

Graded-Relevance Composed Multimodal Retrieval for E-commerce Visual Search at Scale

Anubhav Gupta
Walmart Global Tech
Bangalore, India
anubhav.gupta0@walmart.com

Hrushikesh Mohapatra
Walmart Global Tech
Bangalore, India
Hrushikesh.Mohapatra@walmart.com

Prijith Chandra
Walmart Global Tech
Sunnyvale, United States

Asish Mohapatra
Walmart Global Tech
Bangalore, India

Anuj Garg
Walmart Global Tech
Bangalore, India

Arvind Maan
Walmart Global Tech
Bangalore, India

Sudip Datta
Walmart Global Tech
Bangalore, India

Venkat Bulusu
Walmart Global Tech
Bangalore, India

Sitesh Kumar Jalan
Walmart Global Tech
Bangalore, India

Abstract

Visual search on large e-commerce catalogs must serve both “similarity” queries that ask for items resembling an uploaded image and “modifier” queries that comprise an image and text describing a desired modification (e.g. a color change or style swap). The latter is the setting known as *composed image retrieval* (CIR). Existing CIR methods, however, treat relevance as binary and train on triplets with a single positive target—a poor fit for real catalogs where many candidates partially satisfy a user query and ranking across that partial-match spectrum drives the customer experience. We propose a methodology for training CIR retrievers on graded relevance, consisting of: (i) a VLM to curate training data, generating both queries (object detection + modifier synthesis) and 4-level relevance labels without manual annotation, (ii) an iterative relevance-feedback loop that expands the training set by mining hard negatives from the in-training retriever, and (iii) a hierarchy-aware angular objective to train the retriever directly on the graded labels rather than collapsing them to a binary split. We call this methodology GradCIR and instantiate it on a PaliGemma2 bi-encoder trained on 3.5M graded pairs curated from raw Walmart catalog data. A controlled graded-vs-binary ablation isolates the supervision granularity and shows lift of 4.9%–5.9% in NDCG@10. The same recipe applied to other multimodal encoders lifts early-fusion backbones by up to 8.5% NDCG@10. On the public FashionIQ benchmark, GradCIR (applied to PaliGemma2) reaches 0.6703 average recall when fine-tuned, slightly ahead of the strongest peer-reviewed supervised baseline we compare against, and matching or exceeding all published CLIP-L-class zero-shot CIR methods. The system is deployed in production at Walmart, where it’s serving live visual-search user traffic.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2026/10
<https://doi.org/XXXXXXX.XXXXXXX>

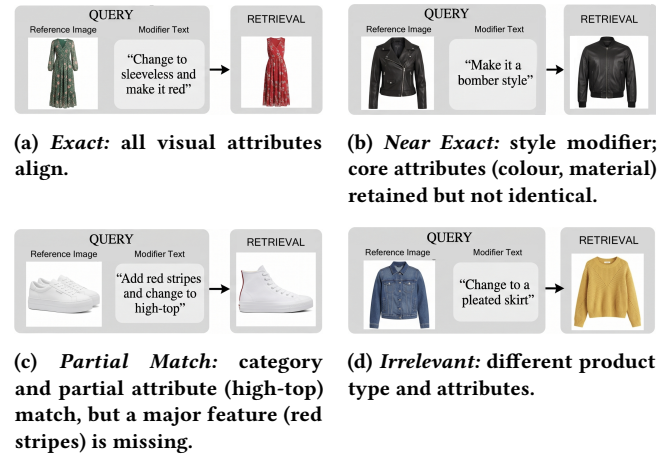


Figure 1: An example of graded CIR relevance labeling over a 4-level ordinal scale.

CCS Concepts

• Information systems → Multimedia and multimodal retrieval.

Keywords

Composed multimodal retrieval, Vision-language models, E-commerce visual search

ACM Reference Format:

Anubhav Gupta, Hrushikesh Mohapatra, Prijith Chandra, Asish Mohapatra, Anuj Garg, Arvind Maan, Sudip Datta, Venkat Bulusu, and Sitesh Kumar Jalan. 2026. Graded-Relevance Composed Multimodal Retrieval for E-commerce Visual Search at Scale. In *Proceedings of* . ACM, New York, NY, USA, 9 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Visual search is at the core of many e-commerce platforms today, allowing customers to search for products using a multimodal query consisting of an image and the accompanying text. A typical visual

search system consists of (1) a *retriever* that takes the user query and returns a set of candidate items from the catalog, and (2) a *ranker* that orders the retrieved candidates by relevance. This paper focuses on the retrieval component, which is typically realised as deep metric learning over high-dimensional embeddings: relevant (query, item) pairs are placed closer in the embedding space than irrelevant ones, and approximate nearest-neighbour (ANN) search retrieves the top- K candidates at inference time.

Many user queries, especially for categories like Fashion, have a visual aspect (colour, appearance, design) which can be difficult to express in words. Most existing systems, however, tackle only the limited use-case of “shop similar items” [17, 12, 53]. Building a robust visual search system at production scale poses several unique technical challenges. The first is *domain shift*: query images and product images come from different distributions, as catalog images are usually clean and high-resolution, whereas user-uploaded images contain noise, clutter, and irrelevant background. The second is *catalog scale*: Walmart’s catalog contains hundreds of millions of items, and retrieving tens of highly relevant items with low latency is critical to a good customer experience. The third is *relevance gradation* (Figure 1): in a real catalog, candidates are rarely cleanly “relevant or irrelevant”. For a query of a grey velvet 3-seater sofa, an identical product is an exact match; the same sofa in beige is a strong near match; a generic grey 3-seater is a partial match; a velvet armchair is irrelevant. Collapsing these four cases into a binary labeling scheme discards the information that makes a retrieval ranking useful.

