TRIAD: MULTIMODAL RETRIEVAL ACROSS ABSTRACTION, CONSTRAINTS, AND STATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal retrieval increasingly supports agentic systems, where user requests can involve cross-level relations, combine multiple constraints, and evolve across dialogue turns. In these settings, a highly similar result may still violate the user's requirements, while retrieved content may directly inform downstream actions without human screening. We introduce TRIAD, a multimodal retrieval benchmark for evaluating how well retrieval systems satisfy these complex, evolving requests. TRIAD comprises three tracks: abstraction, testing bidirectional retrieval between visual instances and concepts at different hierarchical levels; constraints, testing the joint satisfaction of positive and negative conditions; and state, testing the integration of requirements across dialogue turns and the updating of earlier conditions when users revise their requests. The benchmark contains six subsets, each with 1,000 queries and a fixed pool of 1,000 candidates per query, evaluated using a common multi-positive mean average precision (mAP) protocol. Simpler reference tasks and hard negatives grouped by the requirements they violate help characterize specific retrieval difficulties. We evaluate 20 configurations of 17 systems and find that no system leads across all six subsets. Performance declines with hierarchical distance, and negative-only queries generally underperform positive-only queries with the same number of constraints. Across paired tasks expressing the same final intent, multi-turn retrieval yields lower mAP in 39 of 40 configuration–domain comparisons. The benefits of explicit reasoning also vary across models and tasks. These findings highlight the gap between retrieving similar content and satisfying a complete request.

1 INTRODUCTION

Traditional multimodal search presents users with a ranked list of results in response to a text, image, or combined query. Embedding-based retrievers produce this ranking by mapping queries and candidates into a shared space and comparing their similarity. Its traditional use, in which users inspect and filter results ranked primarily by approximate similarity, rests on three premises. *The query is self-contained*: the current input already expresses the full active intent, without requiring the system to determine which earlier requirements still apply. *Overall similarity can stand in for relevance*: candidates are compared mainly by degree of match, without hard requirements whose violation cannot be compensated for by similarity in other respects. *Retrieval ends with human judgment*: users can skip results that appear relevant but fail to meet their actual needs.

In agentic settings, user requests become more detailed and complex and evolve through interaction, so these premises no longer generally hold. A request may seek concepts in a superordinate or subordinate relation, rather than simply similar content (Vendrov et al., 2016). For example, when retrieving a superordinate concept for a sparrow, biologically similar bird species may resemble the sparrow more closely in appearance or fine-grained semantics, yet cannot substitute for a superordinate concept such as *bird* or even *animal*. A request may also impose multiple constraints that must hold jointly: "like this scarf, but blue and not patterned" requires both inclusion and exclusion, together with preservation of attributes that were not requested to change. Similarity in other respects cannot compensate for violating these requirements. A request may further span multiple turns, requiring selective retention or updating of historical state: when a user first asks for orange and later changes to blue, the system must retain conditions from the history that remain

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

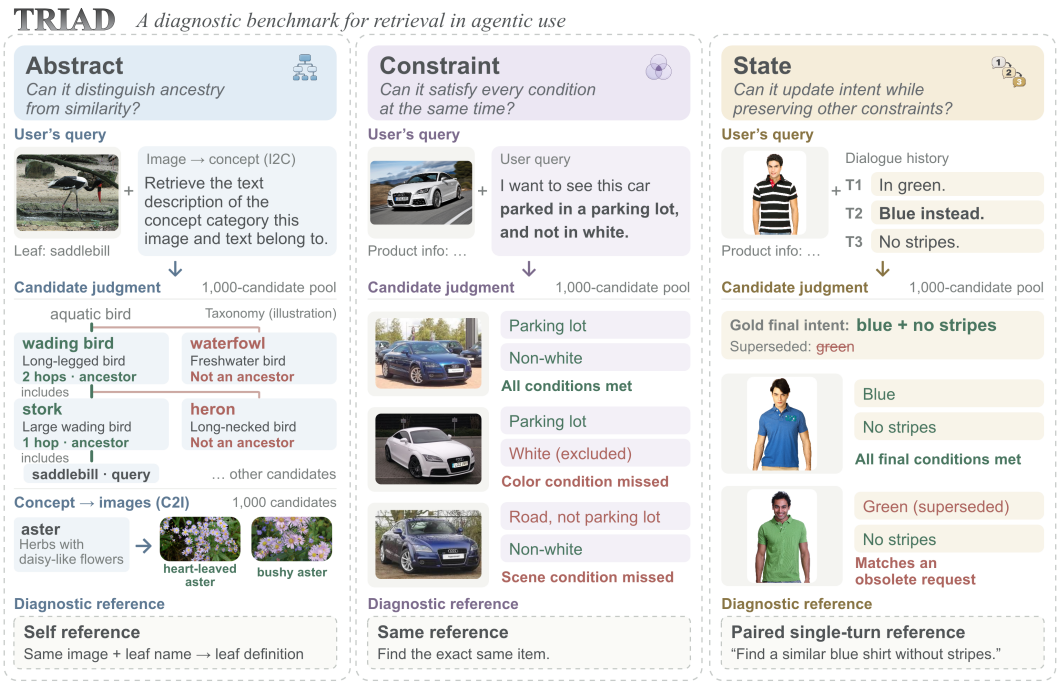

Figure 1: **Overview of TRIAD.** Cross-level, multi-constraint, and multi-turn retrieval test hierarchical relations, constraint composition, and intent updates. Green/red indicate satisfied/violated requirements. Dashed boxes show simpler diagnostic references.

valid while replacing the abandoned value with the new one, rather than assume that the latest turn is self-contained. Meanwhile, retrieved results may feed directly into subsequent reasoning or tool calls, without a user screening each result for errors. Relevance must therefore be judged against the user's full current requirements, rather than overall similarity alone. Although existing benchmarks partially cover multi-constraint and multi-turn retrieval, a unified framework for evaluating these requirements and diagnosing specific model capability gaps is still lacking (Wu et al., 2019; Alhamoud et al., 2025; Guo et al., 2018). Our concern is not to replace similarity-based ranking, but to test whether the relevance it represents responds to the complete request: we define correct results from the request's requirements, measure whether a model ranks them highly, and examine which requirements high-ranked errors violate.

To this end, we introduce **TRIAD**, a multimodal retrieval benchmark designed to go beyond similarity matching and measure how well retrieval systems execute three types of specifications in user requests: cross-level retrieval, multiple constraints, and state updates (Figure 1). ***The abstraction track*** evaluates bidirectional retrieval between visual instances and concepts at different levels, testing retrieval of the correct hierarchical relations beyond direct category matching. ***The constraints track*** combines positive and negative requirements around a reference product, testing whether multiple conditions and attributes that should be preserved are jointly satisfied. ***The state track*** distributes requirements across dialogue turns, testing integration of new conditions and retrieval according to the final intent after a revision. Each track also includes a lower-complexity reference and distractors grouped by the requirements they violate, enabling comparisons between basic matching and complex requests and analysis of high-ranked errors involving hierarchical confusion, constraint violations, or stale values. The abstraction track is split by retrieval direction, and the constraints and state tracks are each instantiated in fashion and cars, yielding six main subsets of 1,000 queries with 1,000 candidates per query. All subsets use a common multi-positive mAP protocol and are reported separately.

We evaluate 20 configurations of 17 systems to examine how representative embedding models from the retrieval literature perform on TRIAD. The results show that 1) instance-to-concept retrieval quality decreases with hierarchical distance; 2) query groups containing negative constraints generally score lower than positive-only groups, with candidates violating only negative require-

ments often becoming high-ranked distractors; and 3) multi-turn versions typically underperform their paired single-turn counterparts expressing the same final intent. Within-model comparisons with and without reasoning further show no uniform gain: its effects vary across models, tasks, and integration methods.

2 Related Work

Broad-Coverage and Instruction-Following Evaluation. Broad-coverage benchmarks bring together data across tasks, domains, and modalities to assess general capabilities across retrieval settings. BEIR and MTEB assemble heterogeneous tasks under a shared evaluation interface with per-task reporting (Thakur et al., 2021; Muennighoff et al., 2023). M-BEIR and the MMEB series extend unified evaluation to instruction-conditioned multimodal retrieval, including video, visual documents, and omni-modal settings (Wei et al., 2024; Jiang et al., 2024; Meng et al., 2025; Huang et al., 2026). MMEB-V3 additionally introduces OmniSET for controlled diagnosis with semantic content aligned across modalities. FollowIR and InstructIR focus on instruction following in text retrieval, examining whether relevance judgments respond to the instruction (Weller et al., 2025; Oh et al., 2024).

Focused Capability Evaluation. Specialized studies examine hierarchical relations, condition composition, or dialogue in greater depth. Hierarchy-aware methods use category or visual part–whole hierarchies to structure representations for classification and retrieval (Vendrov et al., 2016; Barz & Denzler, 2018; Bertinetto et al., 2020; Wang et al., 2025). Composed image retrieval specifies a target through a reference image and modification text (Vo et al., 2019; Wu et al., 2019; Liu et al., 2021). Related evaluations examine multiple positives (Baldrati et al., 2023), conditional similarity (Vaze et al., 2023), instance identity (Psomas et al., 2025), hard negatives (Byun et al., 2025; Mahadev et al., 2026), input robustness (Sun et al., 2023), and multi-turn interaction (Liu et al., 2026). NegBench specifically tests inclusion and exclusion requirements (Alhamoud et al., 2025). Image–text compositionality benchmarks examine compositional generalization, attribute and relation understanding, and fine-grained image–text distinctions (Ma et al., 2023; Yuksekgonul et al., 2023; Hsieh et al., 2023; Thrush et al., 2022). Dialogue retrieval, retrieval-oriented rewriting, multi-turn RAG, and agentic evaluations study information needs and retrieval behavior over interaction histories (Guo et al., 2018; Tan et al., 2019; Zhao et al., 2024; Hu et al., 2026; Katsis et al., 2025; Cheng et al., 2025; Ali et al., 2026; You et al., 2026).

TRIAD complements these studies by bringing cross-level relations, joint constraints, and state updates into a unified multimodal retrieval evaluation framework. Disaggregated results, low-complexity references, and typed hard negatives reveal difficulties in following specific requirements. A dimension-by-dimension comparison with existing benchmarks appears in Appendix F.

3 The TRIAD Benchmark

3.1 Three Capability Axes and Design Principles

TRIAD defines relevance through the relations, conditions, and active intent specified in a request. Its abstraction, constraints, and state tracks evaluate cross-level retrieval, joint constraint satisfaction, and state updates, respectively. Four design principles help locate ranking difficulties: (P1) provide low-complexity references; (P2) screen potential false negatives using hierarchical or attribute rules; (P3) categorize distractors by the requirements they violate; and (P4) audit input shortcuts that could bypass the target capability.

3.2 Tasks and Data

The three tracks yield six main subsets: I2C and C2I for abstraction, and Fashion and Cars for each of the constraints and state tracks. Each subset contains 1,000 queries, each with 1,000 fixed candidates and a labeled positive set. Figure 1 illustrates the inputs and retrieval targets.

3.2.1 MAIN-TASK DATA

Abstraction: Retrieval Across Levels. Using the WordNet 3.0 noun IS-A hierarchy (Miller, 1995) and ImageNet-21K images (Deng et al., 2009), we construct bidirectional cross-level retrieval between instances and concepts: Instance-to-Concept (I2C) and Concept-to-Instance (C2I). I2C asks for an image's ancestor concepts one or two levels above its leaf class; C2I asks for images from descendant classes one or two levels below a concept. Image inputs contain the image and leaf-class name, while concept inputs contain the name and definition. Positives follow WordNet ancestor–descendant relations. The full pools contain 1,317 internal concepts and 15,981 instance images.

Constraints: Multi-Condition Retrieval. We retain 12,946 items from Fashion Product Images (Aggarwal, 2019) and 8,007 entries from Stanford Cars (Krause et al., 2013), constructing queries that seek similar items to a reference while satisfying one to three conditions. Queries include a reference image, product text, and requested changes; candidates include an image and product text. Positives must satisfy all explicit conditions and preserve attributes designated by the construction rules. Positive conditions (P) require an attribute, and negative conditions (N) exclude one; D1–D7 denote 1P, 1N, 2P, 2N, 1P1N, 2P1N, and 1P2N.

State: Multi-Turn Retrieval. We rewrite constraints-track queries as two- or three-turn conversations, preserving the reference item, final intent, and complete positive set. Natural sessions progressively add conditions; conflict sessions introduce earlier values that are explicitly replaced later. Queries use the reference image and text with the full dialogue, while candidate inputs remain unchanged. Positives must satisfy the final active conditions, retaining unrevised requirements and discarding replaced values.

3.2.2 DIAGNOSTIC AND REFERENCE DATA

We use three references to distinguish basic matching from additional requirements: Self retrieves between images and their own leaf concepts in both directions; Same retrieves other images of the same item from an image–text reference; and the paired single turn expresses the complete final intent of a multi-turn query at once, rather than retaining only its last turn. These references help locate difficulties, but their candidate composition need not match the main tasks, so they do not provide a common upper bound.

Hard negatives are typed by the violated requirements. Abstraction uses semantically related candidates with incorrect hierarchical relations. Constraints distinguishes violations of positive conditions, negative conditions, preserved attributes, and both positive and negative conditions, alongside generic visually similar negatives. State introduces revision traps that satisfy an obsolete value but violate the final conditions. Complete construction procedures and statistics appear in Appendix A.

3.3 DATA QUALITY AND VALIDITY

We screen data through image filtering, attribute normalization, and checks of conditions and final intent. Fashion uses two-model agreement on visual attributes; Cars uses single-model annotation with programmatic checks (Appendix A.6). Human sample-based reviews cover 80 queries or sessions and 80 images, distinguishing query issues by their nature and impact on understanding the requirements. Among the sampled images, 507 of 513 determinate visual attributes are annotated correctly, with eight additional attributes marked not visible/undetermined, yielding 98.83% accuracy. Query-review and attribute-verification results appear in Appendix E.1. Some models were trained on source datasets or related category knowledge, so we do not claim complete decontamination; per-system exposure audits appear in Appendix D.3.

3.4 DESIGN DIFFERENCES FROM EXISTING BENCHMARKS

Table 1 summarizes the design differences between TRIAD and three representative benchmarks. MMEB-V2 integrates heterogeneous tasks under a unified evaluation protocol, but score differences across tasks alone do not directly identify specific capability gaps (Meng et al., 2025); TRIAD connects ranking performance to specific difficulties through reference tasks and hard negatives grouped

by the requirements they violate. NegBench primarily evaluates negation understanding through the presence or absence of particular semantic entities (Alhamoud et al., 2025), whereas TRIAD considers entities with multiple attributes, jointly organizing positive attribute requirements, negative exclusions, and implicit preservation constraints from a reference entity to test whether models satisfy requirements across multiple dimensions. ChatSearch focuses on recovering retrieval intent from multi-turn image–text contexts (Zhao et al., 2024); its publicly documented construction does not treat multidimensional constraints distributed across turns, replacement of earlier conditions, and trap interference as explicitly controlled evaluation factors. TRIAD distributes known constraints across turns, explicitly constructs condition accumulation and the replacement of old values by new requirements, and introduces trap candidates that satisfy obsolete conditions but violate the final request, to diagnose whether models retain valid conditions, update superseded requirements, and exclude interference from earlier states.

Table 1: Design comparison with representative benchmarks. Extended criteria and comparisons appear in Appendix F.

Dimension	**MMEB-V2**	**NegBench**	**ChatSearch**	**TRIAD (ours)**
Targets	Classification, QA, retrieval	Object presence / absence	Dialogue-specified images	Instances ↔ higher/lower concepts; products
Conditions	Task-specific; composed retrieval	Positive, negative, hybrid	Descriptions and feedback	1–3 conditions; 7 polarity mixes; preserved attributes
State	Context tasks (e.g., VisDial)	Single-turn	Single/multi-turn; text/image–text	Accumulate/replace; matched single turn
Diagnostics	Task-specific	Negation distractors; presence/absence pairs	Context-format comparisons	Typed violations and traps; Self/Same/single-turn

3.5 Evaluation Protocol and Metrics

We L2-normalize query and candidate vectors and rank by cosine similarity. Across the six main subsets, 65.3%–100.0% of queries have multiple positives; MRR and Hit@K consider only the first positive or whether any is retrieved. We use mAP, averaging query AP equally over full candidate rankings with the complete labeled-positive count as denominator. Metric definitions appear in Appendix B.1; supplementary Top-1 labeled-positive hit rates appear in Appendix C.

4 Experimental Evaluation

We compare existing embedding models on the six subsets, using cross-benchmark rankings and a reranking reference to examine performance and headroom under specific retrieval requirements.

4.1 Models and Experimental Setup

We evaluate 20 configurations of 17 independent systems spanning dual-tower encoders, discriminative embedders, and reasoning embedders (Table 2). Configurations include different sizes within a series, UME-R1 Disc/Gen, and Embed-RL-2B/4B with and without CoT to examine scaling and reasoning effects. All models are evaluated zero-shot without fine-tuning on TRIAD. Models use fixed per-query candidate sets and checkpoint-specific input adaptations. We report mAP separately on the six main subsets, without an aggregate score. Encoding, fusion, and reasoning configurations appear in Appendix B; confidence intervals and statistical tests for paired comparisons appear in Appendices C and D. We additionally use Qwen3-VL-Reranker-8B to rerank candidates from Qwen3-VL-Embedding-8B as a two-stage retrieval reference (Section 4.4).

4.2 Overall Results and Cross-Task Performance

Table 2 summarizes the six main subsets; complete Self and Same reference-task results appear in Appendix C.1. WeMM-Embedding-9B achieves the highest point estimates on C2I and both domains of the constraints and state tracks, while RzenEmbed-v2 leads I2C. The two leading C2I scores differ by only 0.001 mAP.

Table 2: mAP on six main subsets (higher is better). Bold/underlined values mark the best/second-best embedding configurations before rounding. The final reranking reference is excluded from these marks and family comparisons (Section 4.4).

