TIGER-FG: Text-Guided Implicit Fine-Grained Grounding for E-commerce Retrieval

Xinyu Sun *
Kuaishou Technology
Beijing, China
22221189@zju.edu.cn

Huangyu Dai*
Kuaishou Technology
Beijing, China
11931034@zju.edu.cn

Lingtao Mao*
Kuaishou Technology
Beijing, China
mltzju@163.com

Zexin Zheng
Kuaishou Technology
Beijing, China
zhengzx25@mail2.sysu.edu.cn

Zihan Liang
Kuaishou Technology
Beijing, China
liangzih@seas.upenn.edu

Ben Chen†
Kuaishou Technology
Beijing, China
benchen4395@gmail.com

Chenyi Lei†
Kuaishou Technology
Beijing, China
leichy@mail.ustc.edu.cn

Wenwu Ou
Kuaishou Technology
Beijing, China
ouwenweu@gmail.com

Abstract

E-commerce image search often takes a cropped image as the query, while each candidate is represented by full item images and structured text. This image-to-multimodal retrieval setting presents two asymmetries: a *modality disparity* – a visual query must match image–text items, and a *granularity disparity* – a cropped query must be compared with full images containing background context and possible distractors. Detection-based pipelines handle the granularity disparity through explicit localization but incur extra cost and error propagation, whereas CLIP-style encoders avoid detection, but are vulnerable to backgrounds or irrelevant items. To address these limitations, we propose **TIGER-FG**, a **t**ext-guided **i**mplicit fine-grained **g**rounding framework for image-to-multimodal **e**-commerce **r**etrieval. TIGER-FG uses item text as semantic guidance to produce target-focused item representations without object detection for retrieval. We further introduce dual distillation objectives that preserve target-region spatial consistency and query–item similarity structure, yielding more stable and discriminative multimodal representations. In addition, we construct **ECom-RF-IMMR**, a realistic benchmark suite with a 10M-pair training set and two evaluation benchmarks covering standard and cluttered item layouts. TIGER-FG achieves 80.1 Recall@1 on **ECom-RF-IMMR-Normal**, outperforming the strongest baseline by 6.1 percentage points, and reaches 75.2 Recall@1 on **ECom-RF-IMMR-Mosaic**, using only 85.7M query-side parameters and 256-dimensional embeddings. Results on public e-commerce benchmarks further demonstrate its generalization to noisy and one-to-many retrieval scenarios.

1 Introduction

In e-commerce image search, users often start with an image containing a product of interest. Practical systems first localize the target product, typically through an upstream detector that proposes candidate regions for user selection, and then use the resulting cropped region as the visual query. However, each candidate is represented by full item images with structured text such as title, category, and attributes, where the image often contains background context, multiple products, or other distractors. This setting induces the image-to-multimodal retrieval (IMMR) task [Cheng et al., 2023], where a cropped visual query must be matched against scene-level multimodal item candidates.

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

This setting exposes two challenges that are difficult for existing retrieval methods to handle. The first is the *modality disparity*, where a visual query must be represented in a space compatible with item candidates described by both images and structured text. This differs from homogeneous image–image retrieval, where the query and candidate share the same modality. The second is the *granularity disparity*, where a cropped query must be compared with full item images. As a result, global image-level representations can be dominated by visually salient but query-irrelevant content.

Existing approaches address these disparities only partially. One practical paradigm, illustrated in Figure 1a, uses explicit object detection before item encoding. Given a full item image, such methods first detect candidate item boxes, crop the corresponding regions, encode each cropped region, and then compare region embeddings with structured item text to retain the most text-compatible representation for indexing [Cheng et al., 2024, Nan et al., 2025]. A related grounding-based variant, not shown in the figure, replaces detector-based candidate region generation with text-conditioned grounding models such as GroundingDINO [Liu et al., 2024]. It takes the full item image and item text as input, grounds text-relevant regions, and then uses the selected region features for item representation. These explicit-region pipelines can mitigate the granularity disparity, but they introduce a multi-stage indexing process and make the stored representation dependent on detection or grounding quality and region selection. These issues become more pronounced in e-commerce data, where item titles and categories are long, structured, and attribute-dense, and adapting generic detection or grounding models usually requires costly in-domain annotation. Another paradigm adopts dual encoders built on vision–language pretraining, including CLIP [Radford et al., 2021, Yang et al., 2022], BLIP [Li et al., 2022a, 2023], ALIGN [Jia et al., 2021], and recent MLLM-based embedders [Li et al., 2026, Zhang et al., 2024]. These models support efficient ANN retrieval by encoding queries and candidates independently, but they mainly produce global image–text representations and often fail to focus on the query-relevant region when candidate images contain salient backgrounds, multiple objects, or other distractors. Although IMMR has been formalized as an item retrieval problem [Cheng et al., 2023], learning fine-grained multimodal item representations without explicit detection remains underexplored.

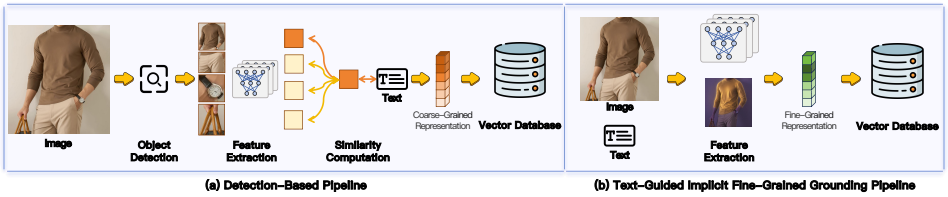


Figure 1: **Pipeline comparison for IMMR.** (a) Detection-based methods localize candidate regions from the full item image and select region features using text. (b) TIGER-FG encodes the full image–text item into a target-focused representation without object detection.

We observe that e-commerce item candidates naturally provide structured text, including titles, categories, and attributes, which often specifies the target item and its discriminative properties. Such text can serve as semantic guidance for learning target-focused item representations from full item images, without relying on explicit detection. Based on this observation, we propose **TIGER-FG**, a text-guided implicit fine-grained grounding framework for IMMR, as illustrated in Figure 1b. TIGER-FG directly produces target-focused item embeddings from full image–text candidates without box prediction during indexing or retrieval. It mitigates the modality and granularity gaps while retaining the efficiency of a dual-encoder retrieval architecture. We further use distillation objectives to preserve target-region structure and query–item similarity relationships.

Experiments on **ECom-RF-IMMR** show that TIGER-FG consistently improves image-to-multimodal item retrieval. On **ECom-RF-IMMR-Normal**, TIGER-FG achieves 80.1 Recall@1, outperforming the strongest baseline by 6.1 points with compact 256-dimensional embeddings. On **ECom-RF-IMMR-Mosaic**, which features multi-item candidate images and stronger cross-item interference, TIGER-FG achieves 75.2 Recall@1. Results on **LookBench** [Gao et al., 2026] further show that TIGER-FG transfers to noisy and one-to-many item retrieval settings. Together with ablation and qualitative analyses, these results show that structured item text, clutter-aware training, and distillation provide effective guidance for detection-free fine-grained grounding in IMMR. Our contributions are summarized as follows:

(1) **Text-guided implicit fine-grained grounding for IMMR.** We propose a detection-free item encoder that uses structured item text as semantic guidance for visual token interaction. Instead of extracting boxes from full images, the encoder directly produces target-focused representations from item candidates. This design reduces deployment cost and improves robustness when candidate images contain background context, multiple products, or other distractors.

(2) **Distillation-enhanced multimodal item representation learning.** We introduce two complementary distillation objectives for multimodal item representation learning. Spatial-relational distillation aligns target-region spatial consistency, while similarity-distribution distillation preserves the global query–item similarity structure. These objectives jointly constrain fine-grained visual structure and global retrieval behavior, improving representation stability and discriminability in IMMR.

(3) **A large-scale benchmark suite for IMMR.** We construct **ECom-RF-IMMR**, a realistic benchmark suite with a 10M-pair training set and two evaluation splits, a standard split and a cluttered mosaic split with cross-category distractors. The evaluation splits provide structured item text and region-level annotations for assessing retrieval under clean and cluttered item layouts. Experiments on ECom-RF-IMMR and two public benchmarks demonstrate the effectiveness of our method.

2 Related Work

Vision–language retrieval has been widely studied with dual-encoder architectures such as CLIP [Radford et al., 2021, Yang et al., 2022] and ALIGN [Jia et al., 2021], and with cross-attention models such as ALBEF [Li et al., 2021]. Subsequent work targets unified multimodal retrieval [Wei et al., 2024, Liang et al., 2025] and MLLM-based embeddings [Zhang et al., 2024, Li et al., 2026]. In parallel, self-supervised models such as DINO [Oquab et al., 2023, Darcet et al., 2023, Siméoni et al., 2025] learn strongly object-centric representations, and recent CLIP variants such as DeCLIP [Wang et al., 2025] and SmartCLIP [Xie et al., 2025] advance local visual representations for dense perception and fine-grained vision–language alignment, respectively. While conventional vision–language retrieval often focuses on symmetric image–text matching under global supervision, these methods do not directly address the region-to-scene asymmetry that defines IMMR, where a cropped query is matched against full-scene multimodal candidates. Detection-based retrieval pipelines provide an alternative by explicitly localizing products before matching, as discussed in Section 1 and Appendix B.

