AutoConcept: Training-Free Concept-Guided Reranking for Metadata-Available Composed Image Retrieval

Tianyu Wang[1]✉ and Tianjiao Wu[2]

[1] School of Computer Science and Technology, Soochow University, Suzhou, China
tywangcs@foxmail.com
[2] INSTITUT NATIONAL DES SCIENCES APPLIQUEES DE LYON, Lyon, France
tianjiao.wu@insa-lyon.fr

Abstract. Composed image retrieval (CIR) retrieves a target image from a reference image and a text modification. This paper studies metadata-available CIR reranking, where a fixed CIR model first returns a candidate pool and gallery metadata is then used for second-stage concept-guided scoring. We introduce AutoConcept, a training-free reranker that converts concept evidence into an interpretable memory. AutoConcept filters noisy concepts, activates query-relevant positive constraints with an auxiliary negative penalty, and combines base retrieval scores with metadata-based concept-candidate alignment through inference-time calibration. On FashionIQ, AutoConcept yields significant early-rank improvements over WeiMoCIR and consistent plug-in gains on LinCIR candidate pools. Metadata-aware controls show that structured concept memory adds signal beyond direct query-text and extracted-attribute matching, while a query-only variant further supports the effectiveness of concept-level reranking. A supplementary real-human concept-label study indicates that the same memory interface can consume participant-provided evidence. These results position AutoConcept as an interpretable concept-memory reranker for product-style CIR galleries with available metadata.

Keywords: Composed image retrieval · Metadata-available reranking · Concept memory

1 Introduction

Composed image retrieval (CIR) asks a system to retrieve a target image given a reference image and a natural-language modification. In fashion and e-commerce retrieval, candidate items are commonly associated with metadata such as titles, attributes, tags, or descriptions. Users may also express concept-level constraints, such as preferred colors, sleeve length, neckline, material, or styles to avoid. Most CIR systems construct one composed query representation and rank the gallery by similarity to that query [28, 9, 26, 6, 22]. A complementary second-stage layer that exposes and applies such concepts can make reranking more controllable and easier to inspect.

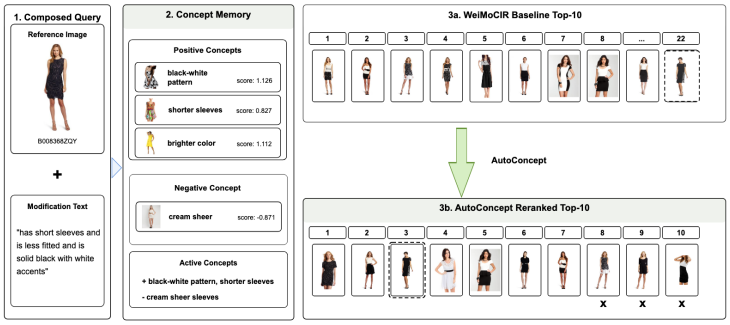

Fig. 1. AutoConcept in action. Given the same composed query and first-stage candidate pool, the explicit concept memory promotes candidates that match query-relevant positive constraints and penalizes candidates matching negative constraints.

This work studies metadata-available concept-guided CIR reranking. A fixed first-stage retriever returns candidate images, and gallery metadata is available only after this candidate pool has been formed. AutoConcept addresses this setting by converting controlled concept evidence into an interpretable memory that guides candidate ordering without training. The central question is whether explicit concept evidence can improve a strong CIR candidate pool beyond direct query-to-metadata or attribute-to-metadata matching.

We propose AutoConcept, a training-free concept-guided reranker. AutoConcept keeps the first-stage CIR model fixed and operates on its top-K candidate pool. It constructs positive concept constraints from controlled evidence, retains a lightweight negative constraint penalty, filters noisy concepts, activates query-relevant concepts with a relevance gate, and combines base retrieval scores with metadata-based concept-candidate alignment. The result is a modular reranking layer that can be attached to different CIR candidate generators.

Our main experiments use FashionIQ with five controlled concept-evidence seeds. AutoConcept improves WeiMoCIR [29] from 0.1125 to 0.1379 Recall@10 and LinCIR candidates [8] from 0.2605 to 0.3009 under the same metadata-available reranking protocol. Metadata-aware controls locate the contribution relative to direct query-to-metadata, extracted-attribute, and composed-text reranking; an aligned-vs-shuffled analysis shows that coherent query-evidence correspondence matters. We further collect a real-human concept-label dataset to test whether the same memory interface can consume participant-provided item-level concept evidence.

The contributions are:

1. We formulate metadata-available CIR reranking with an explicit concept-memory layer over fixed first-stage candidate pools.
2. We give an explicit scoring formulation with positive concept constraints, an auxiliary negative penalty, query-concept relevance gating, and closed-form inference-time score calibration.

