Beyond Global Similarity: Multi-Conditional Retrieval for Fine-Grained Cross-Modal Understanding

Supplementary Material

1. Details of Baselines

For each model used in this paper, Tab. 1 summarizes the parameter size, architecture, whether we use an explicit retrieval instruction, the maximum input context length, and the underlying backbone checkpoint. We group systems into retrievers and rerankers according to their role in our two-stage pipeline.

Tab. 1: Details of retriever and reranker models used in experiments. Size denotes the number of parameters of each model. Architecture indicates the model type used in our setup (all models are operated in decoder-only mode). Instruction specifies whether we prepend a task-specific natural-language prefix when encoding queries (for example, retrieval prompts or templates that produce an `<emb>` token), rather than feeding bare queries. Max length denotes the maximum input length in tokens used for each model in our experiments, typically matching the backbone context window. Backbone identifies the pretrained checkpoint on which each retriever or reranker is built.

2. Examples of MCMR

Tab. 2 presents examples from the five MCMR sub-datasets, covering *upper clothing*, *bottom clothing*, *shoes*, *jewelry*, and *furniture*.

3. Complete Results

Tab. 3 reports the full numerical results of our query-side modality ablation study. For each of the five retrievers evaluated in our main experiments, we provide detailed scores under three query regimes: (i) fused queries containing both image-derived and text-derived constraints, (ii) text-only queries obtained by removing all image-grounded constraints, and (iii) image-only queries obtained by removing all text-grounded constraints. Each subtable reports Recall@K, MRR, and NDCG@K across a broad range of cutoffs, enabling fine-grained inspection of early-rank precision as well as long-range ordering behavior. These full results make clear how different models respond to the removal of cross-modal evidence.

Tab. 4 reports the full numerical results for the query-side compositional constraint study described in §4.3. For each retriever (CORAL, GME, LamRA, LLaVE, and MM-Embed) and for each configuration of the text- and image-derived constraint counts, we list Recall@K, MRR, and NDCG@K in percentage form. We consider matched constraint settings with $k_T{=}k_I \in \{1, 2, 3, 4, 5\}$ while keeping the fused candidate pool and evaluation protocol identical to the main experiments. The main text focuses on the trends for $k_T{=}k_I \in \{2, 3, 4, 5\}$; here we additionally include the 1T+1I configuration for completeness.

4. Prompts

4.1. Prompts for Constructing MMR

Step 1: Image-side Attribute Extraction Fig. 1 shows the prompt template used to generate fine-grained visual attribute descriptions from each product image. The model is instructed to output structured, image-grounded features only, without inferring hidden or functional properties.

Step 2: Text-side Description Generation Fig. 2 illustrates the prompt design for generating textual product descriptions from metadata. It enforces strict separation from visual evidence and standardizes phrasing for key attributes such as price and release date.

Step 3: Universal Leakage Checker Fig. 3 shows the verification prompt used to detect visual-to-text leakage after description generation. It flags any exact or paraphrased overlap between `image_feature` and `text_desc`, as well as category mismatches. This step guarantees that text-side descriptions remain strictly text-grounded before query synthesis.

Step 4: Query Generation Fig. 4 illustrates the prompt used to simulate realistic customer queries. The model combines visual and textual attributes under strict composition rules, ensuring each query naturally expresses cross-modal information while maintaining linguistic diversity and authenticity.

Step 5: Query Check. Fig. 5 presents the verification prompt that filters low-quality or unfaithful queries. The model checks for balanced modality coverage (at least two image-derived and one text-derived matches), valid price and date phrasing, and normalization issues such as unit inconsistencies or brand leakage. Only queries passing all logical gates are retained for dataset construction.

Pointwise. Fig. 6 shows the pointwise prompt used to score query–candidate pairs during reranking.

