CoCo-IR: Contextual Composed Image Retrieval

Shengcao Cao[1,2], Tanmaya Shekhar Dabral[2], Zhongli Ding[2], Madhuri Shanbhogue[2], Kaifeng Chen[3], Zhe Li[2], Mojtaba Seyedhosseini[2], Yu-Xiong Wang[1], and Liang-Yan Gui[1]

[1] University of Illinois Urbana-Champaign
[2] Google DeepMind
[3] OpenAI. Work done at Google DeepMind.

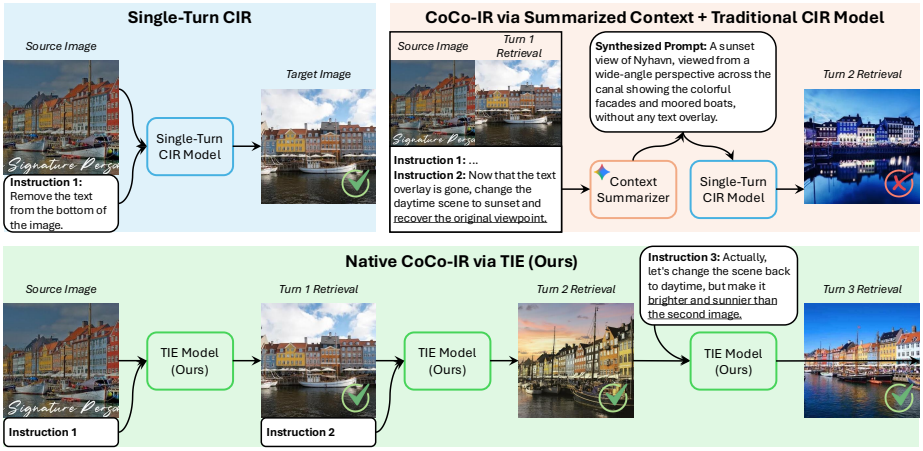

Fig. 1: Comparison between conventional single-turn Composed Image Retrieval (CIR) and our novel Contextual Composed Image Retrieval (CoCo-IR) task, achieved via our proposed TIE framework. **(Top Left)** A standard single-turn CIR model successfully processes a simple, isolated instruction. **(Top Right)** Traditional CIR models are architecturally unable to process multi-turn context with interleaving images and text. Even with a strong external model (*e.g.*, Gemini-2.5-Pro) that summarizes the entire history into a single prompt, context information may inevitably be compressed, leading to a failed image retrieval. **(Bottom)** Our proposed **TIE** natively processes an evolving context across multiple turns, enabling seamless comprehension of context-dependent instructions that refer to specific past images (*e.g.*, "original viewpoint" and "second image"), successfully completing the entire contextual retrieval trajectory within one unified framework.

Abstract. Current instruction-based image retrieval systems are powerful but limited to single-turn interactions, failing to capture the iterative nature of complex, real-world visual searches. To overcome this limitation, we introduce **Contextual Composed Image Retrieval (CoCo-IR)**, a novel task that enables users to progressively refine

search results through interactions. We address this new task by proposing a new model based on a Large Multimodal Model (LMM) that functions as a context-aware reasoner for CoCo-IR. Our model interprets the entire interaction history to generate **Transformable Image Embeddings (TIE)** that evolve across turns. To fuel the model training without expensive human annotations, we develop a fully autonomous, scalable data engine that leverages LMMs to generate high-quality contextual retrieval data, and uses model-guided verification to mine challenging hard negatives. Extensive experiments demonstrate that our approach establishes new state-of-the-art performance: We achieve 39.4 mAP@5 on the challenging single-turn benchmark CIRCO; furthermore, on our new CoCo-IR benchmark, our model maintains robust performance with 44.1 R@1 on 4-turn dialogues, dramatically outperforming existing methods (28.2 4-turn R@1) that fail to handle multi-turn context. Project page: https://CoCo-IR.github.io.

1 Introduction

The quest for effective visual search has rapidly evolved from simple keyword queries to sophisticated instruction-based systems. A prominent example is Composed Image Retrieval (CIR) [28], where a user provides a source image I_{src} and a natural language instruction T to retrieve a target image I_{tgt} that reflects the desired transformation (*e.g.*, "change the car to red"). This paradigm, powered by large-scale web data [4, 24, 33] and multimodal encoders [22, 32, 34], has seen significant progress through improved modality fusion [8, 23] and foundation model reasoning [11].

However, a fundamental limitation of classical CIR is its restriction to *isolated, single-turn* queries. By processing only one query at a time, these models fall short of the natural, iterative, and exploratory nature of real-world visual search. In practice, a user's intent is often too complex or evolving to be captured in a single command. For instance, a user might want to *progressively refine* their search based on intermediate results, *explore a sequence of related but distinct images*, where the intermediate discoveries are just as important as the final target, or even *change their mind* partway through the process (as illustrated in Fig. 1). Existing CIR models cannot support this interactive dynamic; any refinement, exploration, or pivot forces the user to start over. Instead of building upon established visual context, users must attempt to craft a new, complex prompt from scratch, highlighting a critical gap between current capabilities and real-world user needs.

To bridge this gap, we introduce a new task: **Contextual Composed Image Retrieval (CoCo-IR)**. We formulate instruction-based retrieval as an interactive dialogue, where users progressively refine their search over multiple turns. In this framework, the model must interpret each new instruction T_m relative to the *entire* interaction history $H_m = (I_0, T_1, I_1, \ldots, T_{m-1}, I_{m-1})$ that includes the initial source image I_0, previous instructions $T_1, \ldots, T_{m-1}$, and intermediate images $I_1, \ldots, I_{m-1}$. This contextual setting is inherently more flexible and powerful, enabling users to decompose complex visual search goals into

a sequence of simpler, more manageable steps, making the search process more intuitive and effective.

