Journal of Data-centric Machine Learning Research (2026) Submitted 01/26; Published 08/26

CIRCLED: A Multi-turn CIR Dataset with Consistent Dialogues across Domains

Tomohisa Takeda[1] T_TAKEDA@HAL.T.U-TOKYO.AC.JP
Yu-Chieh Lin[2] YUCHIEH.LIN@KIOXIA.COM
Yuji Nozawa[2] YUJI1.NOZAWA@KIOXIA.COM
Youyang Ng[2] YOUYANG.NG@KIOXIA.COM
Osamu Torii[2] OSAMU.TORII@KIOXIA.COM
Yusuke Matsui[1] MATSUI@HAL.T.U-TOKYO.AC.JP
[1] *Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan*
[2] *Kioxia Corporation, Tokyo, Japan*

Reviewed on OpenReview: *https://openreview.net/forum?id=xW9cXbbMJJ*

Editor: Hugo Jair Escalante

Abstract

Existing Multi-Turn Composed Image Retrieval (MTCIR) datasets lack dialogue-history consistency and are restricted to the fashion domain. To address these limitations, we construct CIRCLED by extending FashionIQ, CIRR, and CIRCO. In CIRCLED, the query at each turn progressively approaches the target image. Data are generated via a CIReVL-based retrieval pipeline and curated with multiple filters on retrieval success, turn length, consistency, and information redundancy to ensure quality. In total, we collect 22,608 multi-turn sessions across nine subsets, substantially exceeding Multi-turn FashionIQ (11,505 sessions) in both scale and generality. We further apply multiple baseline methods and quantitatively assess retrieval accuracy on CIRCLED. Our work provides a practical, high-quality benchmark to facilitate future research on multi-turn CIR. The dataset is publicly available at `https://huggingface.co/datasets/tk1441/CIRCLED`, and the code at `https://github.com/mti-lab/circled`.

Keywords: Composed Image Retrieval, Multi-turn Retrieval, Dataset, Vision-Language Models

1 Introduction

Image retrieval aims to find images that match a user's intent in large corpora. Traditional text-to-image (Lu et al., 2021; Li et al., 2023) and content-based (Liu et al., 2016; Cao et al., 2010) methods search with either a text query or an image, but they struggle when intent is ambiguous or nuanced.

Multimodal retrieval addresses this by combining text and images in a shared embedding space with pretrained vision-language models such as CLIP (Radford et al., 2021) and BLIP (Li et al., 2022). Composed Image Retrieval (CIR) further refines user's intent by conditioning on a reference image plus a text modification, preserving visual structure while

©2026 Takeda, Lin, Nozawa, Ng, Torii, and Matsui.

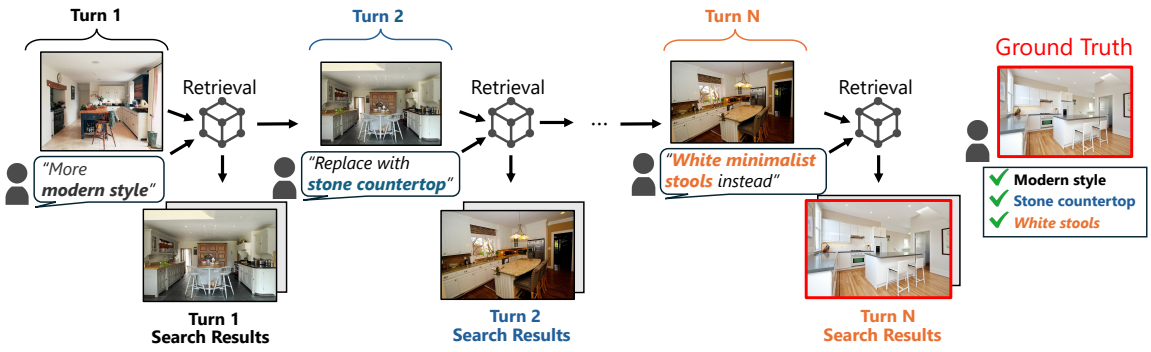


Figure 1: Example of multi-turn CIR search. When users cannot find the image they want, they can search again in the next turn based on the existing search results to find the image they want. The input for each turn is the same as the conventional CIR query, which is an image and text.

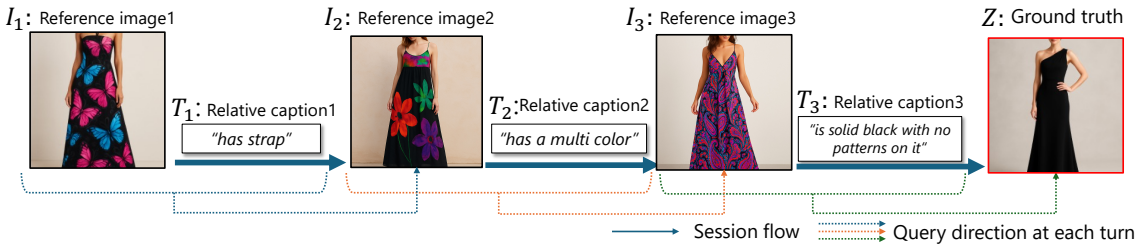
flexibly steering the search (Tian et al., 2023; Zhang et al., 2024; Karthik et al., 2024; Liu et al., 2024).

Nevertheless, most CIR systems use a single-turn setting and give limited consideration to multi-turn usage, i.e., interaction that progressively clarifies the user's intent. This single-turn setup forces users to state their intent in one shot. It performs poorly when user's intent is vague or when users refine requirements while inspecting results.

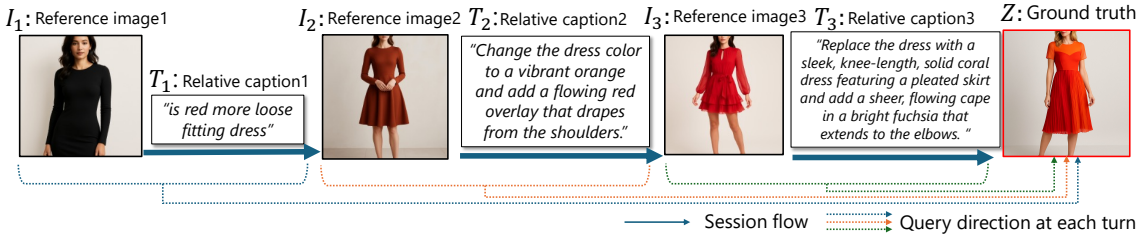
Against this background, multi-turn CIR (Yuan and Lam, 2021; Pal et al., 2023; Chen et al., 2025), which follows the flow illustrated in Fig. 1, has recently attracted attention. In multi-turn CIR, retrieval is performed based on the history of L image-text pairs $\{(I_1, T_1), \ldots, (I_L, T_L)\}$ used in previous turns, thereby more faithfully capturing the user's evolving intent.

A key issue in existing multi-turn CIR is inconsistent dialogue history. At present, the only publicly available dataset is Multi-turn FashionIQ (Yuan and Lam, 2021). It is built by simply concatenating single-turn queries. As shown in Fig. 2a, it does not provide a structure in which each turn progressively approaches the ground truth (GT). For example, even when the GT is a black dress, the dataset may include unrelated text such as "has a strap" or "has a multi color," and assign later-turn reference images that are colorful dresses far from the intended black dress. These intermediate-turn mismatches undermine realistic iterative search because later images and captions diverge from the target. Moreover, prior work has focused on fashion, leaving multi-turn CIR in general domains largely unexplored.

We introduce CIRCLED, a multi-turn dataset with consistent, unidirectional progress toward the GT (Fig. 2b). We design the query sequence $\{(I_1, T_1), \ldots, (I_L, T_L)\}$ so each turn steadily approaches the GT, mirroring real search behavior. Each turn (I_i, T_i) remains aligned with the GT, preserving coherence at both turn and session levels. We construct CIRCLED by extending FashionIQ (Wu et al., 2021), CIRR (Liu et al., 2021), and CIRCO (Baldrati et al., 2023), broadening coverage beyond fashion to general domains. CIRCLED contains 22,608 sessions spanning 2–6 turns and 202,845 images (Table 1), roughly twice that of Multi-turn FashionIQ, and includes 1,078 sessions with 5–6 turns, providing longer interactions.



(a) An example from an existing multi-turn CIR dataset



(b) An example from our multi-turn CIR dataset CIRCLED

Figure 2: Comparison of multi-turn CIR datasets. Solid arrows indicate the progression of a session; dotted arrows indicate which previous inputs a given turn refers to. (a) In the existing Multi-turn FashionIQ, feedback is primarily oriented toward the next turn's image and does not necessarily yield a consistent progression toward the ground truth[1]. (b) In our dataset, the query at each turn consistently points toward the ground truth, yielding a structure that progressively approaches it.

We conduct several baseline experiments on CIRCLED to analyze its characteristics. Furthermore, we assess performance using three metrics to complement aspects that previous multi-turn CIR settings could not evaluate.

The main contributions of this work are as follows:

- CIRCLED: a multi-turn CIR dataset with consistency and progressive information gain, missing in prior work.
- Broader coverage beyond fashion by extending FashionIQ, CIRR, and CIRCO.
- Three metrics (Hits@10, Final Recall@10, AUC) enabling analysis of turn-wise success and early reachability, as well as final accuracy.

The dataset is publicly available at `https://huggingface.co/datasets/tk1441/CIRCLED`, and the construction and evaluation code at `https://github.com/mti-lab/circled` (see Appendix A for documentation).

Dataset	#Sessions by Turn Length						#Images	Categories
	2	3	4	5	6	Total		
Multi-turn FashionIQ (Yuan and Lam, 2021)	8,649	2,216	640	–	–	11,505	16,712	Dress, Shirt, Toptee
CIRCLED (Ours)	14,216	5,424	1,890	772	306	22,608	202,845	Dress, Shirt, Toptee, General

Table 1: Comparison of the number of sessions (=queries) by dialogue length L and covered categories in CIRCLED and Multi-turn FashionIQ. L denotes the number of turns per session; larger L generally indicates higher difficulty. CIRCLED covers $2 \leq L \leq 6$ and includes long sessions with 5–6 turns, whereas Multi-turn FashionIQ is mostly $L \leq 3$ (maximum $L = 4$) and contains no 5–6 turn sessions. *Total* is the total number of sessions, *#Images* is the number of database images, and *Categories* lists the covered categories.

2 Related Work

2.1 Composed Image Retrieval

Composed Image Retrieval (CIR) (Saito et al., 2023; Zhang et al., 2024; Gu et al., 2024; Liu et al., 2024) retrieves a ground-truth image using a reference image together with a text modification. By integrating visual and textual information, CIR enables flexible retrieval refinement in line with the user's intent.

Traditional CIR methods train on triplets (reference image, text, GT image), which demands large-scale datasets and costly training (Baldrati et al., 2022; Zhang et al., 2024). To reduce dependence on labeled data, recent works leverage Vision-Language Models (VLMs) and Large Language Models (LLMs) for zero-shot CIR (Saito et al., 2023; Baldrati et al., 2023; Karthik et al., 2024; Shimada et al., 2026). Most existing CIR methods use a single turn with one reference image and one text. Iterative search is simulated by repeatedly issuing single-turn queries. However, an explicit multi-turn framework that references past retrieval history and continuously tracks user intent has received relatively little attention.

2.2 Multi-turn Composed Image Retrieval

Multi-turn CIR is the task of retrieving the final target (ground truth) using the history of image-text pairs $\{(I_1, T_1), \ldots, (I_L, T_L)\}$ as the query.

Yuan and Lam (2021) pioneered multi-turn CIR by extending FashionIQ (Wu et al., 2021) and releasing Multi-turn FashionIQ (11,505 sessions, 3 categories), the first dataset for this setting. It is built by concatenating single-turn pairs: each turn (I_i, T_i) is designed as a query for the next image I_{i+1}. However, there is no guarantee that intermediate images $\{I_2, \ldots, I_{L-1}\}$ progressively approach the ground truth, making intermediate turn evaluation infeasible. Consequently, existing methods only report final-turn retrieval metrics.