	Abstraction		Constraints		State	
Model	I2C	C2I	Fashion	Cars	Fashion	Cars
Dual-tower encoders						
CLIP ViT-L/14 (Radford et al., 2021)	0.318	0.224	0.185	0.065	0.180	0.062
SigLIP-2-so400m (Tschannen et al., 2025)	0.232	0.278	0.173	0.082	0.144	0.069
jina-clip-v2 (Koukounas et al., 2024)	0.354	0.401	0.123	0.052	0.117	0.050
Discriminative embedders						
WeMM-Embedding-9B (Zhou et al., 2026)	0.524	**0.638**	**0.547**	**0.367**	**0.498**	**0.291**
E5-Omni-7B (Chen et al., 2026)	0.333	0.557	0.364	0.186	0.223	0.101
RzenEmbed-v2 (Jian et al., 2025)	**0.569**	0.610	0.482	<u>0.298</u>	0.441	<u>0.272</u>
Qwen3-VL-Emb.-8B (Li et al., 2026)	0.494	<u>0.637</u>	<u>0.508</u>	0.230	<u>0.454</u>	0.177
Qwen3-VL-Emb.-2B (Li et al., 2026)	0.463	0.507	0.442	0.157	0.212	0.085
OmniEmbed-3B (Xu et al., 2025)	0.384	0.496	0.184	0.084	0.167	0.076
VLM2Vec-V2.0-2B (Meng et al., 2025)	0.333	0.370	0.318	0.182	0.265	0.142
VLM2Vec-2B (Jiang et al., 2024)	0.306	0.348	0.213	0.101	0.191	0.092
VLM2Vec-7B (Jiang et al., 2024)	0.403	0.392	0.251	0.128	0.221	0.107
Embed-RL-2B (no CoT) (Jiang et al., 2026)	0.448	0.465	0.355	0.198	0.308	0.170
Embed-RL-4B (no CoT) (Jiang et al., 2026)	0.518	0.527	0.356	0.227	0.321	0.194
B-Tokens-2B (Sun et al., 2026)	0.453	0.467	0.368	0.169	0.300	0.146
UME-R1-7B (Disc) (Lan et al., 2025)	0.538	0.577	0.436	0.249	0.321	0.195
Reasoning embedders						
UME-R1-7B (Gen) (Lan et al., 2025)	0.537	0.589	0.359	0.174	0.235	0.130
TWN-4B (Zhang et al., 2026)	0.527	0.521	0.264	0.140	0.201	0.092
Embed-RL-2B + T-CoT (Jiang et al., 2026)	0.504	0.519	0.403	0.221	0.374	0.195
Embed-RL-4B + T-CoT (Jiang et al., 2026)	<u>0.538</u>	0.577	0.360	0.281	0.376	0.259
Top-30 reranking (ceiling probe) (Li et al., 2026)	0.607	0.674	0.566	0.327	0.516	0.258

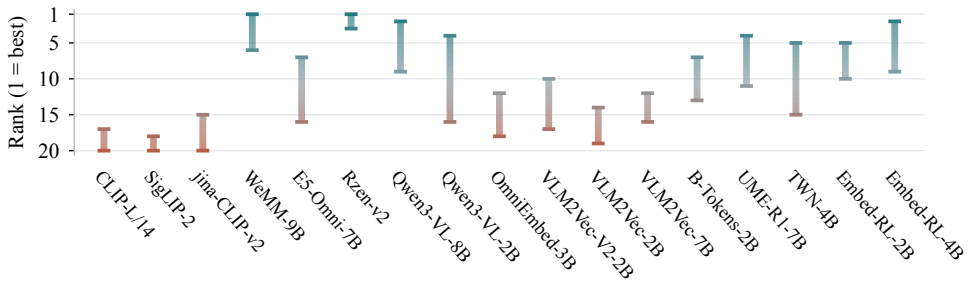

Figure 2: Rank ranges across six subsets; vertical segments connect the best and worst ranks, with lower being better. The 17 displayed configurations follow main-table order, using Gen for UME-R1 and +T-CoT for Embed-RL. Ranks use scores before display rounding among all 20 configurations, with average ranks for ties.

Relative strengths vary with task and domain: the median span between a configuration's best and worst ranks across the six subsets is 5.5. Figure 2 illustrates why a single overall ranking cannot capture adaptation to different retrieval requirements; complete rank-agreement analyses appear in Appendix C.

Under the current input adaptations, the best dual-tower score is below the best discriminative-model score on every subset. Within the Qwen3-VL-Embedding series, 8B outperforms 2B on all six subsets. Across series or generations, however, parameter count does not directly predict performance: VLM2Vec-V2.0-2B exceeds VLM2Vec-7B on constraints and state but falls behind on abstraction. Paired reasoning analyses appear in Section 5.4.

4.3 Ranking Comparison with MMEB-V2

We compare the ranks of the 13 systems shared between the main table and MMEB-V2 on its Image subset and on the six TRIAD subsets (Table 3). The leading order on general evaluation does not fully carry over: RzenEmbed-v2 rises from third on MMEB-V2 Image to first on I2C; UME-R1 (Gen) rises from seventh to third, and Embed-RL-4B/2B (+T-CoT) rise from eighth/ninth to second/fifth. Reasoning embedders' relative strength on I2C may reflect richer query information from reasoning, helping alignment with more distant ancestor concepts.

Table 3: Rank comparison between MMEB-V2 Image and TRIAD (selected systems). Ranks are computed among the 13 systems in Table 2 shared with MMEB-V2; parentheses show positions gained or lost relative to MMEB-V2, and — denotes no change.

Model	MMEB-V2 Image	I2C	C2I	B-Fashion	B-Cars	C-Fashion	C-Cars
WeMM-9B	1	4 (↓3)	1 (—)	1 (—)	1 (—)	1 (—)	1 (—)
Qwen3-VL-8B	2	6 (↓4)	2 (—)	2 (—)	4 (↓2)	2 (—)	5 (↓3)
Rzen-v2	3	1 (↑2)	3 (—)	3 (—)	2 (↑1)	3 (—)	2 (↑1)
Qwen3-VL-2B	4	7 (↓3)	8 (↓4)	4 (—)	10 (↓6)	11 (↓7)	12 (↓8)
B-Tokens-2B	5	8 (↓3)	10 (↓5)	6 (↓1)	9 (↓4)	6 (↓1)	6 (↓1)

Multi-turn tasks also change relative order: Qwen3-VL-Embedding-2B leads B-Tokens-2B on MMEB-V2 Image and Fashion constraints but falls behind on Fashion state. Strength on general image tasks therefore cannot substitute for evaluation of cross-level and multi-turn capabilities. Complete comparisons appear in Appendix F.

4.4 Reranking Reference and Performance Headroom

We use Qwen3-VL-Reranker-8B to jointly score and reorder the Top-30 returned by Qwen3-VL-Embedding-8B, leaving the remaining candidates in place and computing mAP over the full ranking. Relative to the first-stage baseline from the same run, all six subsets improve by 0.037–0.113 mAP. Reranking exceeds the best embedding configuration on four subsets, while WeMM still leads both Cars subsets. Joint scoring improves ranking, but reaches only 41%–77% of oracle@30, which ideally places positives first within the retrieved Top-30. This reference reveals remaining headroom within the fixed retrieved set; positives outside that set still require a better first stage. Complete results and statistics appear in Appendix D.2; Appendix E.2 additionally reports a small-scale comparison of human relevance ratings and reranker scores on identical ten-candidate sets.

5 Diagnostic Analysis

Chapter 4 reports model performance across retrieval demands, but aggregate scores alone do not explain which requirements are harder to satisfy. We combine low-complexity references, typed negatives, and input ablations to examine difficulties in cross-level relations, joint constraints, and multi-turn state, then assess the effect of adding reasoning within the same checkpoint. All analyses follow the scoring protocol in Section 3.5; qualitative examples from all three tracks appear in Appendix D.4.

5.1 Distance and Direction in Cross-Level Retrieval

Cross-level retrieval becomes harder at greater hierarchical distances. Median mAP across 17 configurations is 0.702 for Self, 0.497 for one-level ancestors, and 0.251 for two-level ancestors (Figure 3a). On shared queries, two-level retrieval remains worse in all 17 configurations after removing labeled positives at the other distance. Thus, the distance gap is not solely an occupancy effect (Appendix C.6).

Retrieval direction also changes models' relative strengths. C2I performance varies with concept depth (Figure 3b), while leading I2C systems do not necessarily lead C2I: Rzen moves from rank 1 to 3, whereas WeMM rises from rank 6 to 1 (Figure 3c). Retrieving ancestor concepts from instances and retrieving instances from concepts thus place different demands on models.

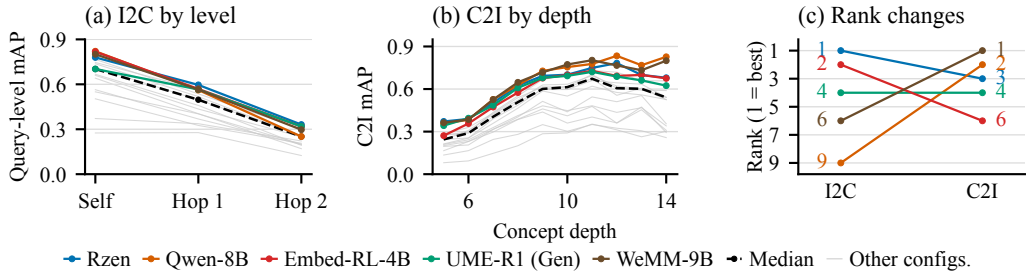

Figure 3: **Hierarchical distance and retrieval direction.** (a) mAP for Self and ancestors one and two levels away; (b) C2I mAP by concept depth; (c) I2C/C2I ranks of five representative configurations. Panels (a,b) show 17 configurations; the black dashed line is the median. Configuration selection and ranking follow Figure 2.

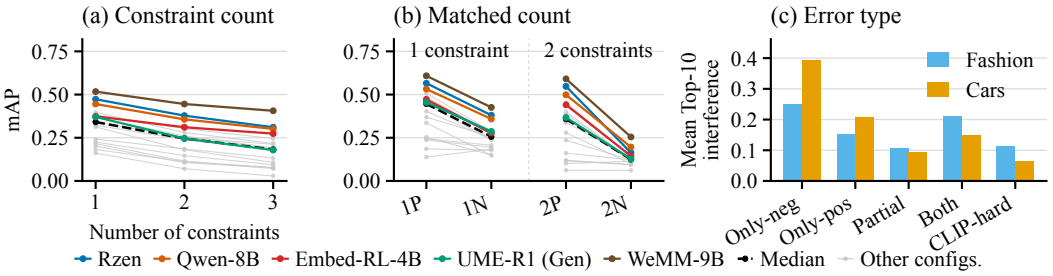

Figure 4: (a) Query-weighted mAP by constraint count; (b) matched-count positive-only and negative-only queries (P/N: positive/negative conditions). Panels (a,b) average Fashion and Cars equally per configuration; dashed black lines mark configuration medians. (c) Top-10 entry rates by negative type, normalized by that type's candidate count and averaged across configurations. The 17 configurations follow Figure 2.

5.2 Joint Constraints and Negation

Averaging mAP equally over Fashion and Cars, all 17 configurations decline as constraints increase from one to three (Figure 4a). With counts matched, negative-only queries score lower for 16 of 17 configurations in both the 1P–1N and 2P–2N comparisons (Figure 4b). The difficulty thus persists at matched counts; domain-specific and category/attribute-standardized results appear in Appendix C.3.

Typed negatives reveal how this difficulty manifests in ranking: candidates satisfying all positive conditions but violating one negative requirement have the highest Top-10 entry rate in 30 of 34 configuration–domain combinations (Figure 4c). Models still tend to prioritize candidates that resemble the reference but satisfy only part of the request.

In modality ablations, both text-only and image-only inputs underperform the full input, indicating the importance of combining both modalities to satisfy joint requirements; complete results appear in Appendix D.1.2.

5.3 Multi-Turn Information Integration and State Updates

We compare single- and multi-turn requests with the same reference item and final intent to examine performance when requirements are distributed across a conversation. mAP decreases in 39 of the 40 comparisons spanning 20 configurations and two domains (Table 4). The overall degradation persists when single-turn requests are rescored on the same state-task candidate sets (Appendix C.4). A breakdown by session type shows that degradation is concentrated in revision sessions, while multi-turn expression without revisions has a smaller overall effect, suggesting that updating superseded conditions is a major difficulty in multi-turn retrieval. Changes for revision and non-revision sessions appear in Table 14 of Appendix C.4.

Table 4: Single-to-multi-turn mAP changes for selected configurations. Δ mAP is multi-turn minus single-turn, computed from unrounded means and rounded to three decimals. UME-R1 uses Gen and Embed-RL uses +T-CoT; all 20 configurations appear in Appendix C.4.

Configuration	Fashion			Cars		
	Single	Multi	Δ mAP	Single	Multi	Δ mAP
WeMM-Embedding-9B	0.547	0.498	−0.049	0.367	0.291	−0.076
RzenEmbed-v2	0.482	0.441	−0.041	0.298	0.272	−0.026
Qwen3-VL-Embedding-8B	0.508	0.454	−0.054	0.230	0.177	−0.053
Qwen3-VL-Embedding-2B	0.442	0.212	−0.230	0.157	0.085	−0.072
Embed-RL-2B + T-CoT	0.403	0.374	−0.030	0.221	0.195	−0.026
Embed-RL-4B + T-CoT	0.360	0.376	+0.016	0.281	0.259	−0.022
UME-R1-7B (Gen)	0.359	0.235	−0.124	0.174	0.130	−0.045

Table 5: mAP before and after reasoning. Each pair reports the non-reasoning baseline followed by Δ mAP (reasoning minus non-reasoning). Full results and tests appear in Appendix C.5.

Model	Abstraction		Constraints		State	
	I2C	C2I	Fashion	Cars	Fashion	Cars
UME-R1-7B (Disc)	0.538	0.577	0.436	0.249	0.321	0.195
Δ mAP	−0.001	+0.011	−0.077	−0.075	−0.086	−0.066
Embed-RL-4B (no CoT)	0.518	0.527	0.356	0.227	0.321	0.194
Δ mAP	+0.020	+0.049	+0.004	+0.055	+0.055	+0.065

Keeping only the last turn does not resolve this difficulty: last-turn-only inputs underperform full history in all 22 configuration–domain comparisons, showing that earlier conditions remain important for recovering the current request. Yet obsolete conditions in revision sessions can also interfere with ranking. Multi-turn retrieval therefore requires not just using history, but retaining valid conditions and updating superseded values. History-truncation and revision analyses appear in Appendix D.1.1.

5.4 Effects and Limitations of Reasoning Configurations

Can explicit reasoning improve retrieval? We compare two pairs with the encoder weights held fixed: UME-R1 switches query encoding from Disc to Gen while leaving the candidate side unchanged, whereas Embed-RL-4B adds external T-CoT on both the query and candidate sides (Table 5).

The evaluated UME-R1 Gen configuration does not yield consistent improvements: the mean I2C difference is near zero, C2I improves by 0.011 mAP, and all four constraints- and state-track subsets decline. Two-sided T-CoT shows a different overall trend. Embed-RL-4B improves significantly on five subsets; its mean difference on the Fashion constraints task is +0.004, with an interval crossing zero.

6 Conclusion

We introduce TRIAD, a multimodal retrieval benchmark for evaluating systems' ability to follow requirements involving cross-level relations, joint constraints, and multi-turn state updates. Experiments reveal persistent difficulties with distant hierarchical relations, negative constraints, and multi-turn formulations, while the benefits of explicit reasoning vary across models and tasks. These findings show that semantic similarity does not imply satisfaction of the complete request; TRIAD provides a unified evaluation and diagnostic framework for studying how specific requirements can be reflected in retrieval rankings. The benchmark currently focuses on external retrieval for agentic systems; future work will extend it to retrieval and updates within their own memory systems. We have also yet to evaluate these capabilities within real agentic systems, where we plan to examine their impact on downstream task completion.

AI USE STATEMENT

We used generative AI to assist with data attribute annotation and query generation. The resulting data underwent sample-based human review, with procedures and results reported in Appendix E.1. We also used generative AI to assist with experimental code development, figure and table preparation, translation, and language editing; all such outputs were reviewed by the authors. The research conception, hypotheses, and experimental design were developed independently by the authors without generative AI assistance. The authors take responsibility for the final content of the paper and its associated research artifacts.

ETHICS STATEMENT

TRIAD is constructed from existing image datasets and WordNet. Queries and dialogues are generated or constructed for evaluation rather than collected from real user interactions. Human quality reviews and relevance ratings were conducted by a single human annotator; these activities did not require ethics approval under departmental requirements. Review procedures and results are reported in Appendix E. Generative AI assistance and human verification are disclosed in the AI Use Statement. The use and subsequent release of source data remain subject to applicable licenses and access conditions.

REPRODUCIBILITY STATEMENT

Appendix A documents the data sources, construction procedures, filtering rules, and candidate sampling methods. Appendix B describes the evaluation metrics, model configurations, runtime environments, and query and candidate input formats. Statistical methods and supplementary experimental settings are provided in Appendices C and D, while human-review and relevance-rating procedures are described in Appendix E. Availability plans for the evaluation code, dataset, and leaderboard are stated in Appendix B.4.

REFERENCES

Param Aggarwal. Fashion product images dataset. Kaggle dataset, 2019. URL https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-dataset. Accessed September 25, 2026.

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip H. S. Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29612–29622, 2025.

Mohammed Ali, Abdelrahman Abdallah, Amit Agarwal, Hitesh Laxmichand Patel, and Adam Jatowt. RECOR: Reasoning-focused multi-turn conversational retrieval benchmark. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2688–2723, 2026.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *IEEE/CVF International Conference on Computer Vision*, pp. 15338–15347, 2023.

Björn Barz and Joachim Denzler. Hierarchy-based image embeddings for semantic image retrieval. *arXiv preprint arXiv:1809.09924*, 2018.

Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Jaeseok Byun, Young Kyun Jang, Seokhyeon Jeong, Donghyun Kim, and Taesup Moon. MA-CIR: A multimodal arithmetic benchmark for composed image retrieval. In *IEEE/CVF International Conference on Computer Vision*, pp. 21342–21352, 2025.

Haonan Chen, Sicheng Gao, Radu Timofte, Tetsuya Sakai, and Zhicheng Dou. e5-omni: Explicit cross-modal alignment for omni-modal embeddings. *arXiv preprint arXiv:2601.03666*, 2026.

Yiruo Cheng, Kelong Mao, Ziliang Zhao, Guanting Dong, Hongjin Qian, Yongkang Wu, Tetsuya Sakai, Ji-Rong Wen, and Zhicheng Dou. CORAL: Benchmarking multi-turn conversational retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1308–1330, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. In *Advances in Neural Information Processing Systems*, 2018.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2023.

Yuan Hu, Zhiyu Cao, Peifeng Li, and Qiaoming Zhu. ReCQR: Incorporating conversational query rewriting to improve multimodal image retrieval. *arXiv preprint arXiv:2603.26669*, 2026.

Haohang Huang, Xuan Lu, Mingyi Su, Xuan Zhang, Ziyan Jiang, Ping Nie, Kai Zou, Tomas Pfister, Wenhu Chen, Wei Zhang, Xiaoyu Shen, and Rui Meng. MMEB-V3: Measuring the performance gaps of omni-modality embedding models. *arXiv preprint arXiv:2604.23321*, 2026.