This new CoCo-IR task demands a paradigm shift in modeling. As our experiments (Tab. 2) demonstrate, prior methods fail to support CoCo-IR because they cannot natively encode a sequence of interleaved images and instructions, *even with a powerful external model that can summarize the context*. To solve this, we propose **Transformable Image Embedding (TIE)**, which unifies multimodal comprehension and embedding generation into a single Large Multimodal Model (LMM) [26]. By enabling end-to-end optimization, TIE inherently understands the evolving context, yielding substantially stronger performance.

To effectively process a sequence of multimodal turns, we enhance the LMM with two key architectural designs. First, in contrast to prior embedding models that extract the hidden state of the final text token, which is inherently biased toward its immediate semantic predecessors, we introduce a specialized ⟨EMB⟩ token. ⟨EMB⟩ acts as a dedicated global information bottleneck, forcing the model to aggregate the full interaction history into a compact, query-aware embedding. Second, to handle the structure of multi-turn interactions, we design a hybrid attention mechanism by employing bidirectional attention *within* each turn for deep multimodal fusion, and causal attention *across* turns to respect the temporal flow of the interaction. The resulting transformable image embeddings accurately capture the cumulative semantic transformation specified by the entire dialogue, not just the most recent instruction.

Fueling this *context-aware* model requires a large-scale multimodal long-context dataset, which, to our knowledge, does not exist yet. Manually annotating such data is prohibitively expensive. We address this by developing a *scalable, LMM-powered data engine* with two key innovations. First, for instruction generation, we anchor the process with clustered real images and leverage **self-reflection**, where an LMM generates a transformation instruction and then scores its own quality and ambiguity, allowing us to filter for high-quality data. Second, to learn fine-grained distinctions, we employ LMMs as **verifiers** for mining challenging hard negatives, which are candidate images that are visually similar to the target but fail the instruction (Fig. 4). This autonomous data engine proves to produce data of high quality in our human verification. Additionally, as demonstrated in single-turn results, we achieve state-of-the-art performance using 14× fewer training samples than prior work [35].

In summary, our key contributions are:

- We introduce CoCo-IR, a new task for *multi-turn contextual composed image retrieval* that better reflects real-world image search.
- We propose TIE, an LMM-based architecture that generates *context-aware, transformable image embeddings* that capture the full dialogue history.
- We develop a scalable data engine that leverages LMM-driven self-reflection and hard-negative verification to create the first large-scale CoCo-IR dataset.
- We demonstrate state-of-the-art performance on both single-turn CIR benchmarks (*e.g.*, 39.4 mAP@5 on CIRCO [2]) and our new multi-turn benchmark (*e.g.*, a robust 44.1 R@1 on 4-turn context where prior methods collapse).

2 Related Work

From Multimodal Encoders to LMMs. Early successes in vision-language pre-training come from dual-encoder models that align image and text representations in a joint space [15, 22, 24, 32, 33]. These models excel at zero-shot cross-modal retrieval but lack the deep fusion necessary to understand complex instructions. Recent efforts propose large multimodal models (LMMs) [1, 12–14, 16, 18] that integrate vision encoders with large language models, enabling more sophisticated reasoning and instruction-following capabilities [5]. Our work builds on this LMM paradigm, as the ability to interpret and understand multiple interactive rounds is fundamental to our contextual retrieval task.

Composed Image Retrieval (CIR). Our work directly extends composed image retrieval [28] (CIR), where a query consists of a source image and a modifying text instruction. A significant body of work has focused on this single-turn task, designing lightweight modality transformers [2, 8, 23] or leveraging strong reasoning from foundation models [11]. To address inherent data bottlenecks, prior work has established static benchmarks [2, 19, 30] and increasingly relied on synthetic data generation pipelines [3, 7, 35, 36]. Our proposed data engine differs significantly from these prior efforts. Specifically, we advance beyond single-step retrieval by generating *multi-turn* interactions, integrating *hard negative mining* to improve embeddings, and applying rigorous *filtering and quality control* to ensure context coherence. IRR [29] and MAI [6] extends CIR to a scenario where the user provides a reference image only at the first turn, and subsequent turns are text-only refinements. In contrast, our formulation supports *interleaved image-text input at every turn*, allowing users to introduce new visual context mid-session. Both IRR and MAI are also limited to the fashion domain, while ours is open-domain using a general-purpose LMM.

3 Approach

In this section, we first formalize our new task, Contextual Composed Image Retrieval (CoCo-IR). We then describe the details of our model architecture, training procedure, and finally our scalable data engine.

3.1 Task Formulation: From Single-Turn to Multi-Turn Contextual Retrieval

We begin by formally introducing the task of **Composed Image Retrieval (CIR)** [28], which serves as the foundation for our multi-turn task. In the standard CIR setting, the goal is to retrieve a specific target image I_{tgt} from a large corpus of images $\mathcal{C}$. The query consists of a source image I_{src} and a free-form natural language instruction T that describes a desired transformation or a semantic relationship between the source and target images. A model is trained to learn a function that maps the composed query (I_{src}, T) and the target image

I_{tgt} into a shared embedding space, such that the distance between the query embedding and the correct target embedding is minimized.

While powerful, this single-turn paradigm does not capture the iterative nature of complex real-world visual searches. We therefore extend this paradigm to a **multi-turn, interactive setting**. This framework allows a user to progressively refine their search through an interactive dialogue, reusing the context from previous interactions to inform subsequent queries.

A multi-turn retrieval session begins similarly to the single-turn case. At the first turn ($t = 1$), the user provides an initial source image I_0 and a text instruction T_1. The retrieval system's task is to return the corresponding target image I_1, among up to top-k predictions, from the corpus $\mathcal{C}$. The key difference emerges in subsequent turns. For any turn $t > 1$, the user provides a new instruction T_t that can *refer to any element in the preceding interaction history*, $H_t = (I_0, T_1, I_1, \ldots, T_{t-1}, I_{t-1})$. The model must then interpret T_t in the context of this history to retrieve the next target image, I_t. This process can continue for an arbitrary number of turns, enabling the decomposition of a complex search intent into a sequence of simpler, context-aware steps.

