This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

CoLLM: A Large Language Model for Composed Image Retrieval

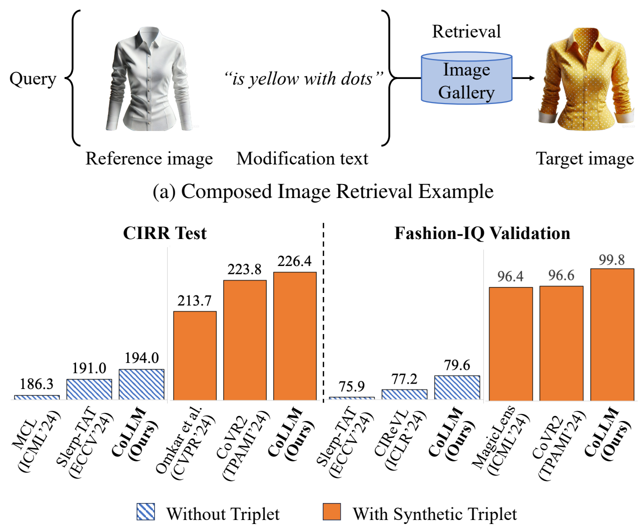
Chuong Huynh[1*] Jinyu Yang[2†] Ashish Tawari[2] Mubarak Shah[2,3] Son Tran[2]
Raffay Hamid[2] Trishul Chilimbi[2] Abhinav Shrivastava[1]

[1] University of Maryland, College Park [2] Amazon
[3] Center for Research in Computer Vision, University of Central Florida
[1]{chuonghm, abhinav}@cs.umd.edu
[2]{viyjy, atawari, sontran, raffay, trishulc}@amazon.com [3]shah@crcv.ucf.edu

Abstract

Composed Image Retrieval (CIR) is a complex task that aims to retrieve images based on a multimodal query. Typical training data consists of triplets containing a reference image, a textual description of desired modifications, and the target image, which are expensive and time-consuming to acquire. The scarcity of CIR datasets has led to zero-shot approaches utilizing synthetic triplets or leveraging vision-language models (VLMs) with ubiquitous web-crawled image-caption pairs. However, these methods have significant limitations: synthetic triplets suffer from limited scale, lack of diversity, and unnatural modification text, while image-caption pairs hinder joint embedding learning of the multimodal query due to the absence of triplet data. Moreover, existing approaches struggle with complex and nuanced modification texts that demand sophisticated fusion and understanding of vision and language modalities. We present CoLLM, a one-stop framework that effectively addresses these limitations. Our approach generates triplets on-the-fly from image-caption pairs, enabling supervised training without manual annotation. We leverage Large Language Models (LLMs) to generate joint embeddings of reference images and modification texts, facilitating deeper multimodal fusion. Additionally, we introduce Multi-Text CIR (MTCIR), a large-scale dataset comprising 3.4M samples, and refine existing CIR benchmarks (CIRR and Fashion-IQ) to enhance evaluation reliability. Experimental results demonstrate that CoLLM achieves state-of-the-art performance across multiple CIR benchmarks and settings. MTCIR yields competitive results, with up to 15% performance improvement. Our refined benchmarks provide more reliable evaluation metrics for CIR models, contributing to the advancement of this important field. Project page is at collm-cvpr25.github.io.

*This work was done during internship at Amazon. † Project lead.



(a) Composed Image Retrieval Example

(b) Recall Sum of state-of-the-art methods on CIR benchmarks

Figure 1. (a) An example of CIR. (b) Recall Sum at {1,10,50} for CIRR and {10, 50} for Fashion-IQ between CoLLM and state-of-the-art (SoTA) models under zero-shot settings. We evaluate two training scenarios: (i) without triplet data and (ii) with synthetic triplet data.

1. Introduction

Composed Image Retrieval (CIR) enhances traditional image retrieval [42, 62] by combining text and image queries, offering greater flexibility in search systems, with applications in e-commerce, fashion, and design [9, 12, 20, 63, 67]. As illustrated in Fig. 1a, CIR retrieves similar items, such as shirts, where the reference image provides a visual basis, and the modification text specifies desired modifications. By expanding the capabilities of conventional image or text-based searches, CIR allows for more nuanced and specific queries. However, this advanced approach also presents significant challenges compared to traditional image retrieval.

Supervised CIR approaches face challenges in data acquisition. They require high-quality CIR triplets (reference image, target image, and modification text), collected through a labor-intensive and expensive process. Consequently, existing CIR triplet datasets are limited in scale and domain coverage, restricting model generalizability. To overcome these limitations, recent CIR methods [14, 23, 28, 58, 64] have adopted zero-shot approaches [36]. These can be broadly categorized into two strategies: (i) leveraging vision-language models [29, 45, 53] (VLMs) for composed query embedding generation, and (ii) generating synthetic triplets for supervised training. VLM-based approaches utilize pre-trained models that already align vision and language features in latent space. These methods rely on large-scale image-caption pairs and follow two main directions, i.e., textual inversion [8, 15, 55] and direct interpolation [21]. Synthetic triplet generation approaches [23, 28, 64] employ Large Language Models (LLMs) [7, 57, 65] to generate modification text, with CompoDiff [14] further utilizing diffusion models [5, 46] to create synthetic reference-target image pairs. Despite the recent advances in CIR, several critical challenges persist:

- **[Data] Limitation-1:** VLM-based methods may hinder efficient and effective composed query embedding learning by relying solely on image-caption pairs.
- **[Data] Limitation-2:** Existing synthetic triplet datasets often suffer from a lack of diversity and unnatural modification text. Moreover, these datasets are either very small or closed-source, impeding research progress in the field.
- **[Model] Limitation-3:** Current methods for composed query embeddings mainly use shallow transformer models or linear interpolation. While these methods are computationally efficient, they lack the ability to capture the full complexity of the composed understanding tasks.
- **[Evaluation] Limitation-4:** Existing CIR benchmarks are often compromised by noise, particularly in the form of ambiguous samples. Ambiguity arises when multiple target images can correctly match a single query, but only one is labeled as the ground truth, ignoring valid matches. Although CIRCO [4] attempts to address this issue, its efforts are limited in scale. This ambiguity in benchmarks hinders meaningful model evaluation and comparison.

