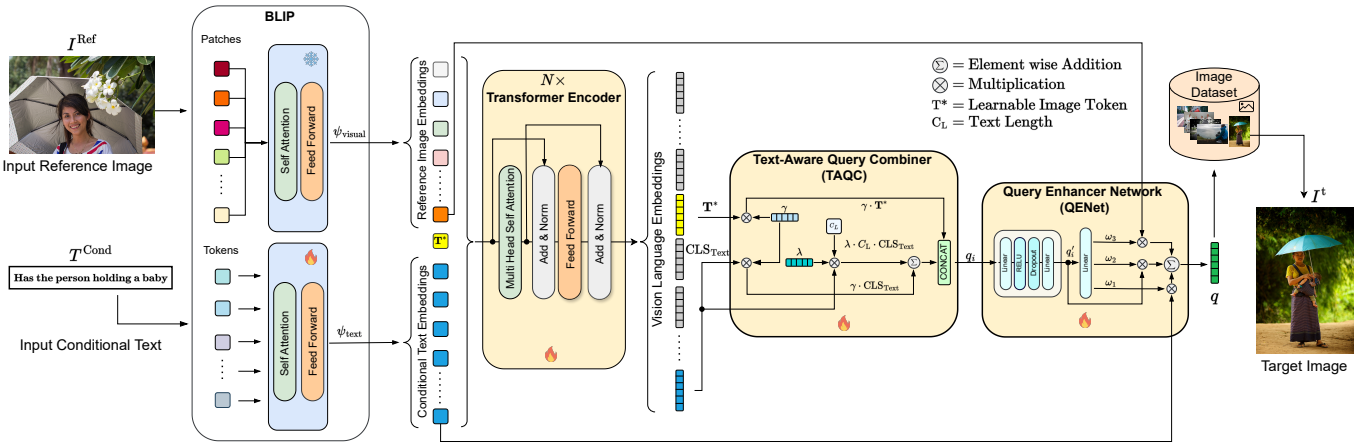

Fig. 2: Overview of the proposed TEOZCIR framework. Given a reference image I^{Ref} and conditional text T^{Cond}, we extract their embeddings using BLIP [27]. Along with a learnable token $\mathbf{T}^*$, these embeddings are passed through a transformer encoder. The token $\mathbf{T}^*$ acts as an enriched visual anchor, guided by the conditional text. The resulting embeddings are processed by the Text-Aware Query Combiner (TAQC) to generate an intermediate query q_i, which is then refined by the Query Enhancer Network (QENet) to produce the final query representation q for retrieval.

III. Method

In this paper, we introduce TEOZCIR, a novel text-oriented approach specifically designed to address the zero-shot CIR task with a multimodal input query $\{I^{Ref}, T^{Cond}\}$, where I^{Ref} represents the reference image and T^{Cond} denotes the conditional text. The approach seamlessly integrates context from both modalities: the reference image provides the visual grounding, while the conditional text imposes textual constraints that guide the retrieval process. The goal is to retrieve a target image that visually aligns with the combined semantic information from both visual and textual inputs through the final query representation q. The overall proposed model is shown in Fig. 2. We first input the image and the conditional text to a BLIP [27] pretrained model keeping the image encoder ψ_{visual} frozen. We augment the output of the BLIP model by concatenating it with a learnable image token, denoted as $\mathbf{T}^*$. Traditionally, a class token of a transformer encoder is used to capture the global representation. In our design, this image token $\mathbf{T}^*$ is designed to jointly learn from both the reference image and the conditional text. During training, $\mathbf{T}^*$ captures global visual information from the input image while being conditioned by the textual content T^{Cond}.

We believe that such an image token $\mathbf{T}^*$ would allow the model to learn crucial information injected from the conditional text. The transformer encoder captures the token-level context to produce visual-language embeddings. We then introduce the Text-Aware Query Combiner (TAQC) module (see Section III-A), which processes the class text token and the learnable image token to devise an enhanced query representation q_i. This representation is semantically aligned visual aspects induced with more textual facet. Such an intermediate query q_i is advanced to the Query Enhancer Network (QENet) (see Section III-B), which amplifies its expressiveness. The refined query q is finally used to retrieve the target image.