Model	Size	Architecture	Instruction	Max length	Backbone
Retriever					
GME-Qwen2-VL-7B-Instruct	7B	Decoder	Yes	32K	Qwen2-VL-7B-Instruct
LLaVE-7B	7B	Decoder	Yes	32K	LLaVA-OneVision-7B
LamRA-Ret-Qwen2.5VL-7B	7B	Decoder	Yes	128K	Qwen2.5-VL-7B
MM-Embed	8B	Decoder	Yes	32K	NV-Embed
CORAL	3B	Decoder	No	128K	Qwen2.5-3B-Instruct
VLM2VEC	4B	Decoder	No	131K	Phi3.5V
MLLM-as-Reranker					
Qwen2.5-VL-7B-Instruct	7B	Decoder	Yes	128K	Qwen2.5-VL-7B-Instruct
Qwen2.5-VL-32B-Instruct	32B	Decoder	Yes	128K	Qwen2.5-VL-32B-Instruct
InternVL3-8B-Instruct	32B	Decoder	Yes	32K	InternVL3-8B-Instruct
Qwen3-VL-4B-Instruct	4B	Decoder	Yes	262K	Qwen3-VL-4B-Instruct
Qwen3-VL-8B-Instruct	8B	Decoder	Yes	262K	Qwen3-VL-8B-Instruct

Table 1. Details of retriever and reranker models used in our experiments.

Prompt for Image Attribute Extraction

You are a meticulous vision annotator.
Task
1. Identify the product category you see in one or two lowercase words.
2. List short appearance descriptors capturing fine-grained, objective visual details of the product only.
Rules
• Output exactly one JSON array: the category first, then the descriptors.
• Aim for 4–10 descriptors; if fewer are certain, output only those (do not guess).
• Each descriptor $\leq$ 8 English words, all lowercase, American spelling.
• Describe only what is clearly visible; do not mention size, price, release date, performance or comfort claims.
• If a word implies function/technology (e.g., waterproof, breathable, thermal, cushioned), skip it unless the exact text is visibly printed.
• Focus on: colors, materials/textures, local patterns/graphics, construction/parts/edges, seams/overlays/stitching, closure/hardware (e.g., zipper, buckle, laces, clasp), shape/silhouette, logos/text when readable.
• If readable logos or text appear, include the exact lowercase text (no quotes). Do not invent brands or hallmarks.
• Exclude background or other items not part of the product (e.g., other garments, props, body parts).
• Skip any feature you are not 100% certain is visible; do not infer hidden details.
Before you output, ensure coverage of at least 4 of these slots if visible: {metal color/finish}, {stone color & shape/cut}, {setting/arrangement}, {band/chain/bracelet details}, {closure/clasp}, {engraving/motif/overlays}, {logos/text/hallmarks}.
Return only the JSON array (no prose).

Figure 1. Prompt for Image Attribute Extraction used in image-side annotation.

5. Human Evaluation Protocol

We conduct a small-scale human study to verify the naturalness and attribute fidelity of the generated queries. We randomly sample 100 target products from all domains. For each product, Annotator A is given the product image, its structured metadata, and the internally extracted fine-grained attributes, and is asked to write a natural-language search query that a user might realistically issue—without access to the model-generated query.

Annotator B then evaluates, under double-blind conditions, both the human-written and model-generated queries.

Category	Query Text	Target Text	Target Image
Upper	I'm looking for a men's jacket in gray with a plaid pattern and green accents, size L. It should be waterproof with a hood and front pockets, made of durable nylon twill that needs hand washing. Prefer something released around 2013 and priced about $200.	"title": "Columbia Men's Whirlibird III Interchange Jacket", "description": "Three jackets in one – a warm, repellent Omni-Heat liner; a waterproof-breathable and critically seam-sealed shell; and a combination of both – giving you ultimate versatility to stay warm, dry, protected and comfortable in fluctuating winter conditions.", "price": 200.0, ...	
Bottom	I'm looking for a pair of men's brown cotton jeans with a slim, straight-leg fit. I'd like a flat front and zipper fly, made from 100% cotton denim that's soft, durable, and machine washable. Perfect for casual or work wear, ideally under $30 and released around 2021.	"title": "World of Leggings Plus Size Spandex Knee High Boy Shorts - Shop 16 Colors", "description": "Our knee high and seamless plus size boy shorts are a one size seamless nylon spandex plus size leg piece that are a must for any woman's leggings wardrobe. They are made with high quality stitching and have fantastic stretch ...	
Shoes	I'm looking for men's brown leather high-top work boots with a lace-up closure and a cushioning OrthoLite footbed. They should have slip-resistant soles and meet ASTM EH safety standards for electrical hazard protection. I'd like a pair under $260.	"title": "Danner Men's Bull Run 8Ẅork Boot", "description": "Built in the USA with a dedication to quality that goes back to 1932, the Bull Run is a utilitarian work boot with a timeless design that stays in style when you punch out. The full-grain leather upper is the perfect blend of strong and soft. Combine that with the sturdy Danner Wedge outsole and you get all-day comfort that lasts.", "price": 259.95, "features": "100% Leather", ...	
Jewelry	I'm looking for a silver-tone bracelet with a Cuban link chain and slide clasp. Please show options in 14k gold over 925 sterling silver, about $450, released around 2022.	"title": "SAVEARTH DIAMONDS 5.80 Ct to 10.90 Ct Lab Created Moissanite Diamond 6MM Width Cuban Link Chain Necklace For Men In 14k Gold Over 925 Sterling Silver, 16ö to 30L̈ength, Color: G-H, Clarity: VVS1", "description": "Jewelry has the power to be this one little thing that can make you feel unique, ...", ...	
Furniture	I'm looking for some cute rustic Halloween decorations for my home, i want a tiered-tray setup with small wooden signs and fall-themed accents, under $20.	"title": "CYNOSA Halloween Decorations Halloween Tiered Tray Decor Fall Decor Hocus Pocus I Smell Children Boo Wooden Signs and Orange Plaid Gnomes Plush Farmhouse Rustic Tiered Tray Decor for Home Table", "description": "Package including: 1 x Black and Orange Check Plaid Gnome; 1 x Round Shape Sign Ï Smell Children; 1 x Square Shape Halloween Themed Sign Öctober 31; 1 x Rectangle Shape Black and Orange Plaid Sign (Boo!); ...	

