NeurIPS 2026

IntentLens

Grounding Underspecified Multimodal Queries for Recommendation via Tool-Augmented Reasoning

Xiao Chen[1], Dong Fang[2], Haitao Li[2], Qing Li[1,3]

[1] The Hong Kong Polytechnic University [2] LightSpeed
[3] The Hong Kong Polytechnic University Shenzhen Research Institute



Interactive Recommendation with Underspecified Multimodal User Intent

With camera-equipped wearables become increasingly popular, a user can pair a reference image with a fuzzy request.

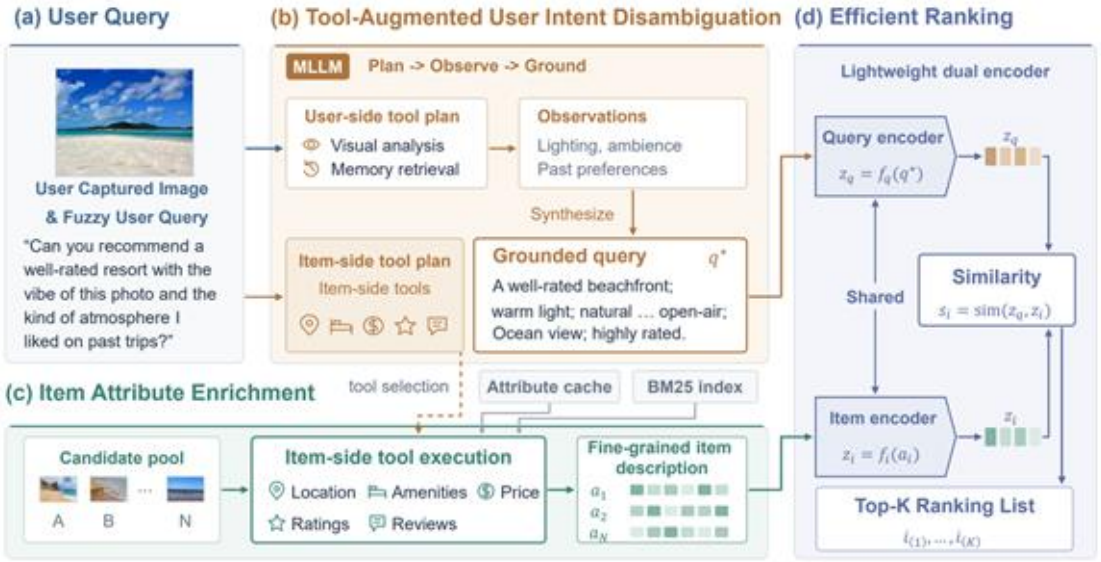
Recommendation setting	How intent is specified
Passive preference matching	Multimodal features enrich item representations for matching against historical interactions.
Fully specified textual requests	The user explicitly states the constraints on candidate items.
Underexplored setting	A single multimodal query leaves visual preferences, personal preferences and item-side constraints implicit.

Task definition

Given a reference image I, a request q, user history H_u and a candidate set C, rank C according to the user's latent preferences.

Intent grounding before candidate ranking

Acquire query-relevant evidence to disambiguate user intent and enrich candidate representations.



1a User intent disambiguation

Plan visual and user-memory tools. Synthesize their observations into a grounded query.

1b Item attribute enrichment

Select item-side tools for this query. Read cached attributes and review evidence in a decoupled pass.

2 Lightweight ranking

A dual-encoder ranker scores the grounded query against each enriched candidate representation.

Counterfactual ranking reward

A frozen dual-encoder ranker evaluates each rollout over the full candidate pool. Candidate representations are fixed.

Query representation	Rank of first relevant item	Reciprocal rank
Clarified query with tool evidence	**2nd**	**0.50**
Same clarified query without tool evidence	4th	0.25
Original request without tool evidence	5th	0.20

Positive evidence credit requires a gain over both no-evidence baselines.

The smaller gain is 0.25. Clipping at 0.10 and weighting by 0.20 gives a ranking reward of 0.52.

Masked GRPO Four rollouts per query. Observation tokens are masked in both policy and KL losses.

Benchmarks for underspecified multimodal queries

2,500 multimodal queries

773 queries in Google Review-MMTool and 1,727 in Yelp-MMTool. Each query has a set of relevant items.

Selected methods	Google MRR	Yelp MRR
BGE Retrieval	25.7	32.2
UniMP	33.9	43.7
GPT-5.4 with IntentLens evidence	41.6	41.5
IntentLens (7B)	**40.2**	**46.7**

IntentLens-7B has the highest non-oracle MRR on Yelp. GPT-5.4 with evidence leads on Google.

Isolating the effect of query grounding

Google Review-MMTool. The dual encoder, candidate pool and candidate-side evidence are held fixed.

Raw query	Predicted-grounded query	Oracle-grounded query
33.6	40.2	44.4

MRR × 100

+6.6 points

Using the predicted-grounded query raises MRR from 33.6 to 40.2.
The oracle-grounded query reaches 44.4 with the same dual encoder.