A. Text-Aware Query Combiner (TAQC)

The Text-Aware Query Combiner (TAQC) module is designed to enhance retrieval performance by generating a semantically aligned context-aware query representation from visual-language embeddings. To facilitate effective fusion between the two modalities, a learnable token $\mathbf{T}^*$ a class token of the image is concatenated between the reference image and conditioned text embeddings. This token acts not only as a boundary between the two modalities but also as a contextual aggregator. From the encoder output, the image token $\mathbf{T}^*$ and class token of the entire sequence CLS_{Text} are extracted. To emphasize the relative importance of edited visual content and textual content from each token, two learnable vector parameters γ and λ are employed. Specifically, γ is used to scale both tokens, enhancing the model's ability to capture the cross modal context; λ in conjunction with C_L, a scalar denoting the number of valid tokens in the conditioned text, dynamically emphasizes the textual component CLS_{Text}. The scalar value C_L accounts for the varying impact of longer and shorter conditional text, when determining contributions. To enrich the textual contributions from different perspectives the γ scaled CLS_{Text} and λ emphasized in conjunction with text length CLS_{Text} are element-wise added. Furthermore, the learnable nature of the weighting parameters allows the model to optimize feature fusion during training, thereby enhancing its generalization ability across diverse queries and data domains. The intermediate query representation q_i is finally generated, by concatenating (see the CONCAT operation in Fig. 2) the two contributions from the image and the conditional text thoroughly weighted in this TAQC module. The intermediate query q_i can be described according to (1):

$$q_i = \text{CONCAT}\left(\gamma \cdot \mathbf{T}^*, \left(\gamma \cdot \text{CLS}_{\text{Text}} + \lambda \cdot C_L \cdot \text{CLS}_{\text{Text}}\right)\right) \quad (1)$$

Particularly, the two concatenated components from (1) offer complementary perspectives of the text, enabling the learnable parameters to dynamically capture and integrate the essential semantics of both the textual and visual contexts.

B. Query Enhancer Network (QENet)

The Query Enhancer Network amplifies the expressiveness of the output from TAQC (see Section III-A), which is passed through a fully connected network to capture more intricate relationships among the features. The output q_i' is then proceeded to the last linear layer, which turns it to 3 weighing parameters $\{\omega_1, \omega_2, \omega_3\}$. These vectors are used to perform element-wise scaling on intermediate query and corresponding feature components, allowing the model to adaptively re-weight different semantic aspects of the query. Such parameters play a critical role in adaptive balancing, re-calibrating, emphasizing and generalization. By dynamically emphasizing relevant features and downplaying less informative ones, QENet helps the model generalize across diverse query types that significantly improves retrieval performance. The final query representation q is formally described in (2):

$$q = \omega_1 \cdot \psi_{\text{text}} + \omega_2 \cdot q_i' + \omega_3 \cdot \psi_{\text{visual}} \tag{2}$$

where ψ_{visual} and ψ_{text} are the BLIP's visual and text encoder outcome respectively.

IV. Experimental Results

This section presents the experimental results for evaluating our approach on the zero-shot CIR task. Implementation details are provided in Sec.IV-A, while datasets and evaluation metrics are described in Sec.IV-B. Results are reported in Sec.IV-C, where we compare our method with state-of-the-art approaches under fair conditions, considering backbone size and total trainable parameters. When ViT-B variants are unavailable, we use the next larger model (e.g., ViT-L) for consistency. Experiments were conducted on three standard CIR benchmarks: CIRR [35], FashionIQ [36], and CIRCO [37]. Ablation studies in Sec. V analyze the contribution of each TEOZCIR component with additional experiments.