3 Methodology

Problem definition. We formalize image-to-multimodal retrieval (IMMR) as follows. The candidate set $\mathcal{G} = \{(\mathbf{I}_j^{\mathrm{p}}, \mathbf{T}_j^{\mathrm{p}})\}_{j=1}^{|\mathcal{G}|}$ consists of multimodal item entries. Each entry pairs a full item image $\mathbf{I}_j^{\mathrm{p}} \in \mathbb{R}^{H_p \times W_p \times 3}$ with a structured item text $\mathbf{T}_j^{\mathrm{p}}$, such as title, category, and attributes. The query is a cropped visual region specified by a box $\mathbf{b}^{\mathrm{q}}$ in a source image $\mathbf{I}^{\mathrm{q}} \in \mathbb{R}^{H \times W \times 3}$, which depicts the target product instance. Given $(\mathbf{I}^{\mathrm{q}}, \mathbf{b}^{\mathrm{q}})$ and a relevance indicator $y_j \in \{0, 1\}$, the task is defined as

$$\Psi_{\mathrm{IMMR}}(\mathbf{I}^{\mathrm{q}}, \mathbf{b}^{\mathrm{q}}) = \{(\mathbf{I}_j^{\mathrm{p}}, \mathbf{T}_j^{\mathrm{p}}) \in \mathcal{G} \mid y_j = 1\}. \tag{1}$$

This task is characterized by two disparities between the visual query and the multimodal item candidates. The *modality disparity* arises because the query is purely visual, whereas each candidate entry is represented by both image appearance and structured text semantics. The *granularity disparity* arises because the query captures a local item region while each candidate image shows a full item scene that may include background context, multiple products, or other distractors. We use the item-side box $\mathbf{b}_j^{\mathrm{p}}$ only as auxiliary supervision during training (§3.2), and at indexing and retrieval time the encoder consumes only $(\mathbf{I}_j^{\mathrm{p}}, \mathbf{T}_j^{\mathrm{p}})$.

Framework overview (Fig. 2). TIGER-FG addresses the two disparities with an asymmetric dual-encoder architecture. The query encoder maps the cropped query region into a visual representation, while the item encoder uses the full item image and structured text to produce a target-focused multimodal representation. Structured text guides item-side visual token interaction, enabling target-focused encoding without explicit box prediction during retrieval.

We next describe the text-guided item representation (§3.1), item-side regularization objectives (§3.2), dual-encoder training with similarity-distribution distillation (§3.3), and dataset construction (§3.4).

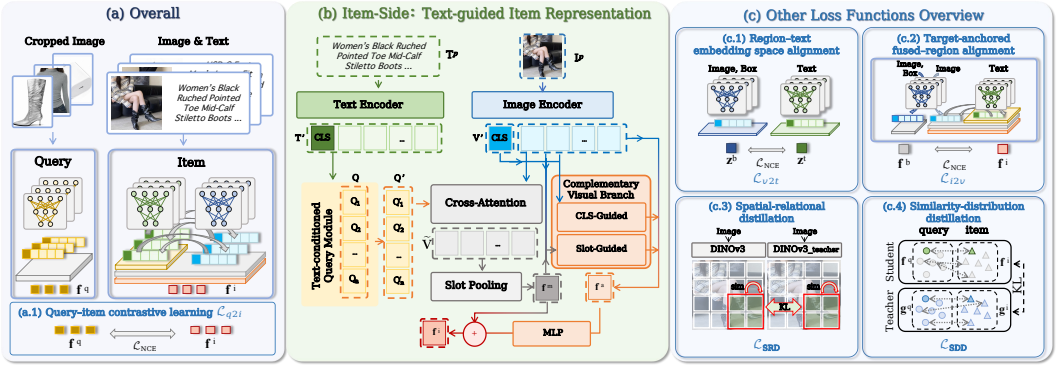


Figure 2: **Overview of the proposed TIGER-FG framework. (a)** Dual-encoder retrieval architecture. **(b)** Text-guided item representation learning. **(c)** Joint training objectives for item representation and query–item alignment.

3.1 Text-Guided Item Representation

For item-side encoding, given a multimodal item $(\mathbf{I}^{\mathrm{p}}, \mathbf{T}^{\mathrm{p}}) \in \mathcal{G}$, a DINOv3 ViT encodes $\mathbf{I}^{\mathrm{p}}$ into visual tokens $\mathbf{V} \in \mathbb{R}^{(N_v+1)\times C_v}$, and a BERT-based text encoder encodes $\mathbf{T}^{\mathrm{p}}$ into textual tokens $\mathbf{T} \in \mathbb{R}^{N_t \times C_t}$. Two linear layers project them into a shared C_u-dimensional space, giving $\mathbf{V}' \in \mathbb{R}^{(N_v+1)\times C_u}$ and $\mathbf{T}' \in \mathbb{R}^{N_t \times C_u}$. For notational simplicity, let $\mathbf{c}' = \mathbf{V}'_{[\mathrm{CLS}]} \in \mathbb{R}^{C_u}$ denote the visual class token, and let $\mathrm{sim}(\cdot,\cdot)$ denote cosine similarity. Let $\mathbf{V}'_{\mathrm{patch}} \in \mathbb{R}^{N_v \times C_u}$ denote the patch-token subset of $\mathbf{V}'$. We omit the item index j in this subsection since the same forward pass applies to all items.

Text-guided cross-attention. Following Fig. 2(b), we maintain a set of N_q learnable query tokens $\mathbf{Q} \in \mathbb{R}^{N_q \times C_u}$, similar to the learnable queries in DETR [Carion et al., 2020] and BLIP-2 [Li et al., 2023]. These query tokens are updated through two cross-attention steps. They first attend to the textual tokens to form text-conditioned queries, which then attend to the visual tokens,

$$\mathbf{Q}' = \mathrm{CA}(\mathbf{Q}, \mathbf{T}'), \qquad \widetilde{\mathbf{V}} = \mathrm{CA}(\mathbf{Q}', \mathbf{V}'_{\mathrm{patch}}), \tag{2}$$

where $\mathrm{CA}(\mathbf{X}, \mathbf{Y})$ denotes cross-attention with $\mathbf{X}$ as queries and $\mathbf{Y}$ as keys and values. Through this two-step interaction, each slot in $\widetilde{\mathbf{V}}$ carries text-conditioned visual evidence. We aggregate the slots with an MLP-based weighting and produce a single text-guided item feature $\mathbf{f}^{\mathrm{m}} \in \mathbb{R}^{C_u}$,

$$\mathbf{f}^{\mathrm{m}} = \sum_{k=1}^{N_q} \pi_k \, \widetilde{\mathbf{V}}_k, \qquad \boldsymbol{\pi} = \mathrm{softmax}\big(\mathrm{MLP}_{\pi}(\widetilde{\mathbf{V}})\big). \tag{3}$$

Complementary visual branch. While the slot module captures text-relevant semantics, it may underweight visual cues that are not explicitly mentioned in the text. We therefore add a parallel branch (the CLS-Guided path in Fig. 2(b)) that pools the visual patch tokens $\mathbf{V}'_{\mathrm{patch}}$ using two complementary anchors, the text-guided feature $\mathbf{f}^{\mathrm{m}}$ and the visual class token $\mathbf{c}'$,

$$\mathbf{s} = \lambda_m \, \mathrm{norm}\big(\mathrm{sim}(\mathbf{f}^{\mathrm{m}}, \mathbf{V}'_{\mathrm{patch}})\big) + \lambda_c \, \mathrm{norm}\big(\mathrm{sim}(\mathbf{c}', \mathbf{V}'_{\mathrm{patch}})\big), \qquad \lambda_m + \lambda_c = 1, \tag{4}$$

where $\mathrm{norm}(\cdot)$ denotes ℓ_2 normalization over the N_v tokens, putting the two similarity vectors on the same scale before fusion. The patches are then pooled by a temperature-controlled softmax over $\mathbf{s}$,

$$\mathbf{f}^{\mathrm{a}} = \sum_{k=1}^{N_v} \frac{\exp(s_k/\tau_p)}{\sum_{\ell=1}^{N_v} \exp(s_\ell/\tau_p)} \, \mathbf{V}'_{\mathrm{patch},k}. \tag{5}$$

A residual MLP combines the semantic and appearance features into the final item embedding,

$$\mathbf{f}^{\mathrm{i}} = \mathbf{f}^{\mathrm{m}} + \mathrm{MLP}(\mathbf{f}^{\mathrm{a}}) \in \mathbb{R}^{C_u}. \tag{6}$$

This embedding is used by the item-side regularizers in §3.2 and the query–item training in §3.3.