We propose CoLLM, an LLM-based CIR approach to address the aforementioned limitations. CoLLM tackles **Limitation-1** by dynamically synthesizing triplets from image-caption pairs, introducing two key components: a reference image embedding synthesis and a modification text synthesis module. The former employs Spherical Linear Interpolation [52] (Slerp) to generate an intermediate embedding between a given image and its nearest neighbor within the training batch, serving as a synthesized reference image embedding. The latter utilizes a pre-defined text interpolation template to generate modification text based on the current caption and that of the nearest image neighbor. This strategy effectively leverages the vast amount of readily available image-caption pairs on the Internet, enabling a model's training in a supervised CIR manner. By doing so, CoLLM overcomes the scarcity of labeled triplet data and paves the way for more robust and scalable CIR models.

To address **Limitation-2**, we introduce Multi-Text CIR (MTCIR), a synthetic dataset of 3.4M image pairs with 17.7M modification texts. MTCIR focuses on two often-overlooked aspects: image diversity and naturalistic modification texts. We curated images from diverse sources to ensure variety. For modification text generation, we employ a two-stage approach using Multi-modal LLM (MLLM) [34, 35] for detailed captioning and LLM for describing inter-caption differences. Uniquely, MTCIR provides multiple short modification texts for each image pair, covering various attributes. This better reflects human query formulation, offering a more realistic, comprehensive training foundation for CIR models.

To overcome **Limitation-3**, CoLLM harnesses the power of LLMs for composed query understanding. It is motivated by the extensive world knowledge embedded in pre-trained LLMs. With their deep semantic understanding, we posit that LLMs are superior to shallow transformers and simple embedding interpolation techniques in comprehending the intricate relationships between reference images and modification texts. By leveraging LLMs, CoLLM aims to capture nuanced semantic connections, enhancing the quality and relevance of composed image retrieval results.

To address **Limitation-4**, we refine two popular CIR benchmarks: CIRR [37] and Fashion-IQ [61]. We use MLLMs to evaluate sample ambiguity in each benchmark and regenerate clear modification text for ambiguous ones. Our pipeline incorporates multiple validation steps to guarantee the enhanced quality of the refined samples.

Our main contributions can be summarized as: (i) We propose a method to synthesize CIR triplets on-the-fly from image-caption pairs, which can even outperform models trained on CIR triplets, eliminating the need for costly CIR datasets. (ii) We collect and will release MTCIR, a new CIR triplet dataset covering 3.4 million image pairs with 17.7 million modification texts. To our knowledge, MTCIR is the largest open-source synthetic CIR dataset. (iii) We introduce an approach to leverage LLMs for composed query understanding, utilizing their instruction-following and embedding generation capabilities. (iv) We refine CIRR and Fashion-IQ to provide more robust evaluation benchmarks for the CIR community. Extensive experiments on popular CIR benchmarks and settings demonstrate the effectiveness of our model innovations and new datasets (Fig. 1b).

2. Related Works

Composed Image Retrieval. Composed Image Retrieval (CIR) has garnered significant attention due to its flexibility in search systems [37, 43, 48]. Zero-shot CIR methods have been extensively explored, with textual inversion [4, 8, 15, 49, 55] emerging as a prominent technique. This approach maps image encoder outputs to text encoder inputs, creating pseudo-word tokens. SlerpTAT [21] emphasizes the text encoder's importance by re-aligning visual and textual embeddings. While most zero-shot CIR models utilize image-caption pairs for training, our work introduces an innovative method to synthesize triplets from these pairs during training. This method enables supervised CIR style training, enhancing the model's understanding of composed queries without relying on real CIR datasets.

Recent works enhance enhancing models' language understanding using Large Language Models (LLMs) [26, 30], primarily leveraging their text generation abilities. MCL [30] composes image-text query in the LLM input space and aligns joint embedding with the target image caption. However, this may introduce noise due to limited visual information in the target image captions. CIReVL [26] employs LLMs to generate captions based on the reference image and modification text, subsequently retrieving the target image using a text-to-image retrieval approach. Our model also utilizes LLMs but differs by directly producing a composed query embedding for target image retrieval, potentially reducing intermediate steps and associated errors.

CIR with Synthetic Datasets. The scarcity of supervised CIR datasets has prompted recent studies to leverage generative models for synthetic triplet creation. CompoDiff [14] employs image editing pipelines [5, 46] to generate target images, though their nature limits performance. Other approaches [23, 28, 58, 59, 64] utilize LLMs to generate modification text for real image pairs. Our work introduces a novel two-stage approach, combining MLLMs [34, 35] for detailed captioning and LLMs [1, 2] for describing inter-caption differences. Distinctively, our synthetic triplets provide multiple concise modification texts for each image pair, encompassing various attributes, in contrast to the conventional single, lengthy modification text. This enables a more nuanced representation of image modifications, potentially improving CIR model performance and versatility.

Large Language Models. Recent advancements in LLMs have expanded their applications beyond text generation to include image understanding [34, 35] and text embedding generation [40, 60, 68]. Large Language Embedding Models (LLEMs) are specialized versions trained using contrastive learning, leveraging LLM knowledge for embedding generation while often disabling text generation capabilities. To enhance text embedding quality, [27, 31, 39] propose removing causal attention in LLMs, thereby improving text information encoding efficiency and training these modified LLMs on text retrieval datasets. This text retrieval capability has been further extended to text-image retrieval by aligning visual and LLM text embedding spaces [22, 25, 38]. These developments demonstrate the potential of LLMs for composed query understanding in CIR tasks. Our work builds upon this foundation, extending LLMs/LLEMs to generate composed query embeddings by incorporating reference images and modification text. This novel approach leverages the multimodal capabilities of LLMs to enhance the performance of CIR systems, leading to more accurate and context-aware image retrieval.

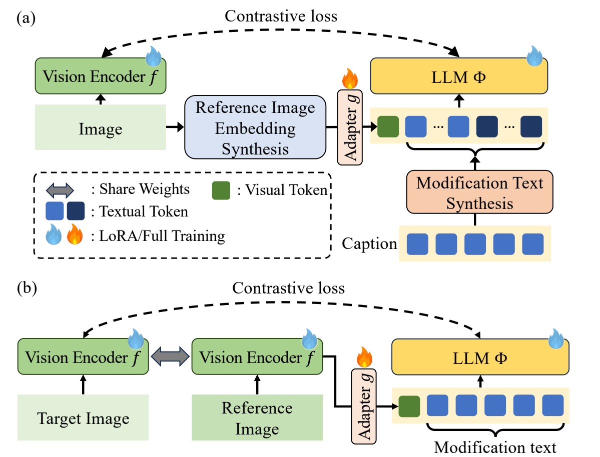


Figure 2. An overview of our model and training strategies when using (a) image-caption pairs and (b) CIR triplets.