Evaluation Metric. To evaluate performance in this new interactive setting, we must bridge the gap between dynamic user behavior and the requirements of standardized evaluation. As shown in Fig. 1, intermediate images in a interactive search often hold independent value or serve as critical stepping stones. In a real-world interactive search, if a user does not immediately see their desired intermediate image in the top-k results, they naturally adapt their subsequent instructions based on what was actually retrieved. However, to construct a rigorous and reproducible benchmark, evaluation trajectories (including all intermediate target images and their corresponding sequential instructions) must be fixed and predefined. Consequently, if the model fails to retrieve the intended ground-truth target within the top-k results at any turn, the retrieval trajectory diverges from the benchmark's established path. When this happens, predefined subsequent instructions that rely on that specific missing context (*e.g.*, in Fig. 1, commanding the system to "make it brighter and sunnier than the *second image*") immediately become ambiguous or incoherent.

Because we cannot simulate on-the-fly user adaptation, our evaluation must measure the model's ability to independently sustain this exact, unbroken chain of correct retrievals. Furthermore, because real-world multi-turn retrieval is often an exploratory process, intermediate discoveries are just as important as the final destination. Therefore, rather than merely evaluating the accuracy of the final result, we introduce the **m-turn Recall@k** metric to explicitly validate the entire trajectory. We define an m-turn query as a success if and only if the ground-truth target image for *every* turn is successfully retrieved within the top-k results for that respective turn. Formally, let the sequence of ground-truth target images for an m-turn interaction be $(I_1^*, I_2^*, \ldots, I_m^*)$, and let $\mathcal{R}_t^k$ be the set of top-k images retrieved by the model at turn t, given the ground-truth history up to step t: $(I_0, T_1, I_1^*, \ldots, I_{t-1}^*, T_t)$. The entire retrieval is marked as correct if and only if $\bigwedge_{t=1}^{m}(I_t^* \in \mathcal{R}_t^k)$. While this cumulative success criterion is strict,

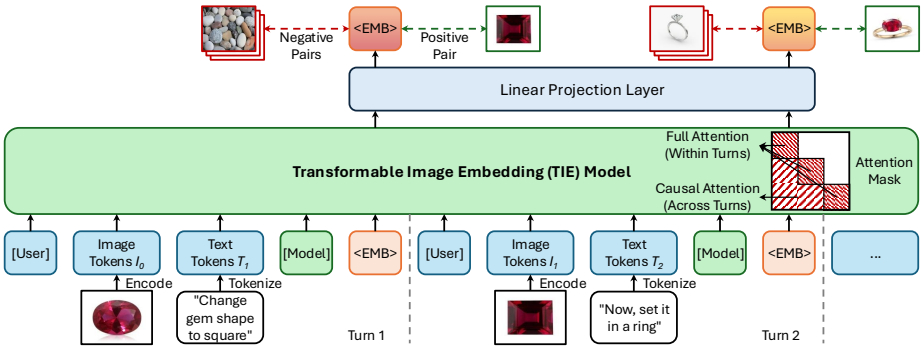

Fig. 2: Overview of our Transformable Image Embedding (TIE) model. TIE processes the multi-turn context as a unified token sequence (bottom), formatting the visual and textual inputs into an interleaved dialogue. A key architectural component is the special ⟨EMB⟩ token at the end of each turn, which acts as a global bottleneck to summarize the full preceding context. As illustrated by the hybrid **Attention Mask** (middle right), the model employs full (bidirectional) attention within each turn for deep multimodal fusion, and causal attention across turns to respect the temporal flow. Finally, the embedding generated by the ⟨EMB⟩ token passes through a linear projection layer and is trained with a contrastive learning objective (top). This loss pulls the embedding toward the positive target image (green dashed arrows) while pushing it away from negative samples in the batch (red dashed arrows).

it is a necessary and logical design choice for a static benchmark: it penalizes off-trajectory compounding errors and ensures the model is rewarded only when it successfully supports the user's entire exploratory journey. Ultimately, m-turn Recall@k provides a robust measure of a model's true contextual understanding over successive interactions.

3.2 Model Architecture and Training

Our Transformable Image Embedding (TIE) model is built upon a Large Multimodal Model (LMM) [26], which is not originally designed for an embedding task. We therefore enhance its architecture and training objective specifically for the task of CoCo-IR, as illustrated in Fig. 2.

Model Architecture. The model is designed to process a sequence of multimodal, interactive turns. In each turn, user inputs, which can be text, images, or both, are encoded into a unified token sequence. Text is processed by a standard tokenizer, while images are encoded into token embeddings using an image encoder [34]. For a query turn, both the image and text instruction are provided to the model. For encoding a target image, only the image is provided as input.

A key modification to the base LMM is our method for generating embeddings. Instead of training the model to autoregressively predict the next token, we introduce a special ⟨EMB⟩ token. When it is the model's turn to produce an embedding, we feed this special token as input, prompting the model to

summarize the entire interactive history seen so far into a single, compact representation. The output embedding is derived by passing the last-layer hidden state corresponding to this ⟨EMB⟩ token through a final linear projection layer. Compared to the previous "last token" solution [10] that is biased toward immediate predecessors, this specialized ⟨EMB⟩ token acts as a dedicated global information bottleneck, encouraging better history aggregation (Tab. 4).

To facilitate better integration of information within a query, we also modify the transformer's attention mechanism. While the attention across different turns remains causal to preserve the temporal sequence, we employ bidirectional (full) attention *within* each turn. This allows the image tokens and text tokens of a query to freely attend to each other, resulting in a more holistic representation.

Contrastive Training. We train our model using a contrastive learning objective based on the InfoNCE loss [21,22]. The goal is to learn an embedding space where a query is close to its corresponding target image while being distant from all other negative images.