3.2 Item-Side Regularization

Given the item embedding defined above, we further regularize the item encoder with two cross-modal alignment objectives and one self-supervised distillation, all defined over a batch of size B. To avoid repetition we abbreviate the InfoNCE loss as

$$\mathcal{L}_{\mathrm{NCE}}(\mathbf{a}\to\mathbf{b};\,\tau) = -\frac{1}{B}\sum_{j=1}^{B}\log\frac{\exp(\mathrm{sim}(\mathbf{a}_j,\mathbf{b}_j)/\tau)}{\sum_{\ell=1}^{B}\exp(\mathrm{sim}(\mathbf{a}_j,\mathbf{b}_\ell)/\tau)}. \tag{7}$$

Region–text embedding space alignment ($\mathcal{L}_{v2t}$).

Following Fig. 2(c.1), we directly align the cropped target region with its text in a shared contrastive space. By aligning the cropped target region with its structured text before fusion, this objective explicitly establishes region–text correspondence in the shared alignment space. The subsequent text-guided cross-attention therefore starts from compatible visual and textual embeddings, rather than learning this correspondence only through the final fused item representation.

Let $\mathbf{I}^{\mathrm{p}}_{j,\mathrm{box}}$ denote the target region of the j-th item and $\mathbf{T}^{\mathrm{p}}_{j}$ its structured text. Before dual-encoder training, we perform CLIP-style image–text contrastive pretraining on our constructed **ECom-RF-IMMR-10M** training set, described in §3.4, with details provided in §4. We then encode $\mathbf{I}^{\mathrm{p}}_{j,\mathrm{box}}$ and $\mathbf{T}^{\mathrm{p}}_{j}$, and project their class-token features into a C_a-dimensional alignment space using the pretrained heads Proj_v and Proj_t. This alignment space is separate from the C_u space used in §3.1,

$$\mathbf{c}^{\mathrm{b}}_j = \mathrm{CLS}(\mathrm{Enc}^{\mathrm{p}}_v(\mathbf{I}^{\mathrm{p}}_{j,\mathrm{box}})),\ \mathbf{z}^{\mathrm{b}}_j = \mathrm{norm}(\mathrm{Proj}_v(\mathbf{c}^{\mathrm{b}}_j)),\ \mathbf{z}^{\mathrm{t}}_j = \mathrm{norm}(\mathrm{Proj}_t(\mathrm{CLS}(\mathrm{Enc}_t(\mathbf{T}^{\mathrm{p}}_j)))). \tag{8}$$

with $\mathbf{z}^{\mathrm{b}}_j, \mathbf{z}^{\mathrm{t}}_j \in \mathbb{R}^{C_a}$ and $\mathrm{norm}(\cdot)$ denoting ℓ_2 normalization. We use a bidirectional InfoNCE loss,

$$\mathcal{L}_{v2t} = \tfrac{1}{2}\big[\mathcal{L}_{\mathrm{NCE}}(\mathbf{z}^{\mathrm{b}}\to\mathbf{z}^{\mathrm{t}};\tau_{v2t}) + \mathcal{L}_{\mathrm{NCE}}(\mathbf{z}^{\mathrm{t}}\to\mathbf{z}^{\mathrm{b}};\tau_{t2v})\big]. \tag{9}$$

Target-anchored fused–region alignment ($\mathcal{L}_{i2v}$).

As shown in Fig. 2(c.2), we use the target-region feature $\mathbf{f}^{\mathrm{b}}_j$ as a visual anchor to keep the fused item embedding focused on the target object. The item-side branch maps the item image and structured text to $\mathbf{f}^{\mathrm{i}}_j = \mathrm{Fuse}(\mathbf{I}^{\mathrm{p}}_j, \mathbf{T}^{\mathrm{p}}_j)$. The visual anchor reuses the box CLS $\mathbf{c}^{\mathrm{b}}_j$ from $\mathcal{L}_{v2t}$ and applies a separate detached projection, $\mathbf{f}^{\mathrm{b}}_j = \mathrm{sg}\big[\mathrm{norm}(\mathbf{W}_r\mathbf{c}^{\mathrm{b}}_j)\big]$, where $\mathbf{W}_r \in \mathbb{R}^{C_u\times C_v}$. Here, $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator. The positive term uses an in-batch InfoNCE objective to pull the fused embedding toward this anchor, defined as $\mathcal{L}^{\mathrm{pos}}_{i2v} = \mathcal{L}_{\mathrm{NCE}}(\mathbf{f}^{\mathrm{i}}\to\mathbf{f}^{\mathrm{b}};\tau_{i2v})$. For the negative term, we draw K texts $\widetilde{\mathbf{T}}^{\mathrm{p}}_{j,k}$ from items whose primary category differs from that of the j-th item, and re-fuse each with the original item image to obtain $\mathbf{f}^{\mathrm{i,neg}}_{j,k} = \mathrm{norm}(\mathrm{Fuse}(\mathbf{I}^{\mathrm{p}}_j, \widetilde{\mathbf{T}}^{\mathrm{p}}_{j,k}))$. We then apply a softplus penalty to the similarity between each mismatched embedding and the visual anchor,

$$\mathcal{L}^{\mathrm{hard}}_{i2v} = \frac{1}{BK}\sum_{j=1}^{B}\sum_{k=1}^{K}\mathrm{softplus}\big(\mathrm{sim}(\mathbf{f}^{\mathrm{i,neg}}_{j,k}, \mathbf{f}^{\mathrm{b}}_j)\big), \qquad \mathcal{L}_{i2v} = \mathcal{L}^{\mathrm{pos}}_{i2v} + \lambda_h\,\mathcal{L}^{\mathrm{hard}}_{i2v}. \tag{10}$$

The cross-category constraint avoids false negatives that arise when a same-category title is visually compatible with the target. The negative term is an additive penalty rather than an in-batch denominator, which avoids competing with the InfoNCE positive on the same partition function.

Spatial-relational distillation ($\mathcal{L}_{\mathrm{SRD}}$). Following Fig. 2(c.3), we distill patch-level spatial relations from a frozen DINOv3 ViT-B/16 teacher into the item-side visual encoder. This encourages the patch-level spatial structure to remain consistent with that of the pre-trained self-supervised backbone during multimodal training. Let $\mathbf{F}^{\mathrm{s}}_j, \mathbf{F}^{\mathrm{t}}_j \in \mathbb{R}^{H_f\times W_f\times C_v}$ denote the spatially reshaped patch tokens of the j-th item from the DINOv3-based encoder on the item-side as the student encoder and the frozen DINOv3 teacher encoder, respectively, with $H_fW_f = N_v$.

Using ROI Align, which resamples the item-box region into a fixed-size feature grid, we obtain $\mathbf{R}^{\{\mathrm{s,t}\}}_j = \mathrm{ROIAlign}(\mathbf{F}^{\{\mathrm{s,t}\}}_j, \mathbf{b}^{\mathrm{p}}_j)$. We then flatten each $\mathbf{R}^{\{\mathrm{s,t}\}}_j$ into h_rw_r patch features and form $\mathbf{A}^{\mathrm{s}}_j = \mathrm{sim}(\mathbf{R}^{\mathrm{s}}_j, \mathbf{R}^{\mathrm{s}}_j)$ and $\mathbf{A}^{\mathrm{t}}_j = \mathrm{sim}(\mathbf{R}^{\mathrm{t}}_j, \mathbf{R}^{\mathrm{t}}_j)$, both in $\mathbb{R}^{h_rw_r\times h_rw_r}$. After converting each row into a

distribution with a row-wise softmax, the student similarity structure is aligned to the teacher with KL divergence,

$$\mathcal{L}_{\mathrm{SRD}} = \frac{1}{B}\sum_{j=1}^{B} D_{\mathrm{KL}}\big(\mathrm{softmax}(\mathbf{A}_j^{\mathrm{t}}) \,\|\, \mathrm{softmax}(\mathbf{A}_j^{\mathrm{s}})\big). \tag{11}$$

Restricting the supervision to the box region keeps the signal on the target object and avoids transferring teacher noise from cluttered backgrounds.