Weijian Jian, Yajun Zhang, Dawei Liang, Chunyu Xie, Yixiao He, Dawei Leng, and Yuhui Yin. RzenEmbed: Towards comprehensive multimodal retrieval. *arXiv preprint arXiv:2510.27350*, 2025.

Haonan Jiang, Yuji Wang, Yongjie Zhu, Xin Lu, Wenyu Qin, Meng Wang, Pengfei Wan, and Yansong Tang. Embed-RL: Reinforcement learning for reasoning-driven multimodal embeddings. *arXiv preprint arXiv:2602.13823*, 2026.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024.

Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. MTRAG: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. *Transactions of the Association for Computational Linguistics*, 2025.

Andreas Koukounas, Georgios Mastrapas, Bo Wang, Mohammad Kalim Akram, Sedigheh Eslami, Michael Günther, Isabelle Mohr, Saba Sturua, Scott Martens, Nan Wang, and Han Xiao. jina-clip-v2: Multilingual multimodal embeddings for text and images. *arXiv preprint arXiv:2412.08802*, 2024.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pp. 554–561, 2013.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. UME-R1: Exploring reasoning-driven generative multimodal embeddings. *arXiv preprint arXiv:2511.00405*, 2025.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Yikun Liu, Jiangchao Yao, Weidi Xie, and Yanfeng Wang. A sanity check on composed image retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2026.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *IEEE/CVF International Conference on Computer Vision*, pp. 2125–2134, 2021.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: Can vision-language foundation models reason compositionally? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10910–10921, 2023.

Rohan Mahadev, Joyce Yuan, Patrick Poirson, David Xue, Hao-Yu Wu, and Dmitry Kislyuk. PinPoint: Evaluation of composed image retrieval with explicit negatives, multi-image queries, and paraphrase testing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. VLM2Vec-V2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025.

George A. Miller. WordNet: A lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, 2023.

Hanseok Oh, Hyunji Lee, Seonghyeon Ye, Haebin Shin, Hansol Jang, Changwook Jun, and Minjoon Seo. InstructIR: A benchmark for instruction following of information retrieval models. *arXiv preprint arXiv:2402.14334*, 2024. URL https://arxiv.org/abs/2402.14334. Also accepted at the 3rd Workshop on Knowledge-Augmented NLP, ACL 2024.

Bill Psomas, George Retsinas, Nikos Efthymiadis, Panagiotis Filntisis, Yannis Avrithis, Petros Maragos, Ondrej Chum, and Giorgos Tolias. Instance-level composed image retrieval. In *Advances in Neural Information Processing Systems*, 2025.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Shitong Sun, Jindong Gu, and Shaogang Gong. Benchmarking robustness of text-image composed retrieval. *arXiv preprint arXiv:2311.14837*, 2023.

Siyu Sun, Jing Ren, Zhaohe Liao, Dongxiao Mao, Xiangyuan Ren, Yiyi Zhang, Haohua Zhao, Weixiong Lin, Jiang Shaohua, Liqing Zhang, and Yuchao Zheng. Bottleneck tokens for unified multimodal retrieval. *arXiv preprint arXiv:2604.11095*, 2026. URL https://arxiv.org/abs/2604.11095.

Fuwen Tan, Paola Cascante-Bonilla, Xiaoxiao Guo, Hui Wu, Song Feng, and Vicente Ordonez. Drill-down: Interactive retrieval of complex scenes using natural language queries. In *Advances in Neural Information Processing Systems*, 2019.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2021.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Sagar Vaze, Nicolas Carion, and Ishan Misra. GeneCIS: A benchmark for general conditional image similarity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6862–6872, 2023.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *International Conference on Learning Representations*, 2016.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval: An empirical odyssey. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Ziwei Wang, Sameera Ramasinghe, Chenchen Xu, Julien Monteil, Loris Bazzani, and Thalaiyasingam Ajanthan. Learning visual hierarchies in hyperbolic space for image retrieval. In *IEEE/CVF International Conference on Computer Vision*, 2025.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. UniIR: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, 2024.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. FollowIR: Evaluating and teaching information retrieval models to follow instructions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, 2025.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion IQ: A new dataset towards retrieving images by natural language feedback. *arXiv preprint arXiv:1905.12794*, 2019.

Mengyao Xu, Wenfei Zhou, Yauhen Babakhin, Gabriel Moreira, Ronay Ak, Radek Osmulski, Bo Liu, Even Oldridge, and Benedikt Schifferer. Omni-Embed-Nemotron: A unified multimodal retrieval model for text, image, audio, and video. *arXiv preprint arXiv:2510.03458*, 2025.

Qijie You, Wenkai Yu, and Wentao Zhang. AgenticRAGTracer: A hop-aware benchmark for diagnosing multi-step retrieval reasoning in agentic RAG. *arXiv preprint arXiv:2602.19127*, 2026.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023.

Longxiang Zhang, Weilong Dai, Guanghao Zhang, Hao Jiang, and Pipei Huang. Think when needed: Adaptive reasoning-driven multimodal embeddings with a dual-LoRA architecture. *arXiv preprint arXiv:2605.14448*, 2026.

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. *arXiv preprint arXiv:2410.18715*, 2024.

Junjie Zhou, Ke Mei, Lei Li, Tianyi Wang, Fengyun Rao, and Jing Lyu. WeMM-Embedding: WeChat multi-modal embedding technical report. *arXiv preprint arXiv:2608.24053*, 2026.

A BENCHMARK CONSTRUCTION AND DATA

This appendix supplements the benchmark construction described in Section 3. We detail the data sources, query and positive construction, candidate sampling, and diagnostic roles for the three tracks, followed by the coverage and quality checks of the final data. Although the tracks target different requests, they follow a common procedure: determine the results that satisfy the request, construct targeted distractors around that requirement, and then fix the candidate set and complete positive set for every query.

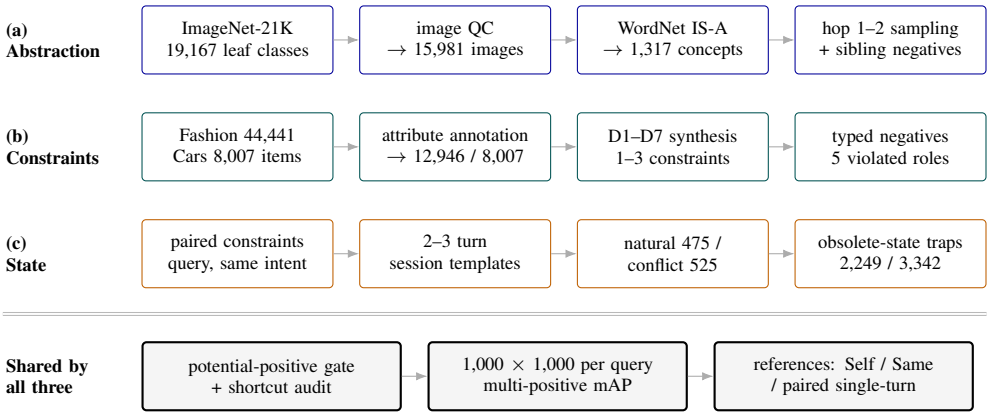

Figure 5: **Construction flow.** The abstraction, constraints, and state tracks start from hierarchical relations, joint conditions, and multi-turn state, respectively, and produce queries, positives, and typed negatives through filtering and annotation. They share potential-positive screening, a fixed 1,000-candidate evaluation protocol, and corresponding lower-complexity references.

Figure 5 summarizes the process. The abstraction track builds bidirectional retrieval between instances and concepts from ImageNet-21K images and WordNet IS-A relations. The constraints track uses product images and attributes to construct queries containing positive and negative conditions. The state track retains the final intent of a constraints-track query and rewrites it as a multi-turn dialogue with accumulation or revision. All three tracks apply image or attribute checks, potential-positive screening, and typed-negative construction.

A.1 ABSTRACTION TRACK CONSTRUCTION

Image and concept pools. The abstraction track is built from ImageNet-21K and the WordNet 3.0 noun IS-A graph. For each of the 19,167 WordNet classes covered by ImageNet-21K, we select ten candidate images, compute image–class-text similarity with Qwen3-VL-Embedding, and retain the highest-scoring image. Deduplication and a minimum-short-side threshold of 224 pixels leave 15,981 instance images.

The WordNet noun graph contains 82,115 synsets and allows multiple hypernym paths. We define hierarchical distance by the shortest hypernym path. Candidate concepts must be one or two levels above a retained leaf class, have depth at least 5, and support at least five retained descendant images. These filters yield 1,317 internal concepts and remove 36 shallow concepts. A further 34 overly broad concepts are excluded from I2C level-2 positive selection but remain in the concept pool. Records contain synset names, definitions, paths, descendants, and lemmas; concept inputs use the format `"{name}. {definition}"`.

Bidirectional queries, positives, and candidates. I2C and C2I each contain 1,000 queries with 1,000 candidates per query. I2C takes an instance image and its leaf-class name as input, with concepts one or two levels above the instance as positives. C2I takes a concept name and definition as input, with images from classes one or two levels below it as positives. Each query retains at most five positives; the remaining candidates are taxonomy-derived sibling hard negatives and random negatives.

Construction first adds positives and taxonomy hard negatives, then fills the remaining positions with random samples. During random filling, I2C excludes the image's own class and its ancestors, while C2I excludes images from strict descendants of the query concept. In the final data, I2C contains 1.77 positives and 9.05 sibling negatives per query on average. C2I fixes five positives per query and contains 150.87 sibling-descendant negatives on average.

Self reference. Self measures basic matching when no cross-level relation is required. In I2C, the target is the instance's own leaf concept; in C2I, the targets are images belonging to that leaf concept. Self uses an independent 1,000-candidate leaf-concept pool. To avoid direct string matching between the concept name and the image leaf label, the official protocol provides only the WordNet definition on the concept side. In the construction-stage I2C Self class-name shortcut check, Qwen3-VL-Embedding-8B obtains an MRR of 0.978 when the concept input includes both its name and definition, compared with 0.815 for definition-only input. The formal protocol therefore uses definition-only concept inputs to avoid direct class-name matching. Each Self query has one positive, so per-query AP equals reciprocal rank and mAP across queries equals MRR.

A.2 Constraints Track Construction

Domain corpora and attribute annotation. The constraints track uses Fashion Product Images and Stanford Cars, two product domains with distinct attribute systems. Fashion begins with 44,441 images. Qwen3.6-27B and Qwen3.5-397B-A17B independently annotate visual attributes under the v8.5 schema. The 13,345 items on which the two annotations agree enter further filtering, and 12,946 items across 21 match pools remain. Cars uses the Stanford Cars test set. Qwen3.5-27B annotates color, body type, view angle, setting, and related attributes; value normalization and programmatic checks leave 8,007 items from 196 model classes.

Attributes have three construction roles. A `constraint` attribute generates an explicitly requested modification; a `modifiable` attribute may vary within a fixed budget; and a `non_constraint_hard` attribute is not explicitly modified but specifies a property that must be preserved. A positive must satisfy all explicit conditions and the required preserved attributes. Typed negatives are generated according to whether they violate a positive condition, a negative condition, or an implicitly preserved attribute. Fashion colors and patterns additionally distinguish coarse families from fine-grained values to avoid treating near-synonyms as negatives.

Queries, positives, and difficulty profiles. Each domain contains 1,000 queries. A query provides a reference image, product text, and one to three requested changes; a candidate likewise contains an image and product attributes. The seven profiles D1–D7 correspond to 1P, 1N, 2P, 2N, 1P1N, 2P1N, and 1P2N, where P requires a value and N excludes one. Their sampling proportions are fixed during construction and are not adjusted to model results.

Query–target pairs are first generated from structured attributes and then naturalized by an LLM. A positive condition states only the target value and does not reveal the reference value; a negative condition may explicitly mention the excluded value. Ordered attributes such as length use comparative expressions such as *longer* and *shorter*. Programmatic checks verify that the query expresses all intended conditions, does not disclose the reference product's original value in positive modification requests, and preserves the requested negations.

Typed negatives and potential-positive screening. Each candidate set contains all validated positives, five diagnostic negative types, and random fillers:

Type	Operational definition
`only_neg_violation`	violates only one requested negative condition
`only_pos_violation`	violates only one requested positive condition
`partial_match`	satisfies explicit conditions but violates a required implicit property
`both_violation`	violates both positive and negative conditions
`clip_hard`	a visually similar general hard negative

The target quotas for the first four roles are 8/8/8/4 in Fashion and 5/5/3/3 in Cars; visually similar negatives and random samples fill the remaining positions. Every candidate passes a shared

potential-positive gate, so only products that fail the complete positive rule can enter as negatives. Final scoring uses every valid positive for the query rather than only the target selected during construction.

Same reference. Same measures same-item retrieval without an attribute modification. A query contains a reference image, product text, and an instruction to retrieve the same item; candidates likewise contain product images and text. Other images of the same product may therefore be positive. Same remains multimodal, multi-positive retrieval, but removes the additional attribute changes. Fashion and Cars each contain 500 queries with independently constructed 1,000-candidate pools and average positive counts of 2.042 and 5.156, respectively.

A.3 State Track Construction

Pairing with the constraints track. The state track rewrites constraints-track queries as two- or three-turn dialogues. Every multi-turn query retains the reference product, final active conditions, target product, and complete positive set of its source query, giving the two forms the same final retrieval intent. The constraints track states the complete request at once, whereas the state track requires the model to recover the currently active conditions from dialogue history.

The state track constructs a separate 1,000-candidate pool for each query. It preserves the source query's typed hard negatives and, for conflict sessions, adds revision traps that satisfy an old value but violate the final request. Mean candidate retention is 0.998565 in Fashion and 0.997076 in Cars. The tracks are therefore highly paired but not strictly same-pool, and the single-to-multi-turn difference measures the complete dialogue-history task change.

Session types and generation. The state track contains natural and conflict sessions. Natural sessions add active conditions over turns. Conflict sessions first introduce a value and then explicitly revise it. Conditions that are not revised remain active; superseded values no longer determine final relevance.

Each domain's 1,000 dialogues contain 525 conflict and 475 natural sessions: 300 two-turn conflict sessions originate from one-condition queries, 225 three-turn conflict sessions from two-condition queries, 225 two-turn natural sessions from two-condition queries, and 250 three-turn natural sessions from three-condition queries. Natural and conflict sessions with the same turn count mention the same number of attribute values, but one value in a conflict session is later superseded. All dialogues are generated by deterministic templates with seed 42 and are not paraphrased by an LLM.

Revision traps and state diagnostics. A `conflict_trap` satisfies the superseded value, violates the final revision, and satisfies the remaining active conditions. Fashion contains 2,249 conflict traps covering 459 of 525 conflict queries; Cars contains 3,342 covering 517 of 525. An `implicit_violation` satisfies explicit conditions but breaks an attribute that should remain unchanged. Other hard negatives retained from the constraints track continue to test general condition satisfaction.

For a fixed turn count, a conflict query has one fewer final active condition than a natural query. We report each session type's AP change relative to its source single-turn query to compare retrieval under history accumulation and revision; stratified results appear in Appendix C.4.

A.4 Final Dataset Statistics

A.4.1 Data Distributions

Each of the six main subsets contains 1,000 queries with 1,000 candidates each. Figure 6 combines concept categories, C2I concept depths, constraint and session composition, and condition-attribute distributions in both domains. The 1,317 internal concepts span 14 WordNet lexicographer categories; artifact, animal, and plant account for 84.43%, while C2I query concepts span depths 5–16. Fashion and Cars have identical constraint and session compositions, so each distribution is shown once as a pie chart.

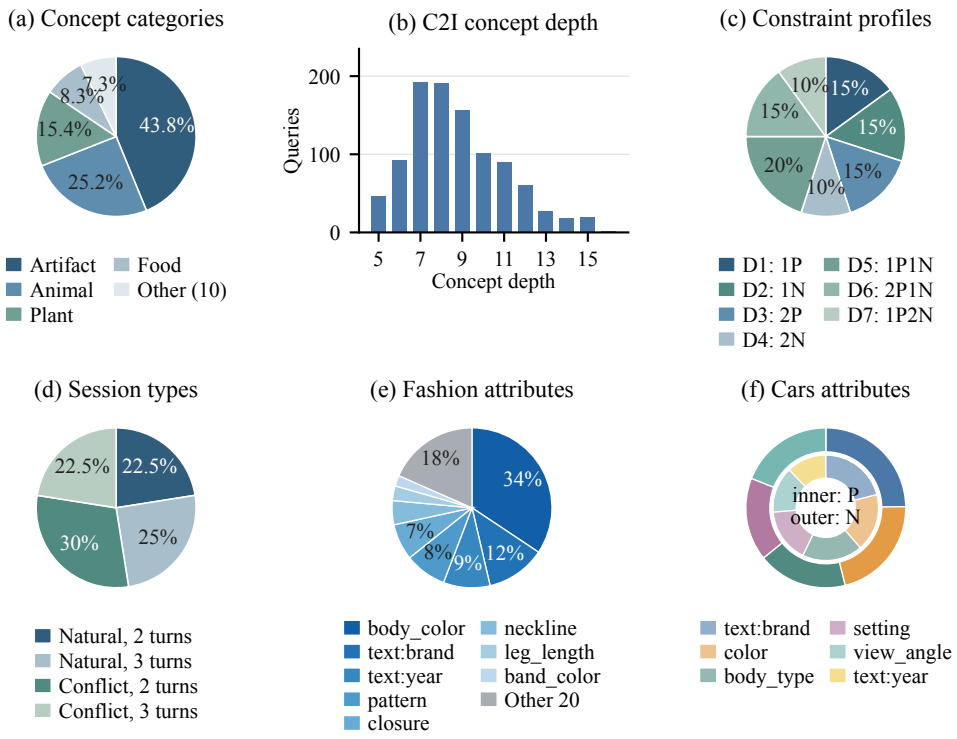

Figure 6: **Data and attribute distributions.** (a) WordNet lexicographer categories of 1,317 concepts, with ten smaller categories grouped as Other; (b) concept depths of 1,000 C2I queries; (c) D1–D7 profiles, where P/N denote positive/negative conditions; (d) natural/conflict sessions crossed with two/three turns. Panels (c,d) each use 1,000 queries per domain, with identical proportions in Fashion and Cars. (e) The 1,950 Fashion conditions: eight fields shown separately and the remaining 20 grouped; shares below 5% are unlabeled. (f) Cars: positive conditions on the inner ring (1,050), negative on the outer ring (900), with consistent field colors; text:year has no negative conditions.