The core of our objective is the loss for a single query at a specific turn, $\mathcal{L}_{t,i}$. For the i-th sample at turn t, the model generates an embedding $q_{t,i}$ from the interactive history $H_{t,i}$ and the new instruction $T_{t,i}$. This query embedding is contrasted against the embedding of the positive target image for that turn, $p_{t,i}$ (extracted from image $I_{t,i}$).

We define a comprehensive set of candidate embeddings $\mathcal{C}_e$ for the entire batch. This set includes all target image embeddings $\{p_{k,j}\}$, all initial source image embeddings $\{s_j\}$, and all mined hard negative embeddings $\{h_{k,j}\}$ across all n samples and their m turns. The positive target $p_{t,i}$ is one of the embeddings within this set $\mathcal{C}_e$.

The loss for the query $q_{t,i}$ is defined as:

$$\mathcal{L}_{t,i} = -\log \frac{\exp(\text{sim}(q_{t,i}, p_{t,i})/\tau)}{\sum_{e \in \mathcal{C}_e} \exp(\text{sim}(q_{t,i}, e)/\tau)}.$$

In this formulation, $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a fixed temperature hyperparameter. The numerator aligns the query with its correct target. The denominator $\sum_{e \in \mathcal{C}_e} \exp(\text{sim}(q_{t,i}, e)/\tau)$ serves as the normalization term, summing over the positive pair and all potential negative pairs in the batch, effectively pushing the query away from all incorrect candidates. Finally, the overall loss is computed by averaging over all samples and turns.

3.3 Scalable LMM-Powered Data Engine

A key challenge in training contextual retrieval models is the lack of large-scale annotated data. To overcome this, we design a fully autonomous, LMM-powered data engine that generates high-quality single-turn and multi-turn training data. This engine operates in a series of stages, from discovering image pairs to mining hard negatives and constructing interaction chains, creating a scalable, model-guided flywheel. We include more specific details of the data engine in the supplementary material.

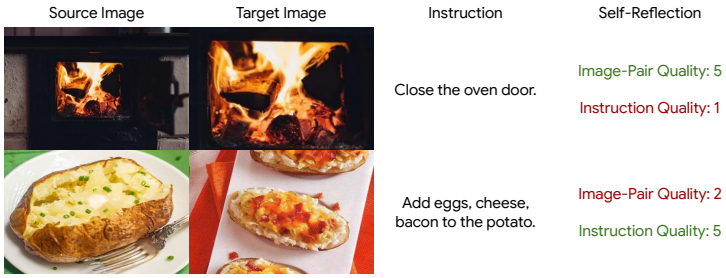

Fig. 3: Self-reflection during instruction synthesis. As part of our data engine, an LMM generates a transformation instruction for a given image pair and then performs self-reflection. It assigns two scores for image-pair quality and instruction quality, respectively. The top example shows a good pair (Quality: 5) but a poor, unrelated instruction (Quality: 1). The bottom example shows a clear instruction (Quality: 5) but a poor image pair (Quality: 2), as too many visual elements are altered. This two-dimensional scoring allows us to filter for high-quality data by selecting samples where both scores are high.

1. Discovering Image Pairs with Transformations. We begin by discovering a large set of image pairs from web images that exhibit meaningful, describable transformations. We employ two complementary strategies in parallel:

- **Feature-Based Discovery:** We extract visual features of images using a pretrained image encoder [9, 22], and group them into clusters. To encourage learning transformations across different visual concepts, each image is associated with its two nearest clusters. Within these expanded clusters, we identify image pairs whose feature similarities fall within a specific range, filtering out pairs that are either too dissimilar to be related by a simple instruction or nearly identical. To ensure diversity, we set a maximum number of pairs to sample from each cluster.
- **Metadata-Based Discovery:** Inspired by MagicLens [35], we also leverage the co-occurrence of images on the same web page as a strong signal of semantic relatedness. We group images by their source URLs and propose pairs that appear together. These pairs are then filtered by a wider feature similarity threshold to capture a broader range of semantic relationships.

2. Synthesizing Instructions with Self-Reflection. For each discovered image pair, we prompt an LMM (*e.g.*, Gemini 2.5 [25]) to generate a corresponding text instruction. The LMM is tasked to first perform a detailed visual comparison of the source and target images. Based on this analysis, it formulates a concise instruction that describes the transformation. Crucially, the process includes a self-reflection step: the LMM scores the quality of both the image pair itself (*"Is the transformation clear and simple?"*) and its own generated instruction (*"Is the instruction unambiguous and effective?"*). This two-dimensional scoring allows us to filter for high-quality data (see examples in Fig. 3).

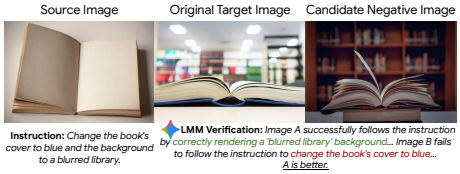

Fig. 4: Verification of hard negative samples. Hard negatives are images that are visually and semantically similar to the ground-truth target (middle) but fail to perfectly satisfy the instruction (left). Here, the candidate negative (right) correctly matches the "blurred library" background but fails the "change the book's cover to blue" command. Our data engine uses an LMM to automatically verify this distinction.

Fig. 5: Construction of multi-turn interaction chains. An LMM rewrites subsequent instructions to be context-aware, transforming disjointed commands into a natural dialogue. The rewritten instruction references "the original image" and "without the hands," and thus can test whether the model understands interaction history.

3. Post-Processing and Data Splitting. We process the generated data to create distinct training and evaluation sets. To rigorously test our model's generalization capabilities, we split the data at the image cluster level, ensuring that all images from a held-out subset of clusters are unseen during training. Furthermore, we apply different quality filters based on the LMM's self-reflection scores. For the training set, we use a more relaxed threshold to maintain data diversity, while for the evaluation set, we enforce the highest possible scores to ensure benchmark quality.

4. Mining Hard Negatives. Effective contrastive learning training requires hard negatives, which are images that are semantically close to the target but do not perfectly satisfy the query. We incorporate a sophisticated, LMM-driven process for mining these negatives.