Pal et al. (2023) construct MT Shoes (4,097 sessions, 10 categories) by concatenating single-turn transactions from the Shoes dataset (Berg et al., 2010) and propose a cascaded memory network to retain historical information. Chen et al. (2025) build FashionMT, a significantly larger dataset with 247,911 sessions, 95 categories, and 1,067,688 images, using LLM-based modification generation. FashionMT introduces "retrospective" settings where

1. For licensing reasons, the images shown in the figure are generated images that are visually similar to images in FashionIQ.

Table 2: Comparison of multi-turn CIR datasets. "Consistent" indicates a consistent rank progression toward the ground truth, within a bounded margin (ε-consistency). "Public" indicates that the constructed multi-turn session data is available for download as of submission. Only CIRCLED is simultaneously consistent, cross-domain, and public.

Dataset	#Sessions	Domain	Consistent	Public
Multi-turn FashionIQ (Yuan and Lam, 2021)	11,505	Fashion	×	✓
MT Shoes (Pal et al., 2023)	4,097	Fashion	×	×
FashionMT (Chen et al., 2025)	247,911	Fashion	×	×
CIRCLED (ours)	22,608	Fashion + General	✓	✓

users may refer back to attributes from previous turns (e.g., "keep the color from turn 2") or roll back to earlier images. Despite its scale and retrospective design, FashionMT evaluates only the final turn and lacks explicit consistency constraints across turns, as intermediate images are not guaranteed to monotonically improve toward the ground truth. Moreover, to the best of our knowledge and as of submission, the constructed multi-turn session data for MT Shoes[2] and FashionMT[3] have not been released for download (only papers and supplementary material are available). In contrast, Multi-turn FashionIQ (Yuan and Lam, 2021) is publicly released.[4] All three datasets remain limited to the fashion domain.

Beyond CIR datasets, interactive and conversational image retrieval has also been studied as a *method*. ChatIR (Levy et al., 2023) and PlugIR (Lee et al., 2024) retrieve images through multi-turn natural-language dialogue with an LLM. UniCVR (Wen et al., 2026) unifies single-turn, multi-turn, and video composed retrieval in a zero-shot framework. These works propose retrieval methods rather than a consistent multi-turn benchmark, which is our focus.

In contrast, our CIRCLED dataset ensures consistent progression toward the ground truth, within a bounded margin, via ε-consistency ($r_{l+1} \leq r_l + \varepsilon$; defined in Sec. 3.3), enabling evaluation at any intermediate turn. This allows us to introduce turn-wise metrics (Hits@10, AUC) that measure retrieval quality throughout the entire multi-turn session, not just the final turn. Furthermore, CIRCLED is publicly available and covers both fashion and general domains (CIRR, CIRCO), offering broader applicability.

Table 2 summarizes how CIRCLED compares with existing multi-turn CIR datasets. It is the only one that is simultaneously consistent, cross-domain, and publicly available.

In a related but distinct setting, Zhao et al. (2025) propose ChatSearch for general conversational image retrieval with free-form multi-modal dialogues. In contrast, our work extends CIR to multi-turn while preserving the structured query format of (reference image, text modification) pairs.

3 Proposed Dataset: CIRCLED

We address two limitations of existing multi-turn CIR datasets: the lack of history consistency and the restriction to the fashion domain. Concretely, we extend three single-turn CIR

2. https://sites.google.com/eng.ucsd.edu/fashionntm
3. https://openreview.net/forum?id=gXyWbl71n1
4. https://github.com/yfyuan01/MultiturnFashionRetrieval

datasets (FashionIQ (Wu et al., 2021), CIRR (Liu et al., 2021), and CIRCO (Baldrati et al., 2023)) to construct a new dataset that features consistent multi-turn dialogues across multiple domains. CIRR (Liu et al., 2021) is a general-domain CIR dataset of real-life images requiring fine-grained distinctions, and CIRCO (Baldrati et al., 2023) is an open-vocabulary CIR dataset built on COCO images with multiple ground truths per query. We process each source dataset *independently* into multi-turn sessions rather than merging them, and we evaluate each subset separately. The general domain (CIRR and CIRCO) accounts for 34.9% of all sessions and the fashion domain (FashionIQ) for 65.1%.

Summary statistics are shown in Table 1. Compared to the existing dataset (Yuan and Lam, 2021), CIRCLED is roughly twice as large in the number of queries and contains about twelve times more database images (largely reflecting the much larger general-domain search spaces of CIRR and CIRCO rather than more annotation per query), and it adds a General category beyond the three fashion categories. It also includes longer dialogues with 5–6 turns.

3.1 Overview

We describe the components of the dataset. Let $\mathcal{X} = \{X_n\}_{n=1}^{N}$ be a database of N images to be searched. We define a *session* as an L-turn query sequence $\{(I_1, T_1), (I_2, T_2), \ldots, (I_L, T_L)\}$, where each I_l is a reference image and each T_l is a relative caption describing the desired change. Each session is paired with a ground truth (GT) image $Z \in \mathcal{X}$. Given a session, a retrieval algorithm S must find Z from $\mathcal{X}$. The final performance of S is computed by averaging results over many sessions. We precompute and release these sessions so that future users can evaluate their multi-turn retrieval algorithms.

We form each session as follows. We choose a GT image $Z \in \mathcal{X}$ from an existing single-turn CIR dataset. We then construct a multi-turn query sequence $\{(I_1, T_1), (I_2, T_2), \ldots, (I_L, T_L)\}$ that progressively approaches Z. For Turn 1, we use the dataset's reference image I_1 and its relative caption T_1. For later turns, we generate $T_2, T_3, \ldots$ using an LLM and select $I_2, I_3, \ldots$ from the retrieval results; details are given in Sec. 4.

A well-formed multi-turn relative caption should (i) reference the current (selected) image, (ii) describe a concrete change toward the goal, (iii) add information not already provided, and (iv) improve the target's rank. The last two requirements correspond to the τ-diversity and ε-consistency properties that we formalize in Sec. 3.3. Turns 2 onward are generated to meet all four requirements. Turn 1 is intentionally different. At Turn 1 the user has not yet seen any candidates and states a vague initial intent, so we reuse the dataset's original single-turn relative caption T_1, a short comparative description. From Turn 2, the user has selected an image from the results and issues a concrete corrective instruction. This asymmetry mirrors the staged shift of intent before and after inspecting results. Keeping the original T_1 also makes Turn 1 a standard single-turn CIR query, which enables direct comparison with existing single-turn methods and grounds the dataset in the original human-annotated CIR data.

This session simulates a realistic search scenario in which the user refines the query while inspecting intermediate results. Specifically, if the first-turn search with (I_1, T_1) does not rank Z sufficiently high, the user selects a desirable image from the current top results as I_2,

Algorithm 1: Single-session evaluation protocol. Apply S to the cumulative history up to turn l to obtain r_l. The dataset fixes the query sequence.

Input: retrieval algorithm S
Output: rank sequence $r_1, \ldots, r_L \in \{1, \ldots, N\}$
for $l \in \{1, \ldots, L\}$ **do**
  $r_l \leftarrow S(\{(I_1, T_1), (I_2, T_2), \ldots, (I_l, T_l)\})$
return $r_1, \ldots, r_L$

and provides an additional description T_2 for I_2. We rerun retrieval with $(I_1, T_1), (I_2, T_2)$, and this process is repeated until Z appears sufficiently high in the ranking.

3.2 Evaluation Protocol

Algorithm 1 outlines the evaluation pipeline for a single session. Let S be the retrieval algorithm under evaluation. We evaluate retrieval performance turn by turn. The first-turn search uses $\{(I_1, T_1)\}$, and the resulting rank of Z is denoted $r_1 \in \{1, \ldots, N\}$:

$$r_1 = S(\{(I_1, T_1)\}). \tag{1}$$

At the second turn, we search using the accumulated information $\{(I_1, T_1), (I_2, T_2)\}$, yielding rank r_2. Repeating this gives, at turn l,

$$r_l = S(\{(I_1, T_1), (I_2, T_2), \ldots, (I_l, T_l)\}). \tag{2}$$

The objective is to obtain a sufficiently small r_l with as few turns (small l) as possible.

In this protocol, the dataset $\mathcal{X}$, each pair (I_l, T_l), and the GT Z are all fixed constants. Importantly, the I_l are predetermined and independent of the algorithm S, as in existing multi-turn CIR datasets. Each session has one ground truth Z (for CIRCO, a set of equivalent target images, of which the best-ranked one defines r_l). There is no per-turn ground truth. The intermediate I_l are fixed query inputs, namely the reference images inspected during the simulated dialogue, and are not retrieval targets. Both the metrics (Sec. 6.1) and ε-consistency are defined on the rank r_l of this single Z under the cumulative history, so they measure the same quantity. Fixing the I_l makes the benchmark reproducible, since every algorithm S is evaluated on the identical query sequence. This yields a controlled, non-interactive benchmark rather than a live user simulation. The I_l were selected by the construction baseline, so the trajectory is fixed, but the same trajectory is used for every method, so the comparison remains fair. If these intermediate choices are poor (e.g., drifting from the GT or repeating information), the reported performance reflects dataset artifacts rather than the quality of S. We guard against this with a strong baseline and the filtering procedures in Sec. 5.1.

The above describes evaluation for a single session. We then apply this protocol to all sessions, compute the metrics defined in Sec. 6.1 (Hits@10, Recall@10, AUC) from the resulting rank sequences $\{r_l\}$, and report the mean over sessions as the final performance.

The protocol in Algorithm 1 generalizes many retrieval problems. When $L = 1$, it reduces to single-turn CIR. When $T_1 = T_2 = \cdots = \varnothing$, it becomes L-turn relevance feedback in image retrieval. When $I_1 = I_2 = \cdots = \varnothing$, it corresponds to L-turn chat-based image

retrieval (Levy et al., 2023) Thus, our protocol subsumes several existing settings as special cases, with multi-turn CIR as a particular instantiation.

3.3 Dataset Characteristics

To realize natural multi-turn dialogues, we define two properties each session must satisfy: ε-consistency and τ-diversity. All sessions in CIRCLED satisfy these properties.

Definition 1 (ε-consistency) *A session is* ε-consistent *w.r.t. a retrieval algorithm* S *if its rank sequence* $r_1, r_2, \ldots$ *satisfies*

$$r_{l+1} \leq r_l + \varepsilon. \tag{3}$$

Here, ε is a small integer margin, and in CIRCLED we set $\varepsilon = 30$. With ε-consistency, we ensure that as turns progress, the GT (within a permitted margin) consistently moves toward higher ranks. If a session violates this property, adding an off-target (I_l, T_l) moves the aggregated query away from Z in the embedding space, so repeating the query sequence can instead push the target image farther down the ranking.

Our dataset is constructed so that all sessions satisfy ε-consistency with respect to the baseline retrieval algorithm described in Sec. 4. This rank progression is not specific to the construction baseline. It also holds under other retrieval models and encoders (Appendix G). By contrast, existing multi-turn CIR datasets do not satisfy this property; in some cases, the GT drifts downward as the dialogue proceeds, making it difficult to reliably evaluate retrieval algorithms.

Definition 2 (τ-diversity) *A session is* τ-diverse *w.r.t. a text encoder* E *if its relative-caption sequence* $T_1, \ldots, T_L$ *satisfies*

$$\max_{1 \leq j < i \leq L} \cos\big(E(T_i),\, E(T_j)\big) \;<\; \tau. \tag{4}$$

Here, E is a text encoder and $\cos(\cdot, \cdot)$ denotes cosine similarity. With τ-diversity, we ensure that each turn contributes novel information. The threshold τ controls how much overlap is tolerated, and we use $\tau = 0.8$ with E a CLIP text encoder. Smaller τ requires more novel information at each turn. Without this constraint, the same or similar relative captions may be repeated across turns, adding no new information and preventing meaningful evaluation of multi-turn retrieval.

