TOWARD DETAIL-INTENSIVE E-COMMERCE IMAGE-TEXT RETRIEVAL: HIERARCHICAL REFINEMENT OF QUERY-RELEVANT REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

E-commerce image-text retrieval (E-ITR) aims to rapidly retrieve relevant products from large-scale image corpora in response to user textual queries, supporting everyday shopping search, recommendation, and assistance. As living standards improve and product choices expand, E-commerce images become increasingly detail-intensive, containing densely packed elements such as product logos and usage instructions. Moreover, real-time response requirements necessitate retrieval from low-resolution E-commerce images, making it more challenging to accurately identify query-relevant products. To address the dilemma, we contribute EComClutter, a detail-intensive E-commerce image-text retrieval database with 94,481 query-image pairs, which are sourced from 35 E-commerce categories (e.g., fashion and electronics). Relevance is annotated through an ensemble of chain-of-thought reasoning outputs from multiple advanced MLLMs, followed by rigorous human verification. Based on EComClutter, we further propose EHiRe, a novel E-ITR framework that hierarchically refines query-relevant representations through Detail Region Rearrangement, Important Tokens Modulation, and Relevant Embedding Selection. EHiRe achieves state-of-the-art retrieval accuracy under low-resolution inference settings. Extensive experiments and comparisons demonstrate the superiority of the proposed method.



Figure 1: A detail-intensive E-commerce image with few query-relevant visual elements and Top-1 retrieval results from EHiRe, Qwen3-VL-Embedding, and GME. EHiRe ranks the query-relevant product first, whereas Qwen3-VL-Embedding and GME do not.

1 INTRODUCTION

E-commerce image-text retrieval (E-ITR) aims to rapidly retrieve relevant products from large-scale image corpora in response to user textual queries. It supports everyday shopping search, recommendation, and assistance. As living standards improve and product choices expand, E-commerce

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.

images become increasingly detail-intensive, with densely packed elements such as product logos and usage instructions. This growth creates a need for image-text retrieval that can identify query-relevant products within these detail-intensive images.

Existing datasets provide foundations for E-commerce image-text retrieval. FashionIQ supports interactive fashion retrieval through natural-language feedback (Wu et al., 2021), while MUGE provides Chinese E-commerce image-text data for retrieval (Yang et al., 2022). MMEB-V2 provides a general retrieval benchmark across images, videos, and visual documents (Meng et al., 2025). On the modeling side, OpenCLIP establishes a shared image-text embedding space through CLIP-style contrastive learning (Cherti et al., 2023). GME is a universal multimodal embedding method (Zhang et al., 2025), while UME-R1 is a reasoning-driven embedding method (Lan et al., 2026).

Despite this progress, existing datasets and methods do not directly address detail-intensive E-commerce image-text retrieval under low-resolution inference. This setting presents 2 related challenges. First, detail-intensive E-commerce images contain densely packed visual and textual elements, making it difficult to identify query-relevant products. Second, real-time response requirements necessitate retrieval from low-resolution E-commerce images, making it more challenging to accurately identify query-relevant products.

Table 1: Representative datasets relevant to general and E-commerce Image-Text retrieval.

Dataset	Year	Domain	Categories	Edge Density ↑
Flickr30K (Young et al., 2014)	2014	General	–	–
MMEB (Jiang et al., 2025)	2025	General	–	–
MMEB-V2 (Meng et al., 2025)	2025	General	–	–
FashionIQ (Wu et al., 2021)	2021	E-commerce	Fashion (3 categories)	0.055
Product1M (Zhan et al., 2021)	2021	E-commerce	Beauty (458 categories)	0.089
MUGE (Yang et al., 2022)	2021	E-commerce	–	0.075
M5Product (Dong et al., 2022)	2022	E-commerce	Mixed (6,232 coarse categories)	0.109
EComClutter	2026	E-commerce	Mixed (35 top-level categories)	**0.131**

To address these challenges, we contribute EComClutter, a detail-intensive E-commerce image-text retrieval database with 94,481 query-image pairs sourced from 35 E-commerce categories, such as fashion and electronics. Relevance is annotated through an ensemble of chain-of-thought reasoning outputs from multiple advanced MLLMs, followed by rigorous human verification. Following Rosenholtz et al. (2007), we use edge density, defined as the proportion of edge pixels in an image, to quantify detail intensity. Table 1 summarizes this measure for representative retrieval datasets.

Based on EComClutter, we further propose EHiRe, a novel E-ITR framework that hierarchically refines query-relevant representations through Detail Region Rearrangement (DRR), Important Tokens Modulation (ITM), and Relevant Embedding Selection (RES). DRR recovers query-relevant visual details for low-resolution retrieval. ITM generates embedding candidates by modulating visual tokens relevant to the query. RES improves the final retrieval ranking by retaining useful embedding candidates. EHiRe achieves state-of-the-art retrieval accuracy under low-resolution inference settings. Figure 1 shows a representative Top-1 retrieval example.