Constraint-attribute coverage. The 12,946 Fashion items contain 56 category-specific fields, of which 28 generate explicit conditions; all six attributes of the 8,007 Cars items appear in queries. Each domain contains 1,950 explicit conditions: 1,050 positive and 900 negative, averaging 1.95 conditions per query. Attribute-use proportions appear in Figure 6(e,f); Table 6 adds field cardinalities and applicable-item counts.

A.4.2 Candidate-Set Statistics

Table 7 reports the mean number of labeled positives and each hard-negative (HN) type per query in the six main subsets. Means are computed over all 1,000 queries in each subset, counting a type as zero for queries in which it is absent.

Table 6: Normalized value counts and applicable-item counts of condition attributes. Use proportions appear in Figure 6(e,f), so use counts are not repeated here. Applicable-item counts reflect the category coverage of structured attributes, not their missingness; for example, heel_type is defined only for footwear. Visual attributes in this table are used to construct and verify retrieval conditions, not supplied as additional product text. Model-visible text fields are listed in Appendix B.3.

Fashion field	Values	Items	Fashion field	Values	Items
band_color	24	736	heel_type	4	593
band_material	4	736	leg_length	4	879
body_color	26	11,840	lens_color	19	370
bra_type	5	286	lens_tint	3	370
buckle_type	4	462	neckline	11	5,079
cap_type	6	103	pattern	19	6,398
closure	13	3,974	rise	2	454
collar	4	144	sleeve_length	3	4,937
dial_color	19	736	sleeve_type	3	286
dial_shape	6	736	sock_length	4	176
dress_length	3	202	strap	3	624
fold_type	4	458	strap_style	8	876
frame_color	19	370	text:brand	334	12,946
frame_shape	6	370	text:year	12	12,946

Cars field	Values	Cars field	Values
text:brand	49	setting	5
color	11	view_angle	4
body_type	7	text:year	2

Table 7: **Mean candidate counts per query.** Counts follow mutually exclusive construction labels; "–" indicates that a type is not used in a subset. Pos.: positives; Sib.: hierarchical-neighbor HN; P-V/N-V: violations of positive/negative conditions only; PN-V: both polarities violated; Imp-V: preserved-attribute violations (`partial_match`); Vis.: visually similar HN; Trap: the exclusive revision-trap type. Trap flags may also overlay existing HN types; including overlays, C–Fashion/Cars means are 2.249/3.342. The table retains original types without double counting.

	A: Abstraction		B: Constraints		C: State	
Type	I2C	C2I	Fashion	Cars	Fashion	Cars
Pos.	1.768	5.000	1.653	1.982	1.653	1.982
Sib.	9.053	150.872	–	–	–	–
P-V	–	–	4.539	3.714	4.539	3.714
N-V	–	–	2.384	2.552	2.384	2.552
PN-V	–	–	1.401	1.350	1.401	1.350
Imp-V	–	–	6.724	2.781	6.724	2.781
Vis.	–	–	5.100	4.603	5.100	4.603
Trap	–	–	–	–	1.146	2.195

A.5 Candidate Similarity and Pool Scaling

Similarity confusability. Using a fixed Qwen3-VL-Embedding-8B encoder, we compare positives, negation-only-violation negatives, and random fillers across all Fashion/Cars query–candidate pairs (Figure 7a). The similarity distributions of negation-only violations and positives overlap substantially, making these negatives more confusable with positives than random fillers.

Pool expansion. For the same 1,000 Fashion constraints queries, we construct strictly nested 1k, 3k, and eligible-full pools, retaining all original positives and typed negatives. Eligible-full is the full-corpus pool after retaining the original candidates and excluding label-matching potential posi-

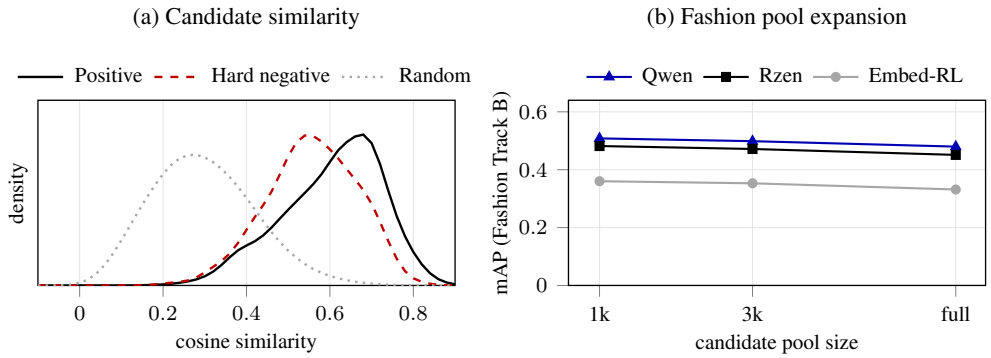

Figure 7: **Candidate similarity and pool scaling.** (a) Cosine-similarity densities pooled across Fashion/Cars under a fixed Qwen3-VL-Embedding-8B: positives, negation-only-violation negatives, and random fillers, with shared Gaussian KDE bandwidth 0.025. (b) Pool-scaling mAP for the same 1,000 Fashion constraints queries; full is the query-specific eligible corpus. Qwen, Rzen, and Embed-RL denote Qwen3-VL-Embedding-8B, RzenEmbed-v2, and Embed-RL-4B + T-CoT.

tives and the query image from added products. The 3k pool uses deterministic sampling with seed 2027; all three models share pool manifests. Eligible-full pools contain 11,956–12,945 candidates per query (mean 12,916).

For the three evaluated models on Fashion constraints, expanding the fixed pool from 1,000 to approximately 13,000 eligible candidates lowers mAP by about 0.03 without changing model ordering (Figure 7b). The main model comparison is therefore stable to this increase in candidate-set size under the present construction.

A.6 Construction Validation and Quality Audit

Each track undergoes checks aligned with its task definition. The abstraction track verifies WordNet relations, shortest-path levels, the blacklist for I2C level-2 positives, image resolution, deduplication, and random-filler rules. The constraints track rechecks positive eligibility, typed-negative roles, and the potential-positive gate against the attribute tables. The state track further verifies that final active conditions match the source single-turn query and that revision traps satisfy the old value while violating the final one.

Attribute annotation uses domain-specific quality control. Fashion retains only items on which two visual-attribute annotations agree, followed by completeness and candidate-construction checks. Cars uses one model followed by value normalization and programmatic validation. We further audit query quality and visual-attribute quality separately: 20 queries from each of the four constraints and state subsets and 40 images from each of Fashion and Cars. Human review judges 507 of 513 assessable visual attributes correct (98.83%), with eight additional attributes marked undetermined. Full procedures and results appear in Appendix E.1.

These checks cover construction rules, query expression, and sampled visual attributes, providing evidence for task consistency and annotation quality in the finalized data.

Extended design and ranking comparisons with existing benchmarks are provided in Appendix F.

B Evaluation Protocol and Implementation Details

This appendix describes the metrics, model adaptations, input organization, and runtime environments used for the main evaluation and diagnostic experiments. Per-model software versions appear in the runtime tables, and task input formats in Appendix B.3. Supplementary analyses of the main evaluation appear in Appendix C, supplementary experiments in Appendix D, and human quality review and relevance ratings in Appendix E. Three interpretive limits that apply throughout are stated once here. The evaluated panel is restricted to publicly available checkpoints, and its comparisons do not establish a performance gap between open models and closed APIs. Cross-system comparisons describe behaviour only and cannot be read causally, since the systems differ in backbone, scale,

training data, and embedding interface. The strength of a family-level claim is limited by member count: only families with at least three members including a same-source control get family-level conclusions, while single-member families yield single-point observations.

B.1 Formal Retrieval and Metric Definitions

This paper uses mean Average Precision (mAP) as the only primary metric. We define the notation as follows: C_q is the query-specific candidate set, P_q is the complete labeled positive set used for scoring, and Q is the evaluated query set; the indicator $r_{q,k}$ equals 1 if the candidate at rank k belongs to P_q, and 0 otherwise. mAP is computed as

$$\begin{gathered}\mathrm{Precision}_q@k = \frac{1}{k}\sum_{j=1}^{k} r_{q,j}, \qquad \mathrm{AP}_q = \frac{1}{|P_q|}\sum_{k=1}^{|C_q|} \mathrm{Precision}_q@k\, r_{q,k}, \\ \mathrm{mAP} = \frac{1}{|Q|}\sum_{q\in Q} \mathrm{AP}_q.\end{gathered} \tag{1}$$

The divisor of AP_q is $|P_q|$, not the ranking length or the number of positives retrieved in a Top-K prefix; multiple positives are scored together within a query rather than converted into independent single-positive trials. AP equals reciprocal rank only for genuinely single-positive controls such as Self; Same remains multi-positive and is scored over its complete positive set.

mAP is our only primary metric. We aggregate at the highest available precision and compute ranks and differences before rounding displayed values. Hard-negative entry rates, dialogue AP changes, and positive coverage@30 serve the corresponding diagnostic analyses and are not combined into an additional aggregate score. Evaluation saves complete per-query rankings for recomputation; each main-task ranking contains 1,000 distinct candidates, excludes the query image itself, and retains the complete annotated positive set.

One convention behind the main table's I2C column deserves a note: that column counts hop-1 and hop-2 ancestor concepts together in P_q over one shared candidate pool, so it is a single number that merges two abstraction distances rather than an average of per-hop results. Per-hop scoring preserves the original candidate pool and full ranking, retaining only the current hop's positives in P_q; candidates from the other hop are not removed and ranks are not compressed. The descriptive values in Section 5.1 aggregate the eligible queries separately for each hop: all 1,000 for hop-1 and the 711 with hop-2 positives for hop-2. Paired hop-1 versus hop-2 tests instead use only the 711 queries shared by both. The descriptive per-hop means and paired differences therefore use different query populations. The main-table I2C value cannot be recovered as a weighted combination of the per-hop values, and the two should not be substituted for each other in comparisons. Appendix C.6 additionally removes labeled positives at the other distance and compresses ranks to assess occupancy effects; this sensitivity analysis does not replace the main-task or original distance-specific scoring.

The benchmark assigns role labels to hard negatives: the five constraints-track types are defined in Appendix A.2, and obsolete-value traps in Appendix A.3. These labels support statistics on each type's entry into top-ranked results: for each type of constraints-track negative, top-10 interference rate is the proportion of negatives of that type ranked at most 10, and the conflict-trap top-10 rate shows whether obsolete instructions reach top-ranked results. Because the conflict-trap rate can show a floor effect once the ranking degrades, it is read across groups of one model rather than across subsets of differing overall difficulty, and it does not enter the primary metric.

The constraints and state tracks are paired by `query_id`. Each query's dialogue cost is its state-track AP minus its corresponding constraints-track AP, averaged within each session-type and turn-count group; negative values indicate degradation under multi-turn expression. AP/mAP changes are reported as signed absolute differences, without division by the baseline score. The two tracks use highly overlapping but non-identical candidate pools, so the difference describes the overall task change from a single-turn request to a multi-turn dialogue.

For scale, random-ranking mAP on the six main subsets ranges from 0.008 to 0.011.

B.2 Evaluated Models and Reasoning Configurations

This section describes the reasoning integration of the evaluated systems and their inference runtime environment. Task fields of queries and candidates are in Appendix B.3.

B.2.1 Reasoning Integration and Configuration Changes

All models encode inputs through their official code and interfaces. B-Tokens uses generative supervision during training but forms its inference-time embedding in a single forward pass, so it is listed as a discriminative rather than a reasoning embedder (Sun et al., 2026). Beyond the official implementations, configuration changes were made only for the following four items:

Modality fusion for dual-tower encoders. CLIP, SigLIP-2, and jina-clip-v2 encode images and text separately. Following MMEB's equal-weight feature addition (Jiang et al., 2024), our implementation adds the two towers' unnormalized projected features with 1:1 weights and then applies L2 normalization; queries and candidates use the same fusion function, and no fusion weight is learned on TRIAD. This is a protocol concession: the cross-family gap in the body describes the behavior of "the dual towers as evaluated plus this fusion protocol" rather than a causal effect of the architecture.

Embedding-token fallback in UME-R1's Gen mode. Gen mode asks the model to generate reasoning text on the query side and uses the last-layer representation of the generation step aligned to the `<gen_emb>` marker as the embedding. A fraction of inputs did not emit this marker, or the marker was rejected by the end-of-output check; in those cases the run falls back to the hidden state of the last usable generation step. Disc mode is unaffected and reads the representation at the `<disc_emb>` position on both query and candidate sides. In addition, UME-R1 receives a 28×28 white placeholder image on pure-text abstraction-track inputs; this is an interface adaptation and introduces no additional dataset imagery.

Substitute external reasoner for Embed-RL. The Embed-RL paper does not release the weights of its reasoner. We therefore generate T-CoT offline with the reasoner's training backbone, Qwen3-VL-8B-Instruct, and then encode with Embed-RL-2B/4B. In the original paper's Embed-RL-2B ablation, using Qwen3-VL-8B-Instruct as the reasoner scores 1.2 percentage points lower on MMEB-V2 image tasks than using the trained reasoner (Jiang et al., 2026, Table 3).

E5-Omni's encoding interfaces. Its official query interface supports only pure text, so image and image–text inputs are all encoded through the document interface, while pure-text abstraction-track items still use the query interface. These interface names do not map one-to-one to the benchmark's query and candidate roles.

UME-R1's Gen mode retains the checkpoint's sampling settings rather than deterministic decoding; inputs include the multimodal token-type metadata required by the interface. TWN loads reasoning and embedding adapters plus a routing gate on Qwen3-VL-4B-Instruct and decides per input whether to generate on either side; because it changes backbone, training, routing, and wrapper together, its difference from other configurations is not a controlled ablation.

B.2.2 Inference Runtime Environment

Embedding encoding uses one NVIDIA A100-SXM4-80GB GPU per job. External T-CoT generation jobs covered by available logs use four GPUs with four-way tensor parallelism. Table 8 consolidates model dependencies and compute precision.

Release validation covers small-sample I2C and Cars constraints tasks for 17 encoding configurations. Jina-CLIP-v2 additionally requires timm 1.0.27; E5-Omni's import chain requires audio libraries including librosa 0.11.0 and lazy_loader 0.5.

Table 8: Software dependencies and experimental precision by model. NI: not installed; –: unrecorded.

Model	PyTorch	Transformers	Flash-Attn	PEFT	NumPy	Precision
CLIP / SigLIP-2	2.6.0	4.52.4	2.7.4	NI	1.26.4	–
jina-clip-v2	2.6.0	4.46.3	2.7.4	NI	1.26.4	–
WeMM-Embedding-9B	2.6.0	5.14.1	2.7.2.post1	NI	1.26.4	BF16
E5-Omni-7B	2.6.0	5.14.1	2.7.2.post1	NI	1.26.4	BF16
RzenEmbed-v2	2.5.1	4.49.0	2.7.4.post1	NI	2.2.6	–
Qwen3-VL-Embedding-8B / 2B	2.6.0	5.14.1	2.7.2.post1	NI	1.26.4	BF16
OmniEmbed-3B	2.6.0	5.14.1	2.7.2.post1	NI	1.26.4	–
VLM2Vec-2B / 7B / V2.0-2B	2.6.0	4.52.3	2.7.2.post1	0.20.0	1.26.4	BF16
B-Tokens-2B	2.6.0	5.14.1	2.7.2.post1	NI	1.26.4	–
Embed-RL-2B / 4B	2.6.0	5.15.0.dev0	NI	0.20.0	2.2.6	BF16
UME-R1-7B (Gen/Disc)	2.6.0	4.52.4	2.7.4	NI	1.26.4	BF16
TWN-4B	2.6.0	4.57.1	2.7.4	0.18.0	1.26.4	BF16
CoT for Embed-RL (Qwen3-VL-8B-Instruct)	2.10.0	4.57.6	2.8.3	NI	2.2.6	BF16
Qwen3-VL-Reranker-8B	2.6.0	5.14.1	2.7.2.post1	NI	1.26.4	BF16

B.3 Query and Candidate Inputs

This section gives the query and doc (candidate) input organization for each track directly. Textual attributes and instructions retain their original English wording; images are supplied separately as visual inputs and are not embedded in the text strings below.

Abstraction. For I2C, the query is an instance image plus its leaf-class name and the doc is the concept text; for C2I, the query is the concept text and the doc is an instance image plus its leaf-class name. Concept text is written as `{name}. {definition}`. The archived shared and UME-R1 interfaces use the following query-side directional instructions; candidates do not receive them:

```
I2C: Retrieve the text description of the concept category this image
  ↪ and text belong to.
C2I: Retrieve images that are instances of the described concept
  ↪ category.
```

These directional instructions are supplied in two ways depending on the model interface: models whose interface exposes a dedicated instruction field (e.g. Qwen3-VL-Embedding and RzenEmbed) receive them through that field rather than inside the content text, whereas models without such a field (the dual-tower encoders and Embed-RL) prepend the same instruction to the query content as a plain-text prefix. Both routes use identical instruction strings. The constraints and state tracks carry no benchmark-level directional instruction, and each model keeps only its own outer wrapper there.

Self uses definition-only concept text: the concept name is omitted from I2C candidates and C2I queries. Its construction is described in Appendix A.1.

Constraints. The query is a reference product image, its reference product text, and a modification request; the doc is a product image and its text attributes. The query text is organized as:

```
Reference product info: {reference_attributes}. User query:
  ↪ {query_text}
```

The doc text is organized as:

```
Product info: {candidate_attributes}
```

Fashion attributes are ordered as gender, brand, year, and season; Cars uses brand and the actual year. Visual attribute labels used to define relevance are not supplied as additional product text. Qwen3-VL retains the outer instruction `Represent the user's input.`, and UME-R1 retains its Gen/Disc wrapper.

State. The doc input is identical to that in the constraints track; the query appends the complete turn-ordered history to the reference product information, organized as:

```
Reference product info: {reference_attributes}. User conversation:
[Round 1] User: {turn_1}
...
[Round T] User: {turn_T}
```

Dialogues with revisions retain both old values and updated requests instead of being replaced with the final state; the model must identify the conditions that remain active from the full input.

Same. Same retains the constraints track's image–text serialization. Fashion and Cars each use six fixed exact-same templates, assigned cyclically in construction order across their 500 queries:

```
Fashion:
Find the exact same item as shown in the picture.
I want this exact product | find it for me.
Looking for this identical piece, the very same one.
Show me this same item, not a similar one | the same.
Need to find the exact match of what's in this photo.
Find me this one | identical product, same everything.
```

```
Cars:
Find this exact same car as shown in the picture.
I want this very car | the identical one.
Looking for exactly this car, the same one.
Show me this same vehicle, not a similar one.
Need to find the exact match of the car in this photo.
Find me this one | the identical car.
```

The same input on each family's encoding path. The task inputs above are identical for every model, but each family encodes them differently. Dual-tower encoders have no native joint image–text input: image and text pass through separate towers and are combined by the pre-fixed 1:1 addition with L2 normalization, so composed queries rely on this fixed fusion rule rather than any weight tuned on TRIAD. In UME-R1's Disc mode both the query and candidate sides read the representation at `<disc_emb>`, whereas Gen mode generates reasoning text on the query side only and then takes the last-layer representation of the generation step aligned to `<gen_emb>`; candidates use Disc encoding in both configurations. Modality fusion, reasoning integration, and interface adaptations used here are described in Appendix B.2.