Each CIRCLED session exhibits a natural and consistent structure with these two properties: the GT rank steadily improves (allowing small fluctuations) as the dialogue progresses, and each turn adds new information. In particular, for sessions with large L, τ-diversity enforces continuous information addition, while ε-consistency ensures stepwise rank improvement, enabling clear visualization of performance differences in history aggregation for long dialogues.

3.4 Dataset Bias Analysis

Since the relative captions in CIRCLED are generated by an LLM, we investigate whether they exhibit different linguistic characteristics compared to existing single-turn CIR datasets

Table 3: Linguistic metrics comparison.

	Fashion		General	
Metric	FashionIQ	CIRCLED	CIRR/CIRCO	CIRCLED
Avg. Length of T_l (words)	5.3	19.0	11.2	19.0
Type-Token Ratio	0.017	0.013	0.020	0.023
Noun Ratio (%)	29.3	45.4	46.4	45.3
Verb Ratio (%)	27.7	14.6	16.3	15.0
Adjective Ratio (%)	28.3	24.3	14.3	19.4

(FashionIQ, CIRR, CIRCO), which were created by humans. We follow prior caption analysis (Wu et al., 2021).

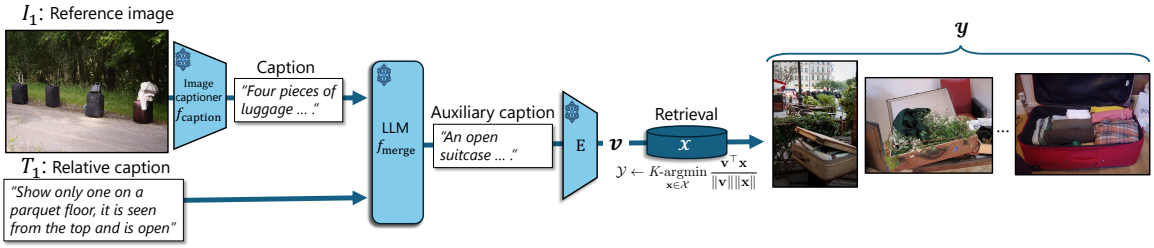
Table 3 compares linguistic metrics. CIRCLED's relative captions are longer but maintain comparable vocabulary diversity. Specifically, CIRCLED's relative captions are significantly longer (average $|T_l| = 19.0$ words vs. 5.3–11.2 in existing datasets). This difference stems from task design: single-turn CIR describes the overall difference between images at once, whereas multi-turn CIR requires explicit step-by-step modification instructions. For example, single-turn may use concise descriptions like "is darker and longer" (4 words), while multi-turn requires specific instructions such as "Replace the blue shirt with a black shirt featuring a centered graphic design" (13 words). The Type-Token Ratio measures vocabulary diversity as the ratio of unique words to total words (Richards, 1987). This metric is comparable between CIRCLED and existing datasets, suggesting that CIRCLED's vocabulary diversity remains competitive. The part-of-speech ratios show the proportion of each word class in the captions. FashionIQ exhibits a balanced distribution (Noun/Verb/Adj $\approx$ 29/28/28%), whereas CIRCLED and CIRR/CIRCO are noun-dominant ($\approx$ 45/15/20%).

FashionIQ and CIRR/CIRCO frequently use comparative expressions such as "more," "darker," and "longer," while CIRCLED prominently features action verbs like "add," "replace," and "change," as the word clouds of frequent terms we provide in Appendix K illustrate. This reflects the difference in format: single-turn describes differences between images, whereas multi-turn provides step-by-step modification instructions. Multi-turn FashionIQ exhibits a similar distribution to FashionIQ because it reuses the original FashionIQ captions without modification.

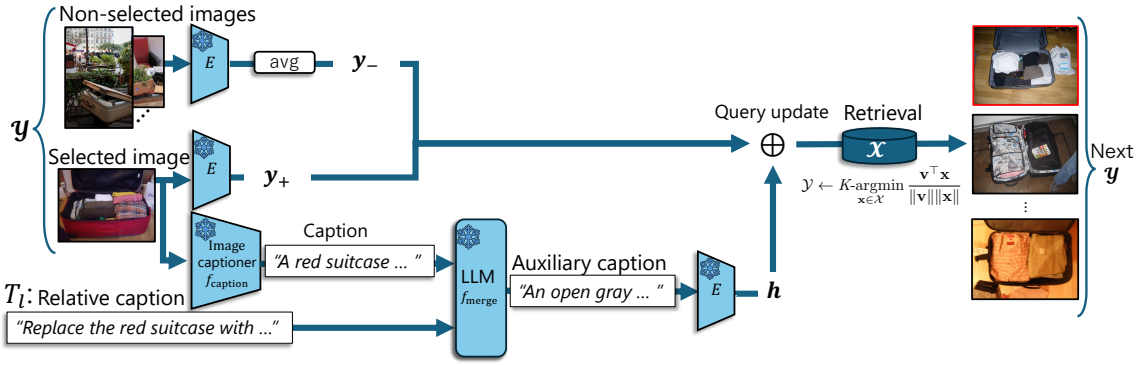
Differences across domains are also observed. In the Fashion domain, both existing datasets and CIRCLED frequently use clothing terms such as "sleeves," "shirt," and "dress," as well as color terms like "black" and "blue." This is because tasks in the Fashion domain focus on fine-grained changes to clothing attributes (color, length, pattern, etc.). In contrast, the General domain features more object and spatial terms such as "dog" and "background." This reflects the prevalence of scene-level modifications in the General domain, such as adding/removing objects or changing backgrounds.

4 Baseline Retrieval Algorithm

We describe the baseline retrieval algorithm S_{base} used to construct $(I_2, T_2), \ldots$. This baseline is a simple yet effective approach to multi-turn CIR and also decides "which image



(a) Search in Turn 1



(b) Search in Turn 2 and beyond

Figure 3: Multi-turn image retrieval pipeline. (a) In Turn 1, we merge the caption generated from the reference image with the relative caption using an LLM and form a retrieval query following CIReVL (Karthik et al., 2024). (b) From Turn 2 onward, we update the query by incorporating features of selected and non-selected images as well as textual information, and then perform image retrieval.

to pick for the next turn" when building the dataset. Let $\mathbf{z} \in \mathbb{R}^D$ denote the D-dimensional feature of the GT image Z, and $\mathbf{x}_i \in \mathbb{R}^D$ the feature of image X_i. For clarity, we redefine $\mathcal{X} = \{\mathbf{x}_n\}_{n=1}^N$ as the set of image features.

4.1 Algorithm Setup

We extend CIReVL (Karthik et al., 2024) to mimic human's multi-turn image search and to synthesize multi-turn CIR sessions. CIReVL requires no training, is encoder-agnostic, and is independent of specific LLMs/VLMs, allowing flexible use with the latest models. Briefly, CIReVL frames CIR as text-only retrieval: it captions the reference image and uses an LLM to fuse that caption with the relative caption into a single natural-language description, which is then embedded and matched against the image index. In this paper, we implement it as follows:

- Image/text embedding encoder: E (BLIP (Li et al., 2022))
- Text merging: $f_{\text{merge}}(\cdot,\cdot)$ (GPT-4o-mini (OpenAI et al., 2024))
- Image-to-caption generation: $f_{\text{caption}}(\cdot)$ (GPT-4o-mini)

- Difference generator: $f_{\text{diff}}(\cdot,\cdot,\cdot)$ (GPT-4o-mini)

We further ensure data quality with the filtering procedures described in Sec. 5.1.

4.2 Retrieval Pipeline

We start retrieval at Turn 1 and repeat it for up to six turns. This six-turn cap is empirical rather than fundamental. Under the construction baseline, the ground truth enters the top-10 within six turns for the large majority of sessions, and only 1,078 sessions (4.8%) require five or six turns. Sessions that still fail after six turns are rare and are removed by the retrieval-success filter (Sec. 5.1). Allowing more turns would mostly add long, hard-to-annotate sessions with diminishing returns and higher generation cost. When the initial image I_1 is far from the target, this manifests as a high Turn-1 rank. Such a session simply needs more refinement turns to bring the GT into the top-10 (the retrieval-success condition). We retain it only if it converges within the six-turn budget, and otherwise filter it out. The cap can be raised at construction time if longer interactions are desired.

4.2.1 Turn 1 ($l = 1$).

As shown in Fig. 3a, Turn 1 follows the standard single-turn CIR setup: given a reference image I_1 and a relative caption T_1, we generate a query and obtain the top-K candidates $\mathcal{Y}$. The steps are:

1. Generate a caption from I_1 using f_{caption}.
2. Merge the generated caption with T_1 using f_{merge}.
3. Encode the merged sentence with the text encoder E to obtain a feature vector $\mathbf{v} \in \mathbb{R}^D$.
4. Use $\mathbf{v}$ to retrieve from $\mathcal{X}$ by cosine similarity and obtain the candidate set $\mathcal{Y} \subset \mathcal{X}$.

If the GT Z is not included in $\mathcal{Y}$, we consider the turn unsuccessful and proceed to the next turn. For CIRCO, where multiple GT variants exist, we judge success if at least one of them appears in $\mathcal{Y}$.

4.2.2 Turn 2 and beyond ($2 \leq l$).

From Turn 2 onward, we use the candidate set $\mathcal{Y}$ obtained in the previous turn and perform the following steps.

Selecting an image. We select from $\mathcal{Y}$ the image y_+ closest to the GT Z, excluding images chosen in earlier turns:

$$\text{sim}(Z, y) = \cos(E(f_{\text{caption}}(Z)),\ E(f_{\text{caption}}(y))) \tag{5}$$

$$y_+ = \underset{y \in \mathcal{Y} \setminus \{I_1, \ldots, I_{l-1}\}}{\text{argmax}} \text{sim}(Z, y). \tag{6}$$

Record this y_+ as the reference image I_l for Turn l. For CIRCO, we determine similarity using the caption of the ground-truth variant Z most similar to the current query.

This step emulates user behavior: scanning the ranked list and picking the item closest to the GT. Image-only selection often echoes the current top ranks and stalls progress, so we select with caption-based textual features.

Generating a relative caption. Next, we generate the relative caption T_l from the selected image y_+, the GT image Z, and the history of past captions $H_{l-1} = (T_1, \ldots, T_{l-1})$ by prompting a VLM:

$$T_l \;=\; f_{\text{diff}}(y_+,\, Z,\, H_{l-1}). \tag{7}$$

We instruct the VLM to describe how Z differs from y_+, avoid repeating H_{l-1}, and add new attributes or viewpoints. For CIRCO, when multiple GT variants exist, we pass the variant Z that is most similar to the current query. This procedure mimics a user who issues iterative, corrective instructions based on the most recently selected image.

Updating the query. Finally, update the query using the following three components:

- $\mathbf{y}_+$: feature of the selected image,
- $\mathbf{y}_-$: mean feature of the non-selected images $\big(\mathcal{Y} \setminus \{y_+\}\big)$,
- $\mathbf{h}$: feature of an auxiliary caption obtained by merging the selected image's caption with the new relative caption.

$$\mathbf{h} = E\big(f_{\text{merge}}\big(T_l,\, f_{\text{caption}}(y_+)\big)\big) \tag{8}$$

Update $\mathbf{v}$ as

$$\mathbf{v} \leftarrow \frac{\mathbf{v} + \alpha\mathbf{y}_+ - \beta\mathbf{y}_- + \gamma\mathbf{h}}{\|\mathbf{v} + \alpha\mathbf{y}_+ - \beta\mathbf{y}_- + \gamma\mathbf{h}\|}. \tag{9}$$

where α, β, γ weight the terms. Adding the selected image and auxiliary text while suppressing non-selected images aligns the update with user intent and curbs drift.