The contributions of this work are three-fold:

- We study E-commerce image-text retrieval (E-ITR) under low-resolution inference settings and contribute EComClutter, a detail-intensive E-commerce image-text retrieval database comprising 94,481 query-image pairs from 35 E-commerce categories with image-query relevance.
- We propose EHiRe, a novel E-ITR framework that hierarchically refines query-relevant representations through DRR, ITM, and RES.
- Through extensive experiments and comparisons, we demonstrate the superiority of the proposed method and its generalization ability.

2 Related Work

2.1 Datasets

Flickr30K pairs images with human-written descriptions for general image-text retrieval (Young et al., 2014). MMEB evaluates multimodal embeddings across multiple tasks, while MMEB-V2 covers images, videos, and visual documents (Jiang et al., 2025; Meng et al., 2025). These general datasets are not organized by E-commerce categories.

FashionIQ supports interactive retrieval across 3 fashion categories using reference images and natural-language feedback (Wu et al., 2021). Product1M focuses on instance-level retrieval across 458 beauty categories (Zhan et al., 2021). MUGE supports Chinese E-commerce image-text retrieval (Yang et al., 2022). M5Product spans 6,232 coarse product categories (Dong et al., 2022).

Table 1 compares the domain, category coverage, and edge density of these datasets. Edge density measures image detail intensity (Rosenholtz et al., 2007). EComClutter covers 35 top-level E-commerce categories and has the highest edge density among the E-commerce datasets in the table.

2.2 Multimodal Embedding Methods

CLIP learns shared image-text embeddings through contrastive learning (Radford et al., 2021). OpenCLIP provides open, reproducible scaling; SigLIP uses a pairwise sigmoid loss (Cherti et al., 2023; Zhai et al., 2023). Together, they represent CLIP-style embedding methods.

Universal multimodal embedding methods support retrieval across multiple input modalities and tasks. Representative methods include GME, VLM2Vec-V2, Qwen3-VL-Embedding, and WeMM-Embedding (Zhang et al., 2025; Meng et al., 2025; Li et al., 2026; Zhou et al., 2026).

UME-R1 uses reasoning to improve multimodal retrieval (Lan et al., 2026). Embed-RL separates reasoning learning from embedding optimization through reinforcement learning (Jiang et al., 2026). Both are reasoning-driven embedding methods and require additional reasoning computation. However, existing methods face challenges in detail-intensive E-ITR under low-resolution inference settings. To address this problem, we propose EHiRe to refine query-relevant representations.

3 Dataset

We introduce EComClutter, a detail-intensive E-commerce image-text retrieval database with 94,481 query-image pairs. These pairs are sourced from 35 E-commerce categories, such as fashion and electronics, and are annotated with three-level (0/1/2) image-query relevance. Figure 2 summarizes its construction through Data Collection and Data Annotation.

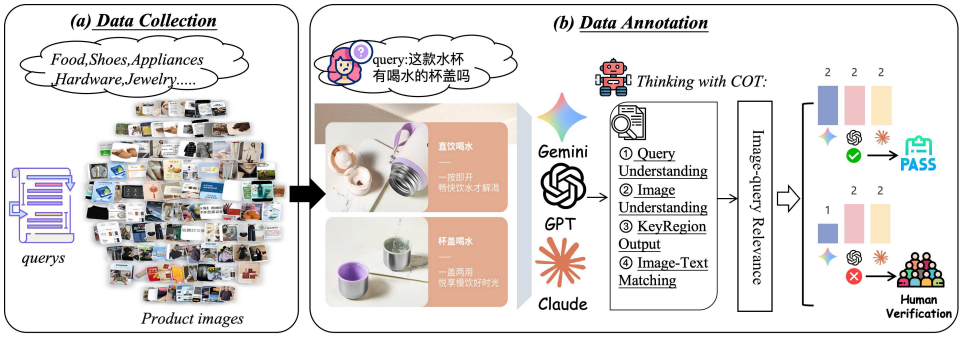

Figure 2: Construction pipeline of EComClutter. (a) Data Collection associates anonymized user textual queries with candidate detail-intensive E-commerce images. The resulting 94,481 query-image pairs cover 35 E-commerce categories, such as fashion and electronics. (b) In Data Annotation, relevance is annotated through an ensemble of chain-of-thought reasoning outputs from multiple advanced MLLMs, followed by rigorous human verification.

3.1 Data Collection

EComClutter is constructed from anonymized user textual queries and candidate detail-intensive E-commerce images from a commercial E-commerce platform. All data were processed under the platform's privacy and data governance requirements. Identifiable faces and sensitive information displayed in product images were masked before inclusion in EComClutter. EComClutter will be publicly released upon publication. Each query is associated with a set of candidate E-commerce images. The resulting 94,481 query-image pairs cover 35 E-commerce categories, such as fashion and electronics.