B.4 Code and Data Availability

The evaluation code, complete dataset, and leaderboard will be made publicly available after the double-blind review process concludes.

C Supplementary Analyses of the Main Evaluation

This appendix supplements the main results with Top-1 labeled-positive hit rates, Self and Same reference tasks, rank agreement, constraint composition, single-to-multi-turn changes and session strata, complete reasoning comparisons, and hierarchy-scoring sensitivity. Input ablations, reranking, and reliability analyses are in Appendix D; pool scaling appears in Appendix A.5. Rzen denotes RzenEmbed-v2, Qwen-8B/2B denotes Qwen3-VL-Embedding-8B/2B, and Omni denotes OmniEmbed-3B; other abbreviations and settings follow the row labels.

Top-1 labeled-positive hit rate. Top-1 hit rate is the fraction of queries whose first-ranked candidate belongs to the existing labeled positive set, with 1,000 queries per main subset. It measures first-result retrieval under the benchmark's relevance labels, not end-to-end agentic task success. Table 9 reports all 20 configurations. For example, WeMM scores 0.464 and 0.318 on the Fashion and Cars constraints tasks, and 0.411 and 0.249 on the corresponding state tasks.

Table 9: Top-1 labeled-positive hit rate on the six main subsets (higher is better). Each value is a proportion over 1,000 queries; configurations and row order follow Table 2.

Model	Abstraction		Constraints		State	
	I2C	C2I	Fashion	Cars	Fashion	Cars
CLIP ViT-L/14	0.318	0.281	0.090	0.016	0.092	0.014
SigLIP-2-so400m	0.233	0.381	0.100	0.035	0.068	0.022
jina-clip-v2	0.328	0.467	0.071	0.017	0.063	0.017
WeMM-Embedding-9B	0.490	0.687	0.464	0.318	0.411	0.249
E5-Omni-7B	0.341	0.627	0.246	0.115	0.110	0.035
RzenEmbed-v2	0.540	0.661	0.403	0.240	0.358	0.210
Qwen3-VL-Emb.-8B	0.475	0.686	0.427	0.171	0.341	0.108
Qwen3-VL-Emb.-2B	0.438	0.588	0.342	0.086	0.092	0.025
OmniEmbed-3B	0.385	0.584	0.095	0.023	0.082	0.020
VLM2Vec-V2.0-2B	0.296	0.468	0.210	0.118	0.155	0.088
VLM2Vec-2B	0.257	0.454	0.115	0.040	0.100	0.046
VLM2Vec-7B	0.360	0.493	0.172	0.083	0.150	0.072
Embed-RL-2B (no CoT)	0.416	0.541	0.254	0.124	0.206	0.101
Embed-RL-4B (no CoT)	0.494	0.617	0.281	0.176	0.233	0.128
B-Tokens-2B	0.430	0.545	0.263	0.083	0.183	0.068
UME-R1-7B (Disc)	0.510	0.636	0.347	0.168	0.213	0.114
UME-R1-7B (Gen)	0.509	0.649	0.259	0.089	0.132	0.051
TWN-4B	0.504	0.614	0.151	0.064	0.093	0.025
Embed-RL-2B + T-CoT	0.477	0.614	0.317	0.167	0.274	0.133
Embed-RL-4B + T-CoT	0.512	0.649	0.281	0.229	0.298	0.199

For query-paired comparisons, we first compute AP differences between the two settings and then use 3,000 query-level bootstrap resamples to report 95% percentile intervals for the mean difference. Paired tests use the two-sided Wilcoxon signed-rank test with Holm correction within each analysis family. The rank-biserial effect size is the difference between positive and negative rank sums divided by their sum; it ranges from -1 to 1, with the sign indicating direction. Each figure or table specifies its reference setting and marker definitions; † in the session-stratified table only indicates that the bootstrap interval excludes zero.

C.1 Results on the Self and Same Reference Tasks

Self retrieves an instance's own leaf concept, or images of that concept in the reverse direction; Same retrieves images of the same reference item without requiring attribute changes. Table 10 reports both tasks for the same 20 configurations as the main table. Their independently constructed candidate sets provide basic-matching references rather than performance upper bounds for the main tasks.

All 20 configurations score higher on both Self directions than on their corresponding cross-level main tasks. The leading model on Same varies by domain: jina-CLIP-v2 leads on Fashion, while RzenEmbed-v2 leads on Cars. Basic category matching and same-item retrieval thus yield different model rankings; further comparisons with cross-level and multi-constraint retrieval appear in Section 5.

C.2 Rank Agreement Across Subsets

Figure 8 shows model-rank agreement across the six subsets. Computed from scores before display rounding for 20 configurations, Kendall τ-b ranges from 0.505 to 0.874 across 15 subset pairs, with

Table 10: mAP on Self and Same (higher is better). Self contains 1,000 queries per direction and Same contains 500 per domain, each with 1,000 candidates. Self has one positive per query, so its mAP equals MRR. Configurations and row order follow Table 2; bold marks the best result in each column.

Configuration	Self		Same	
	I2C	C2I	Fashion	Cars
CLIP-L/14	0.372	0.374	0.246	0.265
SigLIP-2	0.275	0.429	0.184	0.311
jina-clip-v2	0.550	0.506	**0.426**	0.381
WeMM-9B	0.801	**0.809**	0.353	0.560
E5-Omni-7B	0.641	0.738	0.322	0.465
Rzen-v2	0.780	0.768	0.362	**0.682**
Qwen3-VL-8B	**0.821**	0.788	0.341	0.522
Qwen3-VL-2B	0.737	0.692	0.381	0.549
OmniEmbed-3B	0.667	0.655	0.333	0.342
VLM2Vec-V2-2B	0.564	0.552	0.212	0.488
VLM2Vec-2B	0.503	0.485	0.232	0.455
VLM2Vec-7B	0.660	0.593	0.057	0.236
Embed-RL-2B (no CoT)	0.681	0.642	0.391	0.458
Embed-RL-4B (no CoT)	0.783	0.690	0.398	0.394
B-Tokens-2B	0.710	0.665	0.339	0.515
UME-R1-7B (Disc)	0.692	0.737	0.385	0.559
UME-R1-7B (Gen)	0.702	0.741	0.393	0.561
TWN-4B	0.746	0.686	0.357	0.473
Embed-RL-2B + T-CoT	0.766	0.702	0.374	0.472
Embed-RL-4B + T-CoT	0.817	0.748	0.379	0.440

a mean of 0.659; within-track agreement averages 0.747, versus 0.637 across tracks. Rankings also vary by domain: Fashion–Cars agreement is 0.695 for constraints and 0.874 for state, while constraints–state agreement within Fashion and Cars is 0.737 and 0.853, respectively. Relative model strengths therefore vary with both retrieval demands and domains. We retain the six subset scores to expose differences that aggregation can hide; Figure 2 shows each configuration's rank changes across subsets.

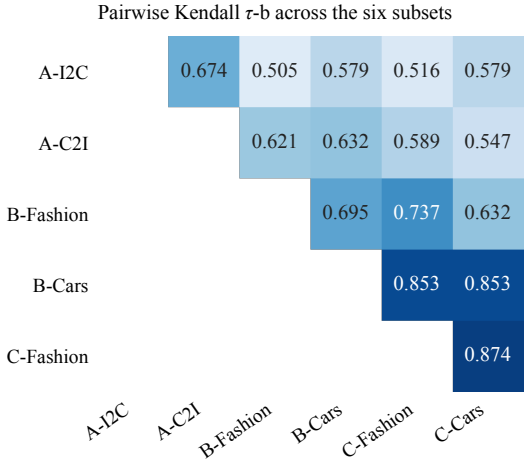

Figure 8: **Rank agreement across the six subsets (Kendall τ-b, 20 configurations, all measured).** Colour depth and cell values are proportional; a lower τ means the model order produced by one subset is less substitutable by the other. The 15 pairs span 0.505–0.874 (mean 0.659), with within- and cross-track means of 0.747 and 0.637; ranks are computed before display rounding with ties averaged.

C.3 Constraint Composition and Retrieval Performance

Table 11 reports mAP for 17 configurations across seven constraint compositions in both domains. The groups contain different queries; below, we examine constraint count, polarity, and category/attribute composition without query-paired tests between groups.

Table 11: mAP across the seven difficulty tiers in both domains. Both Embed-RL configurations use +T-CoT.

	Fashion							Cars						
Model	D1	D2	D3	D4	D5	D6	D7	D1	D2	D3	D4	D5	D6	D7
Rzen	0.68	0.49	0.62	0.25	0.45	0.50	0.23	0.45	0.27	0.47	0.08	0.27	0.30	0.12
Qwen-8B	0.69	0.50	0.66	0.31	0.46	0.51	0.32	0.38	0.22	0.34	0.08	0.20	0.20	0.11
Qwen-2B	0.70	0.41	0.62	0.19	0.41	0.42	0.19	0.32	0.16	0.18	0.06	0.13	0.13	0.07
OmniEmbed-3B	0.30	0.28	0.17	0.15	0.17	0.10	0.09	0.18	0.11	0.07	0.04	0.07	0.05	0.04
VLM2Vec-V2	0.49	0.34	0.40	0.21	0.28	0.29	0.14	0.32	0.15	0.28	0.06	0.15	0.19	0.07
VLM2Vec-2B	0.30	0.28	0.20	0.19	0.21	0.16	0.10	0.19	0.13	0.12	0.06	0.08	0.07	0.03
VLM2Vec-7B	0.43	0.21	0.36	0.15	0.22	0.22	0.11	0.24	0.09	0.20	0.05	0.10	0.12	0.05
Embed-RL-2B	0.59	0.40	0.55	0.22	0.36	0.40	0.18	0.37	0.16	0.36	0.07	0.18	0.24	0.08
Embed-RL-4B	0.54	0.35	0.44	0.20	0.33	0.36	0.22	0.40	0.21	0.44	0.08	0.27	0.31	0.15
UME-R1 Gen	0.58	0.40	0.46	0.20	0.30	0.32	0.14	0.32	0.17	0.28	0.05	0.13	0.15	0.05
B-Tokens	0.58	0.33	0.49	0.18	0.33	0.39	0.18	0.32	0.14	0.23	0.06	0.15	0.16	0.06
CLIP ViT-L/14	0.24	0.26	0.16	0.18	0.18	0.15	0.10	0.14	0.10	0.04	0.04	0.06	0.03	0.03
SigLIP-2	0.32	0.20	0.16	0.14	0.14	0.14	0.07	0.19	0.10	0.07	0.04	0.07	0.04	0.03
jina-clip-v2	0.17	0.28	0.09	0.09	0.12	0.04	0.04	0.11	0.09	0.03	0.03	0.04	0.02	0.02
TWN-4B	0.43	0.37	0.29	0.18	0.23	0.18	0.10	0.31	0.14	0.18	0.06	0.11	0.08	0.05
WeMM-9B	0.70	0.52	0.67	0.34	0.52	0.57	0.40	0.52	0.33	0.51	0.17	0.34	0.40	0.17
E5-Omni-7B	0.62	0.37	0.48	0.20	0.32	0.31	0.16	0.31	0.17	0.30	0.06	0.14	0.18	0.06

Constraint count and cross-domain aggregation. Figure 4a groups D1/D2, D3/D4/D5, and D6/D7 into one, two, and three constraints, comprising 300, 450, and 250 queries per domain, respectively. For each configuration, mAP is first query-weighted within each count group and then averaged equally over Fashion and Cars. Performance decreases at both successive count increases for all 17 configurations; the configuration medians are 0.341, 0.246, and 0.182. Figure 4b uses the same domain averaging for D1/D2 (1P/1N) and D3/D4 (2P/2N); negative-only groups score lower for 16 of 17 configurations in each comparison. Since polarity and attribute composition differ across count groups, the count curves describe difficulty across the constructed query groups rather than a paired intervention on count alone.

Count-matched and composition-standardized comparisons. To examine condition-count and attribute-composition differences, we compare D1/D2 (1P/1N), D3/D4 (2P/2N), and D3/D5 (2P/1P1N). In addition to the complete groups, we form shared strata by product category and explicit attribute-name combination, weighting each stratum by the smaller of its two group counts to equalize category–attribute composition. Negation-bearing groups score lower in 32/34, 32/34, and 28/34 combinations, respectively; after standardization, the counts are 31/34, 30/34, and 28/34. Reference items, attribute values, and candidate composition can still differ, so these comparisons assess the robustness of the observed pattern rather than isolate a causal effect of negation wording.

C.4 State Track Paired Changes

Table 13 reports single-turn and multi-turn mAP and absolute differences for all 20 main-table configurations, covering 40 configuration–domain comparisons. $\Delta\mathrm{mAP} = \mathrm{mAP}_{\mathrm{multi}} - \mathrm{mAP}_{\mathrm{single}}$, computed from unrounded means.

We use one representative configuration from each of 17 independent systems to compare the same query across the constraints and state tracks. Table 14 breaks down the 34 configuration–domain combinations by session type and turn count.

Table 12: Polarity comparisons with matched condition counts and category/attribute standardization. Differences are right minus left. We report the number of negative differences among 34 configuration–domain combinations, their median, and their range. No query-paired or significance claim is made.

Control	Left / right	Right lower	Median Δ	Range of Δ
Count-matched	1P / 1N	32/34	-0.165	$[-0.285, 0.110]$
	2P / 2N	32/34	-0.213	$[-0.423, 0.021]$
	2P / 1P1N	28/34	-0.123	$[-0.211, 0.031]$
Count + category/attributes	1P / 1N	31/34	-0.206	$[-0.326, 0.066]$
	2P / 2N	30/34	-0.217	$[-0.457, 0.035]$
	2P / 1P1N	28/34	-0.131	$[-0.265, 0.033]$

Table 13: Single-to-multi-turn changes for all 20 configurations. Single and multi denote the constraints and state tracks, respectively, with 1,000 paired queries per configuration–domain comparison. Δ mAP is multi-turn minus single-turn, rounded to three decimals. B/C candidate pools overlap highly but are not identical.

Configuration	Fashion			Cars		
	Single	Multi	Δ mAP	Single	Multi	Δ mAP
CLIP-L/14	0.185	0.180	-0.005	0.065	0.062	-0.002
SigLIP-2	0.173	0.144	-0.029	0.082	0.069	-0.013
jina-clip-v2	0.123	0.117	-0.006	0.052	0.050	-0.002
WeMM-9B	0.547	0.498	-0.049	0.367	0.291	-0.076
E5-Omni-7B	0.364	0.223	-0.142	0.186	0.101	-0.084
Rzen-v2	0.482	0.441	-0.041	0.298	0.272	-0.026
Qwen3-VL-8B	0.508	0.454	-0.054	0.230	0.177	-0.053
Qwen3-VL-2B	0.442	0.212	-0.230	0.157	0.085	-0.072
OmniEmbed-3B	0.184	0.167	-0.017	0.084	0.076	-0.008
VLM2Vec-V2-2B	0.318	0.265	-0.053	0.182	0.142	-0.040
VLM2Vec-2B	0.213	0.191	-0.022	0.101	0.092	-0.009
VLM2Vec-7B	0.251	0.221	-0.031	0.128	0.107	-0.020
Embed-RL-2B (no CoT)	0.355	0.308	-0.047	0.198	0.170	-0.028
Embed-RL-4B (no CoT)	0.356	0.321	-0.035	0.227	0.194	-0.032
B-Tokens-2B	0.368	0.300	-0.068	0.169	0.146	-0.024
UME-R1-7B (Disc)	0.436	0.321	-0.115	0.249	0.195	-0.054
UME-R1-7B (Gen)	0.359	0.235	-0.124	0.174	0.130	-0.045
TWN-4B	0.264	0.201	-0.063	0.140	0.092	-0.048
Embed-RL-2B + T-CoT	0.403	0.374	-0.030	0.221	0.195	-0.026
Embed-RL-4B + T-CoT	0.360	0.376	$+0.016$	0.281	0.259	-0.022

Revision sessions show more consistent degradation: all 34 combinations fall below their paired single-turn results in both two- and three-turn conflict groups. Non-revision sessions vary by model. RzenEmbed-v2 improves slightly in both domains' natural groups, and Embed-RL-4B + T-CoT gains 0.040–0.044 mAP in Fashion; however, WeMM-9B still loses 0.026–0.029 in Cars, and VLM2Vec-V2 and TWN also show appreciable losses. Revisions are not the only source of multi-turn cost: models must also integrate valid conditions accumulated in history. More turns do not always increase the loss; RzenEmbed-v2's Fashion conflict change is -0.106 for two turns and -0.060 for three. History length alone therefore does not characterize multi-turn difficulty.

Figure 9 relates obsolete-value interference to degradation. For the selected WeMM-9B, RzenEmbed-v2, and Qwen3-VL-Embedding-8B configurations, queries with a trap in the Top 10 degrade more in both domains than queries without one. On Cars, the two groups' AP changes are -0.153 and -0.055 for WeMM-9B, and -0.089 and -0.005 for RzenEmbed-v2. The gaps are smaller in Fashion, especially for RzenEmbed-v2. Beyond integrating valid conditions, models need to prevent superseded values from continuing to influence ranking.

Table 14: Paired AP change by session type and turn count: state minus the corresponding constraints-track result. The 17 representatives use UME-R1 Gen and Embed-RL-2B/4B + T-CoT, giving 34 configuration–domain combinations with four session groups each. † indicates a 95% query-bootstrap interval excluding zero. Natural and conflict contain different queries and are not paired against each other. The main-text 39/40 count uses all 20 configurations and has a different denominator.