Table 2. Examples of retrieval samples from our MCMR benchmark.

(a) Fused(text+image)

model	R@1	R@5	R@10	R@50	MRR	N@5	N@10	N@50
LLaVE	22.00	38.00	46.00	73.00	29.09	30.53	33.10	39.39
GME-Qwen2VL	9.00	34.00	42.00	64.00	19.50	22.32	24.91	29.81
LamRA-Qwen2.5VL	13.00	31.00	39.00	63.00	21.14	22.73	25.43	30.67
MM-EMBED	20.00	40.00	47.00	64.00	27.58	29.94	32.19	35.88
CORAL	24.00	39.00	45.00	71.00	30.67	32.22	34.11	39.86

(b) Text-only

model	R@1	R@5	R@10	R@50	MRR	N@5	N@10	N@50
LLaVE	2.00	12.00	17.00	30.00	6.54	7.40	9.02	11.98
GME-Qwen2VL	6.00	16.00	24.00	45.00	11.09	11.65	14.08	18.43
LamRA-Qwen2.5VL	2.00	8.00	15.00	27.00	4.77	4.92	7.10	9.68
MM-EMBED	12.00	29.00	33.00	47.00	18.72	20.90	22.18	25.47
CORAL	9.00	16.00	18.00	30.00	12.00	12.82	13.45	16.05

(c) Image-only

model	R@1	R@5	R@10	R@50	MRR	N@5	N@10	N@50
LLaVE	20.00	33.00	41.00	65.00	26.04	27.06	29.55	35.07
GME-Qwen2VL	13.00	25.00	33.00	56.00	18.07	18.91	21.57	26.41
LamRA-Qwen2.5VL	12.00	30.00	38.00	60.00	19.11	20.94	23.57	28.42
MM-EMBED	14.00	28.00	33.00	50.00	19.42	21.09	22.64	26.59
CORAL	10.00	31.00	37.00	54.00	18.82	21.27	23.21	26.81

Table 3. Retrieval on MCMR under three query-visibility regimes (fused / text-only / image-only). Values are percentages.

Each query is assessed on three dimensions using a 1–5 Likert scale: (i) *attribute correctness*, (ii) *cross-modal coverage*, and (iii) *naturalness and clarity*. Annotator B additionally selects which query better matches the target product.