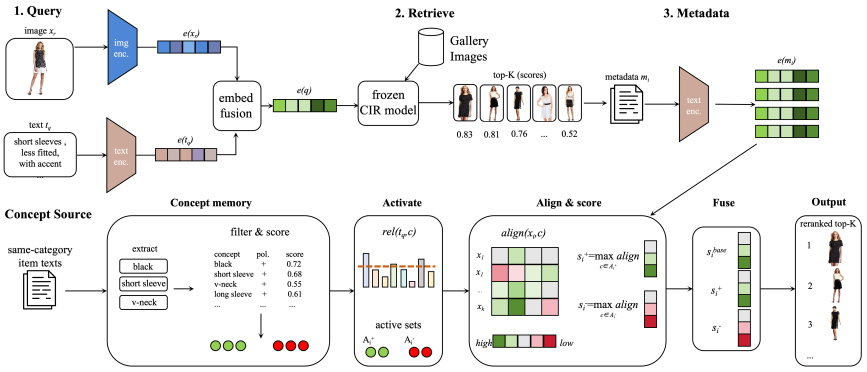

Fig. 2. Overview of the AutoConcept framework. AutoConcept constructs an explicit concept memory from controlled concept evidence, activates query-relevant concepts, and reranks the top-K candidate pool with inference-time score calibration.

3. We evaluate AutoConcept alongside metadata-aware baselines, including query-only evidence, query-to-metadata matching, composed-text metadata matching, and extracted-attribute matching.
4. We show plug-in gains on both WeiMoCIR and LinCIR candidate pools, with aligned-vs-shuffled concept-evidence analysis and a real-human concept-label study.

Figure 2 summarizes the full AutoConcept pipeline. The base CIR model first produces a top-K candidate pool, while controlled concept evidence is converted into a filtered concept memory. At inference time, query-relevant concepts are activated and used to adaptively rerank the candidate pool.

2 Related Work

2.1 Composed Image Retrieval

CIR methods retrieve target images using both a reference image and text feedback. Early systems learn image-text composition modules or attention-based fusion over fashion datasets and real-image benchmarks [28, 9, 26, 6, 21, 3]. Recent zero-shot, training-free, and language-only CIR systems build on pretrained vision-language embeddings such as CLIP, learn lightweight language-side mappings, or use textual inversion and candidate verification [22, 4, 5, 24, 3, 1, 8, 25, 29]. AutoConcept is orthogonal to these retrieval backbones: it adds an explicit concept-memory layer over candidate pools produced by existing CIR methods.

2.2 Training-Free and Plug-in Reranking

Zero-shot and low-training CIR systems reduce training cost by weighted modality fusion, score recombination, textual inversion, language-only training, or candidate verification [24, 3, 1, 8, 25, 29]. This modular view is also related to retrieval-augmented systems, where a frozen generator or scorer is augmented

with external evidence [14, 18, 12]. We use WeiMoCIR as the main training-free CIR baseline and keep the same modular design: the base retriever supplies the candidate pool, and concept memory adjusts the order within that pool, making plug-in evaluation possible. Constraint-based rerankers such as SoFT use MLLMs to derive prescriptive and proscriptive constraints [13]; AutoConcept targets a lighter setting with no per-query MLLM calls, no reranker training, and explicit concept traces.

2.3 Concept Evidence and Controllable Reranking

Reranking is a common way to incorporate external evidence after a strong candidate generator, especially when the control signal should remain inspectable. AutoConcept follows this explicit-control direction for CIR and is related to concept-based interpretability methods such as concept activation vectors, automatic concept explanations, concept bottlenecks, prototype networks, basis decomposition, and concept whitening [15, 7, 16]. It is also related to retrieval and recommendation systems that use memory, profiles, or explicit editable attributes as additional evidence [17, 23, 11, 27, 2, 10]. In AutoConcept, concepts have names, polarity, strength, grounding scores, and evidence items, so the memory can be inspected and diagnosed.

3 Method

3.1 Task Setup

Each query consists of a reference image x_r, modification text t_q, and a target item y in a gallery. A base retrieval model returns a ranked candidate pool $C_q = \{x_i\}_{i=1}^{K}$. In the main FashionIQ experiments, $K = 100$. AutoConcept reranks only this candidate pool.

Each query is associated with a controlled concept-evidence set E_q. AutoConcept converts this evidence into positive concept constraints P_q and a small auxiliary set of negative concept constraints N_q. A concept stores a normalized name, polarity, score, grounding score, and evidence count.