3. Method

This section outlines our methodology, starting with a brief overview of the model architecture, followed by a formal CIR problem definition. We then detail our triplet synthesis strategy, including reference image embedding and modification text synthesis, using image-caption pairs. Lastly, we describe our LLM-based query composition approach.

3.1. Model Architecture

As illustrated in Fig. 2, CoLLM consists of several essential components: (1) a vision encoder $f(\cdot)$ for extracting image features; (2) modules for synthesizing reference image embeddings and modification text; (3) an image adapter $g(\cdot)$ that maps visual features into the language model's semantic space; (4) a LLM $\Phi(\cdot)$ that processes multimodal queries and (5) a projection layer $proj(\cdot)$ (omitted from the figure for simplicity) that maps the LLM output to a suitable representation for retrieval. It is important to note that Fig. 2 (a) and Fig. 2 (b) illustrate architectures designed for two distinct input formats. The former is tailored for image-caption pairs, while the latter is optimized for CIR triplets.

3.2. CoLLM with Image-Caption Pairs as Input

Let $\mathcal{X} = \{(v_i, w_i)\}_{i=1}^N$ denote a set of image-caption pairs, where v_i and w_i represent the image and caption of the i^{th} sample, respectively. To effectively leverage the vast amount of image-caption pairs, we introduce a novel ap-

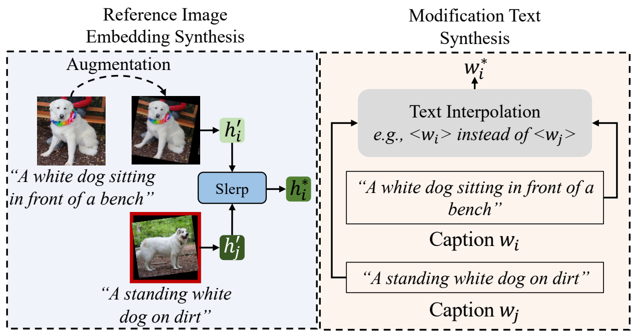


Figure 3. An overview of reference image embedding synthesis and modification text synthesis. The red-framed image represents the nearest neighbor of the augmented image in the training batch.

proach that synthesizes a CIR triplet on-the-fly for each image-caption pair (v_i, w_i). In this process, v_i serves as the target image, while two distinct modules generate the remaining components: (i) a reference image embedding synthesizer and (ii) a modification text synthesizer (Fig. 2a). Notably, we do not generate an actual image for reference image synthesis. Instead, we synthesize the reference image embedding, which is computationally more efficient and allows seamless integration into the CIR pipeline. This formulation enables CoLLM to exploit the rich information in image-caption pairs, transforming them into CIR triplets that can be used for training. The following subsections detail the specific mechanisms for synthesizing the reference image embedding and the modification text.

Reference Image Embedding Synthesis. Our reference image embedding synthesis process is illustrated in Fig. 3 (left). We begin by applying an augmentation $t \sim \mathcal{T}$ to the input image v_i, where $\mathcal{T}$ represents a family of augmentations. The resulting augmented image is denoted as v_i', with its corresponding embedding $\mathbf{h}_i' = f(v_i')$. For $\mathbf{h}_i'$, we identify its in-batch nearest neighbor v_j' with j defined as:

$$j = \underset{j \neq i;\, j \in \{1, \ldots \mathcal{B}\}}{\arg\max}\ \mathrm{sim}(\mathbf{h}_i', \mathbf{h}_j'), \tag{1}$$

where $\mathcal{B}$ is the batch size, $\mathbf{h}_j' = f(v_j')$, and $\mathrm{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T\mathbf{v}/(\|\mathbf{u}\|\|\mathbf{v}\|)$ denotes cosine similarity.

We then employ Spherical Linear Interpolation (Slerp) [52] to synthesize the reference image embedding by interpolating between $\mathbf{h}_i'$ and $\mathbf{h}_j'$. The synthesized reference image embedding is obtained as:

$$\begin{aligned} \theta &= \arccos(\mathbf{h}_i' \cdot \mathbf{h}_j') \\ \mathbf{h}_i^* &= \frac{\sin(\alpha\theta)}{\sin\theta}\mathbf{h}_i' + \frac{\sin((1-\alpha)\theta)}{\sin\theta}\mathbf{h}_j', \end{aligned} \tag{2}$$

where $\alpha \in [0, 1]$ is a hyper-parameter controlling the interpolation strength. The intuition behind this approach is that an image with embedding $\mathbf{h}_i^*$ shares certain visual similarities with the target image v_i. Specifically, a larger α results in $\mathbf{h}_i^*$ more closely resembling $\mathbf{h}_i'$, allowing for fine-grained control over the synthesized embedding.

Modification Text Synthesis. For the augmented image v_i', its nearest neighbor is (v_j', w_j). To generate the modification text w_i^*, we use text interpolation to combine w_i and w_j using pre-defined templates, as illustrated in Fig. 3 (right). The complete set of templates is provided in the supplementary material. During training, we randomly select a template for each sample to ensure diversity in the synthesized modification text. It is worth noting that we do not simply use w_i as the modification text for two primary reasons: (i) w_i alone fails to capture the visual differences between the reference image and the target image. (ii) Using only w_i could lead to model cheating, where it learns to rely solely on w_i for retrieval while ignoring the reference image. Our synthesis approach, in contrast, mimics real-world modification text, which often describes both similarities and differences between the reference and target images. Furthermore, this strategy forces the model to learn composed query embeddings by considering both the reference image and the modification text simultaneously.

By combining the synthesized reference embedding $\mathbf{h}_i^*$ and the interpolated modification text w_i^*, our method generates diverse CIR triplets from image-caption pairs. This approach enables effective training of the CoLLM model on large-scale image-caption datasets, effectively mitigating the dependency on scarce labeled triplet data.

Query Composition with LLM. To construct the composed query, we utilize a pre-trained LLM that processes the synthesized reference image embedding $\mathbf{h}_i^*$, image caption w_i, and the synthesized modification text w_i^*. We define three distinct composed embeddings:

$$\mathbf{c}_i^v = p(\Phi(g(\mathbf{h}_i^*))) \tag{3}$$

$$\mathbf{c}_i^w = p(\Phi(w_i)) \tag{4}$$

$$\mathbf{c}_i = p(\Phi([g(\mathbf{h}_i^*); w_i^*])) \tag{5}$$

where $[;]$ denotes an instruction template that combine two modalities (see supplementary material).