Overall, model-generated queries achieve scores comparable to human-written ones across all criteria, with only a small gap in naturalness. The near-equal preference distribution indicates no strong annotator bias toward either source. These results confirm that the proposed generation pipeline produces high-quality, multi-condition queries suitable for cross-modal retrieval research.

model	$k_T{=}k_I$	Recall@K (%)				MRR (%)	NDCG@K (%)		
		1	5	10	50		5	10	50
CORAL	1t1i	9.00	15.00	22.00	41.00	12.38	12.30	14.61	18.84
	2t2i	15.00	31.00	38.00	57.00	22.09	23.64	25.89	30.11
	3t3i	21.00	36.00	40.00	62.00	27.25	29.06	30.34	35.10
	4t4i	26.00	42.00	48.00	68.00	33.21	34.86	36.77	41.13
	5t5i	33.00	50.00	57.00	79.00	40.19	41.94	44.19	48.93
GME	1t1i	2.00	18.00	20.00	36.00	7.92	10.22	10.87	14.32
	2t2i	10.00	23.00	28.00	46.00	15.75	17.03	18.70	22.67
	3t3i	12.00	32.00	34.00	52.00	19.43	22.32	23.01	27.13
	4t4i	18.00	39.00	42.00	59.00	25.78	28.73	29.73	33.55
	5t5i	21.00	42.00	48.00	68.00	29.40	31.93	33.87	38.36
LamRA	1t1i	5.00	12.00	21.00	42.00	8.51	8.49	11.38	15.94
	2t2i	8.00	19.00	25.00	50.00	13.06	13.90	15.90	21.38
	3t3i	8.00	24.00	31.00	53.00	14.78	16.31	18.67	23.80
	4t4i	12.00	29.00	39.00	63.00	19.48	20.81	24.08	29.41
	5t5i	19.00	35.00	44.00	70.00	26.66	27.84	30.79	36.52
LLaVE	1t1i	5.00	23.00	29.00	50.00	12.55	14.57	16.49	21.14
	2t2i	16.00	32.00	39.00	57.00	21.83	23.49	24.54	29.67
	3t3i	19.00	37.00	43.00	68.00	25.98	28.06	30.04	35.45
	4t4i	21.00	42.00	49.00	69.00	28.93	31.40	33.72	38.10
	5t5i	27.00	50.00	59.00	72.00	35.89	38.50	41.35	44.28
MM-EMBED	1t1i	8.00	18.00	20.00	40.00	12.06	13.36	13.98	18.57
	2t2i	14.00	22.00	27.00	44.00	17.31	18.03	19.55	22.09
	3t3i	19.00	29.00	34.00	49.00	22.96	23.96	25.55	28.80
	4t4i	23.00	40.00	45.00	64.00	30.04	31.96	33.65	37.83
	5t5i	30.00	43.00	48.00	68.00	36.30	37.43	39.15	43.67

Table 4. Query-side modality ablation under compositional constraints ($k_T{=}k_I \in \{1, 2, 3, 4, 5\}$). Values are percentages.

Annotator	**Background**	**English Proficiency**
Annotator A	B.A. in Linguistics	IELTS 7.0
Annotator B	M.S. in Computer Science	TOEFL 99

Table 5. Background and qualifications of annotators involved in the human evaluation study.

Metric	**Human**	**Generated**
Attribute correctness	4.45	4.37
Cross-modal coverage	4.29	4.28
Naturalness & clarity	4.48	4.33
Average score	4.41	4.33
Preference rate	49%	47%

Table 6. Summary of human evaluation comparing human-written vs. generated queries.

Text-side Description Generation

Task: Generate a concise English description for a product from its JSON metadata.
Inputs (use only text fields; the last array is a FORBIDDEN list, not content to describe): - title: title - description: description - features: features - price: price - Date First Available: date - forbidden_visuals (from image_feature; DO NOT mention or paraphrase): image
Hard rules
• Write 2–4 clear sentences, total 80–120 words, single paragraph.
• Use only these fields: title, description, features, price, Date First Available. Ignore images and any visual traits.
• Always mention price and first-availability time with strict phrasing:
– Price: strictly "priced at $X.XX". If absent: "price information not provided".
– Date: strictly "released ¡Month¿ ¡Day¿, ¡Year¿". If absent: "release date not provided".
• Do not include any visual attributes (colors, patterns, graphics, logos, slogans, characters, shape, silhouette, finish, gloss, matte, style cues).
• Treat every item in forbidden_visuals as banned; do not include those words or their paraphrases.
• Focus only on text-sourced facts: product type, use, materials, components, construction, closures, care, durability, explicit sizing, measurements, certifications, standards, packaging, warranty, origin, manufacturing, price, release date.
• Exclude shipping, service, and marketing slogans; compress long size tables into concise ranges when necessary.
• Output only the final paragraph, no notes or reasoning.
Quality gate (internal)
• Remove any overlap with forbidden_visuals or residual visual wording.
• Enforce length 80–120 words; if short and facts exist, add non-visual facts from the text (e.g., care, sizing, certification, origin).
• If fields conflict, prefer "features" over "description". Never invent facts.