Model	Domain	Natural-2	Natural-3	Conflict-2	Conflict-3
Rzen	Fashion	+0.016	+0.003	-0.106†	-0.060†
Rzen	Cars	+0.011	+0.008	-0.063†	-0.051†
Qwen-8B	Fashion	+0.012	+0.002	-0.101†	-0.117†
Qwen-8B	Cars	-0.011	-0.007	-0.091†	-0.098†
Qwen-2B	Fashion	-0.208†	-0.204†	-0.257†	-0.245†
Qwen-2B	Cars	-0.061†	-0.061†	-0.089†	-0.075†
OmniEmbed-3B	Fashion	-0.008†	-0.006†	-0.034†	-0.016†
OmniEmbed-3B	Cars	+0.004	-0.001	-0.020†	-0.009†
VLM2Vec-V2	Fashion	-0.025	-0.047†	-0.088†	-0.038†
VLM2Vec-V2	Cars	-0.011	-0.049†	-0.044†	-0.054†
VLM2Vec-2B	Fashion	-0.007	+0.011	-0.038†	-0.050†
VLM2Vec-2B	Cars	+0.004	-0.005	-0.023†	-0.007
VLM2Vec-7B	Fashion	-0.019	-0.041†	-0.027	-0.035
VLM2Vec-7B	Cars	-0.010	-0.023†	-0.024†	-0.021
Embed-RL-2B + T-CoT	Fashion	-0.002	-0.006	-0.056†	-0.048†
Embed-RL-2B + T-CoT	Cars	+0.002	-0.011	-0.032†	-0.063†
Embed-RL-4B + T-CoT	Fashion	+0.044†	+0.040†	-0.010	-0.005
Embed-RL-4B + T-CoT	Cars	+0.020†	-0.016	-0.029†	-0.061†
UME-R1 Gen	Fashion	-0.073†	-0.091†	-0.167†	-0.154†
UME-R1 Gen	Cars	-0.007	-0.017†	-0.071†	-0.077†
B-Tokens	Fashion	-0.046†	-0.075†	-0.077†	-0.069†
B-Tokens	Cars	+0.007	-0.016†	-0.037†	-0.046†
CLIP ViT-L/14	Fashion	+0.011	-0.009	-0.005	-0.017†
CLIP ViT-L/14	Cars	+0.001	-0.001	-0.007†	-0.001
SigLIP-2	Fashion	-0.016	-0.026†	-0.048†	-0.019
SigLIP-2	Cars	-0.006	-0.004	-0.021†	-0.018†
jina-clip-v2	Fashion	+0.002	-0.002†	-0.016†	-0.007†
jina-clip-v2	Cars	-2.7×10^{-6}	-0.0001†	-0.005†	-0.001†
TWN-4B	Fashion	-0.043†	-0.030†	-0.104†	-0.065†
TWN-4B	Cars	-0.031†	-0.019†	-0.081†	-0.054†
WeMM-9B	Fashion	+0.023†	-0.028	-0.082†	-0.101†
WeMM-9B	Cars	-0.026†	-0.029†	-0.134†	-0.101†
E5-Omni-7B	Fashion	-0.124†	-0.115†	-0.165†	-0.158†
E5-Omni-7B	Cars	-0.075†	-0.069†	-0.094†	-0.098†

The constraints- and state-track pools overlap highly but are not identical, so these paired differences describe overall single-to-multi-turn task changes; same-pool history-input interventions are in Appendix D.1.1.

Single- versus multi-turn retrieval with a shared candidate set. To examine the effect of candidate-set changes on multi-turn degradation, we rescore the complete single-turn requests for all 20 configurations against the corresponding state-task candidate sets and compare them with the multi-turn results from the main evaluation. Fashion and Cars each contain 1,000 paired queries, with unchanged positive sets and AP denominators. Rescored single-turn mAP is 0.0047 lower on average than in the main evaluation. With shared candidate sets, multi-turn retrieval remains worse in 38 of 40 configuration–domain comparisons, with an average difference of -0.0410 mAP. Thus, the overall multi-turn degradation persists after accounting for candidate-set changes.

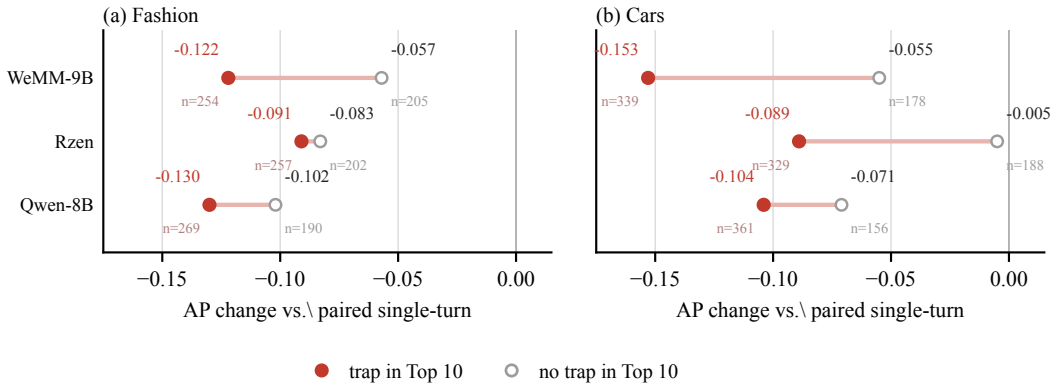

Figure 9: **Paired AP change in conflict sessions, stratified by whether the obsolete-value trap enters the Top 10 (relative to the paired constraints-track single-turn task).** Only the three strongest configurations are shown; red dots are the trap-in-Top-10 stratum and hollow grey dots the trap-free stratum, values are the stratum's mean AP change, and the small numbers below give the session count of each stratum (459 and 517 trap-constructed sessions in Fashion and Cars, respectively). The two strata are descriptive and not a randomized treatment comparison.

C.5 Complete Reasoning Comparisons

Table 15 extends the reasoning comparisons to all six main subsets and four Self/Same references. UME-R1 compares query-side Gen with Disc; Embed-RL-2B/4B compare two-sided +T-CoT with no CoT, holding each encoder's weights fixed. Holm correction is applied separately over the ten tasks in each column, consistent with the paired analysis in the main text.

Table 15: Reasoning-minus-non-reasoning Δ mAP, with 95% intervals from 3,000 query-bootstrap resamples in brackets. † marks $p < 0.05$ for the two-sided Wilcoxon test after Holm correction over ten tasks per column. Each main subset and Self direction contains 1,000 queries; Same contains 500 per domain. B/C denote the constraints and state tracks.

Task	UME-R1-7B	Embed-RL-2B	Embed-RL-4B
I2C	−0.001 [-0.006,+0.004]	+0.056† [+0.045,+0.067]	+0.020† [+0.010,+0.030]
C2I	+0.011† [+0.007,+0.016]	+0.053† [+0.045,+0.061]	+0.049† [+0.041,+0.058]
B–Fashion	−0.077† [-0.090,-0.066]	+0.048† [+0.035,+0.060]	+0.004 [-0.012,+0.020]
B–Cars	−0.075† [-0.084,-0.066]	+0.023† [+0.015,+0.032]	+0.055† [+0.043,+0.066]
C–Fashion	−0.086† [-0.098,-0.074]	+0.065† [+0.054,+0.077]	+0.055† [+0.040,+0.070]
C–Cars	−0.066† [-0.075,-0.057]	+0.025† [+0.018,+0.034]	+0.065† [+0.054,+0.075]
Self–I2C	+0.010† [+0.003,+0.018]	+0.086† [+0.070,+0.102]	+0.034† [+0.021,+0.047]
Self–C2I	+0.004 [-0.004,+0.012]	+0.061† [+0.046,+0.076]	+0.058† [+0.043,+0.072]
Same–Fashion	+0.009† [+0.001,+0.016]	−0.017† [-0.030,-0.005]	−0.019† [-0.033,-0.005]
Same–Cars	+0.003 [-0.002,+0.008]	+0.015† [+0.004,+0.025]	+0.046† [+0.035,+0.057]

C.6 Sensitivity of Hierarchy Scoring

The main task treats one- and two-level ancestors jointly as positives, whereas distance-specific analysis scores each group on the same ranking. Correct ancestors at the other distance can therefore occupy earlier positions. For each distance, we remove labeled positives at the other distance on the 711 shared queries, compress subsequent ranks, and preserve the relative order of all remaining candidates. The two-level-minus-one-level mAP difference remains negative for all 17 configurations; the median configuration-level difference narrows from -0.314 to -0.246. All 17 paired tests remain significant after Holm correction. Positive occupancy explains part, but not all, of the observed gap. No distance-selecting instruction is introduced, so this analysis does not test switching retrieval targets between one- and two-level requests. Table 16 reports every configuration; intervals use 3,000 query-bootstrap resamples and differences are two-level minus one-level.

Table 16: Hierarchy-scoring sensitivity on the same 711 queries. Original scoring retains positives at the other distance in the ranking; adjusted scoring removes only those labeled positives and compresses subsequent ranks. Δ is two-level minus one-level mAP; brackets give the adjusted difference's 95% bootstrap interval. All 17 paired Wilcoxon tests in each scoring family remain significant after Holm correction.

	Original scoring			Other-distance positives removed		
Configuration	Level 1	Level 2	Δ	Level 1	Level 2	Δ [95% CI]
CLIP ViT-L/14	0.392	0.222	-0.171	0.415	0.268	-0.148 $[-0.181, -0.115]$
SigLIP-2-so400m	0.324	0.125	-0.199	0.340	0.150	-0.191 $[-0.220, -0.161]$
jina-clip-v2	0.477	0.200	-0.277	0.498	0.252	-0.247 $[-0.277, -0.216]$
WeMM-Embedding-9B	0.644	0.297	-0.347	0.684	0.415	-0.270 $[-0.297, -0.242]$
E5-Omni-7B	0.484	0.170	-0.314	0.507	0.217	-0.290 $[-0.321, -0.259]$
RzenEmbed-v2	0.655	0.331	-0.324	0.699	0.463	-0.235 $[-0.264, -0.208]$
Qwen3-VL-Emb.-8B	0.648	0.251	-0.397	0.672	0.363	-0.309 $[-0.337, -0.281]$
Qwen3-VL-Emb.-2B	0.603	0.254	-0.349	0.636	0.348	-0.288 $[-0.317, -0.259]$
OmniEmbed-3B	0.546	0.193	-0.354	0.569	0.252	-0.316 $[-0.347, -0.286]$
VLM2Vec-V2.0-2B	0.427	0.198	-0.229	0.452	0.245	-0.208 $[-0.237, -0.179]$
VLM2Vec-2B	0.374	0.206	-0.168	0.397	0.237	-0.160 $[-0.190, -0.129]$
VLM2Vec-7B	0.502	0.249	-0.254	0.531	0.317	-0.214 $[-0.246, -0.184]$
B-Tokens-2B	0.561	0.256	-0.305	0.600	0.333	-0.267 $[-0.298, -0.239]$
UME-R1-7B (Gen)	0.634	0.318	-0.316	0.679	0.441	-0.237 $[-0.266, -0.208]$
TWN-4B	0.625	0.297	-0.328	0.661	0.414	-0.247 $[-0.276, -0.218]$
Embed-RL-2B + T-CoT	0.608	0.302	-0.306	0.648	0.406	-0.242 $[-0.273, -0.212]$
Embed-RL-4B + T-CoT	0.643	0.318	-0.325	0.687	0.440	-0.246 $[-0.275, -0.218]$

D Supplementary Experiments

This appendix uses history truncation, single-modality inputs, and image-removal controls to examine the role of inputs in history integration, joint constraints, and cross-level retrieval, and Top-30 reranking to assess recoverable ranking performance and recall limitations. Additional checks examine statistical resolution, query formulation robustness, and training-data exposure. Input ablations use full input as the reference, whereas reranking uses first-stage retrieval; candidate settings and sample scope are specified in each subsection. Evaluation protocols and implementation details appear in Appendix B; supplementary main-result analyses, model abbreviations, and statistical definitions appear in Appendix C.

D.1 Construct-Validity Ablations

We test history integration by retaining only the last turn, image–text contributions to joint constraints through single-modality inputs, and the role of images in cross-level retrieval through image removal. The three ablations use a fixed panel of 11 configurations listed in the corresponding figures and table, distinct from the 20-configuration main evaluation.

D.1.1 History Integration and Revision Sessions

The last-turn-only control preserves the state-track candidate set and representations, replacing only the full query history with the final turn. Figure 10 also includes the paired constraints-track single-turn results as a task-level reference. Both overall and within natural sessions, last-turn-only is below full history in all 22 configuration–domain combinations, showing the importance of integrating still-active conditions from history. Last-turn-only and full history use exactly the same pools; the paired single-turn pools overlap highly but are not identical.

Revision sessions show a different result: last-turn-only exceeds full history in 10 of the 22 combinations, marked with hollow red squares in Figure 10(c,f). Some configurations benefit from adding history, while others are hurt by history containing superseded values. History can thus supply information that remains valid while also allowing superseded values to influence ranking, offsetting part of the benefit from integrating history.

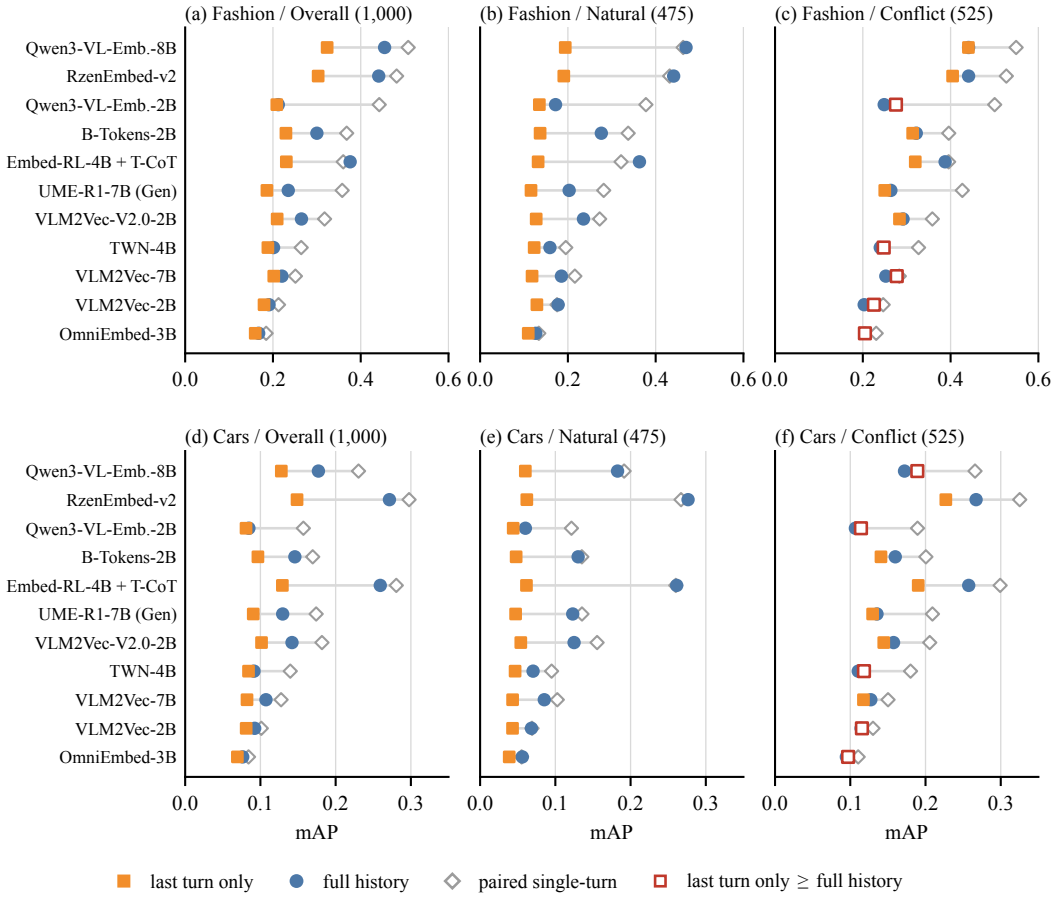

Figure 10: **Three-point ladder over last-turn-only, full-history, and paired single-turn inputs.** 11 ablation-panel configurations × 2 domains × 3 strata; every cell is mAP. The top row is Fashion and the bottom row Cars; the three columns are the overall, natural, and conflict strata, with the query count of each stratum in parentheses. Rows are ordered by the Fashion overall paired single-turn mAP and kept identical across all six panels. Last-turn-only and full-history inputs use exactly the same pool, whereas the paired single-turn reference uses a highly overlapping but non-identical constraints-track pool and is therefore a task-level reference only. Across all 22 cells of the overall and natural columns, last-turn-only is below full history; 10 cells in the conflict column reverse this and are marked with hollow red squares.

The trap-stratified results in Appendix C.4 provide a complementary observation: in the three displayed configurations, queries degrade more when candidates matching obsolete values enter the Top 10. Together with the last-turn-only control, this suggests that history both supplies informa-

tion needed for the current request and can interfere through superseded values; multi-turn retrieval requires information integration and state updating.

D.1.2 IMAGE–TEXT ABLATIONS

The single-modality ablation evaluates 11 configurations on the constraints and state tracks in both domains (44 cells, 1,000 queries). Text-only keeps the condition or dialogue text but removes the reference image and its metadata; `text+refmeta` restores textual reference metadata without the image; image-only keeps the reference image and drops the text. Models needing an image on the text path use the recorded gray-image control.