Training Objective. We employ a contrastive loss $\mathcal{L}_{cl}$ [44], consistent with previous works [56, 59], to compute the loss between query embeddings $\{\mathbf{c}_i^v, \mathbf{c}_i^w, \mathbf{c}_i\}$ and the target image embedding $\mathbf{z}_i = f(v_i)$. The final loss for each sample during pre-training is defined as:

$$\mathcal{L} = \frac{1}{3}\left(\mathcal{L}_{cl}(\mathbf{c}_i^v, \mathbf{z}_i) + \mathcal{L}_{cl}(\mathbf{c}_i^w, \mathbf{z}_i) + \mathcal{L}_{cl}(\mathbf{c}_i, \mathbf{z}_i)\right) \tag{6}$$

This formulation encourages the model to learn discriminative representations for different query types while maintaining consistency with the target image embedding. By combining losses from image-to-image ($\mathbf{c}_i^v$), text-to-image ($\mathbf{c}_i^w$), and composed ($\mathbf{c}_i$) queries, we ensure that the model learns to effectively process and align various input modalities for compositional image retrieval.

Table 1. Comparison of synthetic CIR training datasets.

Dataset	Public	Reference-Target Image Pair				Modification Text Generation			
		Type	Source	# Pairs	# Entities	LLM	LLM input	# Text	# Text/Pair
LaSCo [28]	✓	image	VQAv2 [13]	360K	82K	GPT-3 [6]	question-answer	360K	1
VDG [23]	✓	image	NLVR2 [54], COCO [32]	467K	183K	Vision-LLaMA2	image pair	523K	1.12
WebCoVR [59]	✓	video	WebVid2M [3]	1.6M	131K	Custom LLM	caption	1.6M	1
CC-CoIR [58]	✓	image	CC3M [51]	3.3M	357K	Custom LLM	caption	3.3M	1
SynthTriplets [14]		synth. image	SD [46], IP2P [5]	18.8M	37.6M	OPT-6.7B [65]	-	18.8M	1
MagicLens [64]		image	Web crawled	36.5M	-	PaLM [7]	image metadata	36.5M	1
MTCIR (Ours)	✓	image	LLaVA-558k [34]	3.4M	423K	Sonnet 3 [2]	synth. caption	17.7M	5.18

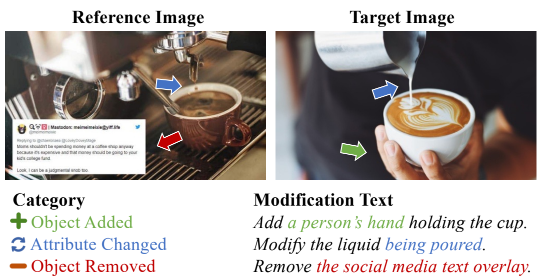


Figure 4. An example from the MTCIR dataset. Each sample contains multiple short texts describing different modifications.

3.3. CoLLM with CIR Triplet as Input

For CIR triplet inputs, our model design is shown in Fig. 2b. Let $\mathcal{X} = \{(v_i^r, v_i^t, w_i)\}_{i=1}^N$ be a set of triplets, where v_i^r, v_i^t, and w_i represent the reference image, target image, and modification text of the i^{th} sample, respectively. The reference embedding is obtained as $\mathbf{h}_i = f(v_i^r)$. The composed embedding is then computed as $\mathbf{c}_i = p(\Phi([g(\mathbf{h}_i); w_i]))$. The training objective is $\mathcal{L} = \mathcal{L}_{cl}(\mathbf{c}_i, \mathbf{z}_i)$, where $\mathbf{z}_i = f(v_i^t)$ denotes the target embedding.

4. Dataset Construction

This section outlines the data construction process for two primary objectives: (1) creating a novel CIR training dataset called Multi-Text CIR (MTCIR) and (2) refining widely used CIR benchmarks, specifically CIRR [37] and Fashion-IQ [61]. We detail the methodologies for each task, highlighting the improvements and innovations.

4.1. Multi-Text CIR (MTCIR) Dataset

Existing CIR triplet datasets [23, 28, 59] lack diversity and contain unnatural modification text. We introduce Multi-Text CIR (MTCIR) to address these limitations.

To enhance image diversity, we source images from the LLaVA-558k dataset [34] filtered from image caption datasets [41, 50, 51] based on noun-phrase frequency for broad concept coverage. We pair images using CLIP [45] visual similarity metrics, following CIRR's grouping approach, yielding 3.4M pairs from approximately 423K images. Leveraging the detailed captions generated by LLaVA-Next-34B [35] for each image in the LLaVA-558k dataset, we employ Claude 3 Sonnet [2]

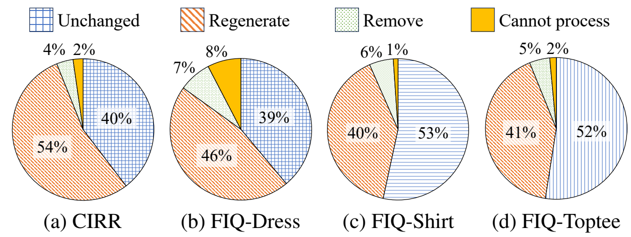


Figure 5. Statistical analysis of the refinement process for CIRR and FashionIQ (FIQ) datasets.

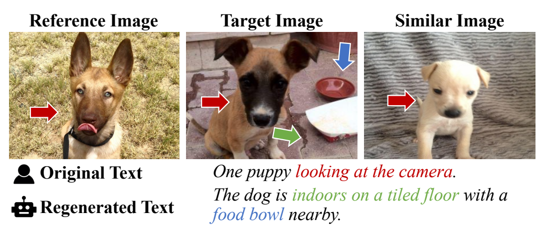


Figure 6. An example from the refined CIRR dataset. The original modification text yields multiple plausible results (shown as similar images to the target on the right). In contrast, the newly generated text accurately captures the distinctive features of the target image, differentiating it from other candidates.

(`20240229-v1:0`) to generate modification text for each image pair, using these captions as input. To control detail levels in Claude outputs, we define six modification text categories in the prompt, inspired by prior research [4, 37]. This approach maintains diverse aspects and prevents repetition. We also incorporate examples from the CIRR dataset to ensure human-like modification text. Detailed prompting strategies are provided in the supplementary material.