3.2 Information Source Protocol

AutoConcept uses information available under the metadata-available reranking protocol. For each evaluation query, concept evidence is constructed before reranking and excludes the ground-truth target identity, target image, target metadata, candidate ranks, and post-retrieval relevance labels. The main controlled evidence samples same-category training-side item texts and converts them into concept constraints with a fixed extraction and filtering pipeline. Candidate metadata is used only at reranking time to score items already returned by the first-stage retriever. We additionally evaluate a query-only evidence variant, where concepts are extracted from the modification text t_q.

Table 1. Information sources in the FashionIQ metadata-available protocol.

Source	Used?	Purpose
Reference image x_r	Yes	first-stage CIR query
Modification text t_q	Yes	query encoding and concept activation
Training-side evidence item text	Yes	controlled concept-evidence construction
Candidate metadata m_i	Yes	second-stage candidate-concept alignment
Target image / target metadata	No	excluded from evidence construction and scoring
Ground-truth target ID	No	excluded from scoring and evidence construction
Candidate rank / relevance label	No	excluded from evidence construction

3.3 Concept Memory Construction

AutoConcept extracts short attribute phrases from concept-evidence item text, embeds phrases with a sentence encoder, clusters near-duplicate phrases, and scores concepts with positive or negative evidence. A compact set of quality-control constraints removes generic names, polarity conflicts, poorly grounded concepts, and weak concepts. The filter keeps frequent but meaningful apparel attributes such as sleeves, collar, buttons, stripes, and V-neck.

3.4 Query-Relevant Concept Activation

For a query q and concept c, query-concept relevance is

$$r(q, c) = \cos(e(t_q), e(c)), \tag{1}$$

where $e(\cdot)$ is the sentence-encoder embedding and c is the concept name. A query-relevance gate activates the concept set:

$$A_q = \{c : r(q, c) \geq \tau\}. \tag{2}$$

AutoConcept can also compute a query-specific gate for inference-time calibration:

$$\tau_q = \text{clip}(\mu(R_q) + \kappa\sigma(R_q), \tau_{min}, \tau_{max}), \tag{3}$$

where $R_q = \{r(q, c) : c \in P_q \cup N_q\}$. The clipping range is fixed before evaluation and shared across runs. Appendix A.1 reports sensitivity around the operating region.

3.5 Candidate-Concept Alignment

For active concept c and candidate item x_i, AutoConcept computes

$$\text{align}(x_i, c) = \cos(e(m_i), e(c)), \tag{4}$$

where m_i is candidate item metadata text. In our experiments, this text comes from the gallery item metadata used uniformly for candidate items; it is not generated from the test target image and does not use the ground-truth target

identity during scoring. This defines a metadata-available reranking protocol; Section 4.3 compares AutoConcept with a direct metadata-text reranker under the same information setting.

Positive concept scores and the auxiliary disliked-concept penalty are aggregated by max pooling:

$$s_q^+(x_i) = \max_{c \in A_q^+} \text{align}(x_i, c), \quad s_q^-(x_i) = \max_{c \in A_q^-} \text{align}(x_i, c), \tag{5}$$

with empty active sets assigned zero contribution. The concept score is

$$s_q^{concept}(x_i) = w_q(q) s_q^{base}(x_i) + w_p(q) s_q^+(x_i) - w_n(q) s_q^-(x_i). \tag{6}$$

The weights are set at inference time from query-specific concept evidence. Let

$$a_q^+ = \max A_q^+, \quad a_q^- = \max A_q^-, \quad a_q = \max(a_q^+, a_q^-), \tag{7}$$

where A_q^+ and A_q^- are the gated query-concept activation scores, and the maximum over an empty active set is zero. We also compute base-retriever confidence from the top-score margin:

$$b_q = \text{clip}\left(\frac{s_{(1)}^{base} - s_{(2)}^{base}}{\delta}, 0, 1\right). \tag{8}$$

Here $s_{(1)}^{base}$ and $s_{(2)}^{base}$ are the largest and second largest base scores in the candidate pool, and δ is a fixed margin reference shared by all experiments. Positive and negative reliability scales are

$$c_q^+ = c_{min} + (1 - c_{min}) \,\text{clip}\left(\frac{a_q^+ - a_{min}}{a_{max} - a_{min}}, 0, 1\right), \tag{9}$$

$$c_q^- = c_{min} + (1 - c_{min}) \,\text{clip}\left(\frac{a_q^- - a_{min}}{a_{max} - a_{min}}, 0, 1\right). \tag{10}$$

The final per-query weights are

$$w_p(q) = w_p^{cap} c_q^+ (1 - 0.5 b_q), \quad w_n(q) = w_n^{cap} c_q^- (1 - 0.5 b_q), \tag{11}$$

$$w_q(q) = w_q^{min} + \max(w_p^{cap} - w_p(q), 0) + \max(w_n^{cap} - w_n(q), 0). \tag{12}$$

The caps are fixed upper bounds, while the actual weights are decided for each query. Thus, when concept evidence is weak or the base retriever is confident, more score mass stays with the base retriever; when active concepts are strong and the base margin is small, positive concepts receive larger weights. The disliked-concept term is intentionally lightweight and acts as a conservative penalty. Image, text, and concept embeddings are L2-normalized before cosine similarity. This closed-form inference-time calibration makes the scoring coefficients query-dependent without training additional parameters.