- **For the Evaluation Set:** To maintain a static and unbiased benchmark, we first identify candidate negatives by finding the nearest neighbors of the source and target images within the entire web image source, using the feature space of the same pretrained image encoder from the first stage. We then use an LMM as a *verifier*, which is prompted to perform a pairwise comparison between the ground-truth target and a candidate negative, judging which better fulfills the instruction while preserving the content of the source image (see an example in Fig. 4). Only candidates judged to be inferior to the ground-truth target are included as hard negatives, ensuring that they are not false negatives. We collect multiple verified negatives for each query to enable more robust evaluation.
- **For the Training Set:** We employ a model-guided approach for augmenting the training data. We first train an initial TIE model without hard negatives. This model is then used to retrieve the closest embeddings for each query,

which serve as candidate hard negatives. These candidates are subsequently verified by the LMM using the same pairwise comparison logic. This creates a data engine that leverages the trained model's capabilities to discover progressively more challenging negatives for subsequent models to be trained. Due to computation limitations, we retain *one hard negative* per training sample, and perform *one iteration* of the model-guided hard negative mining.

5. Constructing Multi-Turn Contextual Data. Finally, we compose the curated single-turn samples into multi-turn interaction chains. We identify sequences where the target image of one sample ($I_0 \rightarrow T_1 \rightarrow I_1$) is the source image of another ($I_1 \rightarrow T_2 \rightarrow I_2$), and concatenate them to form a multi-turn sequence ($I_0 \rightarrow T_1 \rightarrow I_1 \rightarrow T_2 \rightarrow I_2$) (see an example in Fig. 5). We construct interactions of up to *four turns*, because longer sequences are increasingly sparse. To make the dialogue more natural and context-aware, we prompt an LMM to rewrite the instructions for the second turn onwards ($T_2, \ldots, T_m$). The rewritten instructions are designed to reference elements from the preceding context, transforming a series of disjointed commands into a coherent, interactive dialogue.

6. Quality Control and Human Verification. Despite the automated nature of our data engine, the synthesized data maintains high quality due to our rigorous, multi-stage filtering process. To validate this, we conduct human verification on random samples from the evaluation data (representing 0.14%, 0.96%, 3.40%, and 7.65% of single-, two-, three-, and four-turn interactions, respectively) to ensure the retrieval trajectories are logical and correct. Among these samples, we observed an *average validity of 90.5%*, confirming the high quality of our evaluation set. Moreover, although our training dataset is fully synthetic, TIE achieves state-of-the-art performance on established *human-annotated benchmarks* (Tab. 1) with 14× fewer training samples than prior work [35]. This strong generalization demonstrates that our data engine successfully captures natural search behaviors rather than *overfitting to synthetic artifacts or exhibiting self-preference bias*, as models trained on biased data would inherently struggle to transfer to external, human-curated domains.

4 Experiments

In this section, we conduct a comprehensive evaluation of our proposed model. We first detail our experimental setup, and then assess our model's performance on established single-turn CIR benchmarks. Finally, we evaluate its core capability on our newly proposed CoCo-IR task. Implementation details, analysis of data efficiency, and full results on single-turn CIR are included in the supplementary material.

4.1 Experimental Setup

Benchmarks. We evaluate our model on three standard single-turn benchmarks: **Fashion IQ (FIQ)** [30], **CIRR** [19], and **CIRCO** [2]. FIQ is a benchmark specialized for fashion retrieval; CIRR is a widely-used benchmark known

for its realistic, open-ended human-generated instructions; CIRCO provides a large-scale, open-domain challenge, testing retrieval across a diverse set of concepts. To evaluate our primary contribution in the new multi-turn setting, we use our newly proposed **CoCo-IR benchmark**, which features dialogue-based image retrieval sequences of varying lengths (Sec. 3.3).

Baselines. We compare our approach against several state-of-the-art single-turn composed image retrieval methods. These include **MagicLens** [35], which leverages synthetic data from generative models, **CIReVL** [11], which uses context reasoning from foundation models, and other LMM-based embedding models like **E5-V** [10]. It is important to note that these baselines are *not designed for multi-turn interactions* and architecturally cannot take multiple images as input. For the CoCo-IR benchmark, we need to apply some adaptation strategies (see details in Sec. 4.3) so that they can be comprehensively evaluated.

4.2 CIR: Single-Turn Retrieval

We first demonstrate our model's strong performance on the standard, single-turn Composed Image Retrieval (CIR) task. This evaluation indicates that our LMM-based architecture, even when fine-tuned for an interactive setting, excels at the fundamental single-turn retrieval problem. We evaluate on the FIQ, CIRR, and CIRCO benchmarks, using their standard evaluation protocols: FIQ and CIRR results are reported using Recall@k metrics, while CIRCO results are reported using mean Average Precision (mAP).

As shown in Tab. 1, our model establishes a new state of the art. On FIQ, we reach an overall Recall@10 of 40.1, and on CIRR, we achieve a Recall@1 of 38.7, significantly outperforming prior methods. Our performance on CIRCO is even more pronounced, with an mAP@5 of 39.4, showcasing our model's ability to handle fine-grained, open-domain instructions effectively. This strong single-turn performance serves as a crucial foundation, proving that our model's representations are robust before we even evaluate its more complex, multi-turn capabilities. Complete CIR results are included in the supplementary material.

4.3 CoCo-IR: Multi-Turn Retrieval

We evaluate contextual retrieval on our proposed benchmark, which consists of dialogue chains varying from 1 to 4 turns. As defined in Sec. 3.1, we employ the m-turn Recall@k metric. This metric is cumulative and unforgiving: it requires the model to successfully retrieve the correct target within the top-k results at *every single turn* of the sequence. A failure at any intermediate turn constitutes a failure for the entire session.