The dominant line of work on composed image retrieval (CIR) [4, 31, 2, 13, 51, 20, 50] optimizes for the binary case: triplets of (reference image, modifier text, target image) where exactly one target is correct. This works well for the academic CIR benchmarks but is a poor fit for production retrieval over a real catalog, where the ranking quality across a partial-match spectrum is what drives customer engagement. This paper presents *GradCIR*, a methodology for training composed-multimodal retrievers on graded ordinal relevance. The methodology is backbone-agnostic: in the experiments we instantiate it on PaliGemma2 (referred to as *GradCIR-PG2* subsequently) and additionally apply the same recipe to four other contemporary multimodal encoders to confirm that the gains are properties of the methodology rather than of any particular backbone. The main contributions of the paper are:

- (1) A **graded-relevance formulation of CIR** over a 4-level ordinal scale, trained with a hierarchy-aware angular objective. To our knowledge, all prior CIR work collapses relevance to a single positive target. A controlled ablation (§4.5.1) shows that finer grading produces monotonically higher NDCG with backbone, data, and loss held fixed.
- (2) A **manual-annotation-free, VLM-curated data pipeline** that drives both query generation and graded relevance labeling end-to-end. This is coupled with a relevance-feedback loop that grows the training set by mining hard negatives from the in-training retriever. The pipeline produces 3.5M graded pairs on the Walmart catalog.
- (3) **Competitive public-benchmark results.** On FashionIQ, GradCIR-PG2 reaches Avg(R@10, R@50)=0.6703, slightly ahead

of the strongest peer-reviewed supervised baseline we compare against (SPN4CIR, 0.6641), in supervised setting. In zero-shot setting, it matches or exceeds published CLIP-L-class zero-shot CIR methods without any CIR-triplet supervision.

The remainder of the paper is organised as follows. Section 2 discusses related work on visual search, VLM-based embeddings, and composed (multimodal) retrieval. Section 3 describes our proposed data-creation pipeline, model architecture, and graded-relevance training objective. Section 4 reports results on public benchmarks and our internal Walmart Visual Search Test set, followed by ablations and deployment details. Section 5 discusses scope and limitations of our work. Section 6 concludes with a summary and future directions.

2 Related Work

2.1 Visual Search and Multimodal Embeddings

Production visual-search systems at e-commerce scale [17, 12, 53, 5, 47, 48, 35] typically formulate retrieval as image-similarity matching, learning an embedding space over (query image, catalog image) pairs labelled as same-product or not. Two limitations carry over to all of them: relevance is binary, and the query is image-only. Text-aware variants such as VL-CLIP [14] and the four-tower image-text formulation of Zhu et al. [53] loosen the latter by also accepting text input in the query, but still treat relevance as binary.

The underlying embedding models split into two families that we use as baselines and compare in our architecture ablation (§4.5.3). *Late-fusion* encoders such as CLIP [33], ALIGN [22], and SigLIP [40], VL-CLIP [14] align image and text in a shared space via a contrastive loss and combine them at inference by fusing (averaging in most cases) unimodal embeddings. This two-tower design is fast to serve but cannot represent “image is X but text says change attribute Y” interactions that modifier queries require. *Early-fusion* VLMs such as PaliGemma2 [36] (our backbone), Qwen-VL [43, 1], LLaVa [29], Gemini [9], and ModernVBERT [39] jointly process image and text tokens through a transformer, giving richer cross-modal interactions at a higher serving cost. Recent early-fusion retrievers – GME [51], MM-Embed [28], mmE5 [6], Qwen3-VL-Embedding [25] are trained on synthetic (query, item) pairs from web image-caption corpora but, like the industrial systems above, supervise with binary positive/negative labels. GradCIR is, to our knowledge, the first to train an early-fusion VLM retriever on graded relevance for composed multimodal retrieval.

2.2 Composed Image Retrieval

Composed image retrieval (CIR) [11], also referred to as composed multimodal retrieval in recent surveys, formulates the query as an image plus a textual modifier, and asks the model to retrieve an item that satisfies both. *Classical CIR* approaches train an image-text fusion module on top of a frozen vision-language encoder. ARTEMIS [10], PL4CIR [52], TG-CIR [44], CLIP4CIR [4], BLIP4CIR [31], and SPRC [2] are some of the representative methods in this family. SPN4CIR [13] extends SPRC by mining additional positive and negative triplets, and is the current strongest peer-reviewed supervised baseline on FashionIQ that we compare against.

A second line of work scales CIR with *synthetic triplet supervision*: MagicLens [50] mines 36.7M open-ended instruction triplets from

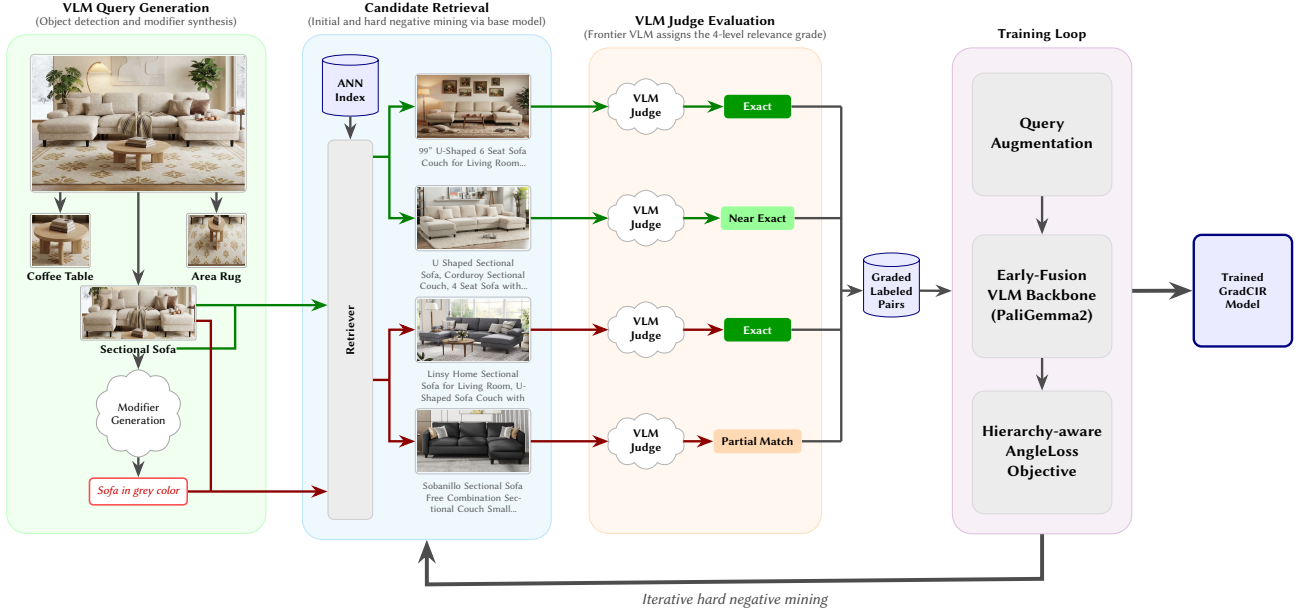


Figure 2: End-to-end architecture of GradCIR.