3.6 Inference-Time Score Interpolation

AutoConcept further sets the outer interpolation coefficient from inference-time confidence so weak concept activation does not override the base retriever. Let $\bar{a}_q$ be the mean activation strength of active concepts for query q. The upper interpolation bound and base value are set from the base margin:

$$\lambda_q^{max} = \lambda_{min} + (\lambda_{max}^{cap} - \lambda_{min})(1 - b_q), \tag{13}$$

$$\lambda_q^{base} = \lambda_{min} + (\lambda_0 - \lambda_{min})(1 - b_q), \tag{14}$$

$$\lambda_q = \mathrm{clip}(\lambda_q^{base} + \gamma \bar{a}_q, \lambda_{min}, \lambda_q^{max}). \tag{15}$$

The interpolation bounds are fixed before evaluation and produce a query-specific coefficient. The final score is

$$s_q(x_i) = (1 - \lambda_q) s_q^{base}(x_i) + \lambda_q s_q^{concept}(x_i). \tag{16}$$

Table 2. AutoConcept reranking procedure.

Step	Operation
1	Take query (x_r, t_q), candidate pool C_q, metadata $\{m_i\}$, and evidence E_q
2	Extract, cluster, score, and filter concept constraints from E_q
3	Compute query-concept relevance $r(q, c)$ and activate concepts
4	Align each candidate metadata text m_i with active concepts
5	Compute positive score, auxiliary negative penalty, and base confidence
6	Infer $w_q(q), w_p(q), w_n(q)$ and λ_q from query-time signals
7	Rerank C_q by the final score $s_q(x_i)$

4 Experiments

4.1 Dataset, Protocol, and Baselines

The main dataset is FashionIQ. We evaluate five controlled concept-evidence seeds and report mean with sample standard deviation when a method depends on constructed evidence. The first-stage baseline is WeiMoCIR:

$$q = 0.20e(x_r) + 0.80e(t_q). \tag{17}$$

Candidate scoring uses candidate image embeddings for first-stage retrieval. Candidate metadata is introduced only in second-stage reranking controls and AutoConcept. The default candidate pool is top-100. We report Recall@10, Recall@50, MRR, and NDCG@10.

We evaluate metadata-aware rerankers on the same candidate pool: query-text to metadata matching, composed reference-text plus query-text to metadata matching, extracted query-attribute to metadata matching, and a resource-matched prescriptive/proscriptive constraint matcher. These baselines locate

Table 3. Main results on two first-stage retrievers. AutoConcept reranks the fixed candidate pool under the same metadata-available protocol.

First-stage	Reranker	R@10	R@50	MRR	NDCG@10
WeiMoCIR	none	0.1125	0.2256	0.0604	0.0680
WeiMoCIR	AutoConcept	0.1379 ± 0.0011	0.2452 ± 0.0018	0.0739 ± 0.0008	0.0844 ± 0.0009
LinCIR	none	0.2605	0.4618	0.1449	0.1637
LinCIR	AutoConcept	0.3009 ± 0.0018	0.4915 ± 0.0023	0.1677 ± 0.0010	0.1912 ± 0.0008

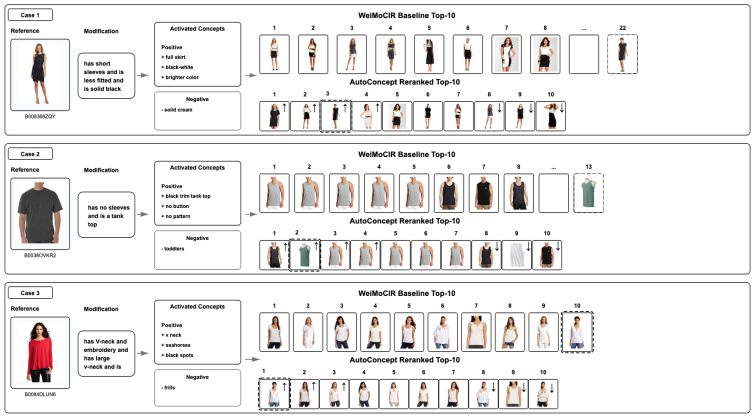

Fig. 3. Qualitative reranking examples. AutoConcept uses active concepts to adjust the top-10 candidate list from WeiMoCIR, improving concept-evidence alignment while retaining the fixed first-stage candidate pool.