4.3 Constructing Multi-turn Sessions

By repeating this process for up to six turns, we extend an original single-turn CIR query $(I_{\text{ref}}, T_{\text{ref}}) \to Z$ into an interactive sequence $(I_{\text{ref}}, T_{\text{ref}}) \to \cdots \to (I_l, T_l) \to \cdots \to Z$. Each (I_l, T_l) is determined by the image-selection and relative-caption generation procedures described above.

5 Dataset Construction

Using the baseline S_{base} defined in the previous section, we extend existing single-turn CIR datasets and construct high-quality multi-turn dialogues. This section details the filtering procedures and resulting statistics.

5.1 Filtering Process

The multi-turn data generated with S_{base} may include failures (the target is never retrieved) or redundant relative captions that add no new information. We therefore apply four filters to remove such cases.

Subsets	Init	△ Succ	△ Multi	△ Rank	△ Text	Final
fdt	10883	5829	1011	86	930	3027
fdv	3722	1468	491	17	386	1360
fst	11300	4656	2067	149	910	3518
fsv	3881	1216	740	54	371	1500
ftt	11544	4705	1698	123	1208	3810
ftv	3732	1011	707	32	476	1506
crt	28225	5847	13328	95	2081	6874
crv	4181	225	2801	4	192	959
cov	220	26	124	2	14	54
Total	**77,688**	**24,983**	**22,967**	**562**	**6,568**	**22,608**

Table 4: Filtering pipeline: per-stage removals (stage-wise decrements) and final counts by subset. Succ: retrieval-success; Multi: multi-turn; Rank: rank-margin; Text: text-redundancy. Subsets: fdt/fdv/fst/fsv/ftt/ftv = FashionIQ Dress/Shirt/Toptee (train/val); crt/crv = CIRR (train/val); cov = CIRCO (val).

	Summary			#Sessions by turn				
Subset	**#Sessions (total)**	**#Images**	**Avg. turns**	**2 turns**	**3 turns**	**4 turns**	**5 turns**	**6 turns**
fashioniq_dress_train	3,027	10,886	2.74	1,583	920	323	144	57
fashioniq_dress_val	1,360	3,653	2.73	745	358	163	64	30
fashioniq_shirt_train	3,518	18,500	2.49	2,376	752	251	101	38
fashioniq_shirt_val	1,500	6,182	2.52	990	333	110	48	19
fashioniq_toptee_train	3,810	15,742	2.62	2,253	1,002	359	126	70
fashioniq_toptee_val	1,506	5,261	2.56	934	392	119	45	16
cirr_train	6,874	16,939	2.50	4,581	1,496	513	219	65
cirr_val	959	2,297	2.38	718	166	41	23	11
circo_val	54	123,385	2.61	36	5	11	2	0
Total	**22,608**	**202,845**	**2.56**	**14,216**	**5,424**	**1,890**	**772**	**306**

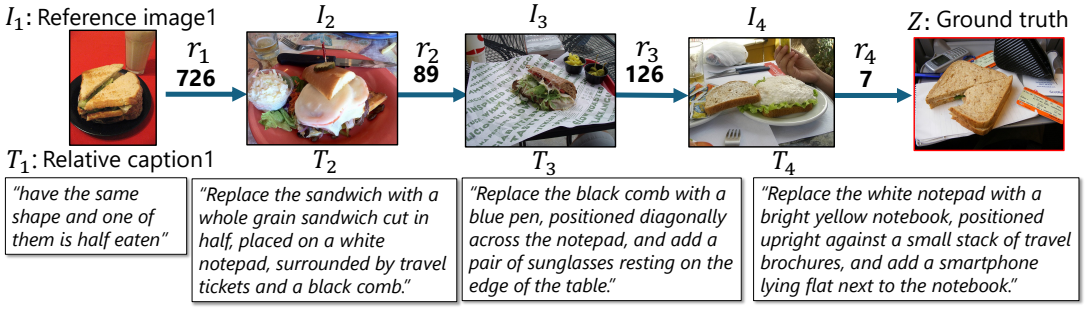
Table 5: Post-filtering statistics: number of sessions, number of images, average turns, and distribution of sessions by turn count for each subset. The Subsets column follows the naming {dataset}_{category (dress, shirt, toptee)}_{split (train or val)}.

1. Retrieval-success filter. We remove any query where the GT Z never appears in the top-K ($K = 10$) candidates $\mathcal{Y}$ during turns 1–6.

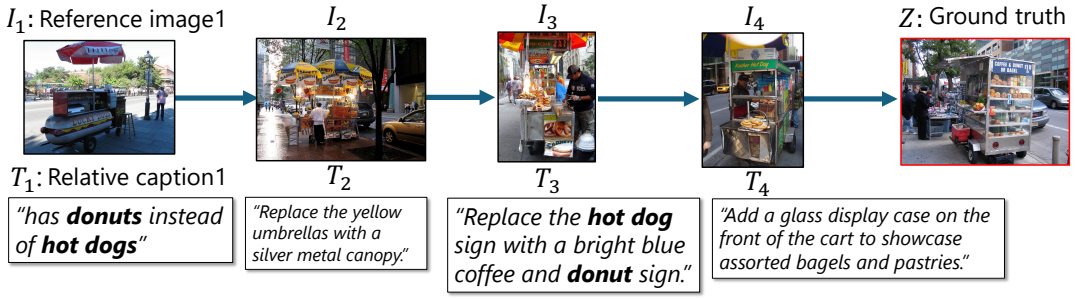
2. Multi-turn filter. We exclude queries that already succeed at Turn 1 (i.e., the original single-turn query), keeping only dialogues that require at least two turns.

3. Rank-margin filter (ε-consistency). We remove any session that has a turn l with $r_{l+1} > r_l + \varepsilon$. We set $\varepsilon = 30$. This enforces ε-consistency and prevents large drops in the GT rank at intermediate turns.

4. Text-redundancy filter (τ-diversity). We remove any session that has a turn i where $\max_{1 \le j < i \le L} \cos\big(E(T_i), E(T_j)\big) \ge \tau$. We set $\tau = 0.8$ and use CLIP (Radford et al., 2021) for E. This enforces τ-diversity and prevents caption redundancy across turns.



(a) Rank-margin filter: mid-turn GT rank drop (ε-consistency violated).



(b) Text-redundancy filter: duplicate captions (τ-diversity violated).

Figure 4: Examples rejected by our filtering. (a) Rank-margin filter ($\varepsilon = 30$). (b) Text-redundancy filter ($\tau = 0.8$).

These four filters ensure both consistency in retrieval accuracy and the incremental addition of information in the relative captions. As summarized in Table 4, four filters (success, multi-turn, rank-margin, and text-redundancy) cut 77,688 queries to 22,608 sessions.

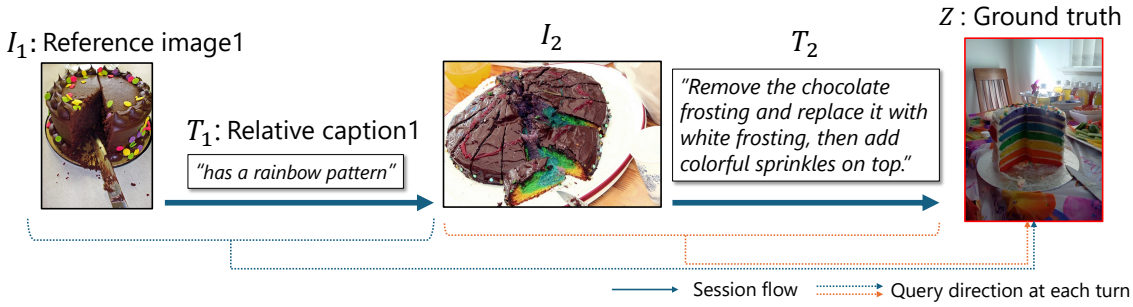
Examples are shown in Fig. 4. Fig. 4a is removed because the GT rank significantly worsens middle dialogue; Fig. 4b is removed because the relative caption repeats earlier content.

In contrast, Fig. 5 shows examples of sessions that passed filtering. These examples demonstrate how retrieval progressively narrows toward the ground truth by combining multimodal information (image and text) at each turn, in both the general domain (`circo_val`) and the fashion domain (`fashioniq_dress_val`).

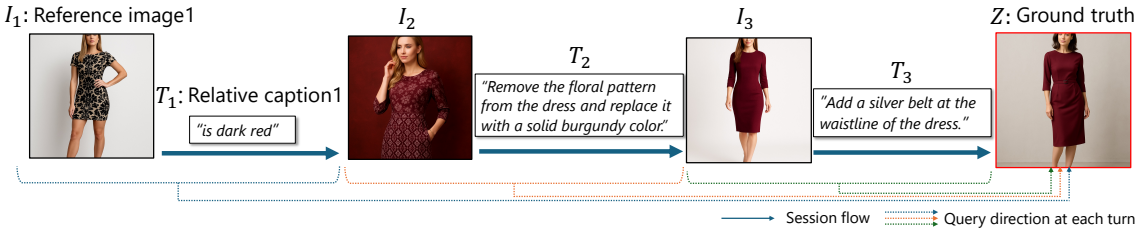
5.2 Statistics

We summarize the key statistics of the filtered dataset. As shown in Table 5, sessions average 2.56 turns and span 2–6 turns. CIRCLED adds a General category and outscales Multi-turn FashionIQ (Yuan and Lam, 2021) in queries, images, and category coverage (Table 1).

Regarding turn-length distribution (Table 5), approximately 63% of sessions (14,216) have 2 turns, 24% (5,424) have 3 turns, and only about 5% (1,078) have 5–6 turns. This distribution is consistent across domains: the proportion of 5–6 turn sessions is 5.1% for Fashion and 4.1% for General, showing no significant difference.



(a) General domain (`circo_val`): A 2-turn session progressively refining cake attributes



(b) Fashion domain (`fashioniq_dress_val`): A 3-turn session progressively adding dress details

Figure 5: Examples of filtered CIRCLED sessions demonstrating gradual progression toward the ground truth. (a) A 2-turn session in the general domain (`circo_val`). (b) A 3-turn session in the fashion domain (`fashioniq_dress_val`). Each turn combines visual (reference image I_l) and textual (relative caption T_l) information while satisfying ε-consistency and τ-diversity[5].

This short-session distribution results from our quality filters, particularly the text-redundancy filter (τ-diversity). As shown in Table 4, the text-redundancy filter removed 6,568 sessions, approximately 12 times more than the 562 sessions removed by the rank-margin filter. That is, most sessions terminate due to τ-diversity violations (generating relative captions similar to previous turns) rather than ε-consistency violations (worsening GT rank). To confirm this, we compared CIRCLED with sessions excluded by the rank-margin and text-redundancy filters ("w/o quality filter" in Table 6; these sessions pass the retrieval-success and multi-turn conditions). As shown in Table 6, the excluded sessions average 3.53 turns—about one turn longer than CIRCLED's 2.54—and contain a much higher proportion of 5–6 turn sessions. As turns progress, fewer new attributes remain to describe differences between the GT and reference images, making redundant relative captions more likely. Allowing longer sessions without filtering would increase turns with little information, degrading dataset quality. Indeed, as detailed in Sec. 5.3, the excluded sessions score lower than CIRCLED on Coherence, Goal Progress, and Redundancy. However, various biases in LLM-based evaluation have been reported, including position bias (sensitivity to presentation order) (Zheng et al., 2023) and self-preference bias (favoring low-perplexity texts) (Wataoka et al., 2024).

2. For licensing reasons, the images shown in the figure are generated images that are visually similar to images in FashionIQ.

Table 6: Session-level evaluation using LLM-as-a-judge (GPT-5-mini). Each metric is rated 1–5 (higher is better). The prompt is provided in Appendix D. "w/o quality filter" includes sessions excluded by filters 3–4 (rank-margin and text-redundancy) but passing filters 1–2.