the web; CompoDiff [15] generates triplets with latent diffusion; CoVR [41] mines triplets from web video captions; CoLLM [20] couples an LLM with CLIP/BLIP backbones and trains on a 3.4M-triplet synthetic dataset (MTCIR). These approaches produce strong zero-shot and supervised CIR models but treat the target as a single binary positive and require an explicit triplet-construction step.

A third line targets *zero-shot CIR* by mapping the visual modality into the text space of a frozen CLIP encoder: Pic2Word [34], SEARLE [3], LinCIR [16], Context-I2W [37], CIReVL [23] (which uses GPT-4 at inference time), and Slerp-TAT [21] are representative methods in this family. These methods avoid CIR-triplet training but inherit the binary-target evaluation protocol.

Graded relevance is the gap. None of the above CIR works, to our knowledge, trains on ordinal graded relevance. Standard CIR datasets (FashionIQ [45], CIRR [30], CIRCO [3]) themselves encode a single positive target per query. In production retrieval over a real catalog this is a poor fit: many candidates partially satisfy a query, and the ordering of those partial matches is what determines the customer experience. GradCIR explicitly targets this regime by training on VLM-assigned 4-level graded labels rather than binary triplets.

3 Methodology

3.1 Problem Formulation and Notation

Let $\mathcal{P} = \{p_i = (v_i, w_i)\}_{i=1}^{N_c}$ denote the product catalog, where v_i and w_i are the image and title of item p_i respectively, and N_c is the total number of items in the catalog. A user query $q = (v^q, w^q)$ is a pair of an image and text. The graded-relevance composed multimodal retrieval problem is to learn an encoder f_θ producing d -dimensional embeddings $e_q = f_\theta(q)$ and $e_p = f_\theta(p)$ such that the ranking induced by $\cos(e_q, e_p)$ over $\mathcal{P}$ respects a 4-level ordinal relevance

label $r \in \{0, 1, 2, 3\}$, corresponding to IRRELEVANT / PARTIALMATCH / NEAREXACT / EXACT.

3.2 Data Preparation

Using large language models to generate high-quality training data has become a popular approach in recent years [42, 7]. We follow a similar approach with vision-language models (VLMs) and curate a graded-relevance dataset from Walmart catalog. Figure 2 shows the end-to-end architecture of GradCIR.

Query generation. For each catalog item, we apply a VLM-based object detector to identify objects from product types of interest, and crop one image per detected product. The cropped image together with the detected object name forms a *similarity query*. A rich catalog image can sometimes contain very large number of products, so we limit ourselves to at most $m=10$ similarity queries per catalog item. To support diverse queries, we also generate *modifier queries* by prompting the same VLM to produce modifier text that alters one or more attributes of the cropped product. Using the same VLM for object detection and modifier-text generation saves compute and ensures the generated modifiers are visually consistent with the query image. We use Gemini-2.5-pro [9] as the VLM for both object detection and modifier-text generation. The final query set Q contains both query types – similarity and modifier queries.

Candidate retrieval and labeling. For each $q \in Q$, we retrieve top- K candidate items $C_q = \{c_1, \dots, c_K\} \subset \mathcal{P}$ using a pretrained multimodal embedding model. A strong VLM judge – Gemini-2.5-pro then labels every pair: $r(q, c_k) = \phi(q, c_k) \in \{\text{EXACT}, \text{NEAREXACT}, \text{PARTIALMATCH}, \text{IRRELEVANT}\}$ (see Figure 1 for representative examples), using the verbatim prompt detailed in Appendix A.

Query augmentation. To improve the model’s robustness to real-world query distributions, we apply several augmentations

Table 1: Query augmentation matrix.

Type	Query Text	Query Image	
		Crop	Whole
Similarity	Blank	✓	✓
	Object name	✓	✓
	Object description	✓	✗
Modifier	Modified attributes	✓	✓
	Name + mod. attributes	✓	✓

per labeled pair: using the uncropped whole image instead of the crop, removing the query text, replacing the object name with a longer object description, or replacing the object name with only the modified attributes. Each augmentation produces a new (query, item) pair sharing the original relevance label. We remove the redundant augmentations, where cropped image is same as original image or object description is same as object name while expanding the dataset. Table 1 enumerates the augmentation matrix.

Iterative hard negative mining. The pretrained retriever used to mine C_q has a bias towards the kind of partial matches it is good at. Relying on it alone leaves the resulting dataset thin on the hard negatives that matter most for a deployed retriever. We therefore run an iterative relevance-feedback loop. Let $\mathcal{D}_t$ denote the labeled dataset at round t (including augmentations) and $f_\theta^{(t)}$ the retriever trained on $\mathcal{D}_t$. At round $t+1$, for each query we mine an additional candidate set with $f_\theta^{(t)}$, deduplicate against previous rounds, label the new pairs with the same VLM judge, and form $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \mathcal{D}_t^{\text{new}}$. We stop when offline NDCG@10 on a held-out query set plateaus. In practice the loop converges in three rounds and produces a final dataset of $\sim 3.5\text{M}$ labeled (query, item) pairs, of which $\sim 2.1\text{M}$ are similarity queries and $\sim 1.4\text{M}$ are modifier queries. The final training set is

$$\mathcal{D} = \{(q, p, \phi(q, p)) \mid q \in \mathcal{Q}, p \in C_q \cup C_q^{\text{HN}}\}$$

where C_q^{HN} is the set of hard negatives mined from the in-training retriever for query q .