The full item-side objective is

$$\mathcal{L}_{\mathrm{item}} = \lambda_{v2t}\mathcal{L}_{v2t} + \lambda_{i2v}\mathcal{L}_{i2v} + \lambda_{\mathrm{SRD}}\mathcal{L}_{\mathrm{SRD}}. \tag{12}$$

3.3 Dual-Encoder Training

The query encoder is a DINOv3 ViT with the same architecture as the item-side visual encoder but trained with independent parameters, taking the cropped query region as input and producing the projected `[CLS]` feature $\mathbf{f}_j^{\mathrm{q}} = \mathrm{MLP}_q(\mathbf{c}_j^{\mathrm{q}}) \in \mathbb{R}^{C_u}$.

Query–item contrastive learning ($\mathcal{L}_{q2i}$). For the query-to-item direction, we augment the in-batch denominator with the same image–text mismatch hard negatives $\{\mathbf{f}_{j,k}^{\mathrm{i,neg}}\}_{k=1}^{K}$ used in $\mathcal{L}_{i2v}$,

$$\mathcal{L}_{q2i}^{\mathrm{hard}} = -\frac{1}{B}\sum_{j=1}^{B}\log\frac{s(\mathbf{f}_j^{\mathrm{q}}, \mathbf{f}_j^{\mathrm{i}})}{\sum_{\ell=1}^{B} s(\mathbf{f}_j^{\mathrm{q}}, \mathbf{f}_\ell^{\mathrm{i}}) + \sum_{k=1}^{K} s(\mathbf{f}_j^{\mathrm{q}}, \mathbf{f}_{j,k}^{\mathrm{i,neg}})}, \tag{13}$$

where $s(\mathbf{a}, \mathbf{b}) = \exp(\mathrm{sim}(\mathbf{a}, \mathbf{b})/\tau_{q2i})$. The reverse direction uses a standard InfoNCE term, and the two are averaged,

$$\mathcal{L}_{q2i} = \tfrac{1}{2}\left[\mathcal{L}_{q2i}^{\mathrm{hard}} + \mathcal{L}_{\mathrm{NCE}}(\mathbf{f}^{\mathrm{i}} \to \mathbf{f}^{\mathrm{q}}; \tau_{i2q})\right]. \tag{14}$$

Similarity-distribution distillation ($\mathcal{L}_{\mathrm{SDD}}$). To regularize the discriminative structure of item representations, we align query-to-item similarity distributions between the student encoder and a frozen MoCo-pretrained image-to-image retrieval ViT, which also appears as a baseline in our experiments. Applying the teacher to the cropped query and item regions yields region-level features $\mathbf{g}_j^{\mathrm{q}}$ and $\{\mathbf{g}_k^{\mathrm{i}}\}_{k=1}^{B}$. We define the student and teacher distributions as

$$\mathbf{p}_j^{\mathrm{s}} = \mathrm{softmax}\big(\big[\mathrm{sim}(\mathbf{f}_k^{\mathrm{i}}, \mathbf{f}_j^{\mathrm{q}})\big]_{k=1}^{B}/\tau_{\mathrm{stu}}\big), \quad \mathbf{p}_j^{\mathrm{t}} = \mathrm{softmax}\big(\big[\mathrm{sim}(\mathbf{g}_k^{\mathrm{i}}, \mathbf{g}_j^{\mathrm{q}})\big]_{k=1}^{B}/\tau_{\mathrm{tea}}\big), \tag{15}$$

and align them with KL divergence,

$$\mathcal{L}_{\mathrm{SDD}} = \frac{1}{B}\sum_{j=1}^{B} D_{\mathrm{KL}}\big(\mathbf{p}_j^{\mathrm{t}} \,\|\, \mathbf{p}_j^{\mathrm{s}}\big). \tag{16}$$

Distilling from the image-to-image teacher transfers relative visual similarity structure, complementing the multimodal and text-guided item representations learned by the student.

Joint optimization objective. We combine two cross-branch losses as $\mathcal{L}_{\mathrm{dual}} = \lambda_{q2i}\mathcal{L}_{q2i} + \lambda_{\mathrm{SDD}}\mathcal{L}_{\mathrm{SDD}}$, and integrate dual-encoder loss with the item-side loss to form the final training objective,

$$\mathcal{L} = \lambda_{\mathrm{dual}}\,\mathcal{L}_{\mathrm{dual}} + \lambda_{\mathrm{item}}\,\mathcal{L}_{\mathrm{item}}. \tag{17}$$

3.4 Dataset Construction

We construct an e-commerce dataset suite for IMMR, including a large-scale training set **ECom-RF-IMMR-10M** and two evaluation benchmarks, **ECom-RF-IMMR-Normal** and **ECom-RF-IMMR-Mosaic**. Training pairs are mined from two complementary sources, pairs of main and auxiliary item images and user image-search click logs. For these main–auxiliary pairs, no user-drawn query boxes exist. We therefore treat the auxiliary view as the query image and use GroundingDINO followed by CLIP text–image alignment filtering to localize the target product on it as a pseudo query box. For click logs, query boxes are directly obtained from user-drawn regions of interest. Item-side boxes are then generated by matching each query crop against GroundingDINO proposals on the item image

using a production image-search embedding model. We further apply Standard Product Unit (SPU) level deduplication and category-balanced sampling to obtain 10M training pairs.

The Normal evaluation set is constructed from held-out click logs using the same mining pipeline, followed by VLM-assisted and human auditing for quality control. To counteract the center bias commonly observed in e-commerce imagery and evaluate robustness under cluttered multi-item layouts, **ECom-RF-IMMR-Mosaic** re-synthesizes the candidate image for each Normal sample. The query image, query box, item title, and item category are kept unchanged, while the candidate image is rebuilt by compositing the ground-truth target crop with cross-category distractor crops on a randomly sampled background. This process creates cluttered scenes with spatial ambiguity and provides a controlled benchmark for evaluating text-guided item grounding. All datasets follow the same schema, $(\mathbf{I}^{\mathrm{q}}, \mathbf{b}^{\mathrm{q}}, \mathbf{I}^{\mathrm{p}}, \mathbf{b}^{\mathrm{p}}, \mathbf{T}^{\mathrm{p}}, \mathbf{c}^{\mathrm{p}})$. Full construction details are provided in Appendix C.

4 Experiments

Implementation details. TIGER-FG is trained on **ECom-RF-IMMR-10M** with a 1:1 mixture of original and Mosaic-augmented samples by default. Both query and item visual encoders are initialized from DINOv3 ViT-B/16 [Siméoni et al., 2025] at 224×224 resolution and trained as independent copies. The text encoder is initialized from the text branch of Chinese-CLIP ViT-B/16 [Yang et al., 2022] and further pretrained on **ECom-RF-IMMR-10M** with CLIP-style image–text contrastive learning. The unified embedding dimension is $C_u = 256$, the item branch uses $N_q = 8$ learnable query tokens, and we use $K = 1$ mismatched-text hard negative per sample. We train for 10 epochs with AdamW, a learning rate of 2×10^{-5}, and batch size 256. For the comparison in Table 1, all baseline models are fine-tuned on **ECom-RF-IMMR-10M** using only original samples without Mosaic augmentation. Matched-Mosaic retraining is reported separately in Appendix F.1. Full hyperparameters and training details are provided in Appendix G.1.

Training hard-example augmentation. We construct hard training signals with two mechanisms that affect the training objective in different ways. *Image–text mismatch* explicitly increases the number of negative item representations. For each item image $\mathbf{I}_j^{\mathrm{p}}$, we sample K mismatched texts from cross-category items and fuse them with the same image to obtain hard negative item embeddings. These negatives are added to both $\mathcal{L}_{i2v}$ and $\mathcal{L}_{q2i}$, encouraging the model to suppress visually plausible but semantically incorrect image–text pairs. *Mosaic augmentation* does not add extra negative entries to the loss. Instead, it re-synthesizes the candidate image of an original query–item pair into a cluttered multi-item scene. Specifically, we composite the target item with several cross-category distractor items on a sampled background, while keeping the query region and its matched item text unchanged. To strengthen in-batch competition, we place the items used in the same Mosaic composition into the same mini-batch. Thus, the model observes multiple candidate entries with nearly identical visual content but different item texts and different matched queries. These samples serve as hard in-batch competitors and force the item encoder to rely on structured text to focus on the correct item, rather than exploiting shared background, layout, or center bias.

Datasets and evaluation. We evaluate on two in-domain benchmarks, **ECom-RF-IMMR-Normal** and **ECom-RF-IMMR-Mosaic**, as described in §3.4. To evaluate generalization, we further adapt two public e-commerce benchmarks, **eSSPR** [Chen et al., 2023a] and **LookBench**, to the IMMR setting, where a cropped image query is used to retrieve image–text item candidates. Results are aggregated across all LookBench subsets, and additional adaptation details are provided in Appendix G.2.

Evaluation metrics. We evaluate retrieval performance using standard retrieval metrics. For **ECom-RF-IMMR-Normal**, **ECom-RF-IMMR-Mosaic**, and **eSSPR**, we report Recall@K with $K \in \{1, 4, 10\}$, and MRR@K and NDCG@K with $K \in \{4, 10\}$. For **LookBench**, we additionally report HitRate@K with $K \in \{1, 4, 10\}$.