Baseline Adaptations. Existing CIR models (*e.g.*, MagicLens [35]) are architecturally limited to single-turn inputs (one reference image and one instruction). Because they inherently cannot process the multi-image, multi-turn history H_m, they cannot be directly evaluated on our benchmark. For a rigorous comparison, we build proxy inputs to adapt these single-turn models to the interactive setting using three distinct strategies (see visualization in the supplementary material):

Table 1: Single-turn retrieval performance on standard benchmarks. This evaluation validates that TIE is a general and powerful retriever, even in a single-turn context. Our model achieves new state-of-the-art performance on CIRR and CIRCO and highly competitive results on FIQ (LinCIR's advantage on FIQ and weakness on CIRCO likely originate from a close linguistic alignment between its training corpus and FIQ; see the supplementary material). To provide a holistic performance comparison across metrics with varying scales, "Rank ↓" reports the average rank of each model across all six metrics.

Model	Backbone	FIQ		CIRR		CIRCO		Rank ↓
		R@10	R@50	R@1	R@5	mAP@5	mAP@10	
CompoDiff [7]	CLIP-G [22]	39.0	51.7	26.7	55.1	15.3	17.7	9.50
E5-V [10]	LLaVA-NeXT-8B [12]	31.8	53.8	33.9	64.1	19.1	20.6	8.75
MagicLens [35]	CLIP-L [22]	30.7	52.5	30.1	61.7	29.6	30.8	8.67
SEARLE [2]	CLIP-G [22]	34.8	55.7	34.8	64.1	13.2	13.9	7.92
CIReVL [11]	CLIP-G [22]	32.2	52.4	34.7	64.3	26.8	27.6	7.83
LDRE [31]	CLIP-G [22]	32.5	55.5	36.1	66.4	31.1	32.2	5.50
MagicLens [35]	CoCa-L [32]	38.0	58.2	33.3	67.0	34.1	35.4	5.00
LinCIR [8]	CLIP-G [22]	**45.1**	**65.7**	35.3	64.7	19.7	21.0	4.83
BGE-VL [36]	CLIP-L [22]	34.6	55.2	38.0	70.3	39.2	40.2	3.75
TIE-4B (Ours)	Gemma3-4B [26]	40.1	60.8	37.7	70.6	37.8	38.4	2.75
TIE-12B (Ours)	Gemma3-12B [26]	40.1	61.0	**38.7**	**70.8**	**39.4**	**40.2**	**1.50**

1. *Concatenated Instructions:* We provide the initial source image (I_0) combined with the concatenation of all text instructions up to the current turn.
2. *Latest Inputs:* We pair the ground-truth image from the previous turn (I_{t-1}) with the current instruction (T_t) as the input.
3. *Summarized Context:* To give the baselines the strongest possible advantage, we use a state-of-the-art LMM (Gemini 2.5 Pro) as an oracle. We prompt the LMM to read the entire history up to the given step (including all intermediate images and text) and synthesize it into a single, optimized text instruction (T'_t) for the retrieval model. This isolates the model's retrieval capability by minimizing the requirement of complex context comprehension.

Results. Tab. 2 demonstrates the profound limitations of adapting single-turn models to multi-turn tasks and the clear superiority of our *native end-to-end* framework. Standard baselines relying on the first two heuristic adaptations collapse rapidly as the interaction depth increases, falling to near-zero recall by the fourth turn. As expected, the *Summarized Context* strategy acts as a strong oracle, significantly elevating the baseline performance. However, even when deliberately advantaged by Gemini 2.5 Pro's summarization capabilities, these baselines still underperform our method. For instance, our 12B model achieves a 4-turn R@1 of 44.1%, whereas the best summarized baseline reaches only 28.2%. This performance gap provides crucial evidence that multi-turn visual dialogue cannot be losslessly compressed into a single text prompt. TIE's ability to natively ingest multiple images and internally maintain dialogue state

Table 2: Multi-turn retrieval performance on our CoCo-IR benchmark. We report the m-turn Recall@k for interactions up to 4 turns, where success requires correct retrieval at *every* turn. Standard CIR models fail even when adapted with various strategies (including Gemini-summarized context). In contrast, our full-context model, TIE, dramatically outperforms all baselines, proving that TIE's end-to-end modeling is necessary for understanding multimodal context. *Note:* Columns evaluate distinct, non-overlapping data splits, so metrics are not monotonic across turns.

Model	Backbone	1-turn		2-turn		3-turn		4-turn	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
First Image + Concatenated Instructions: $(I_0, T_1 + \cdots + T_t)$									
TIPS-SO400M [20]		25.2	51.0	0.4	21.3	0.0	11.0	0.0	5.1
SigLIP-SO400M [34]		31.4	59.3	2.1	27.3	0.0	12.9	0.0	4.9
SigLIP2-SO400M [27]		31.0	59.2	1.6	29.4	0.0	12.1	0.0	3.6
MagicLens [35]	CLIP-L [22]	36.7	57.8	8.1	26.7	1.0	11.4	0.0	2.3
E5-V	LLaVA-NeXT-8B [12]	47.1	68.6	16.2	44.5	6.3	32.6	4.6	30.6
BGE-VL-MLLM-S1	LLaVA-NeXT-7B [17]	60.0	79.5	28.7	58.3	12.5	42.6	4.9	34.7
Last Image + Last Instruction: (I_{t-1}, T_t)									
TIPS-SO400M [20]		25.2	51.0	0.0	24.1	0.0	14.2	0.0	6.7
SigLIP-SO400M [34]		31.4	59.3	0.0	31.5	0.0	18.6	0.0	11.0
SigLIP2-SO400M [27]		31.0	59.2	0.0	33.2	0.0	23.8	0.0	18.6
MagicLens [35]	CLIP-L [22]	36.7	57.8	14.5	35.7	6.8	23.9	6.4	20.7
E5-V	LLaVA-NeXT-8B [12]	47.1	68.6	19.2	46.7	15.6	38.8	13.4	37.2
BGE-VL-MLLM-S1	LLaVA-NeXT-7B [17]	60.0	79.5	36.2	64.1	28.6	51.8	27.0	54.0
Summarized Context: $T'_t = \text{Gemini}(I_0, T_1, I_1, \ldots, T_{t-1}, I_{t-1}, T_t)$									
TIPS-SO400M [20]		25.7	51.9	10.6	22.7	5.2	11.8	1.2	5.9
SigLIP-SO400M [34]		31.0	46.9	11.8	28.6	7.8	23.1	8.8	22.5
SigLIP2-SO400M [27]		51.6	71.4	23.3	48.7	18.0	40.9	16.7	33.7
MagicLens [35]	CLIP-L [22]	43.5	66.5	19.8	47.0	12.7	37.8	10.7	33.0
E5-V	LLaVA-NeXT-8B [12]	50.1	73.1	23.8	53.1	23.5	45.9	21.3	45.4
BGE-VL-MLLM-S1	LLaVA-NeXT-7B [17]	60.8	79.4	35.5	62.7	29.5	54.8	28.2	53.9
Full Context: $(I_0, T_1, I_1, \ldots, T_{t-1}, I_{t-1}, T_t)$									
TIE-4B (Ours)	Gemma3-4B [26]	74.5	89.8	48.3	79.5	34.5	69.3	31.1	65.8
TIE-12B (Ours)	Gemma3-12B [26]	**76.3**	**90.4**	**54.8**	**82.1**	**44.6**	**76.6**	**44.1**	**78.3**