3.2 Data Annotation

Each query-image pair is annotated with three-level (0/1/2) image-query relevance. The values 0, 1, and 2 denote irrelevant, partially relevant, and fully relevant, respectively. Relevance is annotated through an ensemble of chain-of-thought reasoning outputs from multiple advanced MLLMs, followed by rigorous human verification. The ensemble comprises Google Gemini 3.1 Pro Preview, OpenAI GPT-5.6 Sol, and Anthropic Claude Opus 4.8 (Google, 2026; OpenAI, 2026; Anthropic, 2026). Each of the 3 MLLMs independently generates a chain-of-thought reasoning output for the pair and assigns 0, 1, or 2. When all 3 MLLMs assign the same value, the pair passes the consistency check, and the agreed value is retained as its image-query relevance.

To verify the reliability of the consistency check, we randomly sampled 1,000 query-image pairs for which all 3 MLLMs assigned the same value. A group of 20 annotators with E-commerce expertise and task-specific training independently assessed the image-query relevance of each pair. On average, 96% of the annotators assigned the same value as the 3 MLLMs, supporting the reliability of the consistency check.

For pairs with model disagreement, the 20 annotators independently assign an image-query relevance value of 0, 1, or 2. A pair is retained only when more than 90% of the annotators assign the same value; otherwise, it is removed.

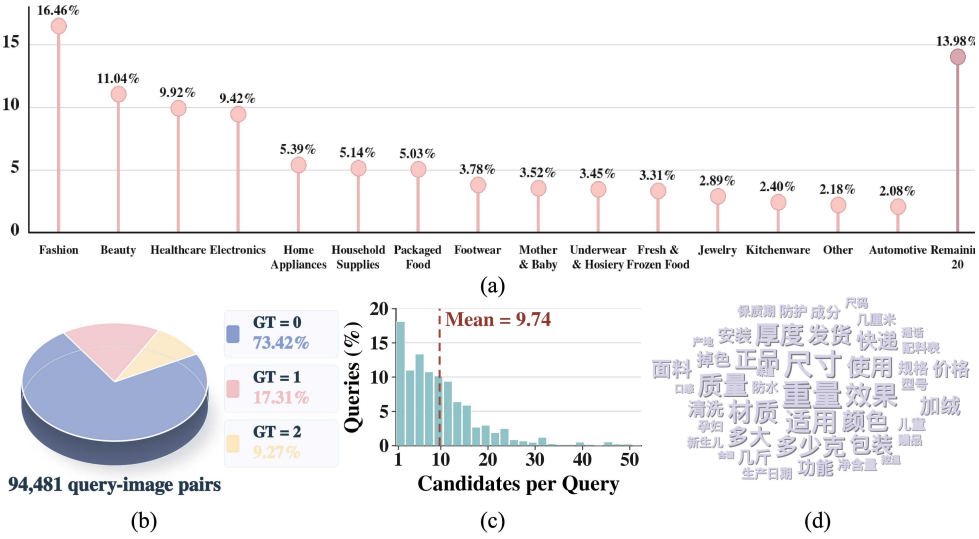

Figure 3: Statistics of EComClutter. (a) Distribution across 35 E-commerce categories. (b) Distribution of image-query relevance. (c) Number of candidate images per user textual query. (d) Frequent themes in user textual queries.

3.3 Dataset Analysis

Figure 3(a) shows that Fashion, Beauty, Healthcare, and Electronics comprise the largest shares, while the remaining categories form a broad long tail. Figure 3(b) reports the relevance distribution across the 94,481 query-image pairs: values 0, 1, and 2 account for 73.42%, 17.31%, and 9.27%,

respectively. Figure 3(c) shows that each user textual query is associated with 9.74 candidate images on average. Figure 3(d) uses font size to indicate frequency; queries about size, weight, and effect are the most frequent.

4 METHOD

Based on EComClutter, we propose EHiRe, an E-ITR framework that hierarchically refines query-relevant representations. Figure 4 presents its overall architecture.

Overall architecture. EHiRe uses the pretrained WeMM multimodal embedding backbone (Zhou et al., 2026) to encode user textual queries and detail-intensive E-commerce images in the candidate set. During training, a frozen high-resolution teacher provides chain-of-thought (CoT)-guided attention and a high-resolution teacher ranking. DRR first recovers query-relevant visual details for low-resolution retrieval. ITM then generates embedding candidates by modulating visual tokens relevant to the query. Finally, RES improves the final retrieval ranking by retaining useful embedding candidates. At inference, retrieval uses the trained backbone without the high-resolution teacher, CoT, ITM, or RES.