4.1 Multimodal Retrieval Performance Comparison

We compare TIGER-FG with three groups of baselines, including CLIP/BLIP-based vision–language models, large-scale multimodal embedding models, and image retrieval models. Following UniIR [Wei et al., 2024], SF and FF denote score-level fusion and feature-level fusion variants for CLIP/BLIP-based models. Results on our constructed benchmarks are reported in Table 1, and results on public benchmarks are reported in Table 2.

Table 1: **Comparison with representative retrieval baselines on the constructed e-commerce benchmarks.** Results are reported on **ECom-RF-IMMR-Normal** and **ECom-RF-IMMR-Mosaic**. Abbreviations of large multimodal embedding models are as follows: BGE-L (*BGE-VL-Large*), GME-2B (*gme-Qwen2-VL-2B-Instruct*), Qwen3-VL-Emb (*Qwen3-VL-Embedding-2B*), and Ops-MM-Emb (*Ops-MM-embedding-v1-2B*). TIGER-FG-RAW uses the same architecture and objectives as TIGER-FG but is trained only on original samples without Mosaic augmentation, while TIGER-FG uses a 1 : 1 mixture of original and Mosaic samples. All values are in %.

Method	Query Param	Dim	ECom-RF-IMMR-Normal							ECom-RF-IMMR-Mosaic						
			Recall			MRR		NDCG		Recall			MRR		NDCG	
			@1	@4	@10	@4	@10	@4	@10	@1	@4	@10	@4	@10	@4	@10
(a) CLIP/BLIP-based Vision-Language Models																
CLIP-SF [Wei et al., 2024]	188.3M	512	66.9	85.7	92.5	74.8	75.8	77.6	79.9	30.0	49.4	61.7	37.8	39.6	40.7	44.9
CLIP-FF [Wei et al., 2024]	188.3M	512	37.2	59.3	72.0	46.0	48.0	49.4	53.7	8.1	18.1	27.6	12.0	13.4	13.5	16.7
BLIP-SF [Wei et al., 2024]	195.4M	768	67.8	84.8	91.0	74.9	75.9	77.4	79.6	14.9	25.8	33.6	19.3	20.4	20.9	23.6
BLIP-FF [Wei et al., 2024]	223.7M	768	68.8	86.0	92.2	76.0	77.0	78.6	80.7	16.1	27.7	35.8	20.7	22.0	22.5	25.2
BGE-L [Zhou et al., 2024]	0.4B	768	69.4	87.0	93.1	76.8	77.8	79.4	81.5	19.3	31.6	39.7	24.3	25.4	26.1	28.8
(b) Large-scale Multimodal Embedding Models																
GME-2B [Zhang et al., 2024]	2.2B	1536	54.9	77.1	86.4	64.1	65.5	67.4	70.6	30.4	50.5	63.1	38.4	40.3	41.5	45.7
Qwen3-VL-Emb [Li et al., 2026]	2.1B	2048	74.0	90.0	94.6	80.8	81.5	83.2	84.8	40.8	59.7	69.1	48.5	49.9	51.3	54.5
Ops-MM-Emb [OpenSearch-AI Team, 2025]	2.2B	1536	72.7	90.3	95.4	80.1	81.0	82.7	84.5	31.9	51.8	63.6	39.9	41.7	42.9	46.9
(c) Image Retrieval Models																
UniEcs [Liang et al., 2025]	198.0M	256	41.4	62.1	73.7	49.8	51.5	52.9	56.8	3.4	7.3	10.9	4.9	5.5	5.5	6.7
OClear [Cheng et al., 2023]	189.2M	256	62.7	83.0	90.8	71.1	72.3	74.1	76.8	14.3	24.7	31.9	18.4	19.5	20.0	22.4
TIGER-FG-RAW	**85.7M**	256	79.4	94.8	98.2	86.0	86.5	88.2	89.4	47.8	64.0	70.7	54.5	55.6	56.9	59.2
TIGER-FG	**85.7M**	256	**80.1**	**95.3**	**98.4**	**86.7**	**87.2**	**88.9**	**90.0**	**75.2**	**93.5**	**97.8**	**83.0**	**83.7**	**85.7**	**87.2**

On **ECom-RF-IMMR-Normal**, TIGER-FG achieves the best performance across Recall, MRR, and NDCG. The gain over strong multimodal embedding baselines shows that text-guided item representation improves the alignment between cropped visual queries and image–text item candidates, even when candidate images are relatively clean. TIGER-FG-RAW also outperforms all baselines on this split, indicating that the proposed architecture and objectives already provide strong item-level alignment without Mosaic training. The performance gap is more pronounced on **ECom-RF-IMMR-Mosaic** under the training protocol of Table 1. This split introduces dense object co-occurrence and stronger visual distractors, causing a substantial degradation for existing models. For example, Qwen3-VL-Emb drops from 74.0 to 40.8 in Recall@1, while TIGER-FG only decreases from 80.1 to 75.2. TIGER-FG-RAW also drops sharply to 47.8 Recall@1, showing that clutter-aware training is critical for robustness under multi-item ambiguity. These results suggest that model scale and generic multimodal pretraining are insufficient for IMMR, where the item representation must bridge both modality and granularity disparities.

Table 2: **Comparison with representative baselines on public benchmarks.** Results are reported on **eSSPR** and **LookBench**.

Method	eSSPR							LookBench								
	Recall			MRR		NDCG		HitRate			Recall			MRR		NDCG	
	@1	@4	@10	@4	@10	@4	@10	@1	@4	@10	@1	@4	@10	@4	@10	@4	@10
(a) CLIP/BLIP-based Vision-Language Models																	
CLIP-SF [Wei et al., 2024]	24.8	97.5	98.6	60.6	60.8	70.2	70.6	36.3	55.4	65.5	16.9	32.4	43.5	44.0	45.6	35.4	38.1
BLIP-FF [Wei et al., 2024]	20.2	**97.7**	**98.9**	58.6	58.8	68.8	69.2	30.5	47.4	57.2	14.9	28.9	38.4	37.3	38.8	30.6	33.2
(b) Large-scale Multimodal Embedding Models																	
Qwen3-VL-Emb [Li et al., 2026]	**28.8**	88.8	93.4	57.5	58.2	65.7	67.2	33.1	51.0	59.2	14.8	29.0	38.6	40.3	41.5	32.2	34.1
(c) Image Retrieval Models																	
UniEcs [Liang et al., 2025]	25.5	93.3	95.9	58.6	59.1	67.7	68.6	28.9	45.6	55.7	12.3	24.7	34.7	35.6	37.2	27.8	30.1
OClear [Cheng et al., 2023]	25.1	96.5	97.8	60.2	60.4	69.7	70.1	29.4	46.4	57.5	14.0	27.9	38.2	36.1	37.8	29.7	32.6
TIGER-FG	26.4	97.1	98.6	**61.3**	**61.6**	**70.7**	**71.2**	**39.8**	**61.3**	**71.0**	**18.5**	**38.8**	**52.2**	**48.6**	**50.0**	**41.5**	**45.0**

The same trend is observed on public benchmarks in Table 2. On **eSSPR**, where images usually contain a single clean item, Recall is already close across methods. TIGER-FG obtains comparable Recall while achieving the best MRR and NDCG, suggesting that relevant items are placed earlier in the retrieved list. On **LookBench**, which contains noisy candidates and one-to-many relevance,

TIGER-FG achieves the best performance across HitRate, Recall, MRR, and NDCG. These results suggest that the learned text-guided item representation generalizes beyond our constructed benchmark and remains effective under ambiguous candidate pools.

Qualitative results are shown in Figure 3. In Figure 3(a), given the same candidate image, TIGER-FG produces concentrated and semantically consistent responses under different text inputs. It distinguishes cup from rack in the first example and category information such as dress from attribute information such as knit in the second. Removing SRD leads to more diffuse responses, indicating that spatial-relational distillation helps preserve localized visual structure. In Figure 3(b), TIGER-FG places query features close to matched item features and forms compact category clusters with clear separation. Although Qwen3-VL-Emb also shows separable clusters, it contains more scattered outliers and query–item mismatches, suggesting weaker fine-grained alignment for IMMR.