Our method generates multiple brief, focused texts for each image pair, each highlighting a unique aspect. This approach captures all changes between two images without producing overly long sentences, better aligning with human-written modification text. These short texts can be combined for training, enhancing dataset diversity. We obtained 17.7M text samples for the 3.4M image pairs, averaging 5.18 short sentences per pair. Fig. 4 illustrates an example, showcasing modification texts corresponding to specific categories and covering diverse attributes.

As shown in Table 1, MTCIR offers several advantages over existing synthetic CIR datasets: (i) Unlike [14], which

Table 2. Performance comparison of various models and visual encoders on CIR benchmarks, evaluated without training on triplet datasets. **Bold** and underlined values denote the best and second-best scores within each vision encoder group. Models incorporating LLMs in their architectures are marked with ⋆. Results reproduced by our team are indicated by ‡.

Method	CIRCO (mAP↑)			CIRR (Rec.↑)			FIQ (Rec.↑)	
	@5	@10	@50	@1	@10	@50	@10	@50
OpenAI CLIP-B/32								
PALAVRA [8]	4.6	5.3	6.8	16.6	58.5	84.0	19.8	37.3
SEARLE [4]	9.4	9.9	11.8	24.0	66.8	89.8	22.9	42.5
Slerp-TAT [21]	9.3	10.3	12.3	28.2	68.8	88.5	23.0	44.0
CIReVL⋆ [26]	**14.9**	**15.4**	**17.8**	23.9	66.0	87.0	**28.3**	**49.4**
CoLLM⋆	12.9	13.2	15.0	**28.6**	**71.8**	**92.7**	24.8	45.2
OpenAI CLIP-L/14								
Pic2World [49]	8.7	9.5	11.3	23.9	65.3	87.8	24.7	43.7
SEARLE [4]	11.7	12.7	15.1	24.2	66.3	88.8	25.6	46.2
LinCIR‡ [15]	13.0	13.9	16.2	24.6	66.9	88.8	26.4	46.6
ContextI2W [55]	13.0	13.8	16.0	25.7	68.6	89.6	27.8	48.9
MCL⋆ [30]	17.8	18.4	21.8	25.9	69.4	91.0	-	-
Slerp-TAT [21]	18.5	19.4	21.4	**30.9**	70.9	89.2	28.3	47.6
CIReVL⋆ [26]	18.6	19.0	21.8	24.6	64.9	86.3	28.6	48.6
CoLLM⋆	**20.3**	**20.8**	**23.4**	29.7	**72.8**	**91.5**	**30.1**	**49.5**
BLIP-L/16								
Slerp-TAT [21]	17.8	18.4	21.1	34.0	72.7	88.9	32.8	53.3
CoLLM⋆	**19.7**	**20.4**	**23.1**	**35.0**	**78.6**	**94.2**	**34.6**	**56.0**

Table 3. Performance of models trained on synthetic triplet datasets. **Bold** indicates the best method overall, while underline highlights the best method within the same vision encoder. Our CoLLM outperforms comparable methods in most metrics.

Method	Dataset	CIRR ↑			FIQ ↑	
		@1	@10	@50	@10	@50
CoCa-L/18 288 × 288						
MagicLens [64]	MagicLens [64]	33.3	77.9	94.4	38.1	58.3
EVA-CLIP ViT-G/14 364 × 364						
CoVR2 [58]	WV-CC-CoVIR [58]	43.7	84.0	**96.1**	38.2	58.4
OpenAI CLIP-L/14 224 × 224						
CompoDiff [14]	SynTrip18M [14]	18.2	70.8	90.3	36.0	48.6
MagicLens [64]	MagicLens [64]	30.1	74.4	92.6	30.7	52.5
CoLLM	MTCIR (ours)	34.7	77.0	93.1	32.9	54.2
BLIP-L/16 384 × 384; fine-tuned on COCO						
CASE [28]	LaSCo [28]	35.4	78.5	94.6	-	-
Omkar et al. [56]	WebCoVR [59]	40.1	78.9	94.7	30.3	46.5
CoLLM	MTCIR (ours)	**45.8**	**84.7**	95.8	**39.1**	**60.7**

uses synthetic images, MTCIR uses real images covering diverse concepts and domains. (ii) MTCIR provides multiple modification texts for each image pair, offering more natural descriptions and training flexibility. (iii) MTCIR is the largest public dataset to advance CIR research.

To prevent the leakage of biometric information, we further process MTCIR at both the image and text levels. For images, we apply face blurring to all instances. For text, we remove any content containing keywords related to human attributes (e.g., skin, hair, gender, age, race).

4.2. Refined CIRR and Fashion-IQ Datasets

Current CIR benchmarks often suffer from label ambiguity, where modification text lacks clarity, potentially matching multiple target images. To address this issue, we propose a refinement method using Claude 3 Sonnet [2] (`20240229-v1:0`) to enhance the validation sets of two well-known benchmarks: CIRR and Fashion-IQ. Each sample in these validation sets comprises a triplet containing a reference image, target image, and modification text.

Our refinement process, applicable to any benchmark, consists of validation, re-generation, and re-validation steps. For each triplet, we first identify the top-3 visually similar images to the target image based on CLIP visual scores, serving as hard negative samples for ambiguity assessment. We employ Claude 3 Sonnet to evaluate each triplet's ambiguity by attempting to identify the target image from these hard negatives. Triplets correctly identified by Claude 3 Sonnet are considered "good" samples and remain unchanged. For "bad" samples, we use Claude 3 Sonnet to re-generate the modification text. The model produces three new modification texts with increasing detail in a hierarchical structure, with each finer text building upon the coarser one. We then repeat the validation process, evaluating the new texts from coarsest to finest and selecting the best one. Any re-generated modification text that passes the validation process replaces the original text and concludes the process. If all generated texts fail the validation, the triplet is removed from the benchmark. Detailed process information is provided in the supplementary material. Besides, following MTCIR, biometric information is also removed from the refined benchmarks.

As illustrated in Fig. 5, up to 8% of samples are omitted due to Claude's processing limitations related to harmful content. Additionally, 4-7% of ambiguous samples that could not be effectively rewritten are removed. More than 40% of the original samples are retained, and 40-55% of triplets are successfully revised. The refined CIRR and Fashion-IQ datasets exhibit less ambiguity (Fig. 6), offering more robust evaluation benchmarks for the CIR community.