A. Implementation details

We follow the same training and evaluation protocol as other zero-shot CIR approaches [33], [34], [38], [40]. Our model is trained on the Laion_combined [34], the only publicly available dataset for ZS-CIR containing 32k automatically constructed triplets. This enables true zero-shot evaluation training solely on synthetic triplets while testing on standard benchmarks. Furthermore, we use the LaSco [38] dataset to conduct an ablation study (see Section V). Our framework is implemented with PyTorch[1] and runs on a single 24GB NVIDIA TITAN RTX. It must be noted that the model requires approximately 7GB of memory, making it suitable for both training and evaluation also on GPUs with limited memory capacity. We adopt the same image pre-processing scheme as in [29]. We use a transformer encoder with $N = 2$ layers and 8 attention heads with dropout 0.1 and other hyperparameter settings from [34]. We use BLIP w/ViT-B [27] to process the multimodal input, initializing from the fine-tuned Image-Text Retrieval (COCO) checkpoint. We freeze the BLIP visual encoder, while keeping the text encoder trainable end-to-end with our proposed model. We train the entire framework with 50 epochs and a batch size of 16. The total trainable parameters are 143M out of 443M total parameters. We use AdamW optimizer with cosine decay. The learning rate of the BLIP's text encoder parameters is initialized with $1e-6$. We used the classification loss [41] formally defined as (3):

$$L = \frac{1}{B}\sum_{i=1}^{B} -\log\left(\frac{\exp\left(\kappa\left(q, \psi^{+}_{\text{visual}}(i)\right)\right)}{\sum_{j=1}^{B}\exp\left(\kappa\left(q, \psi^{+}_{\text{visual}}(j)\right)\right)}\right) \tag{3}$$

where B is the batch size, κ is the cosine similarity and q is the final represented query, $\psi_{\text{visual}}(\cdot)$ is the BLIP visual encoder.

B. Evaluation datasets and metrics

We compare the proposed method with baseline models and recent approaches, with or without the use of powerful LLMs. We show the results in Tab. I and II. We report in Tab. I the total trainable parameters for each approach except for the baselines, to highlight the performance of the models with respect to the trainable parameters. Entry with "–", refers to non-availability of the value in the original paper.

CIRR is a manually annotated open-domain dataset with 36.5k queries over 19k images, aimed at evaluating CIR methods in general settings. Performance is measured in the test set using Recall@K ($K \in \{1, 5, 10, 50\}$), indicating the percentage of target images in the top-K results.

FashionIQ is a fashion-focused retrieval dataset with 30,134 triplets from 77,684 images across three categories: Dress, Shirt, and Toptee. Recall@K ($k \in \{10, 50\}, Avg$) is employed to measure retrieval accuracy in the val set.

CIRCO is an open-domain dataset for zero-shot CIR, with no training split. It features multiple ground truths per query (avg. 4.53), 220 validation and 800 test queries, and a 120k-image gallery from COCO. Evaluation in the test set uses mAP@K $k \in \{10, 25, 50\}$, enabling fine-grained assessment.

C. Results

1) CIRR: From Tab. I, our proposed method TEOZCIR demonstrates state-of-the-art performance equipped with a ViT-B backbone. Compared to MagicLens, our method achieves substantial gains, improving Recall@1 (+12.64), Recall@5 (+11.76), Recall@10 (+9.44) and Recall@50 (+3.74) points. Also, TEOZCIR consistently outperforms the second-best method TransAgg, with improvements of Recall@1 (+1.54), Recall@5 (+1.34), Recall@10 (+1.26) and Recall@50 (+1.33) points across the metrics. These gains are primarily attributed to the proposed TAQC module, which effectively enhances semantic alignment by dynamically weighting critical query features for retrieval. TEOZCIR delivers notable results over projection based methods such as LinCIR, Pic2Word,

[1]Code is available at `https://github.com/miccunifi/TEOZCIR`

TABLE I: Comparison with state-of-the-art approaches in ZS-CIR (Zero Shot Composed Image Retrieval). Best score in bold, second best score underlined. The metrics with "–" refer no value is provided in author's paper.