3.3 Model Architecture

The retriever is a Siamese bi-encoder: a single early-fusion VLM with shared weights encodes queries and items into a common embedding space. Among compact early-fusion options – PaliGemma2 [36], SmolVLM [32], T5Gemma 2 [49], Granite Vision [38] we choose PaliGemma2 for its favourable quality/latency tradeoff at $\sim 3\text{B}$ parameters.

Backbone. PaliGemma2 combines a SigLIP-400M vision encoder with a Gemma 2 2.6B language model, connected via a linear projection that maps patch embeddings into the language model’s token space. Images are encoded at 224×224 resolution into 256 patch tokens, text is tokenized with SentencePiece and capped at 128 tokens. The interleaved image-patch and text tokens are processed jointly by the Gemma 2 decoder stack, yielding a contextualized early-fusion representation at every position.

Retrieval adaptation. Since PaliGemma2 is a generative VLM with a prefix-LM attention pattern (bidirectional over the image and prompt prefix, causal over the generated suffix) and does not

natively expose fixed-length embeddings, we adapt it for retrieval in three ways:

- (1) *Decoding disabled.* We remove the language-modelling head entirely; the model is used purely as an encoder.
- (2) *Last-token pooling.* The hidden state at the last non-padding position serves as the sequence embedding. Because we treat the entire (image, text) input as the bidirectional prefix (no suffix is generated at encoding time), every token, including the final one, attends to all image and text tokens, so its hidden state aggregates the full multimodal context without requiring an additional pooling layer or a dedicated [CLS] token.
- (3) *Role-specific prefixes.* A short prefix token – [QUERY] or [ITEM] – is prepended to the input text, allowing the shared encoder to learn asymmetric representations for the two roles.

We train with Matryoshka representation learning [24], applying nested losses at dimensions $\{128, 256, 512, 1024\}$, so that embeddings can be truncated to a lower dimensionality at deployment time to meet latency constraints without retraining (see §4.6). For a query $q = (v^q, w^q)$ and a candidate item $p = (v^p, w^p)$, encoder f_θ produces embeddings

$$e_q = f_\theta(v^q, w^q), \quad e_p = f_\theta(v^p, w^p),$$

with the relevance score $s(q, p) = \cos(e_q, e_p)$.

3.4 Training Objective

Bi-encoder retrievers are commonly trained with multiple-negatives ranking loss (MNRL) [18] or InfoNCE, which both rely on in-batch negatives and treat relevance as binary: every non-positive item in the batch is an equally bad negative. This is a poor fit for graded relevance, where a NEAREXACT item and an IRRELEVANT item carry very different penalties for being ranked equally with the positive.

We instead use AngleLoss [26, 27], which transforms graded labels into angular constraints in the embedding space. For two training pairs (q_a, p_a, r_a) and (q_b, p_b, r_b) with $r_a > r_b$, AngleLoss enforces $\Delta\theta_{q_a, p_a} < \Delta\theta_{q_b, p_b}$, where $\Delta\theta_{q, p}$ denotes the angle between e_q and e_p . The loss is

$$\mathcal{L}_{\text{angle}} = \log \left[1 + \sum_{(a, b): r_a > r_b} \exp \left(\frac{\Delta\theta_{q_a, p_a} - \Delta\theta_{q_b, p_b}}{\tau} \right) \right], \quad (1)$$

summed over all ordered pairs in the batch with strictly greater relevance, where τ is a temperature hyperparameter. Because the constraint set in Eq. 1 is enumerated over *every pair of distinct grade levels* rather than positive-vs-negative only, the loss exploits all ordering constraints implied by the graded relevance scale (EXACT>NEAREXACT, EXACT>PARTIALMATCH, EXACT>IRRELEVANT, NEAREXACT>PARTIALMATCH, NEAREXACT>IRRELEVANT, PARTIALMATCH>IRRELEVANT). Binary contrastive losses collapse five of these six constraints into one, which is the mechanism we ablate in §4.5.1.

Table 2: Overview of the WVST dataset (Ex.=Exact, NE=NearExact, PM=PartialMatch, Irr.=Irrelevant).

Query Type	# Queries	#Total Cands.	Candidates per Query			
			Ex.	NE	PM	Irr.
Home Vertical						
Similarity	1073	216K	9.55	21.50	128.43	42.82
Modifier	1140	231K	4.57	25.31	120.29	52.17
Fashion Vertical						
Similarity	843	170K	6.03	18.23	118.94	62.59
Modifier	854	171K	4.35	25.12	112.97	60.69

4 Experiments and Results

4.1 Experimental Setup

4.1.1 Implementation Details. We train our primary GradCIR instantiation – GradCIR-PG2 on the 3.5M-pair graded dataset $\mathcal{D}$ described in §3.2. The model is trained for 2 epochs with AdamW optimizer (learning rate 1×10^{-5}) with batch size 512 using LoRA [19] (rank 64, $\alpha=64$, dropout=0.1 applied to all attention and MLP layers) on top of a PaliGemma2 backbone. The objective is AngleLoss (Eq. 1) with $\tau=0.05$. We learn 1024-dimensional Matryoshka embeddings [24] with nested losses at 128, 256, 512 and 1024 for post-hoc dimensionality reduction in production. Training runs on 4xNVIDIA H100-80GB with mixed-precision and gradient checkpointing.

4.1.2 Baselines. We compare GradCIR-PG2 against four contemporary multimodal embedding baselines: FashionCLIP [8], SigLIP2 [40], GME-Qwen [51], and Qwen3-VL-Embedding [25]. All baselines are evaluated both off-the-shelf and after fine-tuning with the GradCIR recipe (same 3.5M graded dataset, identical AngleLoss + LoRA setup); the fine-tuned variants are denoted **+GradCIR** in the tables. For late-fusion baselines (FashionCLIP, SigLIP2) we use the average of image and text embeddings for pre-trained models and weighted average (with learned weights) for fine-tuned models. For early-fusion baselines (GME-Qwen, Qwen3-VL-Embedding) we use the native fused embedding.