Source	N	Avg. Turns	Natural.	Coher.	Progress	Redund.	Overall
CIRCLED (Ours)	463	2.54	3.89±0.71	**4.37**±0.72	**4.56**±0.66	**4.96**±0.20	**4.34**±0.60
w/o quality filter	463	3.53	**4.00**±0.61	4.20±0.74	4.28±0.79	4.87±0.39	4.17±0.63
Multi-turn FashionIQ	336	2.30	3.51±0.72	4.15±0.83	4.08±0.86	4.80±0.53	3.98±0.74
Simple Concat	449	2.84	3.38±0.81	3.47±0.98	3.46±0.99	4.67±0.60	3.44±0.82

5.3 Session Quality Evaluation with LLM-as-a-Judge

To evaluate the quality of generated multi-turn dialogues, we conducted session-level evaluation using LLM-as-a-judge. LLM-as-a-judge has been widely studied as an alternative to human evaluation (Gu et al., 2025), and with appropriate design, it has been reported to show high correlation with human evaluation (Zheng et al., 2023; Badshah and Sajjad, 2025; Thakur et al., 2025).

Based on these findings, we adopted the following design: (1) each session was evaluated independently to avoid ordering bias between comparison targets; (2) we used GPT-5-mini as the evaluation model, different from GPT-4o-mini used for dataset generation, to avoid self-preference bias; (3) each session was rated on five dimensions (naturalness, coherence, goal progress, low redundancy, overall) on a 1–5 scale. We sampled approximately 60 sessions per subset from each source where available (1,711 sessions total). Note that Multi-turn FashionIQ covers only Fashion subsets, and some sessions were excluded due to the Azure OpenAI API content filters. We compared: (1) CIRCLED (proposed dataset), (2) w/o quality filter (sessions excluded by filters 3–4), (3) Multi-turn FashionIQ, and (4) Simple Concat (simple concatenation of single-turn datasets).

As shown in Table 6 (higher is better for all metrics), CIRCLED achieved the highest scores on Coherence, Goal Progress, Redundancy, and Overall. The w/o quality filter set scored highest on Naturalness, but this is attributable to differences in turn-count distribution: it has an average of 3.53 turns, and longer sessions tend to receive higher naturalness scores. When comparing sessions with the same number of turns, there was no significant difference in naturalness between CIRCLED and w/o quality filter. Simple Concat scored lowest on all metrics, confirming that simple concatenation of single-turn datasets results in unnatural multi-turn dialogues. We assess statistical significance with pairwise one-sided Mann-Whitney U tests between CIRCLED and each comparison source (Table 10). CIRCLED significantly outperforms every comparison source on Overall, Coherence, Goal Progress, and Redundancy (all $p < 0.001$). Effect sizes (rank-biserial r) are large against Simple Concat ($r = 0.51$–0.61), small-to-medium against Multi-turn FashionIQ ($r = 0.11$–0.31), and small against the w/o-quality-filter set ($r = 0.08$–0.19). The only non-significant difference is Naturalness against the w/o-quality-filter set, which is consistent with its longer average turn count. We further validate these LLM-judge ratings with a human study on the same 50 sessions (Appendix H). Human annotators agree with the LLM judge at the session level (Spearman $\rho = 0.65$ on Overall), reproduce the same source ranking exactly (Spearman $\rho = 1.00$ over the four source means), and rate CIRCLED highest.

6 Experimental Results on CIRCLED

In this section, we quantitatively and qualitatively evaluate retrieval performance on the proposed multi-turn CIR dataset. We analyze how each turn affects accuracy and the relative utility of text and image information.

6.1 Evaluation Metrics

For evaluation, let $\mathcal{Q}$ denote the set of evaluation sessions (queries). We extend r_l introduced in Eq. (2) to each query $q \in \mathcal{Q}$, and denote by $r_l^{(q)}$ the rank of the (single) GT Z for query q at turn l, computed from the cumulative history up to turn l (the same quantity used in ε-consistency). When multiple ground truth targets exist (as in CIRCO), we take $r_l^{(q)}$ to be the best rank among them. Let $L^{(q)}$ be the session length of query q, and let $L_{\max} = \max_{q\in\mathcal{Q}} L^{(q)}$ be the maximum length in the dataset. We then define

$$\mathrm{hit}_l^{(q)} = \mathbf{1}\big(r_l^{(q)} \le K\big), \quad K = 10, \tag{10}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function: it equals 1 if the GT is in the top-K at turn l and 0 otherwise. We use the following three metrics:

- **Hits@10(turn l)**

$$\mathrm{Hits@10}(l) \;=\; \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \max_{1\le l'\le l} \mathrm{hit}_{l'}^{(q)} \tag{11}$$

  The fraction of queries that reach the top-10 at least once by turn l. This captures how quickly a method succeeds. This value is cumulative; for each query, if $l > L^q$ then $\mathrm{Hits@10}(l) = \mathrm{Hits@10}(L^q)$.

- **Final Recall@10**

$$\text{Final Recall@10} = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \mathrm{hit}_{L^{(q)}}^{(q)}. \tag{12}$$

  Accuracy using all available information at the final turn; used to compare final performance with existing methods.

- **AUC over Hits@10**

$$\mathrm{AUC} \;=\; \frac{1}{L_{\max}-1}\sum_{l=1}^{L_{\max}-1} \frac{\mathrm{Hits@10}(l) + \mathrm{Hits@10}(l+1)}{2}. \tag{13}$$

  The trapezoidal integral of the Hits@10 curve. Larger values indicate reaching the top-10 in fewer turns, thus evaluating both final accuracy and convergence speed.

6.2 Compared Methods

Existing multi-turn CIR methods are difficult to use as fair baselines. MT Shoes (Pal et al., 2023) and FashionMT (Chen et al., 2025) are trained models that release neither their code/pretrained weights nor their constructed datasets, so they can be neither run directly nor retrained, and Multi-turn FashionIQ's method (Yuan and Lam, 2021) is fashion-specific.

A from-scratch re-implementation would not faithfully reproduce them and would make for an unreliable comparison. Instead, we compare reproducible, widely-used CIR methods with publicly available weights, and adapt each to the multi-turn setting in a uniform, training-free way.

Multi-turn adaptation. We do not modify any method. At turn l each method produces a normalized feature $\boldsymbol{f}_l$. We aggregate the per-turn features over the cumulative history with one of three strategies (Latest, Average, and exponentially-decayed Weighted, detailed in Appendix B), and retrieve with the aggregated query under Algorithm 1. This aggregation is identical across methods, so differences reflect the underlying CIR method rather than bespoke multi-turn engineering.

Training-free setting. All baselines are used zero-shot. Text-only and Image-only are plain CLIP similarity. Pic2Word (Saito et al., 2023) and MagicLens (Zhang et al., 2024) use their publicly released pretrained weights. CIReVL (Karthik et al., 2024) composes CLIP retrieval with GPT-4o-mini (OpenAI et al., 2024) captioning and is training-free by design. None are fine-tuned on CIRCLED and the aggregation adds no learnable parameters, so CIRCLED serves as a *zero-shot* evaluation benchmark. We use CLIP (ViT-L/14) as the encoder for all methods.

Baselines: Text-only (relative caption), Image-only (reference images), Pic2Word, CIReVL (GPT-4o-mini), and MagicLens.

6.3 Feature Aggregation Strategies

Let $\boldsymbol{f}_l$ (already normalized) be the feature at turn l. We aggregate them as follows (exponential decay coefficient $\alpha = 0.8$ in experiments):

Latest: $\hat{\boldsymbol{f}} = \boldsymbol{f}_l$
Average: $\hat{\boldsymbol{f}} = \frac{1}{l}\sum_{l'=1}^{l} \boldsymbol{f}_{l'}$
Weighted: $\hat{\boldsymbol{f}} = \frac{1}{C}\sum_{l'=1}^{l} \alpha^{l-l'} \boldsymbol{f}_{l'}, \quad C = \sum_{l'=1}^{l} \alpha^{l-l'}$

We report results using the best aggregation strategy for each subset and method. We provide a detailed analysis of aggregation strategies in Appendix B.

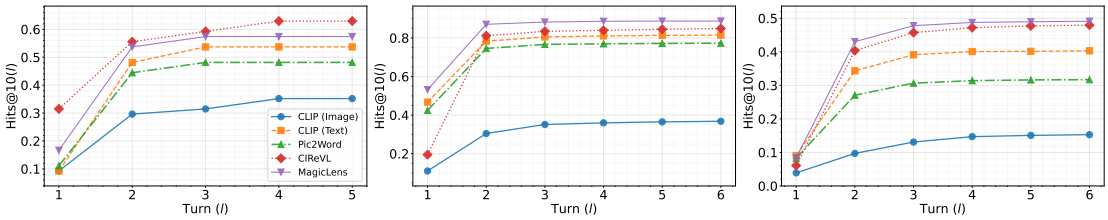
6.4 Quantitative Results

Turn-wise Hits@10 Analysis

Turn-wise Hits@10 results are shown in Fig. 6. On `cirr_val`, MagicLens exceeds 85% but gains little with more turns. On `circo_val` and `fashioniq_dress_val`, accuracy climbs then plateaus near 60% and 50%. Pic2Word trails Text-only, highlighting the value of text.

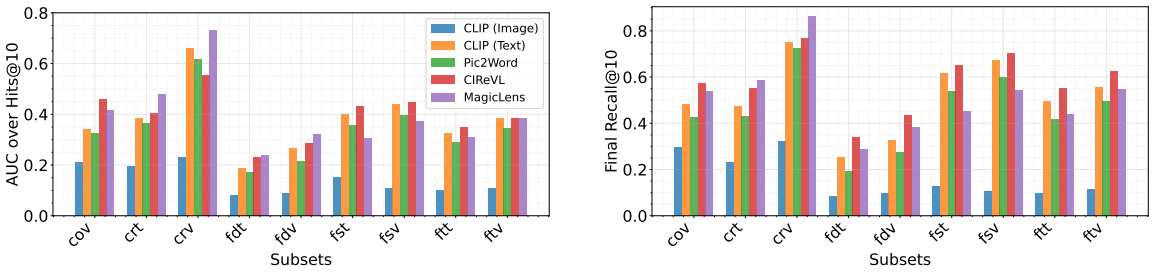
Final Recall@10 and AUC Analysis

Final Recall@10 and AUC results are shown in Fig. 7. CIReVL, Text-only, and MagicLens perform well overall. MagicLens leads on general-domain data, while CIReVL is stronger on fashion. On `cirr_val`, CIReVL beats Text-only on Hits@10 from Turn 2 onward. Its weak first turn lowers AUC, showing that AUC reflects performance across turns. Thus, the preferred method depends on the metric and use case, so we report multiple metrics.



(a) circo_val (b) cirr_val (c) fashioniq_dress_val

Figure 6: Comparison of Hits@10 by turn across subsets.



(a) AUC over Hits@10 (b) Final Recall@10

Figure 7: Final Recall@10 and AUC over Hits@10 across baselines, by subset (codes as in Table 4).

Image-only performs notably worse than Text-only across all subsets, confirming that relative captions carry more discriminative information than reference images alone in multi-turn CIR. The domain-specific performance gap may stem from training data: MagicLens was trained on diverse web-crawled image pairs favoring general scenes, whereas CIReVL's text-based approach better captures fine-grained fashion attributes such as color, pattern, and silhouette.

7 Conclusion

We presented *CIRCLED*, a multi-turn CIR dataset built by extending FashionIQ, CIRR, and CIRCO. It addresses two gaps in prior work: inconsistent dialogue histories and a fashion-only scope. Each turn is designed to steadily approach the ground truth, ensuring coherence at both turn and session levels and enabling study beyond fashion.