5. Experiments

Our proposed CoLLM framework adopts a two-stage training paradigm: pre-training and fine-tuning phases. In pre-training, CoLLM is exposed to a diverse corpus of image-caption pairs, enabling it to understand composed queries comprehensively. Subsequently, in the fine-tuning stage, we leverage our newly curated MTCIR datasets to verify their effective performance on complex multimodal tasks.

5.1. Pre-Training on Image-Caption Pairs

Training and Evaluation Datasets. The pre-training dataset consists of 5 million image-caption pairs, compiled from CC3M [51], LAION [50], and LLaVA-558K [34]. We evaluate our model on three CIR benchmarks: CIRCO [4],

Table 4. Performance of BLIP-L and CoLLM (pre-trained) trained on synthetic triplet datasets. **Bold** values indicate the best score within each group. Models trained on MTCIR demonstrate superior performance compared to those trained on other datasets.

Method	Dataset	CIRR ↑			FIQ ↑	
		@1	@10	@50	@10	@50
BLIP-L [29]	LaSCo [28]	36.6	78.4	94.6	24.8	44.0
	WebCoVR [59]	39.3	78.9	94.7	26.7	43.3
	MTCIR (ours)	**42.4**	**83.1**	**95.9**	**37.9**	**59.2**
CoLLM	LaSCo [28]	43.2	84.1	95.7	38.5	60.1
	MTCIR (ours)	**45.8**	**84.7**	**95.9**	**39.1**	**60.7**

CIRR [37] test set, and Fashion-IQ [61] validation set. All datasets are filtered to exclude biometric information.

Evaluation Metrics. We employ recall on the complete image index set as the evaluation metric for CIRR and Fashion-IQ. For CIRCO, we use mean Average Precision (mAP). We exclude the recall on subset for CIRR due to reliability concerns (see supplementary material for details).

Implementation Details. See supplementary material.

Quantitative Results. Table 2 compares our pre-trained model with other methods not trained on triplet datasets. We evaluate three variants of our architecture, each employing different vision encoders. CoLLM with the CLIP-B encoder achieves the second-best performance on the CIRCO and Fashion-IQ benchmarks while significantly improving on the CIRR benchmark. Although CIReVL [26] outperforms other methods, its use of the closed-source GPT-4 [1] introduces additional costs and increased inference latency, making it an unfair comparison.

For larger vision encoders, our model consistently outperforms both non-LLM (e.g., Slerp-TAT [21]) and LLM (CIReVL and MCL [30]) methods across all benchmarks. Notably, our CLIP-L variant achieves state-of-the-art results on the CIRCO benchmark, while BLIP-L significantly improves CIRR and Fashion-IQ scores. The consistent gains across benchmarks and architectures validate the robustness and generalizability of our triplet synthesis approach. By effectively bridging the gap between abundant image-caption data and scarce triplet annotations, our method paves the way for more scalable and efficient training of CIR models. This innovative strategy represents a significant step forward in addressing one of the critical challenges in the field, Paving the way for larger-scale pre-training and more complex multimodal interactions.

5.2. Fine-tuning on MTCIR

Training Dataset and Benchmarks. Despite the impressive performance achieved using only image-caption datasets, we hypothesize that CoLLM's performance can be further enhanced through fine-tuning on our newly developed MTCIR dataset. For a fair comparison, we also train CoLLM and baseline models on other public datasets, including WebCoVR [59] and LaSCo [28]. Our evaluation focuses on the CIRR [37] test and Fashion-IQ [61] validation sets. We exclude CIRCO [4] due to potential domain overlap between COCO [32] images used in CIRCO and those in LaSCo [28], as well as in the pre-training datasets.

Table 5. Performance of models on refined CIRR and Fashion-IQ validation sets. **Bold** indicates the highest score, while <u>underlined</u> values represent the best metric within the same vision encoder group. CoLLM+MTCIR continues to outperform other configurations on these new benchmarks.

Method	Dataset	Ref. CIRR ↑			Ref. FIQ ↑	
		@1	@10	@50	@10	@50
EVA-CLIP ViT-G/14 364 × 364						
CoVR2 [58]	WV-CC-VIR [58]	56.1	91.1	97.9	54.2	72.9
OpenAI CLIP-L/16 224 × 224						
MagicLens [64]	MagicLens [64]	42.3	86.3	<u>97.0</u>	45.5	68.1
CoLLM	MTCIR (ours)	<u>46.5</u>	<u>87.4</u>	96.1	<u>48.3</u>	<u>68.6</u>
BLIP-L/16 384 × 384; fine-tuned on COCO						
BLIP-L [29]	MTCIR (ours)	53.8	89.9	97.7	54.8	74.3
CoLLM	LaSCo [28]	57.3	92.0	98.1	56.9	75.9
CoLLM	MTCIR (ours)	<u>**60.4**</u>	<u>**92.6**</u>	<u>**98.2**</u>	<u>**57.2**</u>	<u>**76.4**</u>

Implementation Details. See supplementary material.

Quantitative Results. Table 3 presents the evaluation of models trained on publicly available synthetic CIR datasets, including our MTCIR. Fine-tuning CoLLM on MTCIR yields significant gains, improving CIRR@1 by 5% over CoLLM-OpenAI-CLIP-L/14 in Table 2. This improvement reflects MTCIR's enhanced generalizability and our superior data construction. Our method achieves exceptional results on all benchmarks using the same vision encoder compared to previous approaches and models trained on other synthetic datasets. Notably, our fine-tuned CoLLM outperforms MagicLens CoCa-L [64] and CoVR2 [58], despite these models using larger, more advanced vision encoders and being trained on larger datasets. These results underscore the importance of data quality over quantity.

MTCIR vs. Other CIR Datasets. Despite CoLLM's state-of-the-art performance, two questions remain: (i) Can MTCIR benefit other models? (ii) Can CoLLM achieve good performance when fine-tuned on other CIR datasets? To address these questions, we train a baseline BLIP-L [29] model on various CIR datasets, including MTCIR. Both LaSCo and MTCIR experiments are trained for one epoch. As shown in Table 4, we observe that: (i) BLIP-L shows significant improvement when using MTCIR as the training data, and (ii) CoLLM exhibits notable performance degradation when fine-tuned on the LaSCo dataset. These results demonstrate MTCIR's versatility as a plug-and-play dataset, capable of enhancing performance across various models. This underscores MTCIR's potential to make significant contributions to the CIR community.

