AutoConcept

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

Tianyu Wang[1] Tianjiao Wu[2]

[1]Soochow University, Suzhou, China

[2]INSA Lyon, Lyon, France

19 November 2026 Guangzhou, China

Composed image retrieval

Reference image



+

Modification text

"has short sleeves and is less fitted and is solid black with white accents"

⟶

Target image



The target must preserve relevant visual content and satisfy the requested changes.

Why add explicit concepts?

What the first stage provides

- A composed image–text query.
- A fixed top-K candidate pool.
- A similarity score for each image.

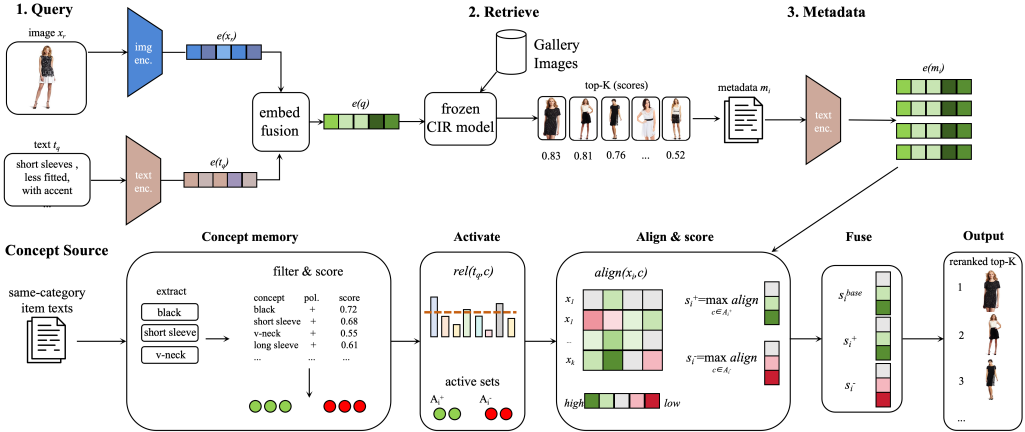
What a product gallery also provides

- Titles, tags, and attribute descriptions.
- Named concepts such as sleeve length, pattern, and neckline.
- Evidence that can be inspected and edited.

Our approach

An explicit concept memory reranks the existing candidates using their metadata, with no reranker training or per-query MLLM calls.

The AutoConcept framework



Building an inspectable concept memory

Construction

1. Extract short attribute phrases.
2. Cluster near-duplicate phrases.
3. Score positive and negative evidence.
4. Filter generic, conflicting, or weak concepts.

Each memory entry records

Field	Meaning
Name	apparel attribute
Polarity	positive / negative
Strength	evidence score
Grounding	visual support
Evidence	supporting items

Query activation and score calibration

1. Query-relevance gate

$$r(q, c) = \cos(e(t_q), e(c))$$

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

Only query-relevant concepts participate.

2. Metadata alignment

$$s_q^{\pm}(x_i) = \max_{c \in A_q^{\pm}} \cos(e(m_i), e(c))$$

3. Concept-guided score

$$s_q^{\text{concept}} = w_q s_q^{\text{base}} + w_p s_q^{+} - w_n s_q^{-}$$

4. Interpolation

$$s_q = (1 - \lambda_q) s_q^{\text{base}} + \lambda_q s_q^{\text{concept}}$$

Query-specific weights use concept activation and the base retriever's top-score margin.

Weak concept evidence or a confident base retriever keeps more weight on the base score.

Evaluation on fixed candidate pools

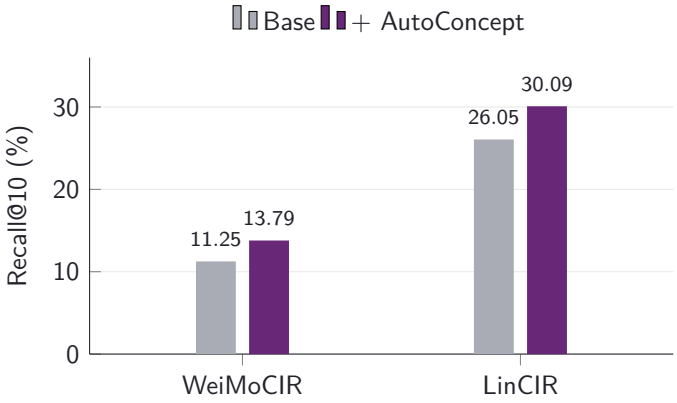
Main setting: FashionIQ

- 6,016 queries, top-100 candidates.
- WeiMoCIR and LinCIR first stages.
- Five controlled concept-evidence seeds.
- Recall@10/50, MRR, NDCG@10.