Figure 4: Overall architecture of EHiRe. EHiRe hierarchically refines query-relevant representations through 3 modules. (a) Detail Region Rearrangement (DRR) recovers query-relevant visual details for low-resolution retrieval. (b) Important Tokens Modulation (ITM) generates embedding candidates by modulating visual tokens relevant to the query. (c) Relevant Embedding Selection (RES) improves the final retrieval ranking by retaining useful embedding candidates.

4.1 DETAIL REGION REARRANGEMENT

Detail-intensive E-commerce images contain densely packed visual and textual elements. Low-resolution inputs make query-relevant visual details harder to recognize. Together, these factors make it more challenging to accurately identify query-relevant products. DRR addresses this challenge by recovering query-relevant visual details for low-resolution retrieval. Query sensitivity and connected-component statistics determine whether this rearrangement is applied to an image. A detailed description is provided in the Appendix. For routed images, DRR retains a resized image as global context and enlarges localized text regions on a canvas of the same size. Unrouted images retain the standard low-resolution input. Both paths preserve the same input size:

$$\widetilde{x}_i = \mathcal{R}_B(q, x_i) = \begin{cases} \mathcal{D}_B(x_i), & r(q, x_i) = 0, \\ \mathcal{C}_B\left[\mathcal{D}_B(x_i), \{\mathcal{U}_B(x_i, b)\}_{b \in \mathcal{B}_i}\right], & r(q, x_i) = 1, \end{cases} \qquad \text{size}(\widetilde{x}_i) = B. \tag{1}$$

For query q and candidate image x_i, $\mathcal{D}_B$ produces the standard low-resolution input, and $\mathcal{B}_i$ contains the localized detail regions. $\mathcal{U}_B$ crops and enlarges a region. $\mathcal{C}_B$ rearranges the resulting images at target size B.

4.2 Important Tokens Modulation

Given a low-resolution input, EHiRe must determine which visual tokens are relevant to the current query. ITM uses the frozen high-resolution teacher and CoT to estimate this relevance. For candidate image x_i, the teacher receives the high-resolution image x_i^H, the query q, and the corresponding CoT c_i. The CoT c_i is taken from the chain-of-thought reasoning outputs produced for the same query-image pair during EComClutter annotation. The high-resolution image and CoT are used only by the frozen teacher during training. ITM aggregates attention from the query and CoT tokens to the high-resolution visual tokens. It then projects the aggregated attention onto the low-resolution token grid:

$$\mathbf{p}_i = \mathrm{Norm}\left(\mathcal{P}_{H\to L}\left[\frac{1}{H|\mathcal{T}_i|}\sum_{h=1}^{H}\sum_{t\in\mathcal{T}_i}\mathbf{A}_{\phi,h}^{t\to V}(x_i^H, q, c_i)\right]\right), \qquad \mathcal{T}_i = \mathcal{T}(q, c_i). \tag{2}$$

The term $\mathbf{A}_{\phi,h}^{t\to V}$ denotes teacher attention. $\mathcal{T}_i$ contains the query and CoT tokens. Attention aggregation and grid projection produce $\mathbf{p}_i$, the query-relevant importance distribution over low-resolution visual tokens.

Token sampling selects visual tokens from high- and low-attention regions according to $\mathbf{p}_i$. Token reweighting then enhances high-attention tokens and suppresses low-attention tokens:

$$\mathcal{S}_{i,a}^{(g)} \sim \mathrm{WSample}_a(\mathbf{p}_i, \rho), \quad \gamma_{i,a}^{(g)} \sim \mathcal{U}(\Gamma_a), \quad m_{i,n}^{(g)} = \prod_{a\in\{h,l\}}\left[1 + \left(\gamma_{i,a}^{(g)} - 1\right)\mathbb{I}_{\{n\in\mathcal{S}_{i,a}^{(g)}\}}\right]. \tag{3}$$

Here, $\mathcal{S}_{i,a}^{(g)}$ is the sampled visual-token set, $\gamma_{i,a}^{(g)}$ is its reweighting factor, and $m_{i,n}^{(g)}$ is the resulting weight for visual token n. $\mathrm{WSample}_a$ denotes weighted sampling from attention region a. The indices h and l denote high-attention enhancement and low-attention suppression, respectively, with $\Gamma_h = \Gamma_+$ and $\Gamma_l = \Gamma_-$. The value ρ is the sampled fraction for each modulation. For each query, token sampling and reweighting produce G embedding candidates.

4.3 Relevant Embedding Selection

The g-th candidate $\mathbf{Z}^{(g)}$ contains the query embedding and the embeddings of all candidate images associated with q. RES evaluates these candidates using a task reward and a teacher reward.