5.3. Re-evaluation on Refined Benchmarks

Ambiguity in the CIRR and Fashion-IQ benchmarks potentially skews the metrics presented in previous tables, particularly for high-performing models. While CIRCO mitigates this issue, it still suffers from domain overlap with train-

Table 6. Ablation studies of proposed components.

(a) Reference image and modification text interpolation.

Image	Text	CIRCO (mAP ↑)	CIRR (Rec. ↑)
		14.7	170.1
	✓	14.5	176.4
✓		39.6	183.3
✓	✓	**52.8**	**194.5**

(b) Unimodal queries in Eq. (3) and Eq. (4).

Image-only	Text-only	CIRCO (mAP ↑)	CIRR (Rec. ↑)
✓		14.5	99.7
	✓	47.1	**196.3**
✓	✓	**52.8**	194.5

(c) Reference image embedding interpolation.

	CIRCO (mAP ↑)	CIRR (Rec. ↑)
Random In-Batch Sample	46.7	182.2
Nearest In-Batch Neighbor	**52.8**	**194.5**

ing images. To assess the impact of ambiguous samples, we re-evaluate the models on our newly refined CIRR and Fashion-IQ benchmarks. Table 5 presents the performance of these models on the revised benchmarks.

Our analysis shows that model rankings remain largely consistent before and after the benchmark refinement. As the modification texts in the refined benchmarks provide additional nuance, we anticipated that models with more robust composed query understanding capabilities would distinguish themselves. Indeed, CoLLM consistently outperforms other models across various settings on the refined benchmarks. Notably, while BLIP-L (MTCIR) trailed CoLLM (MTCIR) by 3.4% (CIRR@1)(Table 4), this performance gap widened to 6.9% (CIRR@1) on the refined CIRR benchmark. This suggests CoLLM better captures nuanced modification text and distinguishes visual differences between source and target images.

5.4. Ablation Studies

We conduct comprehensive ablation studies to evaluate the impact of our proposed components during the pre-training stage. Our experiments use the BLIP-L encoder, trained exclusively on the LLaVA-558k [34] dataset for one epoch. Table 6 presents the performance across various settings. We report the sum of mAP at $k = \{5, 10, 25, 50\}$ for CIRCO and Recall@$k = \{1, 5, 10, 50\}$ for CIRR. **Image and Text Interpolation are Crucial:** As shown in Table 6a, employing Slerp for reference image embedding significantly enhances the model's learning capabilities. Applying interpolation on modification text further improves performance. **Unimodal Queries are Beneficial:** We investigate the necessity of unimodal queries in Eq. (3) and Eq. (4) and observe that omitting unimodal queries during training, especially text-only queries, substantially degrades performance (Table 6b). **Nearest Neighbor is Essential:** For reference image embedding interpolation, one might hypothesize that a random in-batch sample could suffice for interpolation. However, our study demonstrates that

Table 7. Performance of different LLMs with CLIP-L/14 as the visual encoder. **Bold** and <u>underline</u> highlight the best and second best score. * indicates the original LLM for text generation; others are Large Language Embedding Models (LLEMs) fine-tuned for text retrieval.

LLM	CIRCO (mAP↑)			CIRR (Rec.↑)			FIQ (Rec.↑)	
	@5	@10	@50	@1	@10	@50	@10	@50
Stella-Qwen2-1.5B [11]	14.8	15.3	17.4	27.5	69.1	88.0	29.3	49.0
Mistral-7B* [24]	<u>19.8</u>	20.2	22.7	<u>29.6</u>	72.5	<u>91.3</u>	29.5	<u>49.3</u>
E5-Mistral-7B-Inst [60]	19.6	<u>20.3</u>	<u>22.8</u>	29.5	<u>72.7</u>	91.1	<u>29.9</u>	**49.5**
SFR-Embedding-2 [47]	**20.3**	**20.8**	**23.4**	**29.7**	**72.8**	**91.5**	**30.1**	**49.5**

using the nearest image embedding neighbor yields significantly better performance (Table 6c). **LLEMs Outperforms LLMs:** We investigate Large Language Embedding Models (LLEMs), which are fine-tuned for text retrieval compared to their base LLMs. As shown in Table 7, both E5-Mistral-7B-Inst [60] and SFR-Embedding-2 outperform their LLM counterpart, Mistral-7B [24]. The superior performance of LLEMs demonstrates that models specifically tailored for embedding and retrieval tasks can offer substantial advantages in CIR applications compared to general-purpose language models.

6. Limitations and Conclusion

While these advancements significantly improve CIR performance, several areas warrant further investigation. Our work, limited to LLM/LLEMs, could benefit from exploring pre-trained MLLMs for enhanced understanding of CIR tasks. Additionally, our triplet synthesis method generates a single visual token, constraining the model's ability to process detailed image information. Future work should leverage the underutilized text category information in our synthetic datasets to improve model generalization. Lastly, our refined benchmarks, while more reliable, still contain noise from original image pairs, suggesting a need for further refinement of evaluation metrics.

In conclusion, we present novel approaches to Composed Image Retrieval (CIR) that obviate the need for annotated datasets. Our contributions include: (1) an innovative triplet synthesis method utilizing image-caption pairs, (2) a new architecture leveraging LLM's embedding generation capabilities, (3) MTCIR, a diverse, human-aligned synthetic dataset, and (4) refined versions of CIRR and Fashion-IQ benchmarks, enhancing the reliability of evaluation metrics in the field. Our method consistently outperforms existing LLM and non-LLM baselines across popular benchmarks. Notably, MTCIR achieves superior results with only 10-20% of the size of larger datasets, demonstrating a 1-15% increase in Recall. These advancements collectively push the boundaries of CIR, offering more efficient and effective solutions for real-world applications.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3, 7

[2] Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. 3, 5, 6

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. A clip-hitchhiker's guide to long video retrieval. *arXiv preprint arXiv:2205.08508*, 2022. 5

[4] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *ICCV*, 2023. 2, 3, 5, 6, 7, 13, 15

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2, 3, 5

[6] Tom B Brown. Language models are few-shot learners. In *NeurIPS*, 2020. 5

[7] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 2023. 2, 5

[8] Niv Cohen, Rinon Gal, Eli A. Meirom, Gal Chechik, and Yuval Atzmon. "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In *ECCV*, 2022. 2, 3, 6, 15