Information boundaries

- Main evidence: same-category training-side item texts.
- Metadata enters after candidate retrieval.
- Evidence construction excludes target identity, target data, and candidate ranks.

Gains on two first-stage retrievers



+2.54 pp
WeiMoCIR R@10

+4.04 pp
LinCIR R@10

Five-seed means. R@10 SD: 0.11 pp (WeiMoCIR), 0.18 pp (LinCIR).

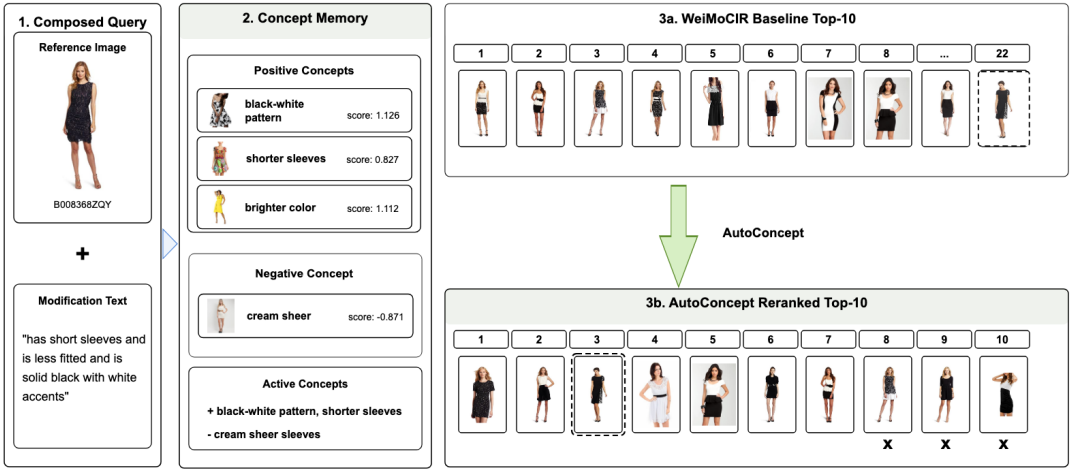
Beyond direct metadata matching

Reranker on the WeiMoCIR pool	R@10 (%)	MRR	NDCG@10
None	11.25	.0604	.0680
Query text to metadata	13.18	.0731	.0822
Extracted attributes to metadata	13.63	.0754	.0850
Constraint matching	13.83	.0763	.0861
Query-only AutoConcept	**14.00**	**.0800**	**.0893**
Controlled-evidence AutoConcept	13.75	.0741	.0845
Reference text + query to metadata	14.30	.0811	.0910

What drives the improvement?

Variant	R@10 (%)
WeiMoCIR	11.25
Full AutoConcept	**13.75**
Positive concepts only	13.80
Stronger negative penalty	13.63
Mean pooling	12.40
Top-k mean pooling	12.70
Dynamic gate, no outer interpolation	13.02

A concept trace explains a change in ranking



Example target: rank 22 to rank 3 within the same candidate pool.

A pilot with human concept labels

Pilot human-label study

Participants label items and select optional concept tags. Deduplicated labels form participant-level memories.

Setting	R@10 (%)
WeiMoCIR	11.25
+ human AutoConcept	11.59

Scope and practical boundaries

- Metadata quality affects alignment.
- Sparse human tags yield modest gains.
- Reranking cannot recover targets outside the candidate pool.

An inspectable, training-free layer for CIR

1. **Explicit concept memory** makes the reranking evidence visible.
2. **Query-relevant activation and calibration** improve early ranks on two fixed retrievers.
3. **Metadata-aware controls** show where concept structure helps and where richer text remains competitive.

Thank you. Questions?