For the g-th candidate, task reward $R_{\mathrm{task}}^{(g)}$ uses three-level (0/1/2) image-query relevance. It contains 2 terms: relevance-aware ordering and relevance separation. The value y_i denotes the image-query relevance of x_i, and $i \succ j$ indicates $y_i > y_j$:

$$R_{\mathrm{task}}^{(g)} = 2\Bigg[\underbrace{(1-\lambda)\left(1 - \frac{\sum_{i\succ j} w_{ij}\sigma_{\tau_r}\left(s_j^{(g)} - s_i^{(g)}\right)}{\sum_{i\succ j} w_{ij} + \epsilon}\right)}_{\text{relevance-aware ordering}} + \underbrace{\frac{\lambda}{|\mathcal{A}_q| + \epsilon}\sum_{(a,b)\in\mathcal{A}_q}\sigma_{\tau_r}\left(\mu_a^{(g)} - \mu_b^{(g)}\right)}_{\text{relevance separation}}\Bigg]. \tag{4}$$

The score $s_i^{(g)}$ is the query-image similarity, and $\mu_l^{(g)}$ is its mean for relevance value l. The weight w_{ij} represents the relevance difference, while $\mathcal{A}_q$ contains the available ordered pairs of relevance values. The coefficient λ balances the 2 task-reward terms, and σ_{τ_r} denotes a temperature-scaled sigmoid. The value ϵ is a small constant for numerical stability.

The teacher reward measures ranking agreement with the high-resolution teacher using Spearman rank correlation:

$$R_T^{(g)} = \rho_S\left(\mathbf{s}^T, \mathbf{s}^{(g)}\right) = \frac{\left\langle \mathrm{rk}(\mathbf{s}^T) - \bar{\mathbf{r}}^T, \mathrm{rk}(\mathbf{s}^{(g)}) - \bar{\mathbf{r}}^{(g)}\right\rangle}{\left\|\mathrm{rk}(\mathbf{s}^T) - \bar{\mathbf{r}}^T\right\|_2 \left\|\mathrm{rk}(\mathbf{s}^{(g)}) - \bar{\mathbf{r}}^{(g)}\right\|_2 + \epsilon}. \tag{5}$$

Here, $\mathbf{s}^{(g)}$ denotes the query-image similarities from the g-th candidate, and $\mathbf{s}^{T}$ denotes those from the high-resolution teacher. rk converts similarities into ranks, and $\bar{\mathbf{r}}^{T}$ and $\bar{\mathbf{r}}^{(g)}$ are the corresponding mean ranks.

Following group relative policy optimization (GRPO) (Shao et al., 2024), task and teacher rewards are standardized separately across the G embedding candidates for each query. This produces task advantage $A^{(g)}_{\text{task}}$ and teacher advantage $A^{(g)}_{T}$. The agreement indicator $\delta_g = \mathbb{I}[A^{(g)}_{\text{task}} A^{(g)}_{T} > 0]$ yields the selection objective:

$$\begin{aligned} \widetilde{A}^{(g)} &= \delta_g \big(A^{(g)}_{\text{task}} + \eta A^{(g)}_{T}\big) + \big(1-\delta_g\big)\beta A^{(g)}_{\text{task}}, \\ \mathcal{L}_{\text{RES}} &= \frac{1}{G(\sigma_Z^2+\epsilon)} \sum_{g=1}^{G} \widetilde{A}^{(g)} \mathcal{D}\Big(\mathbf{Z}_\theta, \mathrm{sg}[\mathbf{Z}^{(g)}]\Big). \end{aligned} \tag{6}$$

η weights the teacher advantage, while β downweights the task advantage when the 2 advantages disagree. $\mathcal{D}$ averages the query and candidate-image mean squared errors. σ_Z^2 sums the mean query- and image-embedding variances across the G candidates. sg denotes stop-gradient. $\mathbf{Z}_\theta$ contains the current embeddings for the query and all candidate images. Minimizing $\mathcal{L}_{\text{RES}}$ pulls $\mathbf{Z}_\theta$ toward $\mathbf{Z}^{(g)}$ when $\widetilde{A}^{(g)} > 0$ and pushes it away when $\widetilde{A}^{(g)} < 0$. This signed update completes Relevant Embedding Selection.

4.4 Training of EHiRe

Training proceeds in 2 stages. First, supervised fine-tuning (SFT) produces an initial E-ITR checkpoint. Second, this checkpoint initializes both EHiRe and the high-resolution teacher. The teacher remains frozen, while ITM and RES optimize the trainable parameters of the EHiRe backbone.