AutoConcept relative to direct metadata and constraint matching. The query-only AutoConcept variant builds concepts from t_q and evaluates the concept-memory scorer with query-local evidence.

Table 3 shows consistent gains on both candidate generators, concentrated in early-rank metrics.

4.2 Significance Testing

We use query-level paired bootstrap tests. Each FashionIQ query is one resampling unit; AutoConcept scores are first averaged over five concept-evidence seeds and then paired with the same query under the base retriever. We resample queries with replacement to compute percentile 95% confidence intervals and two-sided paired bootstrap p-values. The reported p-values are descriptive and are not adjusted for multiple metrics; seed-level robustness is reported through the standard deviations in Table 3.

The LinCIR deltas in Table 3 are also significant with $p < 0.001$ under the same query-level paired bootstrap protocol.

Table 4. Query-level paired bootstrap significance against WeiMoCIR.

Metric	Delta	Relative	95% CI	p-value
R@10	0.0254	+22.54%	[0.0223, 0.0285]	< 0.001
R@50	0.0196	+8.71%	[0.0163, 0.0230]	< 0.001
MRR	0.0135	+22.26%	[0.0118, 0.0151]	< 0.001
NDCG@10	0.0164	+24.16%	[0.0147, 0.0182]	< 0.001

4.3 Metadata-Aware Reranking Controls

Table 5 evaluates AutoConcept alongside metadata-aware rerankers on the same top-100 WeiMoCIR pool. Query-only AutoConcept improves over query-only text-to-metadata, extracted-attribute reranking, and resource-matched constraint matching on early-rank metrics. The reference-text enriched row concatenates reference-item metadata with the modification text and is reported as a strong text-only upper reference under a richer textual signal rather than a same-evidence concept-memory baseline.

Table 5. Metadata-aware reranking controls on FashionIQ seed 1. Query-only AutoConcept uses concepts extracted only from the modification text.

Setting	Metadata	Memory	R@10	R@50	MRR	NDCG@10
WeiMoCIR	No	No	0.1125	0.2256	0.0604	0.0680
Query text → metadata	Yes	No	0.1318	0.2513	0.0731	0.0822
Extracted attributes → metadata	Yes	No	0.1363	0.2591	0.0754	0.0850
Constraint matching	Yes	No	0.1383	0.2603	0.0763	0.0861
Query-only AutoConcept	Yes	Yes	0.1400	0.2630	0.0800	0.0893
Controlled-evidence AutoConcept	Yes	Yes	0.1375	0.2435	0.0741	0.0845
Reference text + query → metadata	Yes	No	0.1430	0.2645	0.0811	0.0910

4.4 Concept-Evidence Alignment

Table 3 is the main controlled concept-evidence setting. Separately, to test whether AutoConcept uses query-relevant concept evidence rather than only category priors, we compare aligned evidence with shuffled evidence. The shuffled control preserves the same domain and category-level concept distribution but breaks query-evidence alignment.

Table 6. Concept-evidence alignment analysis. This setting is separate from the main controlled concept-evidence setting in Table 3.

Variant	Base R@10	R@10	R@50	MRR	NDCG@10
Aligned evidence	0.1125	0.1892 ± 0.0028	0.2715 ± 0.0014	0.1050 ± 0.0010	0.1213 ± 0.0014
Shuffled evidence	0.1125	0.1388 ± 0.0010	0.2469 ± 0.0010	0.0738 ± 0.0008	0.0845 ± 0.0004
Delta	–	+0.0504	+0.0246	+0.0312	+0.0368

The aligned setting is substantially stronger than the shuffled control, supporting the role of coherent query-evidence alignment.

4.5 Real-Human Concept Labels

We also evaluate AutoConcept with a pilot real-human concept-label split. Participants labeled FashionIQ gallery items and selected optional item-level con-

cept tags. Many labeled items have empty or very short original item text, so the selected tags provide external concept evidence beyond rich product metadata. We build participant-level concept memories from the deduplicated labels and rerank the same top-100 WeiMoCIR pool.

Table 7 evaluates whether human item-level judgments can be converted into useful concept memory under the same metadata-available reranking pipeline.

Table 7. Real-human concept-label evaluation on FashionIQ.

Setting	R@10	R@50	MRR	NDCG@10
WeiMoCIR candidate pool	0.1125	0.2256	0.0604	0.0680
WeiMoCIR + human AutoConcept	0.1159	0.2311	0.0624	0.0702
Delta	+0.0034	+0.0056	+0.0019	+0.0022

The participant histories are sparse and are not collected for specific benchmark targets, yet all four target-label metrics improve, showing that AutoConcept can ingest human concept labels under annotation noise. The smaller gain compared with controlled evidence is expected: participants label only a limited subset of items, selected tags may not cover every FashionIQ modifier, and human memories are not constructed around the benchmark target for each query.