4.1.3 Evaluation Datasets. To evaluate the performance of our proposed methodology, we construct a high-quality internal test set, the Walmart Visual Search Test (WVST), and also report results on two standard public benchmarks for fashion retrieval: Street2Shop [17] and FashionIQ [45].

Walmart Visual Search Test (WVST). A high-quality internal test set spanning home and fashion verticals, constructed by stratified sampling of catalog products and user-uploaded review images, cropped with Gemini-2.5-pro. For each query, we construct a labeled candidate pool by taking the union of top- N retrievals from multiple embedding models, then labeling each candidate with one of the four graded-relevance levels using a Gemini-2.5-pro judge. We report results separately for similarity and modifier queries, for the home and fashion verticals. The candidate-label distribution is shown in Table 2.

Street2Shop. A similarity-query benchmark: given a street photo of a clothing item, retrieve the matching shop product. The test split contains 6770 street-image queries and 9802 shop-image candidates. We use the street image plus its category label (e.g., “dress”,

Table 3: Performance on WVST. “+GradCIR” denotes fine-tuning with our graded-relevance recipe.

Model	Home Vertical				Fashion Vertical			
	Adj-R@5		NDCG		Adj-R@5		NDCG	
	Exact	Partial	@5	@10	Exact	Partial	@5	@10
Similarity Queries								
FashionCLIP	0.3507	0.8837	0.7694	0.7740	0.3992	0.9153	0.8178	0.8213
+GradCIR	0.4474	0.9225	0.8267	0.8281	0.4346	0.9266	0.8395	0.8446
SigLIP2	0.4820	0.9004	0.8217	0.8174	0.5156	0.9416	0.8681	0.8630
+GradCIR	0.5525	0.9133	0.8467	0.8429	0.5549	0.9405	0.8761	0.8713
GME-Qwen	0.4049	0.9044	0.8030	0.8066	0.4075	0.9214	0.8228	0.8289
+GradCIR	0.5806	0.9442	0.8770	0.8752	0.5954	0.9524	0.9000	0.8985
Qwen3-VL-Emb	0.5522	0.9359	0.8602	0.8555	0.5956	0.9530	0.8922	0.8883
+GradCIR	0.6406	0.9597	0.9012	0.8984	0.6620	0.9720	0.9203	0.9147
GradCIR-PG2	0.6560	0.9588	0.9066	0.9034	0.6747	0.9721	0.9241	0.9216
Modifier Queries								
FashionCLIP	0.3619	0.8545	0.7674	0.7712	0.4031	0.8826	0.8018	0.8035
+GradCIR	0.4496	0.8939	0.8181	0.8218	0.4642	0.9000	0.8313	0.8339
SigLIP2	0.3167	0.8734	0.7687	0.7721	0.3860	0.9132	0.8198	0.8234
+GradCIR	0.4460	0.8840	0.8025	0.8085	0.4866	0.9151	0.8459	0.8486
GME-Qwen	0.4152	0.8775	0.8001	0.8006	0.4213	0.8926	0.8191	0.8217
+GradCIR	0.4998	0.9133	0.8488	0.8517	0.5153	0.9162	0.8588	0.8613
Qwen3-VL-Emb	0.3909	0.9067	0.8110	0.8118	0.4627	0.9170	0.8421	0.8429
+GradCIR	0.5470	0.9320	0.8677	0.8687	0.5752	0.9377	0.8846	0.8859
GradCIR-PG2	0.5704	0.9401	0.8810	0.8807	0.5908	0.9401	0.8930	0.8942

“top”) as the query and the shop image plus category label as the candidate.

FashionIQ. A modifier-query benchmark: given a reference image and a natural-language modifier (e.g., “make it striped”), retrieve the target item. We evaluate on the validation split following the standard protocol of averaging Recall@10 and Recall@50 across the Dress, Shirt, and Top-tee sub-categories. We use FashionIQ rather than CIRRCIRCO because the latter draw from NLVR2/COCO natural-scene imagery that is out-of-distribution for a catalog-trained model; we discuss this scope choice further in §5.

4.1.4 Evaluation Metrics. For WVST we report two graded-relevance metrics: **Adjusted Recall@5** (Adj-R@5) and **NDCG@{5,10}**, both computed over the labeled candidate pool with NDCG gain values 3 / 2 / 1 / 0 for the Exact / NearExact / PartialMatch / Irrelevant levels respectively. We report Adj-R@5 separately for *Exact* (Exact + NearExact only) and *Partial* (Exact + NearExact + PartialMatch). Adjusted recall normalizes by $\min(K, n_{\text{rel}})$ rather than K , to fairly handle queries with fewer than K relevant items. n_{rel} is the number of relevant items in the candidate pool for the query.

$$\text{Adj-R@K} = \frac{|\text{relevant items in top-}K|}{\min(K, n_{\text{rel}})}.$$

For Street2Shop we report Recall@{1,5,10} following standard practice. For FashionIQ we report Recall@{10, 50} averaged across sub-categories.

4.2 Results on Walmart Visual Search Test

As shown in Table 3, GradCIR-PG2 achieves the best performance on WVST, both in Adj-R and NDCG. It improves over the next-best variant, Qwen3-VL-Emb(+GradCIR), by an average of 2.83% for Adj-R@5 (Exact) and 0.91% for NDCG@10 (average is taken over all query types and for both Home and Fashion categories). The gap is small on similarity queries and widens on modifier queries, where graded supervision matters most.

A second observation addresses whether the gains come from the data or the backbone. Every baseline benefits substantially

Table 4: Retrieval performance on the Street2Shop test split.

Model	R@1	R@5	R@10
FashionCLIP	0.3121	0.5084	0.5932
+GradCIR	0.3329	0.5230	0.5975
SigLIP2	0.5180	0.6739	0.7357
+GradCIR	0.5558	0.7064	0.7718
Qwen3-VL-Emb	0.6443	0.7718	0.8227
+GradCIR	0.6631	0.7886	0.8304
GradCIR-PG2	0.6464	0.7815	0.8266

Table 5: FashionIQ zero-shot CIR Performance.