5 Experiments

5.1 Experimental Setup

Implementation details. We use WeMM as the default backbone and train both stages on the EComClutter training set with AdamW (weight decay 0.01). SFT runs for 2 epochs on 1 NVIDIA A800 GPU with a learning rate of 5×10^{-5} and an effective batch size of 128; its checkpoint initializes EHiRe and the frozen high-resolution teacher. EHiRe runs for 4 epochs on 4 NVIDIA A800 GPUs with a learning rate of 5×10^{-6} and a per-GPU image budget of 256. LoRA uses rank 16 and $\alpha = 32$. We set $(\rho, \Gamma_-, \Gamma_+, G) = (0.375, [0.90, 0.95], [1.05, 1.25], 16)$ for ITM, $(\lambda, \tau_r, \eta, \beta) = (0.5, 5.0, 0.2, 0.2)$ for RES, and the similarity temperature to 0.05.

Evaluation setting. We use an $8{:}1{:}1$ training, validation, and test split for EComClutter. The split is performed by query, so all query-image pairs for the same query remain in the same set. The default low-resolution input is $B = 224 \times 224$; only the cross-resolution evaluation changes it. For each user textual query, the associated candidate E-commerce images form its candidate-image set. Both metrics average over queries with at least 1 fully relevant candidate image (relevance $= 2$). Recall@K measures whether the top-K results contain a fully relevant image. NDCG@K evaluates the complete three-level (0/1/2) image-query relevance against the ideal ranking. We report Recall@1/3/5 and NDCG@1/3/5.

For MMEB-V2, we directly evaluate EHiRe on the official test splits of 6 Image-Text retrieval tasks without any training on MMEB-V2. We use Hit@1, the primary image retrieval metric; Avg is their macro-average. We compare CLIP-style, universal multimodal embedding, and reasoning-driven embedding methods under the same candidate-image sets, input protocol, and metric definitions.

5.2 Performance Evaluation

Main results on EComClutter. Table 2 shows that EHiRe leads the 3 baseline families on all 6 metrics and Avg under the same low-resolution inference settings. Higher Recall and NDCG reflect more accurate retrieval of fully relevant images and ranking by image-query relevance, respectively.

Table 2: Main comparison on EComClutter under low-resolution inference settings. * denotes contrastive fine-tuning without EHiRe. Avg averages 6 metrics; here and in subsequent tables, boldface and underlining mark the best and second-best results.

Method	R@1	R@3	R@5	NDCG@1	NDCG@3	NDCG@5	Avg
OpenCLIP (CVPR 2023)	0.339	0.600	0.729	0.389	0.478	0.538	0.512
SigLIP (ICCV 2023)	0.376	0.627	0.766	0.428	0.519	0.582	0.550
GME (CVPR 2025)	0.583	0.817	0.868	0.628	0.689	0.726	0.719
VLM2Vec-V2 (arXiv 2025)	0.515	0.759	0.875	0.558	0.630	0.693	0.672
Qwen3-VL-Embedding (arXiv 2026)	0.502	0.725	0.854	0.549	0.608	0.665	0.651
UME-R1 (ICLR 2026)	0.566	0.827	0.878	0.609	0.694	0.732	0.718
Embed-RL (ECCV 2026)	0.573	0.810	0.905	0.617	0.692	0.739	0.723
SigLIP* (ICCV 2023)	0.423	0.614	0.798	0.458	0.533	0.602	0.571
VLM2Vec-V2* (arXiv 2025)	0.607	0.784	0.875	0.655	0.684	0.731	0.723
GME* (CVPR 2025)	0.662	0.847	0.902	0.710	0.743	0.777	0.774
UME-R1* (ICLR 2026)	0.671	0.861	0.912	0.711	0.753	0.784	0.782
EHiRe (ours)	**0.759**	**0.912**	**0.953**	**0.793**	**0.813**	**0.846**	**0.846**

Generalization on MMEB-V2. Table 3 shows that EHiRe achieves the best Avg Hit@1 (0.795) across 6 MMEB-V2 Image-Text retrieval tasks, supporting generalization beyond EComClutter.

Table 3: Hit@1 on 6 MMEB-V2 Image-Text retrieval tasks under low-resolution inference settings. Avg is their macro-average.

Method	EDIS	MSCOCO	VisDial	VisualNews	WebQA	Wiki-SS-NQ	Avg
OpenCLIP (CVPR 2023)	0.920	0.593	0.291	**0.765**	0.627	0.552	0.625
SigLIP (ICCV 2023)	0.800	0.720	0.079	0.688	0.580	**0.585**	0.575
GME (CVPR 2025)	**0.939**	0.678	0.440	0.729	0.896	0.517	0.700
VLM2Vec-V2 (arXiv 2025)	0.804	0.725	0.681	0.718	0.893	0.473	0.716
Qwen3-VL-Embedding (arXiv 2026)	0.742	0.692	0.686	0.638	0.528	0.466	0.625
UME-R1 (ICLR 2026)	0.828	0.706	0.639	0.710	0.531	0.488	0.650
Embed-RL (ECCV 2026)	0.734	0.728	0.709	0.628	0.459	0.427	0.614
EHiRe (ours)	0.904	**0.818**	**0.821**	**0.765**	**0.912**	0.552	**0.795**

Evaluation with different backbones. Table 4 applies EHiRe to GME, Qwen3-VL-Embedding, and WeMM. It improves every metric, with average gains of 0.090, 0.166, and 0.113, supporting its applicability to different backbones.