4.6 Ablation Analysis

Table 8 isolates scoring components on the same FashionIQ seed-1 top-100 pool. Max pooling is important because concept matches are sparse. Removing the lightweight negative penalty changes little, confirming that FashionIQ gains are driven mainly by positive concept activation.

Table 8. Ablation analysis on FashionIQ seed 1.

Setting	R@10	MRR	NDCG@10
WeiMoCIR	0.1125	0.0604	0.0680
Full AutoConcept	0.1375	0.0741	0.0845
Positive concepts only	0.1380	0.0742	0.0847
Stronger negative penalty	0.1363	0.0738	0.0840
Mean pooling	0.1240	0.0679	0.0764
Top-k mean pooling	0.1270	0.0695	0.0783
Dynamic gate without outer interpolation	0.1302	0.0703	0.0798

5 Scope and Extensions

AutoConcept is scoped as a second-stage reranker for metadata-available product-style galleries. The main experiments use controlled concept evidence, while real-human annotations provide participant-supplied evidence under natural label noise and sparse item text. The method is most suitable when gallery metadata is available and reasonably aligned with item appearance; missing or inconsistent metadata remains a practical boundary. Broader domains, richer visual grounding, larger human-preference studies, stronger metadata denoising, and heavier

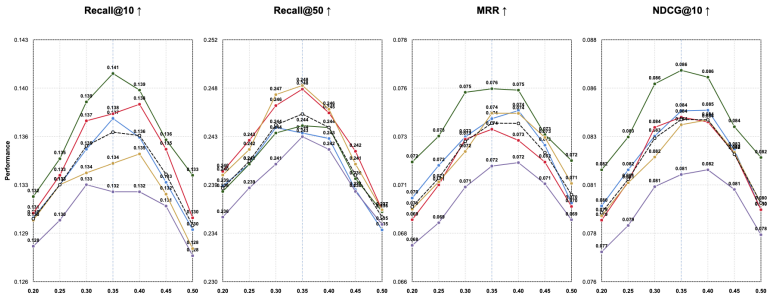

Fig. 4. Gate sensitivity on FashionIQ. AutoConcept remains stable around the operating range used by inference-time gate calibration.

learned verification rerankers are natural extensions under different compute budgets.

6 Conclusion

AutoConcept adds an explicit concept-memory layer to metadata-available composed image retrieval, complementing retrieval-augmented and concept-controlled multimodal systems [19, 20, 10]. It improves both WeiMoCIR and LinCIR candidate pools, compares favorably with direct query and attribute metadata matching, and provides an inspectable memory mechanism. The real-human concept-label study further shows that the same interface can consume participant-provided evidence.

A Additional Evaluation Results

A.1 Gate Sensitivity

Table 9 reports a gate-sensitivity analysis around $\tau = 0.35$. It changes only the query-concept relevance gate and keeps the top-100 candidate pool, controlled concept-evidence setting, concept memory, and calibration bounds fixed. The current implementation can infer τ_q per query from the memory relevance distribution.

The best region is broad. Gates from 0.30 to 0.40 yield similar gains, and all tested values remain above WeiMoCIR. Positive concepts drive most of the observed gain in the current FashionIQ setting. Disliked concepts are retained as a lightweight, inspectable penalty channel rather than as the primary source of improvement.

Table 9. FashionIQ gate sensitivity.

Gate	R@10	R@50	MRR	NDCG@10
0.20	0.1304 ± 0.0013	0.2385 ± 0.0016	0.0697 ± 0.0015	0.0794 ± 0.0015
0.25	0.1328 ± 0.0016	0.2410 ± 0.0016	0.0711 ± 0.0016	0.0811 ± 0.0016
0.30	0.1355 ± 0.0024	0.2442 ± 0.0024	0.0730 ± 0.0017	0.0831 ± 0.0019
0.35	0.1379 ± 0.0011	0.2452 ± 0.0018	0.0739 ± 0.0008	0.0844 ± 0.0009
0.40	0.1363 ± 0.0028	0.2440 ± 0.0015	0.0738 ± 0.0014	0.0840 ± 0.0016
0.45	0.1335 ± 0.0020	0.2397 ± 0.0016	0.0724 ± 0.0010	0.0823 ± 0.0011
0.50	0.1300 ± 0.0022	0.2359 ± 0.0009	0.0703 ± 0.0011	0.0799 ± 0.0014

A.2 Candidate Pool Size

Table 10 varies the first-stage candidate pool size for seed 1. The default top-100 pool gives the strongest AutoConcept result in this analysis. Top-50 limits candidate coverage, while larger pools introduce more metadata noise for the same concept-memory scorer.