Input JSON template

```
{
  "title": {title},
  "description": {description},
  "features": {features},
  "price": {price},
  "Date First Available": {date},
  "image_feature": {image}
}
```

Figure 2. Prompt template used for text-side description generation.

Prompt for Cross-Modal Leakage Detection

Return ONE valid JSON object ONLY, no prose.

Task: Check if any content in image_feature appears in text_desc as an exact phrase or a clear paraphrase or synonym. Use caseinsensitive matching. Treat hyphen and space as equivalent. Handle simple singular or plural.

Category: If the first image_feature item is a generic category , ignore it for leakage counting, but set category_conflict=true only if text_desc clearly names a different category.

Disambiguation: Do not count material specifications like 925 sterling silver" as a match for the visual tone silvertone metal". Be conservative: mark paraphrase only when meaning is the same.

Output JSON schema: "leak": true, false, "matches": ["feature": str, "type": "exact, paraphrase", "evidence": str], "category_conflict": true, false

Input: image_feature: [. . .] text_desc: ". . . "

Return JSON now.

Figure 3. Prompt template for cross-modal leakage detection.

Prompt for Query Generation

You are a real shopper typing in the search box of a large e-commerce site.
Task
Write ONE fluent search query in a natural first-person voice. Use 2-3 sentences and keep the total length between 35-60 words.
Sources
- Description : desc
- Image tags : image

Hard rules — all must be met
- Blend both sources -¿ Use exactly 2-3 visual tags and exactly 2-3 description facts, with no overlaps.
  - If fewer than 3 description facts remain after removing overlaps, use all remaining and do not invent any.
  - If a clear pattern appears in Image tags, specify the base color and the pattern's color, and use exactly one generic pattern term such as "floral print", "graphic motif", "striped pattern". Otherwise mention only the base color.
  - Mention exactly one target size only if present, never list multiple sizes or ranges.
- Price format Add one approximate price using "about" or "under" only if a price appears in the sources, never use the exact price.
- Release time format Add one broad release time using a relative or approximate expression only if a date appears in the sources, never use exact dates.
- Use first-person tone throughout.
- No details beyond the two sources.
- Output only the query text, no bullet points, quotes, or backslashes.

Materials phrasing
- When the Description lists precise fiber percentages, rewrite them into qualitative phrases with no digits and no percent sign, for example "cotton-rich", "with some polyester", "with a touch of elastane", "all-cotton", "polyester blend", "nylon-rich", "wool blend". Each such materials phrase counts as one description fact.

Style hints
- Emphasize fabric, fit, closure, care before marketing fluff.
- Vary phrasing so each query feels unique.
- Keep adjectives objective, for example "lightweight cotton jersey", not "super comfy".

Input JSON
"desc": desc,
"image_feature": image
Output
<query text only>

Figure 4. Prompt template for query generation.

Prompt for Query Quality Judge

You are a strict e-commerce query judge. Output ONE valid JSON object only, no extra text.
Inputs: query, image_feature (list of visual phrases), text_desc (metadata), candidate_price, candidate_year.
1. Coverage Check whether the query is well supported by image feature and text desc. Copy phrases from these sources; do not invent new facts. Mark coverage as OK only if you can find at least two image-based cues and one text-based cue that clearly match the query.
2. Price If the query mentions a price or budget, decide whether candidate_price is roughly consistent with that intent (not obviously too cheap or too expensive). If no clear price phrase or candidate_price is missing, treat price as in_range.
3. Date If the query mentions a year or "new this year", decide whether candidate_year and YEAR_NOW are broadly consistent. Otherwise treat date as in_range.
4. Norm flags and rewrite Set simple flags for unit mixing, brand leakage, overly generic or overly narrow queries. rewrite_minimal may contain a lightly edited version of the query that only fixes clear issues without adding new information.
Return the final JSON object now.

Figure 5. Prompt for query quality judge.

Pointwise Prompt

System.
You are a strict relevance judge. Given a user query and a candidate (image + textual attributes), answer with a single token: True or False. True means the candidate matches the query faithfully. False otherwise. No other words, no punctuation.
User. <image>
Query: {query_text} Candidate: {candidate_text} Does the candidate match the query? True or False

Figure 6. Prompt used for pointwise relevance scoring.