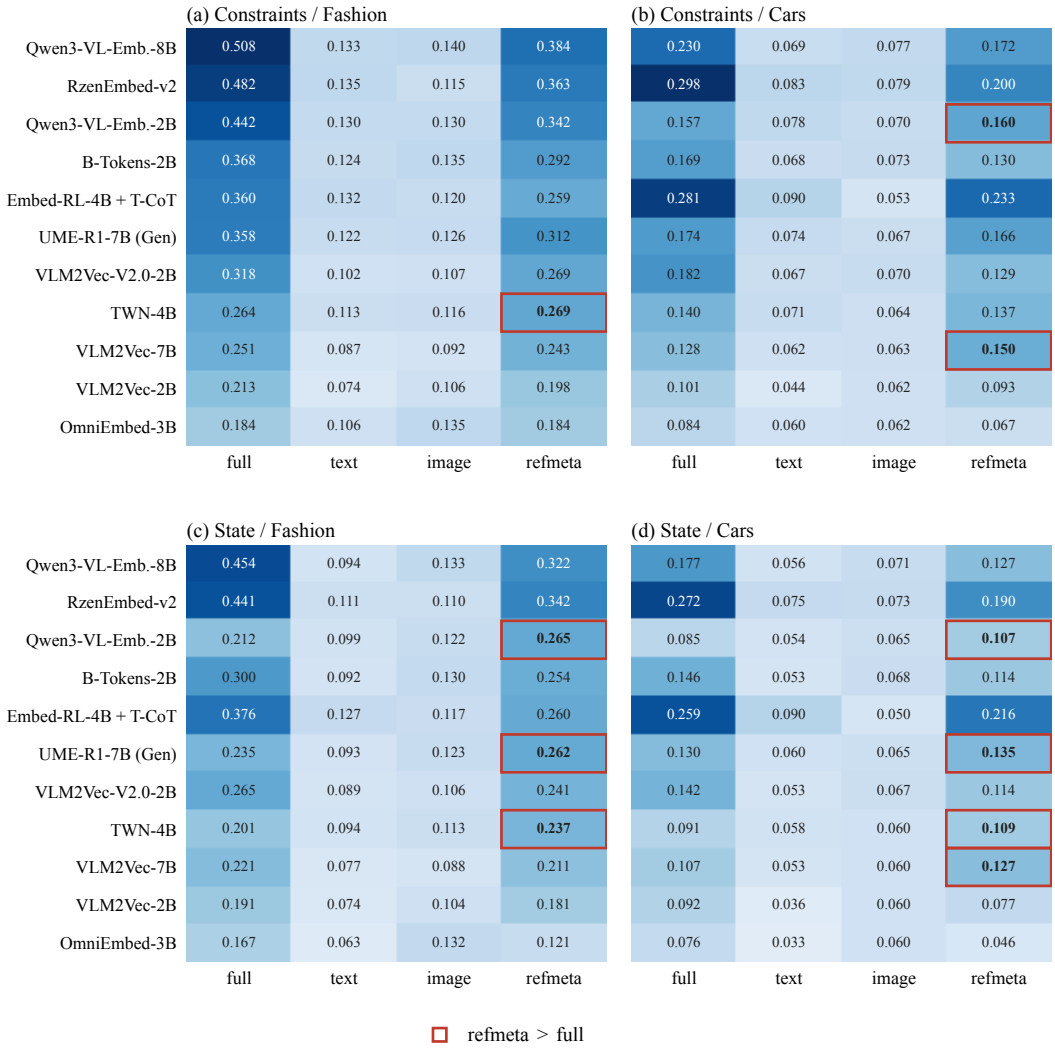

Figure 11: **Complete panel of the query-side single-modality ablation.** 11 ablation-panel configurations × 2 tracks × 2 domains × 4 input variants, i.e. 176 cells of 1,000 queries each. "full" is the complete input, "text" keeps only the condition/dialogue text, "image" keeps only the reference image, and "refmeta" removes the image but restores the reference product text attributes; candidate-side inputs are unchanged throughout. Cells show mAP; color scales are normalized within each domain and are not used for cross-domain comparisons. Red boxes mark the 10 cells where refmeta exceeds full. All four panels share the row order of (a), sorted by its full values, so a configuration's full rank within the other panels may differ. Models that must receive an image also have a gray-image control, which is not shown here.

Figure 11 reports per-cell mAP for all 11 configurations under the four input variants, and three phenomena complement one another. First, no single modality sustains high-level retrieval: text-only

and image-only are below full in all 44 cells; for example, Qwen3-VL-Embedding-8B, RzenEmbed-v2, and Embed-RL-4B + T-CoT lose 0.0433–0.1321 mAP across all 12 model–track–domain combinations when the image is removed while retaining the reference product text, and even the exceptional weaker configurations reach only 39.6%–58.2% of the corresponding subset's best full-input score. Second, the image's contribution is not always positive: `text+refmeta` exceeds full input in 10 cells, all of them in weaker configurations ranked 6–10 of 11 by full-input mAP on the corresponding subset (boxed in red). Third, the absolute loss is bounded by the baseline level: four single-modality comparisons on low-baseline Cars settings lose less than 0.026, simply because there is little room to fall. Taken together, high-level retrieval does require both modalities—no single-modality shortcut preserves it—but the gain from adding the image depends on a model's cross-modal integration: when image and text are poorly coordinated, adding the image can interfere with the representation, which is consistent with the localized gains of weaker configurations under `text+refmeta`. Note that pure text-only removes both the image and the reference product text, so it does not isolate image removal; six paired rank tests among the ten positive mean differences are significant after Holm correction, but mean differences and rank effects do not always share a direction, so these are not collectively described as significant mean-AP gains. This mechanism reading is consistent with the observations but is not a causal conclusion identified by this ablation alone.

D.1.3 The Role of Images in Cross-Level Retrieval

I2C queries provide both an image and its leaf-class name. To examine whether the class name can substitute for the image, we keep the query instruction, query-specific candidate pools, and positive labels fixed, remove only the original query image, and retain the leaf-class name. We compare this variant with the full image-and-name input across 11 configurations on the same 711 queries with positives at both one and two levels. We report overall mAP with positives at both distances scored jointly, rather than using distance-specific score differences as a measure of modality contribution.

Table 17 shows lower overall mAP after image removal in all 11 configurations, with losses of 0.018–0.105; ten configurations show significant differences after Holm correction. Qwen3-VL-Embedding-8B, RzenEmbed-v2, and Embed-RL-4B + T-CoT lose 0.054, 0.089, and 0.070 mAP, respectively. VLM2Vec-2B has the smallest loss, 0.018, which is not statistically significant.

Table 17: I2C with the leaf-class name retained and the original query image removed. Both inputs use the same 711 queries, query-specific candidate pools, and positive labels. mAP scores positives at both distances jointly; Δ is image-removed minus full input. † denotes paired Wilcoxon $p < 0.05$ after Holm correction across the overall, one-level, and two-level contrasts within each configuration. The variant retains the I2C instruction; UME-R1 uses its text interface's blank-image placeholder, without the original query image, and Embed-RL generates external reasoning from the leaf-class name. Means and differences are reported separately at archived precision, so subtraction of displayed values may differ in the last digit.

Configuration	Image + leaf name	Leaf name only	Δ mAP
Qwen3-VL-Embedding-8B	0.5523	0.4980	$-0.0543^{\dagger}$
Qwen3-VL-Embedding-2B	0.5257	0.4699	$-0.0558^{\dagger}$
OmniEmbed-3B	0.4407	0.3506	$-0.0901^{\dagger}$
RzenEmbed-v2	0.6178	0.5290	$-0.0887^{\dagger}$
B-Tokens-2B	0.5002	0.3950	$-0.1052^{\dagger}$
VLM2Vec-2B	0.3470	0.3287	-0.0182
VLM2Vec-7B	0.4578	0.4133	$-0.0445^{\dagger}$
VLM2Vec-V2.0-2B	0.3767	0.3137	$-0.0630^{\dagger}$
TWN-4B	0.5705	0.4660	$-0.1044^{\dagger}$
Embed-RL-4B + T-CoT	0.5993	0.5289	$-0.0704^{\dagger}$
UME-R1-7B (Gen)	0.5973	0.5180	$-0.0793^{\dagger}$

Class names thus provide important information for hierarchical retrieval but do not preserve full-input performance in these configurations: the original image improves I2C cross-level ranking even when the class name is given. This supports an incremental contribution from images in the

evaluated I2C settings, rather than establishing images as necessary for every abstraction retrieval request.

D.2 Reranking Reference and Recall-Limited Upper Bound

We freeze Qwen3-VL-Embedding-8B as the first-stage retriever and use Qwen3-VL-Reranker-8B to jointly score retrieved query–candidate pairs, comparing vector retrieval, actual reranking, and ideal ordering. For every query in the six main subsets, the embedder retrieves Top-30 from the 1,000-candidate pool; the reranker then receives each query–candidate pair with the main protocol's task fields and serialization and reorders only those 30 items. The abstraction track retains its direction-specific I2C/C2I instruction; the constraints and state tracks use the official generic wrapper. The official implementation runs in bfloat16 with FlashAttention-2. This pass covers 1,000 queries and 30,000 scored pairs per subset, 180,000 in total.

Full-pool AP is recomputed after splicing the reranked Top-30 back above the unchanged rank-31–1,000 tail. Positive coverage@30 is the query-average fraction of all positives contained in the first-stage Top-30. Oracle@30 places positives before negatives within the same retrieved set while preserving ranks 31–1,000, giving the ideal upper bound when only the Top-30 can be reordered. Actual reranker performance measures how much ranking quality joint scoring recovers. We recomputed all 6,000 per-query AP values from the saved rankings. The main results sort candidates by descending reranker score, retaining the original first-stage order for ties. To examine numerical sensitivity, we consider two extreme orderings within each tied group only in the sensitivity analysis: positives first and positives last. Across the six subsets, the maximum mAP deviation from the main results is 0.0031, and every subset retains a positive gain over first-stage retrieval.

Table 18 reports first-stage and reranked mAP, paired gains and statistical tests, positive coverage@30 to describe recall, and oracle@30 as the ideal ordering upper bound within that retrieved set. The leading column lists the best configuration for each subset in Table 2 as a descriptive reference; all other values come from the same full run.

Table 18: Top-30 reranking results, paired analyses, and the recall-limited oracle upper bound on all 1,000 queries per subset. Qwen3-VL-Embedding-8B performs first-stage retrieval and Qwen3-VL-Reranker-8B reranks the Top-30. All six subsets show significant improvements under two-sided Wilcoxon signed-rank tests with Holm correction ($p_{\mathrm{Holm}} < 10^{-11}$). Positive coverage@30 is the query-average fraction of positives recalled; oracle@30 ideally orders the same Top-30, providing a recall-limited upper bound. Values other than coverage are full-pool mAP or mAP differences. † denotes the best configuration for that subset in Table 2, used only as a descriptive reference; paired tests compare Qwen first-stage retrieval with its own reranked output. Human relevance ratings on ten-candidate sets appear in Appendix E.2 and do not represent a human ceiling for the full 1,000-candidate pool.

Subset	best pooled system†	Qwen embed. base	Top-30 reranked	reranked − base	positive coverage@30	oracle @30
A–I2C	0.569	0.494	0.607	+0.113	0.865	0.868
A–C2I	0.638	0.637	0.674	+0.037	0.869	0.877
B–Fashion	0.547	0.508	0.566	+0.058	0.884	0.887
B–Cars	0.367	0.230	0.327	+0.097	0.657	0.663
C–Fashion	0.498	0.454	0.516	+0.062	0.891	0.894
C–Cars	0.291	0.177	0.258	+0.081	0.616	0.623

D.3 Resolution and Reliability Checks

This section examines discrimination between models, stability under query rewording, and training-data exposure. Human reviews of queries and visual attributes are reported separately in Appendix E.1.

Statistical Resolution. A benchmark used for ranking must state its resolution: how large a difference lies outside the noise. We therefore resample the frozen query-level AP records at

$n \in \{100, 200, 300, 500, 700, 1000\}$. Figure 12a shows that the median 95% bootstrap half-width over the six subsets falls from 0.060 at $n = 100$ to 0.019 at $n = 1000$ (subset range 0.012–0.022). Figure 12b uses one representative adjacent-ranked pair per subset and asks what constant paired improvement reaches 80% power under a two-sided Wilcoxon test, with Holm correction over all six subsets. The median minimum detectable effect falls from 0.036 to 0.009; at 1,000 queries its full subset range is 0.004–0.019, with the median below the 0.012 median gap of the six selected adjacent pairs. Detectability is pair- and subset-dependent, however; this summary does not establish significance for every adjacent pair or for isolated sub-point differences. The wider marginal confidence intervals in panel (a) do not contradict panel (b): pairing removes query-level variation shared by the two configurations.

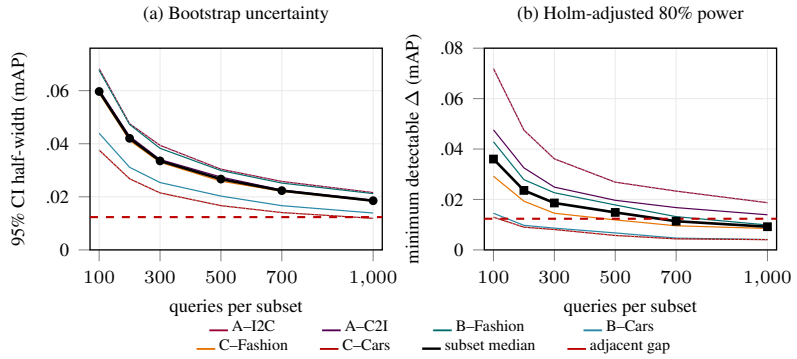

Figure 12: **Resolution of the benchmark.** (a) Median 95% query-bootstrap half-width for each subset, computed with 3,000 resamples; the black line is the median over subsets and the dotted boundaries show their range. All six subsets summarize the 20 configurations in the main table using current query-level records. (b) Minimum detectable constant paired mAP improvement at 80% power, estimated with 1,000 Monte Carlo resamples of a representative adjacent-ranked pair per subset, two-sided Wilcoxon tests, and one six-subset Holm family. The red dashed line in both panels is the median gap (0.012) of the six selected adjacent pairs.

Query Formulation Robustness. We compare three equivalent query formulations using Qwen3-VL-Embedding-8B, RzenEmbed-v2, and Embed-RL-4B + T-CoT. For the abstraction track, we re-encode all 1,000 queries in each retrieval direction using the original instruction and two human-written equivalent alternatives. For the constraints and state tracks, we fix the candidate side and change only the wording of requests or dialogues. Across all 18 model–subset/direction combinations, the mAP range over the three formulations is below 0.026, with maxima of 0.0240 for abstraction and 0.0247 for constraints and state. The largest rank shift is one, with Qwen3-VL and Rzen swapping only on the Fashion state subset under the second wording variant. The main capability conclusions remain stable under equivalent formulations, although absolute scores still vary with wording.

Training-Data Exposure. We audit official papers, model cards, and project pages for 17 independent systems, distinguishing source images (R1), source labels and ontology (R2), and TRIAD-specific query wording (R3), relation and violated-role assignments (R4), and candidate sets and protocol instances (R5; Table 19). Ten systems have confirmed use of source image datasets, and eleven involve ImageNet-1K categories or WordNet relations; dataset-level exposure does not establish that every selected TRIAD image was used in training. Eight systems partially disclose deduplication methods and nine provide no verifiable disclosure; neither supports a blanket decontamination claim.

E5-Omni explicitly uses MMEB training data, which include ImageNet-1K classification training, establishing source-image and category-label exposure (Chen et al., 2026; Jiang et al., 2024). WeMM describes public, weakly supervised web, synthetic, and internal data without an exhaustive source list; its within-batch near-duplicate handling is not benchmark-specific decontamination (Zhou et al., 2026). Unlisted sources cannot therefore be excluded.

TRIAD query wording, positive/negative role assignments, and per-query candidate combinations are newly constructed in this study, remain unpublished before submission, and were not used to train the evaluated checkpoints. R3–R5 in Table 19 therefore indicate unpublished benchmark-specific content. The relevant overlap concerns pre-existing source images, labels, and conceptual knowledge: familiarity with these materials does not imply prior access to complete TRIAD evaluation instances.

Table 19: Public-source exposure audit for 17 systems. ●: confirmed use at the source-dataset level; grey ●: cannot exclude; ○: content constructed in this study, unpublished before submission and not used to train the evaluated models. R1/R2 concern source images and labels/ontology; R3–R5 concern benchmark-specific queries, role assignments, and candidate combinations. Deduplication disclosure does not establish removal of TRIAD overlaps. Source materials for WeMM and E5-Omni were rechecked on September 20, 2026.

System	**R1**	**R2**	**R3**	**R4**	**R5**	Dedup.
CLIP ViT-L/14	●	●	○	○	○	None
SigLIP-2-so400m	●	●	○	○	○	None
jina-clip-v2	●	●	○	○	○	None
RzenEmbed-v2	●	●	○	○	○	Partial
Qwen3-VL-Embedding-8B	●	●	○	○	○	Partial
Qwen3-VL-Embedding-2B	●	●	○	○	○	Partial
OmniEmbed-3B	●	●	○	○	○	None
VLM2Vec-V2.0-2B	●	●	○	○	○	None
VLM2Vec-2B	●	●	○	○	○	None
VLM2Vec-7B	●	●	○	○	○	None
B-Tokens-2B	●	●	○	○	○	None
Embed-RL-2B	●	●	○	○	○	Partial
Embed-RL-4B	●	●	○	○	○	Partial
UME-R1-7B	●	●	○	○	○	Partial
TWN-4B	●	●	○	○	○	Partial
WeMM-9B	●	●	○	○	○	Partial
E5-Omni-7B	●	●	○	○	○	None
Confirmed inclusion	10	11	0	0	0	Partial 8 / None 9

D.4 Qualitative Examples

Figure 13 presents three examples from Qwen3-VL-Embedding-8B, illustrating ranking difficulties in cross-level retrieval, negative-constraint satisfaction, and dialogue-state updates.

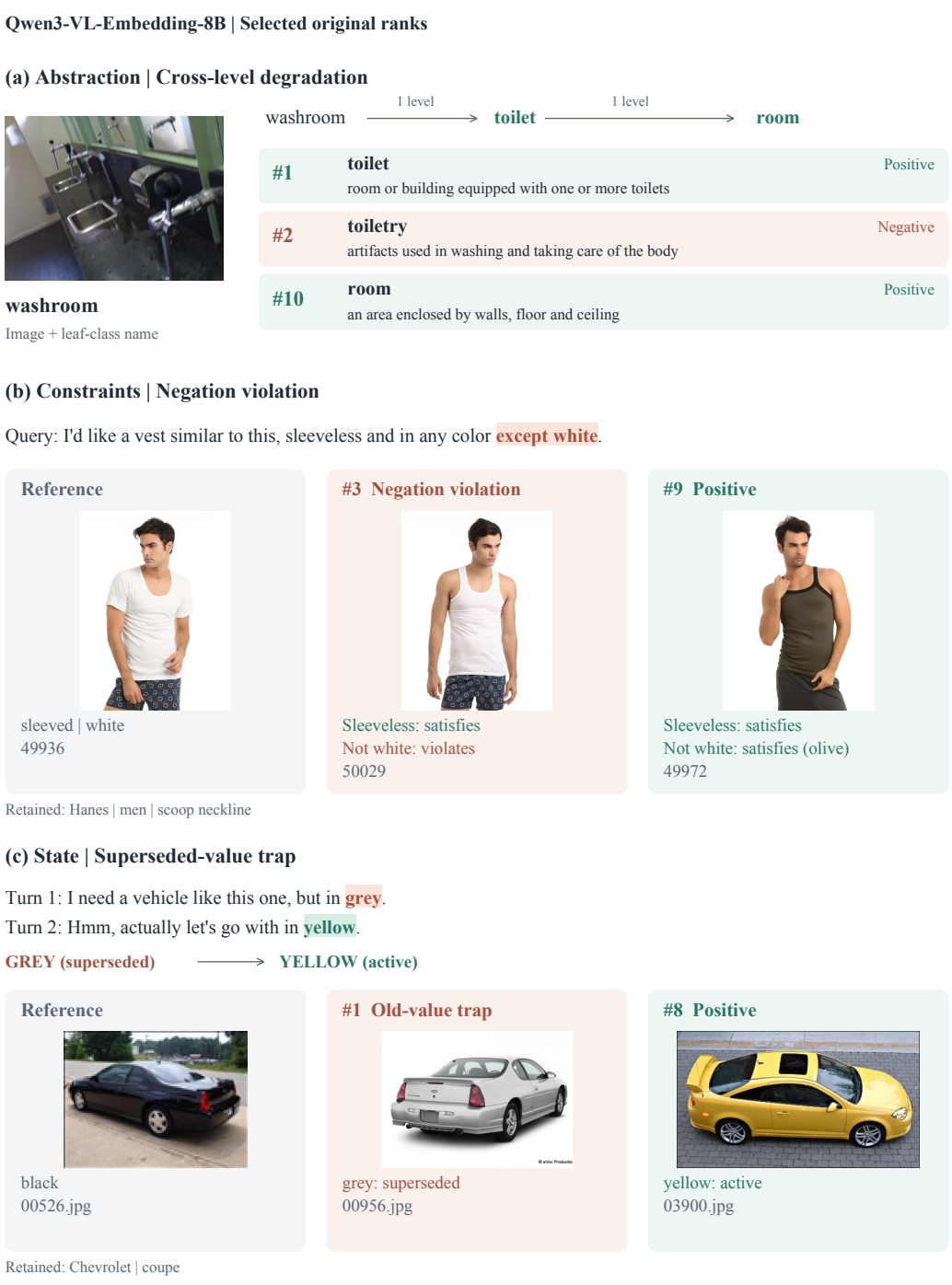

Figure 13: **Qualitative examples across the three tracks.** All panels show selected original candidate ranks from Qwen3-VL-Embedding-8B; green denotes annotated positives and red negatives. (a) Abstraction: the query contains an image and its leaf-class name. Toilet and room are ancestors one and two levels above the query; toiletry is an ordinary negative. Concept definitions are excerpted for display. (b) Constraints: the query requires sleeveless and not white. The white vest violates only the color exclusion, whereas the olive positive satisfies the full request; both retain brand, gender, and neckline. (c) State: the user revises grey to yellow. Both the old-value trap and the yellow positive retain Chevrolet and coupe. Queries and dialogue turns are reproduced verbatim; attributes follow the dataset annotations, and unshown ranks are omitted. Warm highlighting marks the exclusion phrase and superseded value; green highlighting marks the active value after revision.