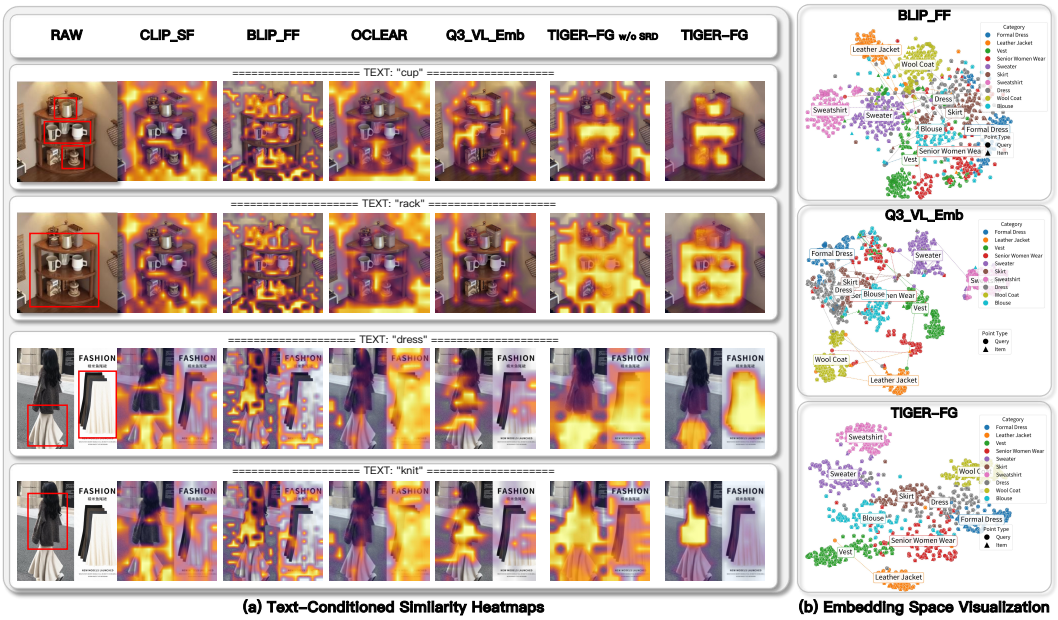


Figure 3: **Heatmap and embedding visualizations.** **(a)** Text-conditioned heatmaps show that TIGER-FG focuses on text-relevant regions. **(b)** Compared with Qwen3-VL-Emb, TIGER-FG forms tighter category clusters and closer query–item alignment.

4.2 Ablation Study

We conduct ablations on **ECom-RF-IMMR-Mosaic**, with results summarized in Table 3. Here CVB refers to the complementary visual branch, SRD to spatial-relational distillation, and SDD to similarity-distribution distillation. The largest drop comes from TIGER-FG-RAW, which uses the same architecture and objectives but is trained only on original samples without Mosaic augmentation. Its Recall@1 decreases from 75.2 to 47.8, showing that clutter-aware training is essential for handling multi-item ambiguity. Replacing DINOv3 with a CLIP-based backbone also causes a clear degradation, confirming the importance of region-sensitive visual representations for IMMR. Removing CVB leads to only a small drop, indicating that the text-guided branch captures the main semantic signal while CVB provides complementary appearance cues. For distillation, removing SDD substantially reduces performance, suggesting that query–item similarity structure from the image-to-image teacher improves retrieval discrimination. In contrast, removing SRD has little effect on ranking metrics, indicating that SRD mainly acts as a grounding-oriented regularizer rather than the primary driver of retrieval accuracy.

4.3 Modality Comparison

We analyze query-side and item-side modality configurations on **ECom-RF-IMMR-Normal**, with results reported in Table 4. The query is represented by a cropped region, while the item side uses the

corresponding image, its associated text, or both. Single-modality settings show clear limitations. The DINOv3 visual baseline achieves 76.7 Recall@1, indicating that region-level visual representations provide a strong retrieval signal. In contrast, using only item text yields 30.1 Recall@1, highlighting its limitations for fine-grained matching. Our model with only item images is lower than the DINOv3 baseline, suggesting that its advantage is not evident in unimodal visual matching. Adding textual information on the item side consistently improves performance. Combining visual and textual signals gives the best results, reaching 80.6 Recall@1 with cropped item images and 80.1 Recall@1 with full item images, and surpassing all single-modality baselines. These results show that text-guided interaction helps align localized query regions with item candidates through both visual evidence and textual semantics.

Table 3: **Ablation Studies.**

Method	ECom-RF-IMMR-Mosaic						
	Recall			MRR		NDCG	
	@1	@4	@10	@4	@10	@4	@10
TIGER-FG	75.2	**93.5**	**97.8**	83.0	83.7	**85.7**	**87.2**
CLIP-backbone	64.6	85.8	93.1	73.5	74.6	76.6	79.1
w/o CVB	74.8	93.3	**97.8**	82.7	83.4	85.4	87.0
w/o SRD	**75.3**	**93.5**	**97.8**	**83.1**	**83.8**	**85.7**	**87.2**
w/o SDD	69.9	91.1	96.8	78.8	79.7	81.9	83.9
TIGER-FG-RAW	47.8	64.0	70.7	54.5	55.6	56.9	59.2

Table 4: **Modality comparison.**

Method	ECom-RF-IMMR-Normal						
	Recall			MRR		NDCG	
	@1	@4	@10	@4	@10	@4	@10
DINOv3 (C.→C.)	76.7	93.8	97.9	84.0	84.6	86.5	87.9
DINOv3 (C.→I.)	72.6	90.4	95.3	80.1	80.9	82.8	84.5
Eco-CLIP (C.→T.)	30.1	57.1	73.3	40.8	43.2	44.9	50.4
TIGER-FG (C.→C.)	27.4	45.2	59.2	34.4	36.5	37.2	41.9
TIGER-FG (C.→I.)	26.0	43.2	56.9	32.8	34.8	35.4	40.0
TIGER-FG (C.→C.+T.)	**80.6**	**95.7**	**98.7**	**87.1**	**87.6**	**89.3**	**90.4**
TIGER-FG (C.→I.+T.)	80.1	95.3	98.4	86.7	87.2	88.9	90.0

5 Conclusion

We study image-to-multimodal item retrieval (IMMR), where a cropped visual query is matched against item candidates represented by full images and structured text. This setting introduces modality and granularity disparities that are not well addressed by standard image–text retrieval models. We propose TIGER-FG, a text-guided retrieval framework that learns target-focused item representations. We further construct **ECom-RF-IMMR**, a large-scale benchmark suite for training and evaluating IMMR under clean and cluttered mosaic item scenes. Experiments on both in-domain and public e-commerce benchmarks show consistent improvements over strong retrieval baselines, especially in multi-item, noisy and one-to-many scenarios. Future work will explore more robust grounding under noisy, sparse, or weakly aligned supervision. We provide a brief discussion of scope and practical limitations in Appendix A.

Acknowledgments and Disclosures

Funding. This work was supported by Kuaishou Technology, which provided computational resources and research infrastructure.

Competing interests. The authors declare no competing financial interests beyond their disclosed affiliations.

References

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pages 9630–9640. IEEE, 2021.

Ben Chen, Linbo Jin, Xinxin Wang, Dehong Gao, Wen Jiang, and Wei Ning. Unified vision-language representation modeling for e-commerce same-style products retrieval. In *Companion Proceedings of the ACM Web Conference 2023*, pages 381–385, 2023a.

Weijing Chen, Linli Yao, and Jin Qin. Rethinking benchmarks for cross-modal image-text retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*, 2023b.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16901–16911, 2024.

Zida Cheng, Chen Ju, Shuai Xiao, Xu Chen, Zhonghua Zhai, Xiaoyi Zeng, Weilin Huang, and Junchi Yan. Category-oriented representation learning for image to multi-modal retrieval. *arXiv preprint arXiv:2305.03972*, 2023.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21252–21262, 2022.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024. URL `https://arxiv.org/abs/2401.08281`.

Chao Gao, Siqiao Xue, Jiwen Fu, Tingyi Gu, Shanshan Li, Fan Zhou, et al. Lookbench: A live and holistic open benchmark for fashion image retrieval. *arXiv preprint arXiv:2601.14706*, 2026.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, et al. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 531–550, 2025.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 2021.

Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. Learning instance-level representation for large-scale multi-modal pretraining in e-commerce. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11060–11069, 2023.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr: Modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL `https://arxiv.org/abs/2104.12763`.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 2022a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 2023.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022b. URL `https://openaccess.thecvf.com/content/CVPR2022/html/Li_Grounded_Language-Image_Pre-Training_CVPR_2022_paper.html`.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Chen Keqin, 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.

Zihan Liang, Yufei Ma, ZhiPeng Qian, Huangyu Dai, Zihan Wang, Ben Chen, Chenyi Lei, Yuqing Ding, and Han Li. Uniecs: Unified multimodal e-commerce search framework with gated cross-modal fusion. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 2025.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*, 2024.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, 2024.

Xinyu Nan, Lingtao Mao, Huangyu Dai, Zexin Zheng, Xinyu Sun, Zihan Liang, Ben Chen, Yuqing Ding, Chenyi Lei, Wenwu Ou, et al. Unidgf: A unified detection-to-generation framework for hierarchical object visual recognition. *arXiv preprint arXiv:2511.15984*, 2025.