Table 4: Results with different backbones on EComClutter. Superscripts give gains over Original; Avg is the mean of 6 metrics.

Backbone	Setting	R@1	R@3	R@5	NDCG@1	NDCG@3	NDCG@5	Avg
GME (CVPR 2025)	Original	0.583	0.817	0.868	0.628	0.689	0.726	0.719
	EHiRe (ours)	$0.724^{+0.141}$	$0.882^{+0.065}$	$0.916^{+0.048}$	$0.761^{+0.133}$	$0.763^{+0.074}$	$0.809^{+0.083}$	$0.809^{+0.090}$
Qwen3-VL-Embedding (arXiv 2026)	Original	0.502	0.725	0.854	0.549	0.608	0.665	0.651
	EHiRe (ours)	$0.728^{+0.226}$	$0.889^{+0.164}$	$0.926^{+0.072}$	$0.765^{+0.216}$	$0.778^{+0.170}$	$0.813^{+0.148}$	$0.817^{+0.166}$
WeMM (arXiv 2026)	Original	0.614	0.807	0.871	0.659	0.702	0.744	0.733
	EHiRe (ours)	$\mathbf{0.759}^{+0.145}$	$\mathbf{0.912}^{+0.105}$	$\mathbf{0.953}^{+0.082}$	$\mathbf{0.793}^{+0.134}$	$\mathbf{0.813}^{+0.111}$	$\mathbf{0.846}^{+0.102}$	$\mathbf{0.846}^{+0.113}$

Cross-resolution evaluation. Table 5 evaluates EHiRe at 112^2, 224^2, and 336^2 input resolutions. EHiRe achieves the best results on all 6 metrics at each resolution, showing consistent cross-resolution performance.

Table 5: Cross-resolution results on EComClutter. Each cell lists $112^2/224^2/336^2$.

Method	R@1	R@3	R@5	NDCG@1	NDCG@3	NDCG@5
SigLIP (ICCV 2023)	0.332 / 0.376 / 0.373	0.593 / 0.627 / 0.631	0.759 / 0.766 / 0.739	0.385 / 0.428 / 0.422	0.483 / 0.519 / 0.514	0.554 / 0.582 / 0.566
Qwen3-VL-Embedding (arXiv 2026)	0.420 / 0.502 / 0.515	0.654 / 0.725 / 0.746	0.814 / 0.854 / 0.861	0.479 / 0.549 / 0.560	0.559 / 0.608 / 0.635	0.626 / 0.665 / 0.688
Embed-RL (ECCV 2026)	0.485 / 0.573 / 0.620	0.763 / 0.810 / 0.827	0.844 / 0.905 / 0.908	0.536 / 0.617 / 0.667	0.622 / 0.692 / 0.722	0.662 / 0.739 / 0.764
EHiRe (ours)	**0.634 / 0.759 / 0.780**	**0.834 / 0.912 / 0.922**	**0.905 / 0.953 / 0.963**	**0.677 / 0.793 / 0.809**	**0.728 / 0.813 / 0.838**	**0.765 / 0.846 / 0.865**

5.3 Ablation Studies

Table 6 compares EHiRe with WeMM SFT and 3 module ablations: w/o RES, w/o ITM, and w/o DRR. The full EHiRe leads all 6 metrics, demonstrating the contribution of each module to its hierarchical design.

Table 6: Component ablations on EComClutter. Avg is the mean of 6 metrics.

Setting	R@1	R@3	R@5	NDCG@1	NDCG@3	NDCG@5	Avg
WeMM SFT (arXiv 2026)	0.685	0.862	0.907	0.725	0.746	0.771	0.783
EHiRe (ours) w/o RES	0.722	0.886	0.924	0.758	0.768	0.794	0.809
EHiRe (ours) w/o ITM	0.728	0.880	0.922	0.764	0.784	0.811	0.815
EHiRe (ours) w/o DRR	0.726	0.879	0.919	0.760	0.779	0.813	0.813
EHiRe (ours)	**0.759**	**0.912**	**0.953**	**0.793**	**0.813**	**0.846**	**0.846**

5.4 Visual Analysis

Figure 5 compares the top-3 candidate E-commerce images retrieved by EHiRe, Qwen3-VL-Embedding, Embed-RL, and SigLIP for the same user textual query. Only EHiRe ranks the fully relevant candidate first, illustrating its use of query-relevant details in this example.