Model	Backbone	R@10	R@50	Avg
Pic2Word [34]	CLIP-L/14	0.2470	0.4370	0.3420
SEARLE [3]	CLIP-L/14	0.2560	0.4620	0.3590
LinCIR [16]	CLIP-L/14	0.2640	0.4660	0.3650
Context-l2W [37]	CLIP-L/14	0.2780	0.4890	0.3835
Slerp-TAT [21]	CLIP-L/14	0.2830	0.4760	0.3795
CIReVL (GPT-4) [23]	CLIP-L/14	0.2860	0.4860	0.3860
CoLLM [20]	CLIP-L/14	0.3010	0.4950	0.3980
Slerp-TAT [21]	BLIP-L/16	0.3280	0.5330	0.4305
CoLLM [20]	BLIP-L/16	0.3460	0.5600	0.4530
GradCIR (zero-shot)	PaliGemma2	0.3036	0.4964	0.4000

Table 6: FashionIQ supervised CIR Performance (fine-tuned on the FashionIQ training split).

Model	Backbone	R@10	R@50	Avg
ARTEMIS [10]	w/o VLP	0.2605	0.5029	0.3817
ComqueryFormer [46]	w/o VLP	0.2937	0.5536	0.4237
PL4CIR [52]	CLIP	0.3902	0.6300	0.5101
TG-CIR [44]	CLIP-ViT-B/16	0.3981	0.6306	0.5144
CLIP4CIR [4]	CLIP-RN50×4	0.4252	0.6560	0.5406
BLIP4CIR [31]	BLIP-base	0.4349	0.6731	0.5540
SPRC [2]	BLIP-2	0.5492	0.7497	0.6485
SPN4CIR [13]	BLIP-2	0.5637	0.7645	0.6641
GradCIR (fine-tuned)	PaliGemma2	0.5684	0.7721	0.6703

from the GradCIR recipe – FashionCLIP, SigLIP2, GME-Qwen, and Qwen3-VL-Embedding gain 24.3% of Adj-R@5 (Exact) on average over their off-the-shelf counterparts, validating the data-pipeline contribution independently of backbone choice. The remaining gap between GradCIR-fine-tuned baselines and GradCIR-PG2 reflects the backbone’s suitability for graded ranking.

4.3 Results on Street2Shop

Table 4 reports the retrieval performance on Street2Shop dataset. GradCIR-PG2 performs closely to Qwen3-VL-Emb(+GradCIR) (within 0.4–1.7 points across all K), and improves over other baseline models by large margins. This confirms that the graded-relevance training transfers well to the binary-target Street2Shop evaluation protocol, and that our training strategy consistently improves performance for similarity queries.

4.4 Results on FashionIQ

We evaluate GradCIR-PG2 on FashionIQ in two settings: *zero-shot* (Walmart-trained model used as-is) and *fine-tuned* (additionally

Table 7: Effect of label granularity on retrieval quality.

Supervision	Sim. Queries		Mod. Queries	
	NDCG@5	NDCG@10	NDCG@5	NDCG@10
Binary (2-level)	0.8762	0.8698	0.8370	0.8376
3-level	0.8775	0.8783	0.8412	0.8450
4-level	0.9154	0.9125	0.8870	0.8875

Table 8: Effect of including modifier queries in training.

Training Data	Sim. Adj-R@5		Mod. Adj-R@5	
	Exact	Partial	Exact	Partial
Similarity only	0.6258	0.9530	0.3016	0.9004
Sim. + Modifier	0.6654	0.9655	0.5806	0.9401

fine-tuned on the FashionIQ training split with $r=64$ LoRA for 5 epochs).

Zero-shot. Without any fashion-CIR fine-tuning, GradCIR (Table 5) reaches average Recall of 0.40, matching or exceeding every published CLIP-L-class zero-shot CIR method. Larger, more recent, BLIP-L variants (CoLLM-BLIP-L, Slerp-TAT-BLIP-L) remain ahead. The results demonstrate that catalog-only graded supervision transfers competitively to the binary-CIR evaluation protocol, with no triplet construction step required.

Supervised. With FashionIQ fine-tuning (Table 6), GradCIR reaches average recall of 0.6703, ahead of the strongest peer-reviewed supervised baseline we compare against, SPN4CIR (0.6641), reaching the state-of-the-art among the supervised CIR methods on FashionIQ.

4.5 Ablations

4.5.1 Graded vs. Binary Supervision. Table 7 isolates the effect of label granularity: all rows share the same backbone (PaliGemma2) and training data (3.5M pairs); only the number of relevance levels changes. NDCG is averaged across Home and Fashion. Moving from binary to 4-level supervision yields an improvement of 4.5%–6% in NDCG@5 and 4.9%–5.9% in NDCG@10 across both query types. The intermediate 3-level setting (which merges NearExact and PartialMatch into a single grade) already improves over binary, but a substantial gap to 4-level remains. This confirms that distinguishing NearExact from PartialMatch—precisely the discrimination that graded relevance adds—is essential for top-of-ranking quality.

4.5.2 Effect of Modifier-Query Training. Table 8 compares training on similarity queries alone vs. training on both query types. Adj-R@5 is averaged across Home and Fashion verticals. We observe that training the model on modifier queries in addition to similarity queries makes the model more robust, and improves its performance on similarity queries as well. This is because modifier queries provide additional supervision on fine-grained image–text interactions that are also relevant for similarity-query understanding. Including modifier queries improves similarity-query Adj-R@5 by 6.3% for exact match retrieval and 1.3% for partial match retrieval, and modifier-query Adj-R@5 by a much larger 92.5% for exact match and 4.4% for partial match. This validates the importance of including modifier queries in training to achieve strong performance on both query types.

4.5.3 Architecture: Late vs. Early Fusion. The fine-tuned-baseline rows of Table 3 double as an architecture ablation. On similarity queries, the late-fusion SigLIP2 closes most of the gap to early-fusion baselines after fine-tuning. On modifier queries, it does improve with fine-tuning, but still remains well below the early-fusion baselines. Similarly, FashionCLIP lags behind even after fine-tuning, and the gap to early-fusion models is larger than for SigLIP2. This is consistent with earlier observations ([34, 20]) that late fusion struggles to model the complex image-text interactions required for modifier-query understanding, even with strong fine-tuning.