OpenSearch-AI Team. Ops-mm-embedding. `https://huggingface.co/OpenSearch-AI/Ops-MM-embedding-v1-2B`, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. Declip: Decoupled learning for open-vocabulary dense perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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

Shaoan Xie, Lingjing Lingjing, Yujia Zheng, Yu Yao, Zeyu Tang, Eric P Xing, Guangyi Chen, and Kun Zhang. Smartclip: Modular vision-language alignment with identification guarantees. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

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.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022.

Licheng Yu, Jun Chen, Animesh Sinha, Mengjiao Wang, Yu Chen, Tamara L Berg, and Ning Zhang. Commercemm: Large-scale commerce multimodal representation learning with omni retrieval. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 4433–4442, 2022.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11782–11791, 2021.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024.

Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. Megapairs: Massive data synthesis for universal multimodal retrieval. *arXiv preprint arXiv:2412.14475*, 2024.

A Limitations

TIGER-FG is designed for image-to-multimodal item retrieval. In this work, we instantiate and evaluate it in the e-commerce setting, where each candidate is typically represented by full item images paired with structured item text. This setting provides a practical testbed for studying cross-modal and granularity disparities in fine-grained item retrieval. Further evaluation may consider item entries with different text fields, category systems, and language styles.

Our benchmarks focus on item-level retrieval under clean and cluttered visual layouts. They are intended to provide controlled and reproducible evaluation of retrieval performance, rather than to simulate all factors in online serving, such as user personalization or live traffic dynamics.

Finally, our training recipe uses item boxes only for auxiliary supervision and dataset construction. These boxes are not required during indexing or retrieval, where TIGER-FG directly encodes full image–text item candidates. Exploring weaker or automatically checked signals may further simplify data construction.

B Extended Related Work

B.1 Vision–language representation learning

Vision–language retrieval learns a shared embedding space for visual and textual inputs. Large-scale contrastive models such as ALIGN [Jia et al., 2021] and CLIP [Radford et al., 2021, Yang et al., 2022] show strong transferability across multimodal retrieval tasks. Subsequent work improves cross-modal alignment with stronger interaction or training signals: ALBEF [Li et al., 2021] introduces cross-attention, BLIP [Li et al., 2022a] uses bootstrapped captioning and filtering, BLIP-2 [Li et al., 2023] connects frozen image encoders and language models with a lightweight querying transformer, and FILIP [Yao et al., 2022] strengthens fine-grained alignment with token-level late interaction. More recent studies, including UniIR [Wei et al., 2024] and UniECS [Liang et al., 2025], further explore unified embedding spaces for heterogeneous retrieval.

These methods provide strong general-purpose retrieval foundations, but they do not directly address the asymmetric matching problem in e-commerce IMMR. Related multimodal retrieval studies also emphasize fine-grained product matching [Zhan et al., 2021, Dong et al., 2022, Yu et al., 2022, Chen et al., 2023a, Jin et al., 2023]. In our setting, a cropped visual query must be matched to the corresponding item region inside a full image–text candidate. This requires item-level discrimination and fine-grained visual grounding beyond coarse image–text relevance [Chen et al., 2023b].

B.2 Region-aware visual representation learning

Another line of work studies local and object-level visual representations. Self-supervised models in the DINO family [Caron et al., 2021, Oquab et al., 2023, Darcet et al., 2023, Siméoni et al., 2025] often produce spatially meaningful patch features, making them useful priors for region-aware representation learning without dense supervision. Recent CLIP-based variants, such as DeCLIP [Wang et al., 2025] and SmartCLIP [Xie et al., 2025], further improve local visual representations for dense perception and fine-grained vision–language alignment, respectively. In parallel, e-commerce pretraining and retrieval benchmarks emphasize instance-level and fine-grained item discrimination [Zhan et al., 2021, Dong et al., 2022, Jin et al., 2023].

These works are closely related to fine-grained representation learning, but most of them are developed for generic vision–language settings or for symmetric item matching. Our e-commerce setting involves asymmetric retrieval, where the query is already object-centric, whereas the candidate image may contain multiple objects. We therefore use DINO-style object-centric features as a frozen teacher for spatial-relational distillation, together with structured item text as task-specific guidance for implicit region selection [Yu et al., 2022, Chen et al., 2023a].

B.3 Grounding-based retrieval pipelines

Industrial e-commerce retrieval often handles the granularity disparity through an explicit detection or grounding stage. Grounded vision–language models such as MDETR [Kamath et al., 2021] and

GLIP [Li et al., 2022b] learn phrase-region or text-region alignment from detection-style supervision. Open-vocabulary detectors and grounding models, including YOLO-World [Cheng et al., 2024] and Grounding DINO [Liu et al., 2024], can localize candidate objects from textual prompts. UniDGF [Nan et al., 2025] further combines detection with fine-grained recognition, but still relies on explicit object localization.

This pipeline design introduces additional computation and can propagate localization errors to region filtering and post-hoc matching. It also faces domain-transfer challenges in e-commerce, where item titles, categories, and attributes are often structured, attribute-dense, and domain-specific. TIGER-FG removes this explicit stage by using structured item text as a semantic cue for soft, patch-level region selection inside the retrieval encoder.

B.4 MLLM-based multimodal embeddings

Recent MLLM-based embedding models extend retrieval to more flexible multimodal inputs, including images, text, and their combinations. Models such as GME [Zhang et al., 2024], MM-Embed [Lin et al., 2024], Ops-MM-Embedding [OpenSearch-AI Team, 2025], jina-embeddings [Günther et al., 2025], and Qwen3-VL-Embedding [Li et al., 2026] show strong retrieval ability under heterogeneous multimodal inputs. They typically build on large multimodal backbones and are trained with broad mixtures of single-modal, cross-modal, and fused-modal data, which improves generalization across retrieval tasks.

Compared with earlier dual-encoder retrieval models, MLLM-based embedders offer stronger generality and support more flexible input formats. Their large backbones and broad objectives make them closer to general-purpose multimodal retrieval systems than lightweight domain-specific item encoders. We include them as representative large-scale multimodal embedding baselines, while keeping TIGER-FG compatible with offline item encoding and standard vector retrieval systems [Douze et al., 2024].

C Dataset Construction Details

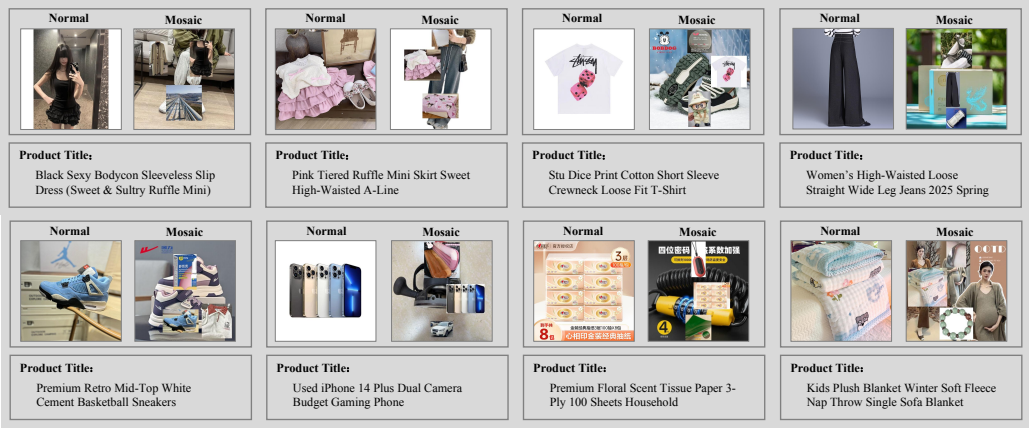


Figure 4: **Paired examples from ECom-RF-IMMR-Normal and ECom-RF-IMMR-Mosaic.** Each cell shows, for the same item text, the original Normal item image (left) and its Mosaic re-synthesis (right). The Mosaic image pastes the Normal target crop, together with cross-category distractors, onto a random background at random scales and locations.

We construct the ECom-RF-IMMR suite—the training set **ECom-RF-IMMR-10M** together with the evaluation benchmarks **ECom-RF-IMMR-Normal** and **ECom-RF-IMMR-Mosaic**—under a single schema $(\mathbf{I}^{\mathrm{q}}, \mathbf{b}^{\mathrm{q}}, \mathbf{I}^{\mathrm{p}}, \mathbf{b}^{\mathrm{p}}, \mathbf{T}^{\mathrm{p}}, \mathbf{c}^{\mathrm{p}})$, with two shared subroutines doing most of the work: CANDIDATEMINING extracts the query-side region, and ITEMBOXPAIRING localizes its counterpart on the item side. Algorithm 1 gives the full pipeline; the rest of this section walks through the parts that need elaboration.