We evaluated several baselines and found that the textual modality is the most informative and that the best-performing method depends on the domain. We position CIRCLED as a *controlled benchmark*. Its rank-progression structure holds across the retrieval methods we evaluate, not only the construction baseline. By construction, it is defined relative to an iterative-refinement retrieval algorithm, rather than a simulation of unconstrained human search. CIRCLED provides a practical dataset and an evaluation framework for future research on multi-turn CIR.[6]

6. The license information for the images used in this paper is provided in Appendix L.

Broader Impact Statement

This work provides a standardized benchmark for multi-turn composed image retrieval, contributing to the development of interactive visual search systems. Potential applications include conversational product search in e-commerce and AI agents that progressively explore visual information to fulfill user requests. By offering a consistent, cross-domain benchmark, CIRCLED highlights several design challenges it poses for future CIR methods. These methods must aggregate and model dialogue history across turns, remain robust beyond the fashion domain, and move from a fixed evaluation trajectory toward closed-loop interactive retrieval. In terms of scalability, the baselines we study are training-free. At each turn they perform a single CLIP forward pass followed by nearest-neighbor search over the subset database, requiring no per-dataset fine-tuning, while construction uses a bounded, $O(L)$ number of LLM (GPT-4o-mini) calls per L-turn session. Table 7 summarizes these costs and the main constraints. Our dataset uses images from existing public datasets (FashionIQ, CIRR, CIRCO), thereby avoiding privacy concerns associated with new image collection. However, since the relative captions are generated by an LLM (GPT-4o-mini), they may inherit linguistic biases present in the model, and the captions are English-only. In addition, the evaluation trajectory is fixed rather than interactive. To mitigate the bias risk, we fully disclose the dataset construction process (prompts, filtering criteria, etc.), enabling the research community to verify and address potential biases.

Acknowledgments and Disclosure of Funding

This work was not supported by any specific grant from funding agencies. The authors declare no competing interests.

References

Sher Badshah and Hassan Sajjad. Reference-guided verdict: LLMs-as-judges in automatic evaluation of free-form QA. In *Proceedings of the 9th Widening NLP Workshop*, pages 251–267, 2025.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *CVPR*, pages 21466–21474, 2022.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *ICCV*, pages 15338–15347, 2023.

Tamara L. Berg, Alexander C. Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, pages 663–676, 2010.

Yang Cao, Changhu Wang, Zhiwei Li, Liqing Zhang, and Lei Zhang. Spatial-bag-of-features. In *CVPR*, pages 3352–3359, 2010.

Yanzhe Chen, Zhiwen Yang, Jinglin Xu, and Yuxin Peng. Mai: A multi-turn aggregation-iteration model for composed image retrieval. In *ICLR*, 2025.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only training of zero-shot composed image retrieval. In *CVPR*, pages 13225–13234, 2024.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL `https://arxiv.org/abs/2411.15594`.

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *ICLR*, 2024.

Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. Interactive text-to-image retrieval with large language models: A plug-and-play approach. In *ACL*, 2024.

Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Chatting makes perfect: Chat-based image retrieval. In *NeurIPS*, pages 61437–61449, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742, 2023.

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*, pages 2125–2134, 2021.

Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. Bi-directional training for composed image retrieval via text prompt learning. In *WACV*, pages 5753–5762, 2024.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, pages 1096–1104, 2016.

Xiaopeng Lu, Tiancheng Zhao, and Kyusong Lee. VisualSparta: An embarrassingly simple approach to large-scale text-to-image search with weighted bag-of-words. In *ACL*, pages 5020–5029, 2021.

OpenAI et al. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

A. Pal, S. Wadhwa, A. Jaiswal, X. Zhang, Y. Wu, R. Chada, P. Natarajan, and H. I. Christensen*. Fashionntm: Multi-turn fashion image retrieval via cascaded memory. In *ICCV*, pages 11323–11334, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

Brian Richards. Type/token ratios: What do they really tell us? *Journal of child language*, 14(2):201–209, 1987.

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. In *CVPR*, pages 19305–19314, 2023.

Ryotaro Shimada, Yu-Chieh Lin, Yuji Nozawa, Youyang Ng, Osamu Torii, and Yusuke Matsui. Towards vision-free cir: Attribute-augmented scoring and llm-based reranking for zero-shot composed image retrieval. In *SIGIR Workshop (SynthIR)*, 2026.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 404–430, 2025.

Yuxin Tian, Shawn Newsam, and Kofi Boakye. Fashion image retrieval with text feedback by additive attention compositional learning. In *WACV*, pages 1011–1021, 2023.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in llm-as-a-judge. In *NeurIPS Workshop on Safe Generative AI*, 2024.

Haokun Wen, Xuemeng Song, Haoyu Zhang, Xiangyu Zhao, Weili Guan, and Liqiang Nie. Unicvr: From alignment to reranking for unified zero-shot composed visual retrieval, 2026. URL `https://arxiv.org/abs/2604.20318`.

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. In *CVPR*, pages 11307–11317, 2021.

Yifei Yuan and Wai Lam. Conversational fashion image retrieval via multiturn natural language feedback. In *SIGIR*, pages 839–848, 2021.

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.

Zijia Zhao, Longteng Guo, Tongtian Yue, Erdong Hu, Shuai Shao, Zehuan Yuan, Hua Huang, and Jing Liu. Chatsearch: A dataset and a generative retrieval model for general conversational image retrieval. *Pattern Recognition*, 167:111696, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, volume 36, pages 46595–46623, 2023.

Appendix A. Dataset Documentation

This section provides documentation for the CIRCLED dataset following the recommendations of "Datasheets for Datasets" (Gebru et al., 2021).

A.1 Access

The dataset and code are publicly available:

- **Dataset**: `https://huggingface.co/datasets/tk1441/CIRCLED`
- **Code**: `https://github.com/mti-lab/circled`

The dataset consists of multi-turn retrieval sessions in JSON format. Each session consists of a sequence of (reference image ID, relative caption) pairs, paired with a ground truth image ID. CIRCLED does not distribute the images themselves; it provides only image IDs and annotations (relative captions and session structures). Users must download the original images from FashionIQ (Wu et al., 2021), CIRR (Liu et al., 2021), and CIRCO (Baldrati et al., 2023) separately, following each dataset's terms of use.

A.2 License

The CIRCLED dataset annotations (relative captions and session structures) are released under the CC BY 4.0 license. The underlying images are subject to their original licenses:

- FashionIQ images: Subject to Amazon's terms of use
- CIRR/CIRCO images: Subject to original Flickr licenses (CC BY, CC BY-NC, etc.)

A.3 Hosting and Maintenance Plan

The dataset is hosted on Hugging Face Datasets, which provides long-term, reliable hosting with version control. The code repository is maintained on GitHub. We commit to maintaining the dataset for at least five years and will respond to issues and pull requests. Any updates or corrections will be versioned and documented in the repository.

A.4 Author Responsibility Statement

The authors confirm that:

- We bear all responsibility in case of violation of rights.
- The dataset annotations do not contain personally identifiable information.
- The relative captions were generated using GPT-4o-mini and reviewed through automated filtering to ensure quality and appropriateness.

A.5 Image Provenance

All images in CIRCLED originate from three existing, publicly available CIR datasets. These are FashionIQ (Wu et al., 2021) (fashion product images), CIRR (Liu et al., 2021)

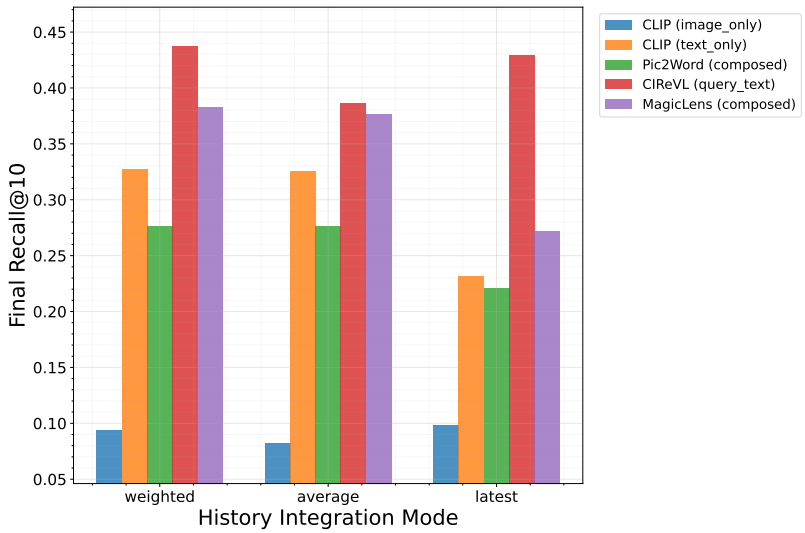


Figure 8: Example results for history-integration strategies. On the dataset extended from fashioniq_dress_val, we compare how different aggregation schemes affect each baseline's performance at each turn.

(real-life images from NLVR2), and CIRCO (Baldrati et al., 2023) (COCO images). We do not collect any new images. CIRCLED contributes the multi-turn session structure and the LLM-generated relative captions on top of these sources. Per-image license information and, where required, generated stand-in images for figures are detailed in Appendices E and L, and each session records the source dataset and original image identifiers so that provenance can be traced back to the source datasets.

A.6 Ethical Considerations

Because CIRCLED reuses images from established public datasets and adds no new image collection, it introduces no new privacy concerns beyond those of the source datasets, and the annotations contain no personally identifiable information. The relative captions are produced by an LLM (GPT-4o-mini) and may reflect linguistic or cultural biases present in that model. We therefore fully release the construction prompts and filtering criteria so that such biases can be audited and improved by the community. As with any interactive retrieval resource, downstream systems trained or evaluated on CIRCLED (e.g., e-commerce search) should be checked for fairness across user groups and product categories. Use of the dataset is subject to the terms of the underlying source datasets.

Appendix B. Effect of History Integration Methods

We evaluate three ways of aggregating single-turn CIR features into a multi-turn method (Latest, Average, and Weighted) and report the results in terms of Final Recall@10. As shown in Fig. 8, on the dataset extended from fashioniq_dress_val, the Weighted mode (which places larger weights on later turns) achieves the highest performance. This aligns

with the design of our dataset, where the query progressively approaches the target image in later turns, making that information particularly valuable.

Moreover, Weighted outperforms Latest (which uses only the final turn), indicating that leveraging the entire history yields more accurate retrieval than relying solely on the last turn. An exception arises for image-only features, where Latest is slightly better, suggesting that the most recently selected reference image carries the most salient visual signal in that setting.

Appendix C. Prompts

We list the prompts used in this work. All prompts are fed into GPT-4o-mini.

C.1 Caption Generation

The prompts for generating image captions vary by dataset domain.