Method	Backbone	# Trainable Params	CIRR				FashionIQ			CIRCO		
			Recall@K				Recall@K (val)			mAP@K		
			K=1	K=5	K=10	K=50	K=10	K=50	Avg	K=10	K=25	K=50
Image-only	ViT-B		6.89	22.99	33.68	59.23	5.90	13.37	9.64	1.60	2.12	2.41
Text-only	ViT-B		21.81	45.22	57.42	81.01	18.70	36.84	27.77	2.67	2.98	3.18
Captioning	ViT-B		12.46	35.04	47.71	77.35	13.98	28.62	21.30	5.77	6.44	6.85
PALAVRA [31]	ViT-B	176M	16.62	43.49	58.51	83.95	19.76	37.25	28.51	5.32	6.33	6.80
SEARLE [37]	ViT-B	165M	24.00	53.42	66.82	89.80	22.89	42.53	32.71	9.94	11.13	11.84
SEARLE-OTI [37]	ViT-B	165M	24.27	53.25	66.10	88.84	22.44	42.34	32.39	7.83	8.99	9.60
CASE [38]	ViT-B	-	35.40	65.78	78.53	94.63	-	-	-	-	-	-
Pic2Word [13]	ViT-L	429M	23.90	51.70	65.30	87.80	24.70	43.70	34.20	9.51	10.64	11.29
LinCIR [32]	ViT-L	442M	25.04	53.25	66.68	-	26.28	46.49	36.39	13.58	15.00	15.85
Context-I2W [39]	ViT-L	496M	25.60	55.10	68.50	89.80	27.80	48.90	38.35	14.62	16.14	17.16
TransAgg [34]	ViT-B	-	38.10	68.42	79.08	93.51	32.07	53.26	42.67	-	-	-
MagicLens [33]	ViT-B	166M	27.00	58.00	70.90	91.10	26.30	47.40	36.85	**23.80**	**25.80**	**26.70**
TEOZCIR (our)	ViT-B	147M	**39.64**	**69.76**	**80.34**	**94.84**	**34.69**	**56.32**	**45.51**	15.72	17.34	18.19

TABLE II: Comparison with LLM powered state-of-the-art approaches in ZS-CIR (Zero Shot Composed Image Retrieval). Best score in bold, second best score underlined. The metrics with "–" refer no value is provided in author's paper.

Method	Backbone	LLM	CIRR				FashionIQ			CIRCO		
			Recall@K				Recall@K (val)			mAP@K		
			K=1	K=5	K=10	K=50	K=10	K=50	Avg	K=10	K=25	K=50
SEIZE [11]	ViT-B	GPT-3.5-turbo	27.47	57.42	70.17	-	28.94	49.86	39.40	**19.64**	**21.55**	**22.49**
CIReVL [12]	ViT-B	GPT-3.5-turbo/GPT-4	23.94	52.51	66.00	86.95	28.29	49.35	38.82	15.42	17.00	17.82
LDRE [14]	ViT-B	GPT-3.5-turbo	25.69	55.13	69.04	89.90	24.81	45.63	35.22	18.32	20.21	21.11
TEOZCIR (our)	ViT-B	✗	**39.64**	**69.76**	**80.34**	**94.84**	**34.69**	**56.32**	**45.51**	15.72	17.34	18.19

SEARLE, SEARLE-OTI and Context-I2W. In particular, compared to Context-I2W, our method achieves approximately (+14.00) point gains on all metrics (see Tab. I). This underscores the model's strong generalization capabilities. In Tab. II TEOZCIR exceeds several recent LLM-powered methods, including CIReVL, SEIZE, and LDRE, avoiding LLMs. Our model achieves the highest Recall@1 of 39.64, outperforming SEIZE by a margin of (+12.17) points, which is also consistent with the other metrics, with an average of (+12.34). These findings highlight TEOZCIR's efficiency and robustness, even trained on a dataset of only 32k triplets.