4.6 Deployment

GradCIR-PG2 is deployed as the retrieval backbone of Walmart’s visual-search system, currently serving a partial traffic slice. This is the first production visual-search system at Walmart.

Serving architecture. At inference time, every catalog item is pre-indexed by its GradCIR-PG2 embedding truncated to 256 dimensions as described in §3.3. User queries are encoded online by the same model and retrieved via approximate nearest-neighbour (ANN) search. The existing text-based reranker is applied on top of the retrieved candidates to produce the final ranking. The full pipeline operates within the production latency SLA.

5 Scope and Limitations

GradCIR targets composed image retrieval *over a structured product catalog* where every candidate is a single sellable item. Two consequences shape our evaluation choices and scope:

Why not evaluate on CIRR or CIRCO? The CIRR and CIRCO benchmarks draw query and candidate images from NLVR2 and COCO respectively, consisting of natural scenes with multiple objects, and free-form composition. A retriever trained on catalog images is fundamentally out-of-distribution on these benchmarks, and the resulting numbers measure adaptation effort more than retrieval quality. FashionIQ and Street2Shop, by contrast, draw from a fashion catalog and align with our training distribution, which is why we report these as our public benchmarks.

What we do not solve. We do not address (i) multi-item compositional queries (“find a sofa like this and a rug like that”), which require a more expressive query representation than a single image-text pair; or (ii) explicit attribute-level controllability at inference time, which would benefit from explicit attribute conditioning. Extending the graded-relevance formulation to these settings is left to future work.

Reproducibility. All architecture choices, hyperparameters, prompts, augmentation details, and recipe details required to reproduce the methodology on any comparable proprietary catalog are provided in the paper. The FashionIQ and Street2Shop results use public data and standard public evaluation protocols.

6 Conclusion

We presented GradCIR, a graded-relevance methodology for composed multimodal retrieval in e-commerce visual search. The central methodological contribution is the use of 4-level ordinal relevance labels, assigned at scale by a VLM judge and exploited by a hierarchy-aware angular objective. The graded-vs-binary ablation isolates this single design choice and shows that finer relevance

granularity is what closes the ranking-quality gap on graded NDCG metrics. The full pipeline curates 3.5M labeled pairs without manual annotation through an iterative relevance-feedback loop. On the internal benchmarks, GradCIR outperforms all baselines on almost every metric, with the largest gains on the more fine-grained graded NDCG metrics. On public benchmarks, GradCIR is competitive with the strongest baselines. The system is deployed in production at Walmart, where it’s serving live visual-search user traffic. Future work includes extending the graded formulation to multi-item compositional queries and exploring attribute-level conditioning for more fine-grained relevance modeling.

A Labeling Prompt

The verbatim prompt used by the VLM judge $\phi(\cdot, \cdot)$ to assign relevance labels is reproduced below.

```
>>SYSTEM
You are an expert in product search and relevance evaluation for an
e-commerce company. You are meticulous, detail-oriented, and have a deep
understanding of user intent in an online shopping context.

Your task is to evaluate the relevance of an e-commerce product (candidate)
to a visual search query. The query includes a product image (which may be a
partial or obstructed view) and a query text. The query text can explain or
override few attributes of the product, the non-overridden attributes (visual
or otherwise) remain part of the query. The candidate includes a product
image and some product details such as title, description or structured
attributes (e.g., size, brand, color).

Step-by-step Instructions:
1. Extract the specifications (attributes, features, and all other important
aspects) from both the query and candidate.
Examples include:
* Intended use, product type, color, shape, size, material, style, brand,
intended age, gender, theme, occasion, etc.
* Cardinal attributes related to usability such as: number of drawers,
shelves, windows, doors, parts, pieces, etc.

2. Compare the query and candidate based on:
* Visual similarity (overall appearance, shape, size, style).
* Specification alignment.
* Whether the candidate would fulfill the intent behind the query.
* Focus on the e-commerce product aspects, not on the lighting, angle,
background etc.

3. Assign one of the following relevance labels to the candidate:
1. Exact: Matches the query exactly in terms of all specifications and
visual appearance.
2. NearExact: Completely satisfies the intent behind the query, but may
contain minor difference in appearance or specification (e.g., different
brand or slightly varied style or different shade of same color or slanted
vs straight legs of a coffee table).
3. PartialMatch: A candidate with same product type and intended use.
Attributes such as color, style, size, material, etc. can be different
from the query.
4. Irrelevant: Either Product type or intended use is different from the
query, or candidate is intended for a different age group or gender or
size.

Examples:
* Query: Product A in red color
  Candidate: Variant of product A in crimson color
  Label: NearExact
* Query: Product A
  Candidate: Very similar to product A with minor visual differences but
  additional features
  Label: NearExact
* Query: Product A in red color
  Candidate: Variant of product A with only difference being green color
  Label: PartialMatch

Final Output Format:
Explanation: <Brief explanation under 100 words>
Label: <Exact | NearExact | PartialMatch | Irrelevant>

>>USER
Query:
Image: <query image> <query text>
Candidate:
Image: <candidate product image> <candidate product title>
```