Table 10. Candidate-pool sensitivity on FashionIQ seed 1.

K	Base R@10	AC R@10	Base MRR	AC MRR
50	0.1125	0.1144	0.0594	0.0609
100	0.1125	0.1375	0.0604	0.0741
200	0.1127	0.1165	0.0608	0.0631
500	0.1129	0.1180	0.0614	0.0645

This sensitivity suggests that AutoConcept is most effective when the first-stage retriever supplies a reasonably precise candidate pool. For weaker retrievers that require very large pools, stronger metadata denoising or visual verification would be needed before reranking.

A.3 Fashion200K Secondary Check

Fashion200K provides a secondary fashion-domain check constructed from item text and sampled candidate pools. Table 11 reports three-seed means. AutoConcept improves all four metrics over the same WeiMoCIR candidate generator.

Table 11. Fashion200K secondary retrieval check.

Setting	R@10	R@50	MRR	NDCG@10
WeiMoCIR	0.6560 ± 0.0339	0.9144 ± 0.0060	0.3458 ± 0.0097	0.4081 ± 0.0167
WeiMoCIR + AutoConcept	0.7049 ± 0.0167	0.9355 ± 0.0013	0.3709 ± 0.0203	0.4404 ± 0.0199

A.4 Runtime and Latency

Table 12 reports milliseconds per query on an Apple M5 MacBook Air with 24GB unified memory, batch size one, PyTorch/MPS where available, and precomputed CLIP image embeddings. In the cached setting, AutoConcept adds 0.3199 ms/query on FashionIQ and 0.1928 ms/query on Fashion200K, corresponding to 5.58% and 6.18% of first-stage retrieval latency. The higher Fashion200K AC full cost comes from uncached text-encoder setup and metadata encoding over a smaller query set; these embeddings are precomputable in the reranking use case.

Table 12. Runtime comparison between the first-stage WeiMoCIR retriever and the AutoConcept second-stage overhead. The AutoConcept loop column isolates the concept-scoring loop after embeddings are available.

Dataset	Queries	Gallery	Mean K	Base retriever	AC full	AC loop
FashionIQ	6016	74381	100.00	5.7290	1.8080	0.3199
Fashion200K	413	40000	62.00	3.1184	6.2910	0.1928

B Implementation Details

B.1 Candidate Pool and Hyperparameters

The main FashionIQ experiment reranks the top-100 candidates from the base retriever. The same top-100 pool is used for WeiMoCIR, AutoConcept, and metadata-aware controls. Concept names, query text, and metadata text are encoded with all-MiniLM-L6-v2; CLIP ViT-B/32 is used for image embeddings and CLIP-based grounding. Quality-control rules and calibration bounds are fixed before evaluation and shared across seeds, candidate generators, and metadata-aware controls; no query labels, target ranks, or relevance labels are used for tuning.

Table 13. Reproducibility settings.

Setting	Role
WeiMoCIR query fusion	fixed weighted fusion of reference image and modification text
First-stage candidate scoring	image-side scoring; candidate metadata enters only in reranking
Candidate pool	top-100 candidates unless stated otherwise
Concept/text encoder	all-MiniLM-L6-v2 for concept names, queries, and metadata
Image encoder and grounding	CLIP ViT-B/32 image embeddings and CLIP-based grounding
Quality control	fixed filtering rules shared across seeds and candidate generators
Score calibration	bounded query-adaptive weights and interpolation, fixed before evaluation
Tuning data	no query labels, target ranks, or relevance labels used for tuning

For K candidates, embedding dimension d, and active concept sets A_q^+ and A_q^-, the reranking loop scales as $O(K(|A_q^+| + |A_q^-|)d)$. The relevance gate keeps the active concept set small.

C Dataset and Concept-Evidence Construction Details

FashionIQ is the main CIR benchmark. Each query contains a reference image, two modification captions, and a target image. We construct controlled concept-evidence sets from training-side item texts; the query-only control extracts concept names only from the modification text.

Fashion200K is a secondary fashion-domain robustness check with composed queries from item text and sampled candidate pools.

The aligned analysis uses controlled attribute families such as color, printed patterns, sleeve length, neckline, length, and fit. It is separate from the main protocol and excludes the target image, target metadata, candidate ranks, and ground-truth target identity. The shuffled control keeps the evidence distribution but shuffles query-evidence mappings within category.

The real-human study uses a separate collection interface. Participants label FashionIQ items and select optional tags from a fixed vocabulary covering color, pattern, sleeve, neckline, length, fit, and style. Labels are deduplicated, converted into the same concept-memory schema, and stored as a separate human-data split.

D Concept Memory and Filtering Details

Table 14. Main concept filter rules.