FashionIQ Dress:

```
Describe this dress in 1-2 sentences. Focus on color, style, length,
 and key design features.
```

FashionIQ Shirt:

```
Describe this shirt in 1-2 sentences. Focus on color, style, collar,
 sleeves, and key design features.
```

FashionIQ Toptee:

```
Describe this top in 1-2 sentences. Focus on color, style, neckline,
 sleeves, and key design features.
```

CIRR:

```
Describe this image in 1-2 sentences. Include main objects, people,
setting, and notable features.
```

CIRCO:

```
Describe this image in 1-2 sentences. Include main objects,
composition, and notable features.
```

C.2 Relative Caption Generation

The prompt for generating relative captions includes history of previous suggestions to ensure diversity. The retry mechanism is activated when the generated caption is too similar to existing ones.

```
IMPORTANT -- Previous changes have already been suggested:
- "{relative caption1}"
- "{relative caption2}"
...
Your task is to identify a COMPLETELY DIFFERENT visual change.
Focus on aspects that have NOT been mentioned before.
```

```
[Only if retrying]
RETRY #{n}: The previous suggestion was too similar to existing ones
.
Please provide a MORE DISTINCTIVE and DIFFERENT instruction.
Consider completely different visual aspects like:
- Different objects or people
- Different colors or lighting
- Different actions or poses
- Different background elements
- Different clothing or accessories

You will see two images.
**Image 1**: This is the REFERENCE image that needs to be modified.
**Image 2**: This is the TARGET image showing the desired result.

Write exactly ONE imperative instruction to transform Image 1 into
Image 2.
Requirements:
1. Start with a verb
2. Be extremely specific about colors, positions, or actions
3. Avoid relative terms like "left" or "right"
4. Do not use quotes or explanatory text
5. Focus on a single, clear change
6. Must be different from previous suggestions
```

C.3 Auxiliary Caption Generation

The prompt for merging the reference image caption with the relative caption to create an auxiliary caption for retrieval.

```
A user is performing image retrieval. The user provides a reference
image
caption and a modification for the retrieved image to refine the
search.
Generate a new query reflecting this modification. Only return the
refined query.

Reference image caption: {reference_image_caption}
Modification: {relative_text}
```

Appendix D. LLM-as-a-Judge Prompt for Session Evaluation

We use GPT-5-mini as an LLM-based judge to evaluate multi-turn retrieval sessions. The system prompt is shown below.

```
You are an expert annotator for multi-turn image retrieval dialogs.

You will see a complete retrieval session:
```

```
- A sequence of (CURRENT image, USER UTTERANCE) pairs showing the
dialog progression
- A TARGET image (the final goal Z)

Your job is to evaluate the ENTIRE SESSION as a coherent dialog for
image retrieval.

Rate the session on these 5 dimensions, each from 1 (very bad) to 5
(excellent):
1. SESSION NATURALNESS: Are the utterances consistently human-like
throughout the session?
2. COHERENCE / CONSISTENCY: Is there logical flow without
contradictions or abrupt changes?
3. GOAL-DIRECTED PROGRESS: Does each turn move closer to the target
image Z?
4. REDUNDANCY (higher = better): Is there low repetition of
information?
5. OVERALL: Holistic quality of the entire session
```

Appendix E. Image Generation for FashionIQ Data

For licensing reasons, in this paper we use generated images in the figures about FashionIQ dataset. The images are generated by GPT-4o image generation function, using text prompts that describe the images in FashionIQ dataset.

Appendix F. Scalability and Constraints

Table 7 summarizes the computational cost and the main constraints of CIRCLED, supporting the discussion in the Broader Impact statement.

Appendix G. Construction-Model Bias Analysis

A concern is that CIRCLED's rank progression might be specific to the construction baseline (CIReVL with a BLIP encoder), favoring its embedding geometry. We assess whether the progression holds under a retrieval setup that does *not* use the construction encoder. For every session we compute, under several models, the ε-consistency satisfaction rate (fraction of sessions with $r_{l+1} \leq r_l + \varepsilon$ at all turns, $\varepsilon = 30$) and the median ground-truth rank per turn. All quantities are derived from the baseline evaluations already reported in Sec. 6 and require no additional retrieval. Crucially, we include CIReVL with a CLIP encoder (the same construction algorithm with a *different* encoder than the BLIP used for construction).

Table 8 reports the result under a fixed CLIP encoder. CIReVL attains 92.6%, clearly exceeding the non-CIReVL methods (80–83%) and Image-only (63.8%). Since CLIP is not the encoder used during construction, this shows that the rank progression is governed by the iterative query-update algorithm and holds under a different embedding space, rather than being tied to the construction encoder.

Fig. 9 shows the per-turn median rank. The Turn-1→2 improvement is universal. Across all text-using models, the median rank drops from several hundred to single or low-double

Table 7: Scalability characteristics and constraints of CIRCLED.

Aspect	Summary
Construction cost	Bounded LLM (GPT-4o-mini) calls per session, namely one caption (f_{caption}), one merge (f_{merge}), and one difference per added turn (f_{diff}), i.e., $O(L)$ calls for an L-turn session.
Query / evaluation cost	Training-free. Each turn is a single CLIP forward pass plus nearest-neighbor search over the subset database, with no per-dataset fine-tuning, and the history aggregation adds no learnable parameters.
Search-space scale	Retrieval scales with the per-subset database size (Table 1) using standard approximate nearest-neighbor indexing.
Constraints	Captions are English and LLM-generated, so they may carry linguistic bias. The evaluation trajectory is fixed and non-interactive. Some fashion source images are subject to link rot and the source datasets' terms of use.

Table 8: ε-consistency satisfaction rate under a fixed CLIP encoder (*not* the BLIP encoder used for construction), over all 22,608 CIRCLED sessions ($\varepsilon = 30$). CIReVL clearly exceeds the non-CIReVL methods, showing the progression is driven by the iterative query-update algorithm and holds under a different encoder.

Model (CLIP encoder)	ε-consistency
CIReVL	92.6%
Text-only	82.9%
MagicLens	82.2%
Pic2Word	80.5%
Image-only	63.8%

digits. Session length is determined by how many turns the construction baseline required, so longer sessions are intrinsically harder. The Turn-1 median rank generally grows with length (e.g., Text-only medians of 225, 514, 688, 650, and 951 for 2-, 3-, 4-, 5-, and 6-turn sessions). On these harder long sessions, non-CIReVL models improve at Turn 2 and then maintain the rank (e.g., for 5-turn sessions, Text-only goes $650 \to 30 \to 40 \to 52 \to 46$) rather than collapse, and over 80% remain ε-consistent. Image-only retrieval shows little progression at any length, confirming that the textual modification drives the progression. We therefore position CIRCLED as a controlled benchmark whose progression structure holds across the retrieval methods we evaluate, not only the construction baseline.

Appendix H. Human Evaluation of Session Quality

To validate the LLM-as-a-judge, we conducted a human study on a 50-session subset of those it evaluated, using the same five dimensions and scoring rubric. The annotation was carried

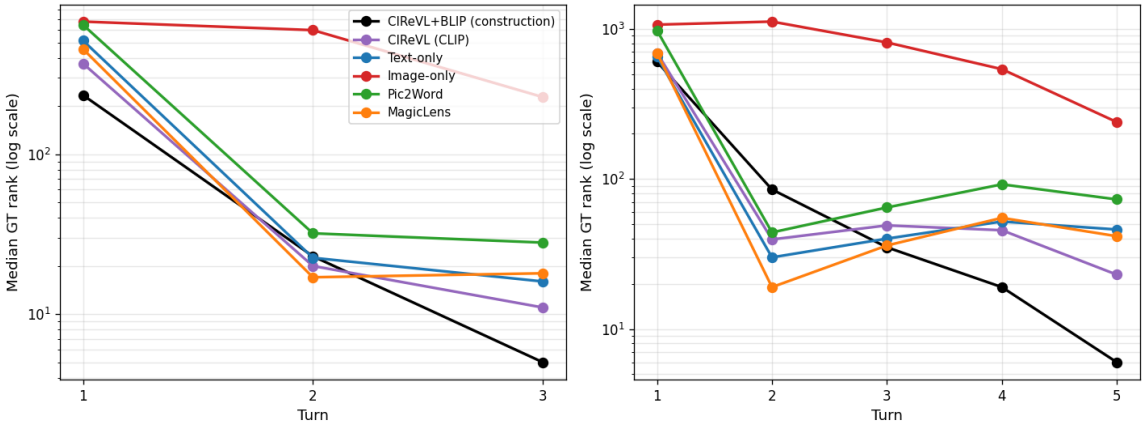


Figure 9: Median ground-truth rank by turn under different retrieval models (log scale, pooled over subsets), for 3-turn (left, $N = 5{,}424$, 24.0% of data) and 5-turn (right, $N = 772$, 3.4%) sessions. All text-using models improve sharply at Turn 2. CIReVL (both encoders) continues to improve, while non-CIReVL methods maintain rank on the harder long sessions. Image-only shows little progression.

out by five researchers across two institutions (four male and one female), all independent of both the GPT-4o-mini used for generation and the GPT-5-mini used for judging. They had seen CIRCLED sessions during dataset construction, so we blinded the study to keep this familiarity from favoring CIRCLED. The 50 sessions were sampled with a fixed seed, balanced across the four sources (13/13/12/12), and shuffled, and each appeared under an anonymous identifier with the source method and the judge's ratings withheld from the annotators' kit. Matching the judge's zero-shot setting, annotators saw only the scoring rubric without worked examples, and they rated separately without discussing individual sessions. Table 9 reports the per-source means averaged over annotators. CIRCLED ranks first on all five dimensions. At the session level, the human ratings agree with the LLM judge (Spearman $\rho = 0.65$ on Overall, over the 50 sessions). Notably, the annotators reproduce the same source ranking as GPT-5-mini exactly (Spearman $\rho = 1.00$ over the four source means), with strong concordance among annotators (Kendall's $W = 0.81$) and a significant Friedman test ($\chi^2 = 12.1$, $p < 0.01$). Thus human annotators reproduce the judge's ranking at both the session and source levels, indicating the conclusions are not an artifact of self-preference between the generator and the judge.

Appendix I. Sensitivity to the Filtering Thresholds

We analyze the dataset's sensitivity to the two filtering thresholds by re-running the full filtering pipeline while varying one threshold and fixing the other (the adopted $\varepsilon = 30$, $\tau = 0.8$ reproduces the 22,608 sessions of Table 1). As shown in Table 11, the rank-margin threshold ε has a negligible effect ($\pm 0.6\%$ over $\varepsilon \in [10, 100]$). The dataset is robust to its choice, consistent with the rank-margin filter removing only 562 sessions overall. In contrast, the diversity threshold τ governs dataset size, changing it $3.6\times$ over $\tau \in [0.7, 0.9]$.

Table 9: Human evaluation of session quality (five annotators, 50 blind sessions, same rubric as the LLM judge). Mean rating per source (1–5, higher is better), best per column in bold. Human ratings agree with GPT-5-mini at the session level (Spearman $\rho = 0.65$ on Overall) and reproduce the same source ranking exactly (Spearman $\rho = 1.00$ over the four source means). Sources differ significantly (Friedman $p < 0.01$).

Source	Natural.	Coher.	Goal Prog.	Non-redund.	Overall
CIRCLED (ours)	**4.25**	**4.15**	**4.40**	**4.45**	**4.22**
w/o quality filter	4.12	4.05	4.05	4.09	3.97
Multi-turn FashionIQ	3.57	3.62	3.72	4.23	3.65
Simple Concat	3.07	3.03	3.10	3.65	3.05

Table 10: Statistical significance of CIRCLED's advantage over each comparison source (one-sided Mann-Whitney U test on the 1,711 session ratings of Table 6). We report p-values and rank-biserial effect sizes r (+ favors CIRCLED). Here, n.s. denotes not significant.

Comparison	Metric	p	r
vs. w/o quality filter	Overall	1.6×10^{-5}	+0.14
	Coherence	1.0×10^{-4}	+0.13
	Goal Progress	5.7×10^{-9}	+0.19
	Redundancy	3.5×10^{-6}	+0.08
	Naturalness	n.s.	−0.09
vs. Multi-turn FashionIQ	Overall	1.5×10^{-12}	+0.26
	Coherence	9.3×10^{-5}	+0.14
	Goal Progress	5.5×10^{-17}	+0.31
	Redundancy	6.3×10^{-9}	+0.11
	Naturalness	9.1×10^{-11}	+0.24
vs. Simple Concat	Overall	1.4×10^{-58}	+0.58
	Coherence	5.2×10^{-44}	+0.51
	Goal Progress	9.0×10^{-64}	+0.61
	Redundancy	1.4×10^{-22}	+0.23
	Naturalness	5.0×10^{-19}	+0.31

It controls the trade-off between scale and per-turn novelty. Our choice ($\varepsilon = 30$, $\tau = 0.8$) lies at a balanced operating point.

Appendix J. Choice of the Construction Generator (GPT-4o-mini)

CIRCLED's construction uses GPT-4o-mini for the caption, merge, and difference generators ($f_{\text{caption}}, f_{\text{merge}}, f_{\text{diff}}$). The LLM judge in Sec. 5.3 is a different model, GPT-5-mini. To justify this choice we compare, on identical images, the captions produced by GPT-4o-mini and two open alternatives (BLIP2 and LLaVA). Verified against the source images, GPT-4o-mini accurately reports the fine-grained attributes relevant to composed retrieval

Table 11: Sensitivity of the final dataset size to the filtering thresholds, varying one while fixing the other. The adopted setting ($\varepsilon = 30$, $\tau = 0.8$) reproduces the 22,608 sessions of Table 1. ε has little effect, while τ governs dataset size.