References

- [1] Shuai Bai et al. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.
- [2] Yang Bai, Xinxiang Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Mong, and Chun-Mei Feng. 2024. Sentence-level prompts benefit composed image retrieval. In *International Conference on Learning Representations*. Vol. 2024, 30531–30548.
- [3] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15338–15347.
- [4] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Effective conditioned and composed image retrieval combining CLIP-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [5] Sean Bell, Yiqun Liu, Sami Alsheikh, Yina Tang, Edward Pizzi, M Henning, Karun Singh, Omkar Parkhi, and Fedor Borisjuk. 2020. Groknet: unified computer vision model trunk and embeddings for commerce. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2608–2616.
- [6] Haonan Chen, Liang Liu, Hongjin Su, Zihan Sun, Liang Wang, Nan Yang, Furu Wei, and Zhicheng Liu. 2025. mmE5: improving multimodal multilingual embeddings via high-quality synthetic data. *arXiv preprint arXiv:2502.08468*.
- [7] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4, 5.
- [8] Patrick John Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhaes, Diogo Goncalves, Ciro Greco, and Jacopo Tagliabue. 2022. Contrastive language and vision learning of general fashion concepts. *Scientific Reports*, 12, 1, 18958.
- [9] Gheorghe Comanici et al. 2025. Gemini 2.5: pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- [10] Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csuska, and Diane Larlus. 2022. Artemis: attention-based retrieval with text-explicit matching and implicit similarity. *arXiv preprint arXiv:2203.08101*.
- [11] Longye Du, Shuaiyu Deng, Ying Li, Jun Li, and Qi Tian. 2025. A survey on composed image retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21, 7, 1–27.
- [12] Ming Du et al. 2022. Amazon shop the look: a visual search system for fashion and home. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2822–2830.
- [13] Zhangchi Feng, Richong Zhang, and Zhijie Nie. 2024. Improving composed image retrieval via contrastive learning with scaling positives and negatives. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*.
- [14] Ramin Giahhi, Kehui Yao, Sriram Kollipara, Kai Zhao, Vahid Mirjalili, Jianpeng Xu, Topojoy Biswas, Evren Korpeoglu, and Kannan Achan. 2025. V1-clip: enhancing multimodal recommendations via visual grounding and llm-augmented clip embeddings. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, 482–491.
- [15] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Heejae Jun, Yoohoon Kang, and Sangdoo Yun. 2023. CompoDiff: versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*.
- [16] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. 2024. Language-only training of the zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13225–13234.
- [17] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. 2015. Where to buy it: matching street clothing photos in online shops. In *Proceedings of the IEEE international conference on computer vision*, 3343–3351.
- [18] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.
- [19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: low-rank adaptation of large language models. *ICLR*, 1, 2, 3.
- [20] Chuong Huynh, Jinyu Yang, Ashish Tawari, Mubarak Shah, Son Tran, Rafay Hamid, Trishul Chilimbi, and Abhinav Shrivastava. 2025. Collm: a large language model for composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3994–4004.
- [21] Young Kyun Jang, Dat Huynh, Ashish Shah, Wen-Kai Chen, and Ser-Nam Lim. 2024. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. In *European conference on computer vision*. Springer, 239–254.
- [22] Chao Jia et al. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*. PMLR, 4904–4916.
- [23] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. 2024. Vision-by-language for training-free compositional image retrieval. In *International Conference on Learning Representations*. Vol. 2024, 16926–16941.
- [24] Aditya Kusupati et al. 2022. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35, 30233–30249.
- [25] Mingxin Li et al. 2026. Qwen3-vl-embedding and qwen3-vl-reranker: a unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*.
- [26] Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.
- [27] Xianming Li and Jing Li. 2024. Aoe: angle-optimized embeddings for semantic textual similarity. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1825–1839.
- [28] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2024. MM-Embed: universal multimodal retrieval with multimodal LLMs. *arXiv preprint arXiv:2411.02571*.
- [29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36, 34892–34916.
- [30] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2125–2134.
- [31] Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. 2024. Bi-directional training for composed image retrieval via text prompt learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 5753–5762.
- [32] Andrés Marafioti et al. 2025. Smolvlm: redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*.
- [33] Alec Radford et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [34] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 2023. Pic2word: mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19305–19314.
- [35] Devashish Shankar, Sujay Narumanchi, HA Ananya, Pramod Kompalli, and Krishnendu Chaudhury. 2017. Deep learning based large scale visual recommendation and search for e-commerce. *arXiv preprint arXiv:1703.02344*.
- [36] Andreas Steiner et al. 2024. Paligemma 2: a family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.
- [37] Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. 2024. Context-i2w: mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 6. Vol. 38, 5180–5188.
- [38] Granite Vision Team et al. 2025. Granite vision: a lightweight, open-source multimodal model for enterprise intelligence. *arXiv preprint arXiv:2502.09927*.
- [39] Paul Teietche, Quentin Macé, Max Conti, Antonio Loison, Gautier Viaud, Pierre Colombo, and Manuel Faysse. 2025. Modernvbert: towards smaller visual document retrievers. *arXiv preprint arXiv:2510.01149*.
- [40] Michael Tschannen et al. 2025. Siglip 2: multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.
- [41] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. 2024. Covr: learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 6. Vol. 38, 5270–5279.
- [42] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11897–11916.
- [43] Peng Wang et al. 2024. Qwen2-vl: enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- [44] Haokun Wen, Xian Song, Xueming Yang, Yinwei Zhan, and Liqiang Nie. 2023. Target-guided composed image retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*.
- [45] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion IQ: a new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [46] Yahui Xu, Yi Bin, Jiwei Wei, Yang Yang, Guoqing Wang, and Heng Tao Shen. 2023. Multi-modal transformer with global-local alignment for composed query image retrieval. *IEEE Transactions on Multimedia*, 25, 8346–8357.
- [47] Fan Yang, Ajinkya Kale, Yury Bubnov, Leon Stein, Qiaosong Wang, Hadi Kiapour, and Robinson Piramuthu. 2017. Visual search at ebay. In *Proceedings*

- of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2101–2110.
- [48] Andrew Zhai, Hao-Yu Wu, Eric Tzeng, Dong Huk Park, and Charles Rosenberg. 2019. Learning a unified embedding for visual search at pinterest. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2412–2420.
- [49] Biao Zhang et al. 2025. T5gemma 2: seeing, reading, and understanding longer. *arXiv preprint arXiv:2512.14856*.
- [50] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhui Chen, Yu Su, and Ming-Wei Chang. 2024. Magiclens: self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*.
- [51] Xin Zhang et al. 2024. Gme: improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*.
- [52] Yida Zhao, Yuqing Song, and Qin Jin. 2022. Progressive learning for image retrieval with hybrid-modality queries. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- [53] Xinliang Zhu et al. 2024. Bringing multimodality to amazon visual search system. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6390–6399.