allows it to handle multi-turn interactions, proving that end-to-end modeling is essential for contextual retrieval.

4.4 Ablation Study

We conduct a comprehensive ablation study to validate the effectiveness of our data engine and model design choices and summarize results in Tab. 3 and 4. We analyze the data efficiency and scaling in the supplementary material.

Data Source and Hard Negatives. In Tab. 3, we evaluate the impact of different data sources and the inclusion of hard negatives. We employ two strategies for

Table 3: Ablation study on the data source. We compare the effectiveness of feature-based (Feat.) and metadata-based (Meta.) discovery strategies, and the impact of mining hard negatives (HN). Mixing sources and adding hard negatives yields the best performance. "Rank ↓" reports the average rank across all four metrics (lower is better).

DS	HN	FIQ R@10	CIRR R@1	CIRCO mAP@5	CoCo-IR 1-turn R@1	Rank ↓
Feat.		31.6	36.9	33.2	65.7	4.50
Feat.	✓	38.0	36.4	**38.0**	74.5	2.75
Meta.		34.2	36.6	32.7	60.7	4.75
Meta.	✓	38.3	34.3	33.8	56.4	4.50
Mix		33.7	**38.0**	34.8	67.7	3.00
Mix	✓	**40.1**	37.7	37.8	**75.4**	**1.50**

Table 4: Ablation study on model design choices. We analyze the impact of Bidirectional Attention (BA), Single-side Contrastive loss (SC), and the Special Embedding Token (ET). The full model combining all three designs achieves the highest accuracy. "Rank ↓" reports the average rank across all four metrics (lower is better).

BA	SC	ET	FIQ R@10	CIRR R@1	CIRCO mAP@5	CoCo-IR 1-turn R@1	Rank ↓
	✓	✓	39.7	33.9	36.6	74.2	3.25
✓		✓	39.6	35.9	37.4	**75.6**	2.25
✓	✓		39.2	37.2	35.9	75.5	3.00
✓	✓	✓	**40.1**	**37.7**	**37.8**	75.4	**1.50**

discovering training pairs: feature-based similarity (Feat.) and metadata-based co-occurrence (Meta.). While both strategies independently yield strong results, mixing them in a balanced ratio provides a noticeable performance boost. This suggests that the two sources offer complementary signals, leading to greater diversity in the training data. Furthermore, the integration of LMM-verified hard negatives (HN) yields a substantial improvement on most benchmarks. This confirms that training with difficult negatives is crucial for forcing the model to learn fine-grained instruction following.

Model Design Choices. We further examine three key architectural designs in Tab. 4: Bidirectional Attention (BA) within turns, the use of a Single-side Contrastive loss (SC), and the special `<EMB>` Token (ET). First, employing bidirectional attention proves superior to pure causal attention, as it allows for better information aggregation between image and text tokens within a specific turn. Second, we compare our objective function against a symmetric contrastive loss commonly used in CLIP training [22]. We observe that the simpler, non-symmetric formulation (denoted as SC) works better for our retrieval task as it prioritizes distinction of candidates. Finally, using a dedicated `<EMB>` token (ET) to summarize the interaction history outperforms using the last token's hidden state, demonstrating the effectiveness of explicitly learning a compact context representation.

Data Efficiency and Scale. In Tab. 5, we analyze the impact of training data size on model performance. Our approach is substantially more data-efficient than prior state-of-the-art methods. The strong baseline MagicLens [35] relies on a large synthetic dataset of 36.7M pairs to achieve its performance. In contrast, our TIE-4B model already surpasses MagicLens on all three single-turn CIR benchmarks (FIQ [30], CIRR [19], and CIRCO [2]) when trained on only 320K samples, corresponding to more than a 100× reduction in data volume. As we further scale the training set to 640K and 1.28M samples, performance

Table 5: Ablation study on data scale. We compare TIE-4B trained on varying dataset sizes against the state-of-the-art MagicLens [35]. Remarkably, with just 320K samples (less than 1% of the baseline's data size), our model already surpasses the baseline across all three single-turn CIR benchmarks. This underscores the high quality of our autonomously generated data and the data efficiency of our architecture. Scaling data size and model size together brings further performance gain (Table 1).