[9] Yashar Deldjoo, Fatemeh Nazary, Arnau Ramisa, Julian Mcauley, Giovanni Pellegrini, Alejandro Bellogin, and Tommaso Di Noia. A review of modern fashion recommender systems. *ACM Computing Surveys*, 2023. 1

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 13

[11] Zhang Dun. Stella en qwen2 1.5b v5, 2024. 8

[12] Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *CVPR*, 2022. 1

[13] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 5

[14] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *Transactions on Machine Learning Research*, 2024. 2, 3, 5, 6, 16

[15] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only training of zero-shot composed image retrieval. In *CVPR*, 2024. 2, 3, 6, 15

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 13

[17] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. spacy: Industrial-strength natural language processing in python. 2020. 13

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 15

[19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 13

[20] Chuong Huynh, Yuqian Zhou, Zhe Lin, Connelly Barnes, Eli Shechtman, Sohrab Amirghodsi, and Abhinav Shrivastava. Simpson: Simplifying photo cleanup with single-click distracting object segmentation network. In *CVPR*, 2023. 1

[21] Young Kyun Jang, Dat Huynh, Ashish Shah, Wen-Kai Chen, and Ser-Nam Lim. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. In *ECCV*, 2024. 2, 3, 6, 7, 14, 15

[22] Young Kyun Jang, Junmo Kang, Yong Jae Lee, and Donghyun Kim. Mate: Meet at the embedding–connecting images with long texts. *EMNLP Findings*, 2024. 3

[23] Young Kyun Jang, Donghyun Kim, Zihang Meng, Dat Huynh, and Ser-Nam Lim. Visual delta generator with large multi-modal models for semi-supervised composed image retrieval. In *CVPR*, 2024. 2, 3, 5

[24] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 8

[25] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024. 3

[26] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. *ICLR*, 2024. 3, 6, 7, 15

[27] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024. 3

[28] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and quality assessment for composed image retrieval. *AAAI*, 2024. 2, 3, 5, 6, 7, 14, 15, 17, 18

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2, 7, 14, 17, 18

[30] Wei Li, Hehe Fan, Yongkang Wong, Yi Yang, and Mohan Kankanhalli. Improving context understanding in multimodal large language models via multimodal composition learning. In *ICML*, 2024. 3, 6, 7

[31] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text

embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023. 3

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5, 7, 15

[33] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 13

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 3, 5, 6, 8, 14, 15

[35] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2, 3, 5, 13

[36] Yikun Liu, Jiangchao Yao, Ya Zhang, Yan-Feng Wang, and Weidi Xie. Zero-shot composed text-image retrieval. In *BMVC*, 2023. 2

[37] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*, 2021. 2, 3, 5, 7, 13, 14, 17

[38] Feipeng Ma, Hongwei Xue, Guangting Wang, Yizhou Zhou, Fengyun Rao, Shilin Yan, Yueyi Zhang, Siying Wu, Mike Zheng Shou, and Xiaoyan Sun. Multi-modal generative embedding model. *arXiv preprint arXiv:2405.19333*, 2024. 3

[39] Gabriel de Souza P Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. Nv-retriever: Improving text embedding models with effective hard-negative mining. *arXiv preprint arXiv:2407.15831*, 2024. 3

[40] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022. 3

[41] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *NeurIPS*, 2011. 5

[42] Khoi Pham, Chuong Huynh, Ser-Nam Lim, and Abhinav Shrivastava. Composing object relations and attributes for image-text matching. In *CVPR*, 2024. 1

[43] Bill Psomas, Ioannis Kakogeorgiou, Nikos Efthymiadis, Giorgos Tolias, Ondřej Chum, Yannis Avrithis, and Konstantinos Karantzalos. Composed image retrieval for remote sensing. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, 2024. 3

[44] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. In *CVPR*, 2023. 4

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 5, 13, 14

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 5

[47] Meng Rui, Liu Ye, Joty Shafiq Rayhan, Xiong Caiming, Zhou Yingbo, and Yavuz Semih. Sfr-embedding-2: Advanced text embedding with multi-stage training, 2024. 8, 14

[48] Amrita Saha, Mitesh Khapra, and Karthik Sankaranarayanan. Towards building large scale multimodal domain-aware conversation systems. In *AAAI*, 2018. 3

[49] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. *CVPR*, 2023. 3, 6, 15

[50] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5, 6, 14

[51] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 5, 6, 14

[52] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, 1985. 2, 4

[53] Swetha Sirnam, Jinyu Yang, Tal Neiman, Mamshad Nayeem Rizve, Son Tran, Benjamin Yao, Trishul Chilimbi, and Mubarak Shah. X-former: Unifying contrastive and reconstruction learning for mllms. In *European Conference on Computer Vision*, pages 146–162. Springer, 2024. 2

[54] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, 2019. 5

[55] Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *AAAI*, 2024. 2, 3, 6, 15

[56] Omkar Thawakar, Muzammal Naseer, Rao Muhammad Anwer, Salman Khan, Michael Felsberg, Mubarak Shah, and Fahad Shahbaz Khan. Composed video retrieval via enriched context and discriminative embeddings. *CVPR*, 2024. 4, 6, 16

[57] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[58] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. CoVR-2: Automatic data construction for composed video retrieval. *IEEE TPAMI*, 2024. 2, 3, 5, 6, 7, 14, 16, 18

[59] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. CoVR: Learning composed video retrieval from web video captions. *AAAI*, 2024. 3, 4, 5, 6, 7, 15, 16, 17

[60] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*, 2023. 3, 8

[61] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *CVPR*, 2021. 2, 5, 7, 17

[62] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 1

[63] ZHOU Yuqian, Chuong Huynh, Connelly Barnes, Elya Shechtman, Sohrab Amirghodsi, and Zhe Lin. Machine-learning models for distractor segmentation with reduced user interactions, 2024. US Patent App. 18/307,353. 1

[64] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. In *ICML*, 2024. 2, 3, 5, 6, 7, 16, 18

[65] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 2, 5

[66] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1724–1732, 2024. 13

[67] Xinliang Zhu, Sheng-Wei Huang, Han Ding, Jinyu Yang, Kelvin Chen, Tao Zhou, Tal Neiman, Ouye Xie, Son Tran, Benjamin Yao, et al. Bringing multimodality to amazon visual search system. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6390–6399, 2024. 1

[68] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023. 3