2) FashionIQ: Along the stronger performance on CIRR, TEOZCIR achieved pronounced gains on FashionIQ. Compared to TransAgg, which is the second strongest, our model records significantly better results on FashionIQ, with improvements of (+2.62) points in Recall@10 and (+3.06) points in Recall@50 (see Tab. I). This indicates that TEOZCIR performed well in varied datasets, excelling in FashionIQ which is domain specific. This stronger performance gap on FashionIQ demonstrates the TEOZCIR's ability to capture fine-grained attribute shifts in fashion items. We outperform both MagicLens and Context-I2W by large margins (+8.92) and (+7.42) points in Recall@50 and (+8.39) and (+6.89) points in Recall@10 respectively, while using fewer trainable parameters. From Tab. II, TEOZCIR surpasses LLM-based methods, such as SEIZE, achieving a significant improvement of (+5.75) points in Recall@10 and (+6.46) points in Recall@50, without relying on LLMs. These results reinforce the strength of our architecture in modeling compositional queries, even in domain-specific settings like FashionIQ.

3) CIRCO: We report the CIRCO results in the right section of Tab. I. Our model achieved substantial gains in mAP metrics, ranking as second best compared with all the considered models. TEOZCIR shows a consistent improvement of approximately (+1.1) points in each metric compared to Context-I2W. MagicLens demonstrates reasonable performance over our model. However, it is crucial to consider factors, such as the number of trainable parameters (see Tab. I) and the scale of the training dataset. Despite being trained on a fraction of the data, TEOZCIR demonstrates strong performance. MagicLens is built on a significantly larger model and trained with proprietary dataset of 36.7 million high-quality triplets (not publicly available). Instead, our model was trained on only 32K samples, which is 0.087% (approximately $1/10th$ of 1%) of MagicLens's training data. Nonetheless, TEOZCIR achieves comparable performance, highlighting the effectiveness of our lightweight design and the impact of the Text-Aware Query Combiner (TAQC) and Query Enhancer

Network (QENet) modules. When compared to LLM-based models (see Tab. II), TEOZCIR surpassed CIReVL and delivered competitive results against SEIZE and LDRE. The results on CIRCO dataset underscore the strength and generalization capability of our approach without relying on LLMs. Given this, it is reasonable to expect that TEOZCIR would achieve even better performance if trained on a larger dataset, further amplifying the benefits of our proposed modules.

V. Ablation Studies

We carry out targeted ablation studies on CIRR and FashionIQ to analyze the contributions of each TEOZCIR's component, the diverse weighting mechanisms and the importance of our image token $\mathbf{T}^*$ and the training dataset used.

A. Effect of each component

We evaluate how each core component of TEOZCIR contributes to its overall performance. As shown in Tab. III, it is the combination of TAQC and QENet that truly stands out consistently achieving the highest Recall@K scores on both CIRR and FashionIQ, and even surpassing existing state-of-the-art models. This success comes from how well the two modules complement each other, and working together to build a more powerful and refined query representation.

TABLE III: Ablation study of each component contribution.

TAQC	QENet	CIRR				FashionIQ		
		Recall@K (test)				Recall@K (val)		
		K=1	K=5	K=10	K=50	K=10	K=50	Avg
✓	✗	21.89	42.91	55.04	81.86	20.88	36.42	28.65
✗	✓	38.09	68.01	78.61	93.04	33.27	54.02	43.64
✓	✓	**39.64**	**69.76**	**80.34**	**94.84**	**34.69**	**56.32**	**45.51**

B. Effect of diverse weighting mechanisms

We examine the impact of two weighting mechanisms that modulate the importance of text and image modality (see Fig. 3 (a) and (b) respectively). Results reported in Tab. IV show that both weighting modalities are needed for better performance.

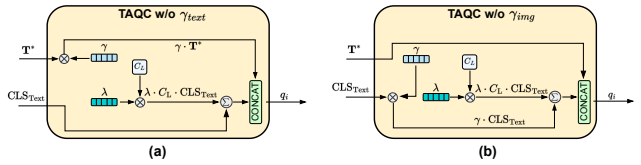

Fig. 3: Ablation study: **(a)** $q_i = \gamma \cdot \mathbf{T}^* \parallel (\text{CLS}_{\text{Text}} + \lambda \cdot C_L \cdot \text{CLS}_{\text{Text}})$ **(b)** $q_i = \mathbf{T}^* \parallel (\gamma \cdot \text{CLS}_{\text{Text}} + \lambda \cdot C_L \cdot \text{CLS}_{\text{Text}})$, where $\parallel$ is the CONCAT operation in short. Text-Aware Query Combiner (TAQC) combines image and text features, where scaling factors γ and λ dynamically adapt the contributions of the learnable image token $\mathbf{T}^*$ and the text class token CLS_{Text}.