Tianyu Wang tywangcs@foxmail.com

Full main results and statistical evidence

First stage	Reranker	R@10	R@50	MRR	NDCG@10
WeiMoCIR	None	.1125	.2256	.0604	.0680
WeiMoCIR	AutoConcept	.1379	.2452	.0739	.0844
LinCIR	None	.2605	.4618	.1449	.1637
LinCIR	AutoConcept	.3009	.4915	.1677	.1912

WeiMoCIR comparison	Delta	95% bootstrap CI	p
R@10	.0254	[.0223, .0285]	$< .001$
R@50	.0196	[.0163, .0230]	$< .001$
MRR	.0135	[.0118, .0151]	$< .001$
NDCG@10	.0164	[.0147, .0182]	$< .001$

Gate and candidate-pool sensitivity

Gate	R@10 (%)	MRR
0.20	13.04 ± 0.13	.0697
0.25	13.28 ± 0.16	.0711
0.30	13.55 ± 0.24	.0730
0.35	13.79 ± 0.11	.0739
0.40	13.63 ± 0.28	.0738
0.45	13.35 ± 0.20	.0724
0.50	13.00 ± 0.22	.0703

K	Base R@10	AC R@10
50	.1125	.1144
100	.1125	.1375
200	.1127	.1165
500	.1129	.1180

The gate has a broad useful region around 0.30–0.40.
Top-100 performs best in the reported pool-size check.

Runtime with and without cached embeddings

Milliseconds per query	Base retrieval	AC full	AC cached loop
FashionIQ	5.7290	1.8080	.3199
Fashion200K	3.1184	6.2910	.1928

- Cached loop overhead: 5.58% and 6.18% of base latency.
- Full cost includes text-encoder setup and metadata encoding.

Bounded inference-time calibration

$$b_q = \mathrm{clip}\left(\frac{s^{\mathrm{base}}_{(1)} - s^{\mathrm{base}}_{(2)}}{\delta}, 0, 1\right), \qquad c^{\pm}_q = c_{\min} + (1 - c_{\min})\,\mathrm{clip}\left(\frac{a^{\pm}_q - a_{\min}}{a_{\max} - a_{\min}}, 0, 1\right)$$

$$w_p = w^{\mathrm{cap}}_p c^{+}_q (1 - 0.5 b_q), \qquad w_n = w^{\mathrm{cap}}_n c^{-}_q (1 - 0.5 b_q)$$

$$w_q = w^{\min}_q + \max(w^{\mathrm{cap}}_p - w_p, 0) + \max(w^{\mathrm{cap}}_n - w_n, 0)$$

$$\lambda^{\max}_q = \lambda_{\min} + (\lambda^{\mathrm{cap}}_{\max} - \lambda_{\min})(1 - b_q), \quad \lambda^{\mathrm{base}}_q = \lambda_{\min} + (\lambda_0 - \lambda_{\min})(1 - b_q)$$

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

$a^{\pm}_q$: maximum gated activation. $\bar{a}_q$: mean active-concept activation.
All caps and bounds are fixed before evaluation. No reranker parameter training.

Human-label pilot: all four metrics improve

Setting	R@10	R@50	MRR	NDCG@10
WeiMoCIR	.1125	.2256	.0604	.0680
+ human AutoConcept	.1159	.2311	.0624	.0702
Reported delta	+.0034	+.0056	+.0019	+.0022

- Participants supply item-level labels and optional concept tags.
- Deduplicated labels form participant-level memories.
- Sparse histories and limited modifier coverage constrain the gains.

Secondary check and implementation details

Fashion200K: a secondary fashion-domain check

Item-text queries and sampled candidate pools, three-seed means.

Method	R@10	R@50	MRR	NDCG@10
WeiMoCIR	.6560	.9144	.3458	.4081
+ AutoConcept	.7049	.9355	.3709	.4404

Encoders and computational cost

- all-MiniLM-L6-v2: concept names, queries, and metadata.
- CLIP ViT-B/32: image embeddings and visual grounding.
- Cached reranking: $O(K(|A_q^+| + |A_q^-|)d)$.

Selected references

1. Wu, H. et al. *Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback.* CVPR 2021.

2. Wu, R. D., Lin, Y. Y. and Yang, H. F. *Training-Free Zero-Shot Composed Image Retrieval via Weighted Modality Fusion and Similarity.* arXiv:2409.04918, 2024.

3. Gu, G. et al. *Language-Only Training of Zero-Shot Composed Image Retrieval.* CVPR 2024. (LinCIR)