Rule	Action	Criterion
Generic concept name	drop	fixed generic-term stoplist
Positive/negative conflict	drop	same normalized name in both memories
Low CLIP grounding	drop	below the fixed grounding quality constraint
Low concept strength	drop	below the fixed concept-strength constraint
Low evidence	drop	evidence count < 1
High global frequency	keep	disabled in main filter
Single weak evidence	keep	disabled in main filter
Minimum positive memory	restore	at least one positive if available

The filter removes obvious generic or weak concepts while keeping visually meaningful apparel attributes such as sleeve length, neckline, buttons, stripes, and V-neck.

E Explanation Diagnostics and Ethics

AutoConcept can list active concepts, evidence items, and candidate-side matches, exposing how memory entries participate in reranking.

AutoConcept stores explicit named concepts rather than opaque user embeddings, supporting inspection, deletion, and editing. The real-human study uses voluntary item-level labels without requesting sensitive user attributes; larger deployments should avoid protected-attribute inference and expose memory-retention controls.

Acknowledgments. We thank the anonymous reviewers for their constructive comments and the participants who contributed item-level concept labels.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agnolucci, L., Baldrati, A., Bertini, M., Del Bimbo, A.: iSEARLE: Improving textual inversion for zero-shot composed image retrieval. arXiv preprint arXiv:2405.02951 (2024)

2. Alaluf, Y., Richardson, E., Tulyakov, S., Aberman, K., Cohen-Or, D.: MyVLM: Personalizing VLMs for user-specific queries. In: Computer Vision – ECCV 2024. Lecture Notes in Computer Science, vol. 15071, pp. 73–91. Springer (2025). https://doi.org/10.1007/978-3-031-72624-8_5
3. Baldrati, A., Agnolucci, L., Bertini, M., Del Bimbo, A.: Zero-shot composed image retrieval with textual inversion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
4. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: Conditioned and composed image retrieval combining and partially fine-tuning CLIP-based features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 4959–4968 (2022)
5. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: Effective conditioned and composed image retrieval combining CLIP-based features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21466–21474 (2022)
6. Chen, Y., Gong, S., Bazzani, L.: Image search with text feedback by visiolinguistic attention learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3001–3011 (2020)
7. Ghorbani, A., Wexler, J., Zou, J., Kim, B.: Towards automatic concept-based explanations. In: Advances in Neural Information Processing Systems. vol. 32 (2019)
8. Gu, G., Chun, S., Kim, W., Kang, Y., Yun, S.: Language-only training of zero-shot composed image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13225–13234 (2024)
9. Han, X., Wu, Z., Huang, P.X., Zhang, X., Zhu, M., Li, Y., Zhao, Y., Davis, L.S.: Automatic spatially-aware fashion concept discovery. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1463–1471 (2017)
10. Hao, H., Han, J., Li, C., Li, Y.F., Yue, X.: RAP: Retrieval-augmented personalization for multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14538–14548 (2025)
11. He, X., Liao, L., Zhang, H., Nie, L., Hu, X., Chua, T.S.: Neural collaborative filtering. In: Proceedings of the 26th International Conference on World Wide Web. pp. 173–182 (2017)
12. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data **7**(3), 535–547 (2021)
13. Jung, Y., Cho, S., Min, H.s., Choi, S.: Soft filtering: Guiding zero-shot composed image retrieval with prescriptive and proscriptive constraints. arXiv preprint arXiv:2512.20781 (2025)
14. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. pp. 6769–6781 (2020)
15. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., Sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors. In: Proceedings of the 35th International Conference on Machine Learning. pp. 2668–2677 (2018)
16. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: Proceedings of the 37th International Conference on Machine Learning. pp. 5338–5348 (2020)
17. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**(8), 30–37 (2009)

18. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems. vol. 33, pp. 9459–9474 (2020)
19. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th International Conference on Machine Learning (2023)
20. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in Neural Information Processing Systems **36** (2023)
21. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2125–2134 (2021)
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: Proceedings of the 38th International Conference on Machine Learning. pp. 8748–8763 (2021)
23. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: BPR: Bayesian personalized ranking from implicit feedback. In: Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. pp. 452–461 (2009)
24. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19305–19314 (2023)
25. Tang, Y., Yu, J., Gai, K., Zhuang, J., Xiong, G., Hu, Y., Wu, Q.: Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5180–5188 (2024)
26. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6439–6448 (2019)
27. Weston, J., Chopra, S., Bordes, A.: Memory networks. In: International Conference on Learning Representations (2015)
28. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11307–11317 (2021)
29. Wu, R.D., Lin, Y.Y., Yang, H.F.: Training-free zero-shot composed image retrieval via weighted modality fusion and similarity. arXiv preprint arXiv:2409.04918 (2024)