C. Effect of Image Token

We evaluate the importance of the learnable image token $\mathbf{T}^*$ in TEOZCIR. As shown in Tab. V, its inclusion consistently improves Recall@K compared to the variant without it.

TABLE IV: Ablation study of diverse weighting mechanisms.

Method	CIRR Test				FashionIQ		
	Recall@K (test)				Recall@K (val)		
	K=1	K=5	K=10	K=50	K=10	K=50	Avg
TEOZCIR w/o γ_{text}	36.96	68.07	79.95	94.15	33.46	54.21	43.83
TEOZCIR w/o γ_{img}	37.93	69.25	80.03	94.72	34.21	55.46	44.84
TEOZCIR (our)	**39.64**	**69.76**	**80.34**	**94.84**	**34.69**	**56.32**	**45.51**

TABLE V: Ablation study of the image token $\mathbf{T}^*$ contribution.

Method	CIRR Test				FashionIQ		
	Recall@K (test)				Recall@K (val)		
	K=1	K=5	K=10	K=50	K=10	K=50	Avg
TEOZCIR w/o $\mathbf{T}^*$	38.89	68.32	79.83	94.01	33.27	54.02	43.64
TEOZCIR (our)	**39.64**	**69.76**	**80.34**	**94.84**	**34.69**	**56.32**	**45.51**

D. Effect of different available training dataset

Finally, we train TEOZCIR on the other available dataset LaSco [38] and evaluate its zero-shot performance. As shown in Tab. VI, the drop in generalization confirms the effectiveness of our training dataset choice.

TABLE VI: Ablation study of different training dataset.

Method	CIRR Test				FashionIQ		
	Recall@K (test)				Recall@K (val)		
	K=1	K=5	K=10	K=50	K=10	K=50	Avg
TEOZCIR w LaSco [38]	32.76	64.55	76.89	92.80	31.66	52.54	42.10
TEOZCIR (our)	**39.64**	**69.76**	**80.34**	**94.84**	**34.69**	**56.32**	**45.51**

VI. Conclusion and Future Work

In this paper, we present a novel multimodal retrieval model named TEOZCIR for the zero-shot CIR task. Our model revolves around fusing image and text contextual embeddings, by enriching visual interpretation with induced enhance textual context. The experimental results in CIRR and FashionIQ, show that our model successfully outperformed the existing compared approaches, increasing the retrieval accuracy across all the metrics. We demonstrated the importance of textual context in the final query representation along with visual context. It is evident in the CIRCO dataset the model performed reasonably better, surpassing several existing methods while relying only on a training dataset of size 32k samples. We believe that TEOZCIR's performance on the CIRCO dataset could further improve with a larger, well-defined training set. Alternative weighting mechanisms can be explored and analyzed in future work. Additionally, inspired by the improvements of powerful LLM models, we plan to investigate the integration of LLMs into our framework.

References

[1] B. Barz and J. Denzler, "Content-based image retrieval and the semantic gap in the deep learning era," *CoRR*, vol. abs/2011.06490, 2020. [Online]. Available: https://arxiv.org/abs/2011.06490

[2] A. Miech, J.-B. Alayrac, I. Laptev, J. Sivic, and A. Zisserman, "Thinking fast and slow: Efficient text-to-visual retrieval with transformers," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9826–9836.

[3] Y. Gao, M. Wang, H. Luan, J. Shen, S. Yan, and D. Tao, "Tag-based social image search with visual-text joint hypergraph learning," in *Proc. of the ACM International Conference on Multimedia (ACM MM)*, 2011.