Cross-level degradation. For the washroom query, the model ranks the immediate ancestor toilet first but places the next ancestor, room, tenth. The contextually related yet non-ancestral concept toiletry appears second. Here, toilet denotes a room or building equipped with toilets, not a plumbing fixture. This example shows that successful matching at a nearby level need not extend to a broader valid concept.

Negation violation. The constraints-track query asks for a similar garment that is sleeveless and not white. The white vest at rank 3 satisfies the sleeve requirement but violates the color exclusion, whereas the olive positive satisfying both appears at rank 9. Both items also satisfy the retained brand, gender, and neckline conditions. This example shows a partially compliant candidate outranking a fully compliant one: satisfying the positive condition does not compensate for violating the negative condition.

Superseded-value trap. In the state-track example, the user first requests a grey car and then changes the color to yellow. The final intent retains Chevrolet and coupe while replacing grey with yellow. Nevertheless, a trap candidate matching the superseded grey value ranks first, while the highest-ranked positive matching the final yellow value appears eighth. The ranking thus favors a withdrawn value over the active requirement, illustrating the difficulty of distinguishing historical requests from the current intent.

E Human Quality Review and Relevance Ratings

This appendix reports human sample-based reviews of query expression and visual attributes, as well as human relevance ratings on small candidate sets.

E.1 Query Quality and Visual Attribute Accuracy

Sampling and scope. We sample queries and images separately with seed 2027. Query review covers B–Fashion, B–Cars, C–Fashion, and C–Cars, with 20 queries or sessions per subset and 80 in total. Sampling balances category coverage and an approximately even allocation across D1–D7; each state-track domain contains 10 natural and 10 conflict sessions. These samples therefore examine different query types rather than estimate corpus-wide quality in proportion to the population.

Visual-attribute review samples 40 images from each of the Fashion and Cars domain pools, without duplicate sampling across B/C. Fashion covers eight coarse categories with five images each, yielding 361 attributes; Cars covers seven body types with five or six images each, yielding 160 attributes. Visual verification excludes source-dataset textual metadata, and the abstraction track's concept taxonomy is also outside this sample review. A single human annotator performs the human reviews and relevance ratings; these activities do not require ethics approval under departmental requirements.

Query review and issue classification. We examine query expression using the query text, reference image, and construction conditions, and assess how detected issues affect understanding of the requirements. Among the 80 queries or sessions, no issue is found in 67; eight raise wording or intent concerns that, on inspection, do not affect understanding, mainly involving unnatural phrasing; five may affect understanding. Table 20 summarizes the results by subset.

Table 20: Human query-quality review of 20 tasks per subset. The three mutually exclusive categories distinguish whether an issue is found and its impact on understanding the requirements.

Subset	No issue found	Concerns without impact on understanding	May affect understanding
B–Fashion	17	0	3
B–Cars	17	3	0
C–Fashion	16	2	2
C–Cars	17	3	0
Total	67	8	5

All five records that may affect understanding are from Fashion: two concern year or new/old-style wording, two concern product-category names and source metadata, and one concerns the attribute's

target part. For example, a watch session uses *not in blue* in both turns, but the construction conditions refer to the band and dial separately, without the query distinguishing these parts. In contrast, expressions such as *running cap cap type* are unnatural but remain understandable. This classification distinguishes wording issues from issues that may affect task understanding; the original queries and data labels have not been modified on this basis.

Visual attributes. For each displayed attribute annotation, we use the image to judge it correct, incorrect, or not visible/undetermined. Determinate accuracy uses correct plus incorrect judgments as its denominator; the not-visible/undetermined rate uses all attributes. Among the sampled images, 507 of 513 determinate visual attributes are annotated correctly, with eight additional attributes marked not visible/undetermined, yielding 98.83% accuracy. The not-visible/undetermined items account for 1.54% of all 521 attributes (Table 21). These results characterize visual-attribute annotation quality in the sampled images, not the accuracy of queries or candidate relevance labels.

Table 21: Human visual-attribute verification of 40 images per domain: 361 attribute judgments for Fashion and 160 for Cars. Determinate accuracy excludes not-visible/undetermined items; their rate uses all attributes as the denominator.

Domain	Correct	Incorrect	Not visible/ undetermined	Determinate accuracy	Not visible/ undetermined rate
Fashion	352	3	6	99.15%	1.66%
Cars	155	3	2	98.10%	1.25%
Total	507	6	8	98.83%	1.54%

E.2 Human Relevance Rating on Small Candidate Sets

Samples and rating setup. We select 20 tasks from each of the six main subsets, showing ten candidates per task. The annotator completes all 120 tasks, yielding 1,200 five-level ratings. Sampling uses difficulty-stratified quotas and requires archived reranker scores for all ten candidates. I2C includes 14 queries with positives both one and two levels away and six with positives only one level away. The C2I depth bins 5–7, 8–10, 11–13, and 14–16 contain 7, 9, 3, and 1 queries. Each constraints/state subset allocates 3, 3, 3, 2, 4, 3, and 2 tasks to D1–D7. Stratified quotas use seed 2027; within-stratum replacements satisfying score completeness use seed 20260909.

Each task retains all dataset gold candidates (at most five) and fills the remaining slots with the highest-ranked non-gold candidates in the reranker's natural Top 30, then shuffles their presentation. All gold candidates in the final comparison sample are recalled in the natural Top 30. The interface displays query images and text, concept names and definitions where applicable, product metadata, and full dialogues, with Chinese structured-condition aids. Model scores, original ranks, and gold labels are hidden.

Five grades and expected AP. Grades are 1 (definitely irrelevant), 2 (probably irrelevant), 3 (uncertain), 4 (probably relevant), and 5 (definitely relevant). Rankings sort grades in descending order and average uniformly over all permutations within tied grades. Grade 3 remains a middle tier rather than replacing the benchmark's binary relevance labels. For query q, let P_q denote the original gold set. For each grade g, let n_g and m_g count its candidates and gold candidates, and let N_g and M_g denote the corresponding counts in higher grades. Then

$$\mathbb{E}[\mathrm{AP}_q] = \frac{1}{|P_q|} \sum_{g:m_g>0} \sum_{j=1}^{n_g} \frac{m_g}{n_g} \frac{M_g + 1 + (j-1)(m_g-1)/(n_g-1)}{N_g + j}, \tag{2}$$

where the within-grade correction is zero for $n_g = 1$. Expected mAP averages the query-level expected AP. Across 31 tasks, 76 candidates receive grade 3 and are included according to the same ranking rule.

Same-pool results. The reranker sorts scores in descending order, breaking exact ties by base rank. Human and reranker AP use identical ten-candidate and gold sets, and all 20 tasks in each subset enter the comparison. Table 22 shows lower human mean AP on both abstraction subsets

and higher human mean AP on all four constraints/state subsets. This measures assisted judgment on small candidate sets; gold-inclusive sampling conditioned on complete recall does not establish a human upper bound for the full 1,000-candidate pool.

Table 22: Human and reranker comparison on identical ten-candidate sets. One human annotator rates all six subsets, with 20 score-complete tasks per row. Differences are human expected mAP minus reranker mAP; entries are descriptive means.

Subset	n	Human mAP	Reranker mAP	Difference
A-I2C	20	0.5512	0.6288	-0.0776
A-C2I	20	0.7280	0.7657	-0.0377
B-Fashion	20	0.7021	0.5725	$+0.1297$
B-Cars	20	0.6610	0.4369	$+0.2241$
C-Fashion	20	0.6750	0.5108	$+0.1641$
C-Cars	20	0.7343	0.4160	$+0.3183$

F Supplementary Comparisons with Existing Benchmarks

F.1 Design Criteria and Extended Comparison

Table 23 extends the main-text comparison along six dimensions. Partial indicates that an operation is covered but organized or scored differently from the criterion used here.

Table 23: Extended design comparison with related benchmarks.

Benchmark	Cross-level retrieval	Condition control	State updates	Multi-positive scoring	Typed negatives	Low-complexity reference
M-BEIR	—	Partial a	—	Partial b	—	—
MMEB	—	Partial a	Partial d	—	—	—
MMEB-V2	—	Partial a	Partial d	Partial b	—	—
CIRCO	—	Partial a	—	✓	—	—
GeneCIS	—	Partial a	—	—	✓	—
NegBench	—	✓	—	—	Partial c	✓
MA-CIR	—	✓	—	—	Partial c	✓
ChatSearch	—	Partial a	Partial d	—	—	—
TRIAD	✓	✓	✓	✓	✓	✓

The six criteria examine cross-level retrieval with positives defined by explicit ancestor/descendant relations; explicit control of condition polarity or operation composition; retention, replacement, and interference of historical conditions; ranking of multiple positives; negatives typed by violated requirements; and low-complexity references for the target operation.

Cross-level criterion. This column requires positives to be defined by ancestor/descendant relations in a category hierarchy; ordinary image–class-name matching alone does not qualify. The criterion does not require a query to select an exact hierarchical distance.

a. Condition organization. Editing subtasks in M-BEIR and the MMEB family, CIRCO, and ChatSearch may involve multiple conditions; GeneCIS distinguishes focusing on or changing objects/attributes. TRIAD further groups one to three conditions by positive/negative composition and uses preserved attributes and typed violations to analyze joint satisfaction (Wei et al., 2024; Jiang et al., 2024; Baldrati et al., 2023; Vaze et al., 2023; Zhao et al., 2024).

b. Multi-positive scoring. M-BEIR allows multiple correct targets, but its main recall convention checks for any correct hit. Some MMEB-V2 document tasks use multiple relevant evidence items and NDCG@5, whereas image/video tasks use Hit@1. CIRCO and TRIAD both use multi-positive mAP; TRIAD applies it to complete candidate rankings across all six main subsets (Wei et al., 2024; Meng et al., 2025; Baldrati et al., 2023).

c. Hard negatives and references. NegBench tests negation through affirmative/negative alternatives and object-present/absent images; MA-CIR constructs controlled negatives for negation, replacement, and addition. Both provide corresponding simplified references. GeneCIS distinguishes errors in reference content from errors in requested conditions. TRIAD organizes typed negatives and diagnostics around hierarchical relations, condition violations, and obsolete states (Alhamoud et al., 2025; Byun et al., 2025; Vaze et al., 2023).

d. Multi-turn state. MMEB's VisDial subtask uses dialogue context, and ChatSearch recovers retrieval intent from multi-turn image–text interactions. TRIAD distributes the same final requirements across turns and introduces subsequently replaced values to distinguish active-condition integration from stale-value interference (Jiang et al., 2024; Zhao et al., 2024); task coverage follows the official image-task list.

F.2 Ranking Data, Configuration Mapping, and Computation

We use official MMEB-V2 Image-Overall(V2) scores read on September 19, 2026. This aggregate covers the original 36 image tasks and excludes the subsequently added MCMR. We compare the more closely aligned image portion rather than Overall-V2, which also includes video/document tasks, and do not directly compare absolute scores under different metrics. Official leaderboard, scoring implementation.

The shared set contains 13 systems with one representative configuration each. UME-R1 uses Gen; Embed-RL-2B and Embed-RL-4B correspond to our +T-CoT configurations. The latter use external reasoning from Qwen3-VL-8B-Instruct on TRIAD, so this is a descriptive comparison of public configurations and our implementations, not a claim that every component of the reasoning pipelines is identical. The main-text reversals use only direct-encoding models; the sensitivity analysis below also excludes the two external-reasoning configurations.

Table 24 combines configuration mapping with complete rankings; differing leaderboard names are listed in its caption.

Both benchmarks rank only these 13 systems from highest to lowest score. MMEB uses the displayed two-decimal scores; TRIAD uses mAP before display rounding, with average ranks for ties. The main text displays selected rows but retains all six subsets; ranks are not recomputed over a smaller displayed set.

For model m and subset t, define rank change as Δr(m,t)=r_MMEB(m)-r_TRIAD(m,t). A positive value denotes a higher position on TRIAD and a negative value a lower position; for example, Rzen's I2C change is 3-1=+2, displayed as 1 (↑2). Spearman ρ is the Pearson correlation of the two complete rank vectors. Changes describe point-estimate rankings and do not establish statistical significance for every pairwise rank difference.

F.3 Complete Rankings and Changes

The full table preserves both agreement and reversals. WeMM remains first on five subsets, all except I2C; Qwen3-VL-Embedding-8B remains second on C2I, B-Fashion, and C-Fashion. TRIAD does not reverse the general ranking wholesale but changes relative order under some retrieval requirements. For example, Qwen3-VL-Embedding-2B leads B-Tokens on MMEB but trails it on both state tasks; E5-Omni-7B likewise leads VLM2Vec-V2.0-2B on MMEB but trails it on both state tasks. This is consistent with the need to integrate complete active requirements in multi-turn expression.

F.4 Rank Correlations and Configuration Sensitivity

Each column recomputes ranks within its own shared set rather than retaining ranks from the other column. Removing the two external-reasoning configurations raises correlations overall, but the main-text I2C reversal of Rzen against WeMM/Qwen3-VL-Embedding-8B and the single-/multi-turn reversal of Qwen3-VL-Embedding-2B against B-Tokens remain. Correlation magnitudes therefore depend on the evaluated configurations, whereas these pairwise reversals do not depend on including external reasoning.

Table 24: Complete rankings within the same 13-system set. Each cell gives rank and change relative to MMEB image ranking; B denotes constraints and C denotes state. Leaderboard mapping: RzenEmbed-v2 maps to RzenEmbed-v2-7B, B-Tokens-2B to Btoks, and E5-Omni-7B to e5-omni-7B. VLM2Vec-7B/2B map to VLM2Vec-V1-Qwen2VL-7B/2B, VLM2Vec-V2.0-2B to VLM2Vec-V2.0-Qwen2VL-2B, and OmniEmbed-3B to omni-embed-nemotron-3b. Other entries use the model names shown; UME-R1 uses Gen and Embed-RL uses +T-CoT.

Model/configuration	MMEB image	I2C	C2I	B Fashion	B Cars	C Fashion	C Cars
WeMM-Embedding-9B	1	4 (↓3)	1 (—)	1 (—)	1 (—)	1 (—)	1 (—)
Qwen3-VL-Embedding-8B	2	6 (↓4)	2 (—)	2 (—)	4 (↓2)	2 (—)	5 (↓3)
RzenEmbed-v2	3	1 (↑2)	3 (—)	3 (—)	2 (↑1)	3 (—)	2 (↑1)
Qwen3-VL-Embedding-2B	4	7 (↓3)	8 (↓4)	4 (—)	10 (↓6)	11 (↓7)	12 (↓8)
B-Tokens-2B	5	8 (↓3)	10 (↓5)	6 (↓1)	9 (↓4)	6 (↓1)	6 (↓1)
E5-Omni-7B	6	11 (↓5)	6 (—)	7 (↓1)	6 (—)	9 (↓3)	10 (↓4)
UME-R1-7B (Gen)	7	3 (↑4)	4 (↑3)	9 (↓2)	8 (↓1)	8 (↓1)	8 (↓1)
Embed-RL-4B + T-CoT	8	2 (↑6)	5 (↑3)	8 (—)	3 (↑5)	4 (↑4)	3 (↑5)
Embed-RL-2B + T-CoT	9	5 (↑4)	7 (↑2)	5 (↑4)	5 (↑4)	5 (↑4)	4 (↑5)
VLM2Vec-7B	10	9 (↑1)	11 (↓1)	11 (↓1)	11 (↓1)	10 (—)	9 (↑1)
VLM2Vec-V2.0-2B	11	12 (↓1)	12 (↓1)	10 (↑1)	7 (↑4)	7 (↑4)	7 (↑4)
VLM2Vec-2B	12	13 (↓1)	13 (↓1)	12 (—)	12 (—)	12 (—)	11 (↑1)
OmniEmbed-3B	13	10 (↑3)	9 (↑4)	13 (—)	13 (—)	13 (—)	13 (—)

Table 25: Spearman correlations between MMEB-V2 image rankings and TRIAD rankings.

TRIAD subset	Shared systems (n=13)	Without external T-CoT (n=11)
I2C	0.582	0.727
C2I	0.775	0.809
B-Fashion	0.934	0.991
B-Cars	0.681	0.800
C-Fashion	0.703	0.800
C-Cars	0.560	0.718

These results support related overall ability with complementary differences under specific requirements, not universally low correlation or fully independent abilities. We report descriptive correlations for this updated comparison and do not transfer the previous ten-system analysis's confidence intervals, p-values, or significance conclusions to the 13-system set. Explanations involving joint constraints and state updates also draw on the main diagnostics rather than treating rank shifts between two leaderboards as causal evidence.