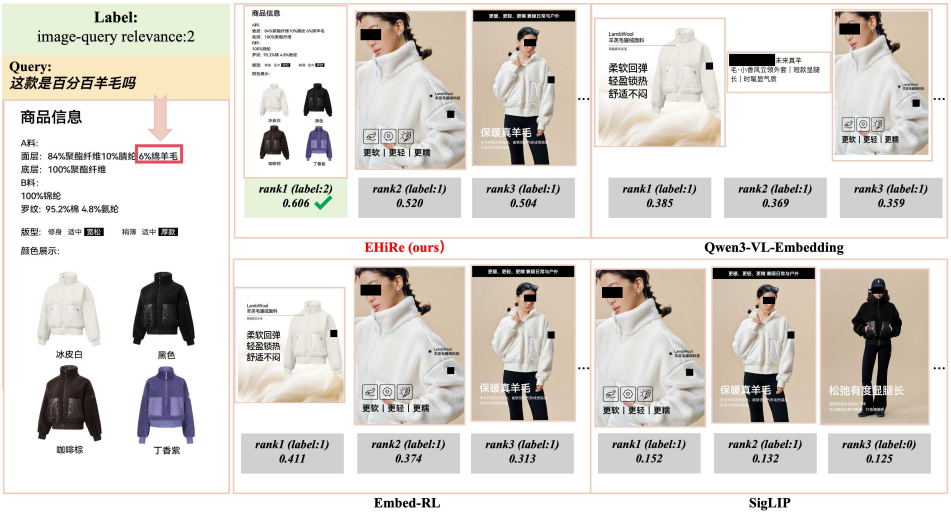

Figure 5: Qualitative retrieval comparison on EComClutter. Only EHiRe (ours) ranks the fully relevant candidate first in this example.

6 Conclusion

In this paper, we study E-commerce image-text retrieval (E-ITR) for detail-intensive E-commerce images under low-resolution inference settings. These images contain densely packed visual and textual elements. Low-resolution inputs make query-relevant products more difficult to identify in these images. To address these challenges, we contribute EComClutter, a detail-intensive E-commerce image-text retrieval database. Based on EComClutter, we further propose EHiRe, a novel E-ITR framework that hierarchically refines query-relevant representations. EHiRe achieves state-of-the-art retrieval accuracy under low-resolution inference settings. Extensive experiments and comparisons demonstrate the superiority of the proposed method. While these results are encouraging, identifying query-relevant products from detail-intensive E-commerce images under low-resolution inference settings remains challenging. We hope our contributions inspire the community to study this problem with new methods and broader perspectives.

AI USE STATEMENT

Generative AI tools were used to assist in drafting sections of this paper. MLLMs were also used in the relevance annotation pipeline described in Section 3.2. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final content of the paper.

REFERENCES

Anthropic. Claude Opus 4.8, 2026. URL https://platform.claude.com/docs/en/models/opus-4-8/overview. Accessed September 24, 2026.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023. doi: 10.1109/CVPR52729.2023.00276.

Xiao Dong, Xunlin Zhan, Yangxin Wu, Yunchao Wei, Michael C. Kampffmeyer, Xiaoyong Wei, Minlong Lu, Yaowei Wang, and Xiaodan Liang. M5Product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21220–21230, 2022. doi: 10.1109/CVPR52688.2022.02057.

Google. Gemini 3.1 Pro Preview, 2026. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview. Accessed September 24, 2026.

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. In *European Conference on Computer Vision*, pp. 39–56, 2026. doi: 10.1007/978-3-032-36839-3_3. URL https://doi.org/10.1007/978-3-032-36839-3_3.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=rJvJXZb0W.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. UME-R1: Exploring reasoning-driven generative multimodal embeddings. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2511.00405.

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. URL https://arxiv.org/abs/2601.04720.

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. URL https://arxiv.org/abs/2507.04590.

OpenAI. GPT-5.6 Sol, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol. Accessed September 24, 2026.

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. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Ruth Rosenholtz, Yuanzhen Li, and Lisa Nakano. Measuring visual clutter. *Journal of Vision*, 7(2): 17, 2007. doi: 10.1167/7.2.17. URL https://doi.org/10.1167/7.2.17.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11302–11312, 2021. doi: 10.1109/CVPR46437.2021.01115.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese CLIP: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022. URL https://arxiv.org/abs/2211.01335.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006/.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision*, pp. 11941–11952, 2023. doi: 10.1109/ICCV51070.2023.01100.

Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. Product1M: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *IEEE/CVF International Conference on Computer Vision*, pp. 11762–11771, 2021. doi: 10.1109/ICCV48922.2021.01157.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9274–9285, 2025. doi: 10.1109/CVPR52734.2025.00866. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Bridging_Modalities_Improving_Universal_Multimodal_Retrieval_by_Multimodal_Large_Language_CVPR_2025_paper.html.

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. URL https://arxiv.org/abs/2608.24053.