[4] A. Alfarrarjeh, C. Shahabi, and S. H. Kim, "Hybrid indexes for spatial-visual search," in *Proc. of the on Thematic Workshops of ACM Multimedia*, 2017, pp. 75–83.

[5] R. Hinami and S. Satoh, "Discriminative learning of open-vocabulary object retrieval and localization by negative phrase augmentation," *arXiv preprint arXiv:1711.09509*, 2017.

[6] J. Kim, E. Cho, S. Kim, and H. J. Kim, "Retrieval-augmented open-vocabulary object detection," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[7] F. Shao, Y. Luo, L. Zhang, L. Ye, S. Tang, Y. Yang, and J. Xiao, "Improving weakly supervised object localization via causal intervention," in *Proc. of the ACM International Conference on Multimedia (ACM MM)*, 2021, pp. 3321–3329.

[8] C. Tan, G. Gu, T. Ruan, S. Wei, and Y. Zhao, "Dual-gradients localization framework for weakly supervised object localization," in *Proc. of the ACM International Conference on Multimedia (ACM MM)*, 2020.

[9] X. Tian, J. Liu, Z. Zhang, C. Wang, Y. Qu, Y. Xie, and L. Ma, "Hierarchical walking transformer for object re-identification," in *Proc. of the ACM International Conference on Multimedia (ACM MM)*, 2022, pp. 4224–4232.

[10] L. Agnolucci, A. Baldrati, M. Bertini, and A. Del Bimbo, "isearle: Improving textual inversion for zero-shot composed image retrieval," *arXiv preprint arXiv:2405.02951*, 2024.

[11] Z. Yang, S. Qian, D. Xue, J. Wu, F. Yang, W. Dong, and C. Xu, "Semantic editing increment benefits zero-shot composed image retrieval," in *Proc. of the ACM International Conference on Multimedia (ACM MM)*, 2024, pp. 1245–1254.

[12] S. Karthik, K. Roth, M. Mancini, and Z. Akata, "Vision-by-language for training-free compositional image retrieval," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2024.

[13] K. Saito, K. Sohn, X. Zhang, C.-L. Li, C.-Y. Lee, K. Saenko, and T. Pfister, "Pic2word: Mapping pictures to words for zero-shot composed image retrieval," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[14] Z. Yang, D. Xue, S. Qian, W. Dong, and C. Xu, "LDRE: LLM-based divergent reasoning and ensemble for zero-shot composed image retrieval," in *Proc. of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 80–90.

[15] S. Jain, K. Pulaparthi, and C. Fulara, "Content based image retrieval," *Int. J. Adv. Eng. Glob. Technol*, vol. 3, no. 10, pp. 1251–1258, 2015.

[16] M. Lux, "Content based image retrieval with LIRe," in *Proc. of the ACM International Conference on Multimedia (ACM MM)*, 2011, pp. 735–738.

[17] G. Tolias, Y. Kalantidis, and Y. Avrithis, "Symcity: Feature selection by symmetry for large scale image retrieval," in *Proc. of the ACM International Conference on Multimedia (ACM MM)*, 2012, pp. 189–198.

[18] J. Cai, Z.-J. Zha, W. Zhou, and Q. Tian, "Attribute-assisted reranking for web image retrieval," in *Proc. of the ACM International Conference on Multimedia (ACM MM)*, 2012, pp. 873–876.

[19] Q. Hu, J. Wu, J. Cheng, L. Wu, and H. Lu, "Pseudo label based unsupervised deep discriminative hashing for image retrieval," in *Proc. of the ACM International Conference on Multimedia (ACM MM)*, 2017, pp. 1584–1590.

[20] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo, "Effective conditioned and composed image retrieval combining clip-based features," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 21 466–21 474.

[21] G. Zhang, S. Wei, H. Pang, and Y. Zhao, "Heterogeneous feature fusion and cross-modal alignment for composed image retrieval," in *Proc. of the ACM International Conference on Multimedia (ACM MM)*, 2021, pp. 5353–5362.