Method	Data	FIQ R@10	CIRR R@1	CIRCO mAP@5
MagicLens CoCa-L [35]	36.7M	38.0	33.3	34.1
TIE-4B (Ours)	320K	38.5	34.8	36.7
TIE-4B (Ours)	640K	38.8	35.5	**38.7**
TIE-4B (Ours)	1.28M	40.0	36.3	37.9
TIE-4B (Ours)	2.56M	**40.1**	**37.7**	37.8

improves consistently, demonstrating the high quality and data efficiency of our autonomously generated data. We also observe that the gains begin to saturate beyond 1.28M samples for the 4B model, suggesting that performance is becoming bottlenecked by model capacity rather than data availability. This interpretation is consistent with Table 1, where scaling from TIE-4B to TIE-12B yields further improvements. Taken together, these results suggest that our LMM-driven data engine provides strong and scalable training signals, while additional gains may be unlocked by pairing this data with larger models.

5 Conclusion

In this work, we introduce **CoCo-IR**, a new task for multi-turn, contextual composed image retrieval that moves beyond the static, single-turn limitations of existing methods. We propose the **Transformable Image Embedding (TIE)** model, an LMM-based architecture that processes the full dialogue history to generate evolving embeddings that capture the user's cumulative, iterative intent. A key enabler is our fully autonomous **data engine**, which leverages LMM-driven self-reflection and hard-negative verification to create the first large-scale, high-quality dataset for this task. Our experiments demonstrate TIE's effectiveness. It achieves state-of-the-art performance on traditional single-turn benchmarks like CIRR and CIRCO, while also dramatically outperforming all baselines on our new multi-turn CoCo-IR task. Where prior methods collapse, TIE's ability to understand multi-turn context allows it to handle complex, iterative refinement, providing a robust foundation for the next generation of interactive visual search.

Acknowledgements. This work was supported in part by NSF under Grants 2106825 and 2519216, the DARPA Young Faculty Award, the ONR Grant N00014-26-1-2099, and the NIFA Award 2020-67021-32799.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: A visual language model for few-shot learning. In: NeurIPS (2022)
2. Baldrati, A., Agnolucci, L., Bertini, M., Del Bimbo, A.: Zero-shot composed image retrieval with textual inversion. In: ICCV (2023)
3. Brooks, T., Holynski, A., Efros, A.A.: InstructPix2Pix: Learning to follow image editing instructions. In: CVPR (2023)
4. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A.V., Bradbury, J., Kuo, W., Seyedhosseini, M., Jia, C., Ayan, B.K., Ruiz, C.R., Steiner, A.P., Angelova, A., Zhai, X., Houlsby, N., Soricut, R.: PaLI: A jointly-scaled multilingual language-image model. In: ICLR (2023)
5. Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., Chang, M.: Can pre-trained vision and language models answer visual information-seeking questions? In: EMNLP (2023)
6. Chen, Y., Yang, Z., Xu, J., Peng, Y.: MAI: A multi-turn aggregation-iteration model for composed image retrieval. In: ICLR (2025)
7. Gu, G., Chun, S., Kim, W., Jun, H., Kang, Y., Yun, S.: CompoDiff: Versatile composed image retrieval with latent diffusion. TMLR (2024)
8. Gu, G., Chun, S., Kim, W., Kang, Y., Yun, S.: Language-only training of zero-shot composed image retrieval. In: CVPR (2024)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
10. Jiang, T., Song, M., Zhang, Z., Huang, H., Deng, W., Sun, F., Zhang, Q., Wang, D., Zhuang, F.: E5-V: Universal embeddings with multimodal large language models. arXiv preprint arXiv:2407.12580 (2024)
11. Karthik, S., Roth, K., Mancini, M., Akata, Z.: Vision-by-language for training-free compositional image retrieval. In: ICLR (2024)
12. Li, B., Zhang, K., Zhang, H., Guo, D., Zhang, R., Li, F., Zhang, Y., Liu, Z., Li, C.: LLaVA-NeXT: Stronger LLMs supercharge multimodal capabilities in the wild (May 2024), https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/
13. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
14. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)
15. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: NeurIPS (2021)
16. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR (2024)
17. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: LLaVA-NeXT: Improved reasoning, OCR, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/

18. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
19. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. In: ICCV (2021)
20. Maninis, K.K., Chen, K., Ghosh, S., Karpur, A., Chen, K., Xia, Y., Cao, B., Salz, D., Han, G., Dlabal, J., Gnanapragasam, D., Seyedhosseini, M., Zhou, H., Araujo, A.: TIPS: Text-image pretraining with spatial awareness. In: ICLR (2025)
21. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
23. Saito, K., Sohn, K., Zhang, X., Li, C.L., Lee, C.Y., Saenko, K., Pfister, T.: Pic2Word: Mapping pictures to words for zero-shot composed image retrieval. In: CVPR (2023)
24. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5B: An open large-scale dataset for training next generation image-text models. In: NeurIPS (2022)
25. Team, G.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
26. Team, G.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
27. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmsen, J., Steiner, A., Zhai, X.: SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
28. Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.J., Fei-Fei, L., Hays, J.: Composing text and image for image retrieval - an empirical odyssey. In: CVPR (2019)
29. Wei, H., Wang, S., Xue, Z., Chen, S., Huang, Q.: Conversational composed retrieval with iterative sequence refinement. In: ACM MM (2023)
30. Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., Feris, R.: Fashion IQ: A new dataset towards retrieving images by natural language feedback. In: CVPR (2021)
31. Yang, Z., Xue, D., Qian, S., Dong, W., Xu, C.: LDRE: LLM-based divergent reasoning and ensemble for zero-shot composed image retrieval. In: SIGIR (2024)
32. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: CoCa: Contrastive captioners are image-text foundation models. TMLR (2022)
33. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. In: CVPR (2022)
34. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV (2023)
35. Zhang, K., Luan, Y., Hu, H., Lee, K., Qiao, S., Chen, W., Su, Y., Chang, M.W.: MagicLens: Self-supervised image retrieval with open-ended instructions. In: ICML (2024)
36. Zhou, J., Xiong, Y., Liu, Z., Liu, Z., Xiao, S., Wang, Y., Zhao, B., Zhang, C.J., Lian, D.: MegaPairs: Massive data synthesis for universal multimodal retrieval. In: ACL (2025)