ε (with $\tau = 0.8$)		τ (with $\varepsilon = 30$)	
ε	#sessions	τ	#sessions
10	22,485	0.70	7,932
20	22,561	0.75	14,066
30	**22,608**	**0.80**	**22,608**
50	22,649	0.85	26,507
100	22,677	0.90	28,380



(a) General (COCO `000000148864`)



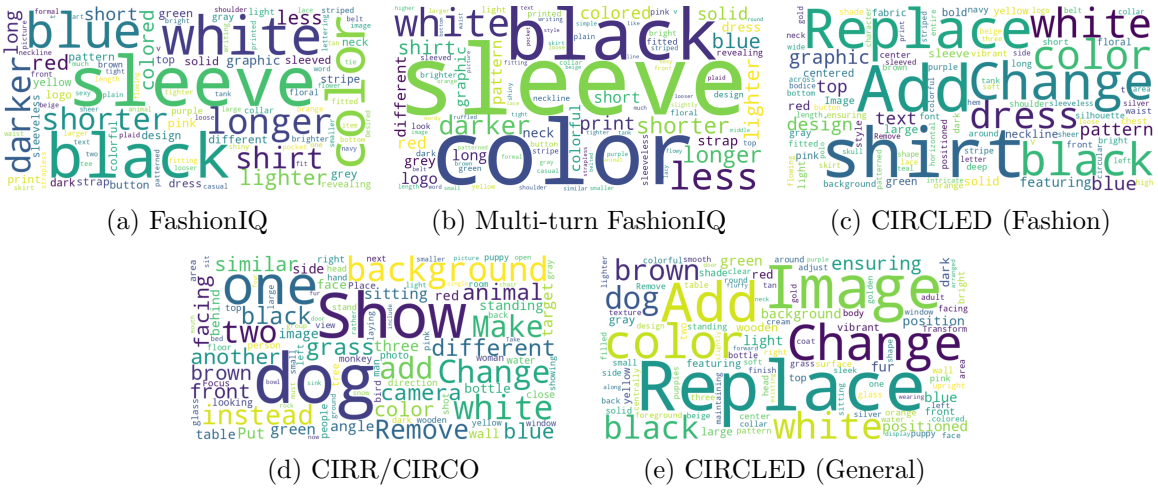
(b) Fashion (FashionIQ `B00CBBV3ZC`, generated)

Figure 10: Source images for the caption comparison in Table 12, both already used in this paper (Fig. 5). The fashion image is the generated stand-in used in this paper (Appendix E, with the original subject to the source dataset's terms). The COCO image is the original (CC BY-NC).

in both the general and fashion domains (Fig. 10, Table 12). For the rainbow cake it gives the correct layer-color order and setting, and for the burgundy dress it correctly separates the floral bodice from the geometric skirt. By contrast, BLIP2 frequently hallucinates a different scene, and LLaVA produces short, subjective captions that, for fashion, often describe the model rather than the garment. In addition, only an instruction-following model can serve $f_{\text{merge}}/f_{\text{diff}}$, which require text integration and difference description rather than plain captioning. Combined with its low cost and wide availability, these properties make GPT-4o-mini a suitable construction generator.

Table 12: Captions generated by GPT-4o-mini and two alternatives (BLIP2, LLaVA) on the images in Fig. 10 (one general-domain, one fashion), already used in this paper. GPT-4o-mini accurately captures the fine-grained, retrieval-relevant attributes. BLIP2 hallucinates a different scene, and LLaVA gives short or subjective captions. Verified against the source images.

General domain, sliced rainbow cake (COCO `000000148864`, Fig. 10)	
GPT-4o-mini	"A vibrant rainbow cake with distinct layers of red, orange, yellow, green, blue, and purple, covered in white frosting and sprinkles, on a silver cake board with a festive table setting." *(accurate)*
BLIP2	"the front of a train station with some people waiting for trains" *(hallucinated)*
LLaVA	"A rainbow cake celebration awaits!" *(correct topic, no detail)*
Fashion domain, burgundy dress (FashionIQ `B00CBBV3ZC`, Fig. 10)	
GPT-4o-mini	"Burgundy dress with a floral pattern on the upper half and a geometric design on the skirt, with fitted three-quarter sleeves and a mid-thigh length." *(accurate)*
BLIP2	"white dress with red and blue stripes, black stars and butterflies" *(hallucinated)*
LLaVA	"woman posed against a red background, blonde hair, glamour" *(describes the model, not the garment)*



(a) FashionIQ (b) Multi-turn FashionIQ (c) CIRCLED (Fashion)

(d) CIRR/CIRCO (e) CIRCLED (General)

Figure 11: Word clouds of relative captions. Top row: Fashion domain (FashionIQ, Multi-turn FashionIQ, CIRCLED). Bottom row: General domain (CIRR/CIRCO, CIRCLED). FashionIQ and CIRR/CIRCO are single-turn CIR datasets, Multi-turn FashionIQ is a multi-turn dataset, and CIRCLED extends single-turn datasets to multi-turn.

Appendix K. Word Clouds of Relative Captions

Fig. 11 complements the linguistic metrics of Table 3 with a qualitative view of the vocabulary used in each dataset. The clouds are computed with the default stopword list of the Python WordCloud library.

Figure	Images	Image ID
Fig. 1	Turn1 reference image, search result	000000190235, 000000037054
Fig. 1	Turn2 reference image, search result	000000037054, 000000537111
Fig. 1	TurnN reference image, search result	000000537111, 000000138205
Fig. 1	Ground truth	000000138205
Fig. 2 (b)	Reference image 1, 2, 3, ground truth	B004EHON6W, B004XJI4I4, B008D5AYOG, B0093K54X6
Fig. 2 (a)	Reference image 1, 2, 3, ground truth	B007XLHLHE, B008I5Q3CS, B007IXDC7U, B007IXDESW
Fig. 3 (a)	Reference image, search results (from left)	000000112439, 000000518636, 000000324422, 000000264963
Fig. 3 (b)	Non-selected images (from top), selected image	000000518636, 000000324422, 000000264963
Fig. 3 (b)	Search results (from top)	000000460329, 000000346912, 000000247950, 000000264963
Fig. 4 (a)	Reference image 1, 2, 3, 4, ground truth	000000253571, 000000219002, 000000448025, 000000332048, 000000166818
Fig. 4 (b)	Reference image 1, 2, 3, 4, ground truth	000000447096, 000000271037, 000000230153, 000000195608, 000000478514
Fig. 5 (a)	Reference image 1, 2, ground truth	000000179795, 000000318881, 000000148864
Fig. 5 (b)	Reference image 1, 2, 3, ground truth	B008ZB39FY, B00CBBV3ZC, B0091S4Q9S, B008HQYFB4
Fig. 10	General image, fashion image (reused from Fig. 5)	000000148864, B00CBBV3ZC

Table 13: List of Image IDs used in Figures

Appendix L. Image Licenses

The licenses of the images used in this paper are summarized in Tables 13 and 14.

Table 14: Image ID, URLs, and Licenses. If an image is noted "generated image" in its License column, it is generated by gpt-4o image generation function, using a text prompt that describes the image on the URL noted in its Image URL column.

Image ID	Image URL	License
000000190235	http://farm7.staticflickr.com/6171/6207331658_b318513022_z.jpg	CC BY 2.0
000000037054	http://farm7.staticflickr.com/6007/6198289264_029a6e88e2_z.jpg	CC BY 2.0
000000537111	http://farm9.staticflickr.com/8064/8211635300_deba3583bf_z.jpg	CC BY 2.0
000000138205	http://farm9.staticflickr.com/8175/8071171348_de3c9af840_z.jpg	CC BY-NC-SA 2.0
000000112439	http://farm2.staticflickr.com/1137/962695681_0be4bcd0f8_z.jpg	CC BY-NC-SA 2.0
000000518636	http://farm4.staticflickr.com/3460/4018979209_8dc1cf8ffd_z.jpg	CC BY-NC-SA 2.0
000000324422	http://farm4.staticflickr.com/3244/2655366225_5b0754fb6e_z.jpg	CC BY-NC-ND 2.0
000000264963	http://farm3.staticflickr.com/2334/2383999861_75d27c265e_z.jpg	CC BY-NC-SA 2.0
000000460329	http://farm4.staticflickr.com/3182/3025936416_f1e5421a1b_z.jpg	CC BY-NC-SA 2.0
000000346912	http://farm4.staticflickr.com/3323/3513955326_d9801f2b83_z.jpg	CC BY 2.0
000000247950	http://farm5.staticflickr.com/4136/4743534549_c7a3612428_z.jpg	CC BY-NC-SA 2.0
B004EHON6W	http://ecx.images-amazon.com/images/I/41DXkZHTOjL.SX342.jpg	generated image
B0093K54X6	http://ecx.images-amazon.com/images/I/31r71PVOUEL.SX342.jpg	generated image
B004XJI4I4	http://ecx.images-amazon.com/images/I/31VgYvG6qbL.SX342.jpg	generated image
B008D5AYOG	http://ecx.images-amazon.com/images/I/416DhASM95L.SX342.jpg	generated image
B007XLHLHE	http://ecx.images-amazon.com/images/I/41LQcUsGHXL.SY445.jpg	generated image
B008I5Q3CS	http://ecx.images-amazon.com/images/I/41rprim9AFL.SX342.jpg	generated image
B007IXDC7U	http://ecx.images-amazon.com/images/I/41UxaR--QlL.SY445.jpg	generated image
B007IXDESW	http://ecx.images-amazon.com/images/I/31BrgjaaIZL.SY445.jpg	generated image
000000253571	http://farm1.staticflickr.com/228/471397100_afd0fe517a_z.jpg	CC BY-NC 2.0
000000219002	http://farm7.staticflickr.com/6038/6215343038_6bae3b45cb_z.jpg	CC BY-NC-SA 2.0
000000448025	http://farm2.staticflickr.com/1019/1035634033_503438c7ea_z.jpg	CC BY-NC-ND 2.0
000000332048	http://farm3.staticflickr.com/2320/2073978153_7d320747e4_z.jpg	CC BY-NC 2.0
000000166818	http://farm3.staticflickr.com/2054/2247395963_7cb97cbf8d_z.jpg	CC BY-NC-SA 2.0
000000447096	http://farm6.staticflickr.com/5001/5255997294_7df1664c69_z.jpg	CC BY-NC-SA 2.0
000000271037	http://farm1.staticflickr.com/62/195329469_b95cf37cc0_z.jpg	CC BY 2.0
000000230153	http://farm3.staticflickr.com/2405/2005336437_2e7b5da6db_z.jpg	CC BY-NC-SA 2.0
000000195608	http://farm9.staticflickr.com/8302/7782408182_25d279cd27_z.jpg	CC BY 2.0
000000478514	http://farm8.staticflickr.com/7252/7782273360_221f840574_z.jpg	CC BY 2.0
000000179795	http://farm3.staticflickr.com/2737/4114913419_2736255889_z.jpg	CC BY-NC-SA 2.0
000000148864	http://farm8.staticflickr.com/7080/7404128930_7b4c76a4e2_z.jpg	CC BY-NC 2.0
000000318881	http://farm6.staticflickr.com/5528/10038641393_8f1ce796d0_z.jpg	CC BY-NC-ND 2.0
B008ZB39FY	http://ecx.images-amazon.com/images/I/41YqRey8zgL._SX342_.jpg	generated image
B008HQYFB4	http://ecx.images-amazon.com/images/I/31b%2BheNk5nL._SX342_.jpg	generated image
B00CBBV3ZC	http://g-ecx.images-amazon.com/images/G/01/x-locale/brands/lifestyle-assets/5370858011._CB335605943_SR150,160_.jpg	generated image
B0091S4Q9S	http://ecx.images-amazon.com/images/I/31kZNo52IkL._SX342_.jpg	generated image