[22] H. Wen, X. Zhang, X. Song, Y. Wei, and L. Nie, "Target-guided composed image retrieval," in *Proc. of the ACM International Conference on Multimedia (ACM MM)*, 2023, pp. 915–923.

[23] X. Jiang, Y. Wang, M. Li, Y. Wu, B. Hu, and X. Qian, "Cala: Complementary association learning for augmenting comoposed image retrieval," in *Proc. of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 2177–2187.

[24] Y. Yang, M. Wang, W. Zhou, and H. Li, "Cross-modal joint prediction and alignment for composed query image retrieval," in *Proc. of the ACM International Conference on Multimedia (ACM MM)*, 2021, pp. 3303–3311.

[25] F. Zhang, M. Yan, J. Zhang, and C. Xu, "Comprehensive relationship reasoning for composed query based image retrieval," in *Proc. of the ACM International Conference on Multimedia (ACM MM)*, 2022, pp. 4655–4664.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. of the International Conference on Machine Learning*. PMLR, 2021.

[27] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. of the International Conference on Machine Learning*. PMLR, 2022.

[28] Y. bai, X. Xu, Y. Liu, S. Khan, F. Khan, W. Zuo, R. S. M. Goh, and C.-M. Feng, "Sentence-level prompts benefit composed image retrieval," in *Proc. of the International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=m3ch3kJL7q

[29] A. Baldrati, M. Bertini, T. Uricchio, and A. D. Bimbo, "Composed image retrieval using contrastive learning and task-oriented clip-based features," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023.

[30] Z. Liu, W. Sun, Y. Hong, D. Teney, and S. Gould, "Bi-directional training for composed image retrieval via text prompt learning," in *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5753–5762.

[31] N. Cohen, R. Gal, E. A. Meirom, G. Chechik, and Y. Atzmon, ""this is my unicorn, fluffy": Personalizing frozen vision-language representations," in *Proc. of the European Conference on Computer Vision*. Springer, 2022.

[32] G. Gu, S. Chun, W. Kim, Y. Kang, and S. Yun, "Language-only training of zero-shot composed image retrieval," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[33] K. Zhang, Y. Luan, H. Hu, K. Lee, S. Qiao, W. Chen, Y. Su, and M.-W. Chang, "MagicLens: Self-supervised image retrieval with open-ended instructions," in *Proc. of the International Conference on Machine Learning*, vol. 235. PMLR, 21–27 Jul 2024. [Online]. Available: https://proceedings.mlr.press/v235/zhang24an.html

[34] Y. Liu, J. Yao, Y. Zhang, Y.-F. Wang, and W. Xie, "Zero-shot composed text-image retrieval," in *Proc. of the British Machine Vision Conference (BMVC)*. BMVA, 2023. [Online]. Available: https://papers.bmvc2023.org/0381.pdf

[35] Z. Liu, C. Rodriguez-Opazo, D. Teney, and S. Gould, "Image retrieval on real-life images with pre-trained vision-and-language models," in *Proc. of the IEEE/CVF International Conference on Computer Vision*, 2021.

[36] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris, "Fashion IQ: A new dataset towards retrieving images by natural language feedback," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11 307–11 317.

[37] A. Baldrati, L. Agnolucci, M. Bertini, and A. Del Bimbo, "Zero-shot composed image retrieval with textual inversion," in *Proc. of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 338–15 347.

[38] M. Levy, R. Ben-Ari, N. Darshan, and D. Lischinski, "Data roaming and quality assessment for composed image retrieval," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, Mar. 2024, pp. 2991–2999. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/28081

[39] L. Ventura, A. Yang, C. Schmid, and G. Varol, "CoVR-2: Automatic data construction for composed video retrieval," *IEEE Transactions on Pattern Analisys and Machine Intelligence*, 2024.

[40] G. Gu, S. Chun, W. Kim, H. Jun, Y. Kang, and S. Yun, "Compodiff: Versatile composed image retrieval with latent diffusion," 2024. [Online]. Available: https://arxiv.org/abs/2303.11916

[41] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays, "Composing text and image for image retrieval-an empirical odyssey," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6439–6448.