Query-side sources. Query tuples come from (i) product catalogs, where each product ships a main view $\mathbf{I}^{\mathrm{p}}_{\mathrm{main}}$ and an auxiliary view $\mathbf{I}^{\mathrm{p}}_{\mathrm{sub}}$, and (ii) user image-search click logs, which natively contain user-drawn query boxes $\mathbf{b}^{\mathrm{q}}$. For the product source we first drop perceptually identical main/sub pairs via a perceptual hash, then treat the auxiliary view as the query and use GroundingDINO conditioned on the title $\mathbf{T}^{\mathrm{p}}$ to propose a region, keeping only those whose CLIP text–image similarity exceeds a category-aware threshold τ_{clip} (typically in the range of 0.25–0.35 in our e-commerce setting). This yields a localized $\mathbf{b}^{\mathrm{q}}$ whose content is consistent with the title.

Item-side box assignment. Item images frequently depict full scenes or multi-object compositions, so we do not assume a trivial one-to-one correspondence. For every item image $\mathbf{I}^{\mathrm{p}}$, we generate open-vocabulary proposals via GroundingDINO, embed both the query crop and each proposal through an online representation model $\mathcal{E}$, and select the proposal $\mathbf{b}^{\mathrm{p}\star}$ whose embedding is closest to that of the query crop. We then filter by this query–item cosine similarity s: pairs with $s > \tau_{\mathrm{high}}{=}0.97$ are discarded as the two crops are nearly identical, and pairs with $s < \tau_{\mathrm{low}}{=}0.80$ are discarded as the two crops are unlikely to depict the same product. In our dataset construction pipeline, $\mathcal{E}$ is used only for offline box assignment and filtering and does not share training samples, supervision, parameters, or checkpoints with TIGER-FG or its distillation teachers.

Training-set post-processing. We deduplicate at the SPU level: if multiple pairs share the same item SPU, we keep only one and drop the rest. This ensures distinct pairs in a batch correspond to distinct products, so they can serve as valid in-batch negatives for contrastive learning. After category-balanced sampling, we obtain 10M pairs as **ECom-RF-IMMR-10M**.

Normal evaluation set. **ECom-RF-IMMR-Normal** is built mainly from held-out user click logs, where the query is a user-uploaded photo and the item is the product the user clicked on. The training set and the Normal evaluation set are disjoint, with no duplicate samples shared between them. After the same mining pipeline, every candidate is passed through a verifier $\mathcal{V}$ (VLM pre-filtering followed by human auditing) to remove residual noise from automated mining. We then apply category-balanced sampling to obtain 100K final pairs as **ECom-RF-IMMR-Normal**.

Category coverage and diversity. A natural concern is whether category-balanced sampling produces a benchmark that is genuinely diverse, or one that simply over-represents a few large verticals. To answer this, we analyze the taxonomy of **ECom-RF-IMMR-Normal** along three axes (Figure 5). Each sample carries a four-level category path $\mathbf{c}^{\mathrm{p}}{=}(c_1, c_2, c_3, c_{\mathrm{leaf}})$, with 100K samples spanning 76 L1, 809 L2, 5,903 L3, and 10,859 leaf categories in total. Panel (a) shows the L1 distribution: no single L1 exceeds 5%, and the top verticals span apparel, home, hardware, fresh food, electronics, and appliances—going well beyond the fashion- or general-object-centric setups of most existing benchmarks. Panel (b) shows within-L1 diversity: L2 counts stay in the single or low double digits across most verticals while leaf counts often run into the hundreds (e.g., *Home & Living* has 9 L2 and 796 leaves; *Hardware & Tools* has 24 L2 and 803 leaves), indicating that the bulk of the fine-grained variation lives at the leaf level. Panel (c) makes the long tail explicit: covering 80% of samples requires only 40 of 76 L1 but 5,647 of 10,859 leaves—roughly half of L1 versus over half of leaves, showing that L1 is close to uniform while the leaf distribution is heavily long-tailed, so the benchmark genuinely stresses long-tail retrieval.

Mosaic evaluation set. **ECom-RF-IMMR-Mosaic** is derived entirely from **ECom-RF-IMMR-Normal**: the query image $\mathbf{I}^{\mathrm{q}}$, the query box $\mathbf{b}^{\mathrm{q}}$, and the item title/category $(\mathbf{T}^{\mathrm{p}}, \mathbf{c}^{\mathrm{p}})$ are all kept verbatim, and only the item image and its box are re-synthesized. We build two pools directly from Normal: a *background pool* $\mathcal{I}_{\mathrm{bg}}$, consisting of the full item images, and a *distractor pool* $\mathcal{I}_{\mathrm{dist}}$, consisting of the item target crops together with their categories. For each Normal tuple with target crop $\mathbf{I}_{\mathrm{tgt}}{=}\,\mathrm{crop}(\mathbf{I}^{\mathrm{p}}, \mathbf{b}^{\mathrm{p}})$, we sample (i) a background $\mathbf{I}_{\mathrm{bg}} \sim \mathcal{I}_{\mathrm{bg}}$ drawn from a different Normal sample whose SPU and category both differ from those of the current tuple, to prevent leakage from the background itself, and (ii) up to $k{=}4$ distractor crops from $\mathcal{I}_{\mathrm{dist}}$ whose categories differ from $\mathbf{c}^{\mathrm{p}}$. The target and distractors are then pasted onto the background at random scales and positions, with placements rejected and resampled to keep overlaps between objects to a minimum. Each Normal tuple yields exactly one Mosaic tuple, giving a final set of 100K pairs. The transformed target box is obtained from the synthesis process. Randomizing the surrounding context with cross-category distractors introduces controlled multi-object interference and removes the center bias of e-commerce

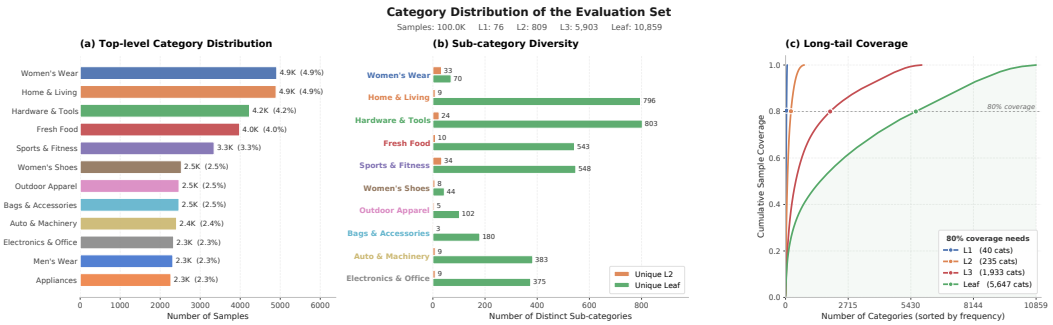


Figure 5: **Category distribution of ECom-RF-IMMR-Normal.** (a) Top-level (L1) distribution showing a balanced mix of verticals, with no single L1 exceeding ∼5% of samples. (b) Within-L1 diversity, measured by the number of distinct L2 and leaf categories under each of the top L1 verticals. (c) Long-tail cumulative coverage: reaching 80% of samples requires 40 L1, 235 L2, 1,933 L3, and 5,647 leaf categories, highlighting the benchmark's long-tail nature. Overall statistics: 100K samples over 76 L1 / 809 L2 / 5,903 L3 / 10,859 leaf categories.

imagery, forcing models to ground the query and title to the correct object rather than to positional or saliency priors.

Qualitative examples of Normal vs. Mosaic. Figure 4 juxtaposes item images of eight representative samples from **ECom-RF-IMMR-Normal** and their re-synthesized counterparts in **ECom-RF-IMMR-Mosaic**, while the query image, query box, and item title are held fixed. In Normal, the target product typically dominates the frame and is roughly centered, so a naive global-matching model can often succeed by relying on the dominant object or positional priors. In contrast, Mosaic embeds the same target crop into a heterogeneous scene that also contains one or more cross-category distractors (e.g., a handbag next to a mini skirt, a sneaker next to a T-shirt, or a hairdryer next to a tissue-paper pack), at varying scales and spatial locations. This explicitly breaks the center bias and forces retrieval to rely on the item title $\mathbf{T}^{\mathrm{p}}$ to disambiguate the target.

Algorithm 1 Unified Data Construction Pipeline for the ECom-RF-IMMR Benchmarks

Require: Product catalog $\mathcal{P}$, where each $p \in \mathcal{P}$ provides $(\mathbf{I}^{\mathrm{p}}_{\mathrm{main}}, \mathbf{I}^{\mathrm{p}}_{\mathrm{sub}}, \mathbf{T}^{\mathrm{p}}, \mathbf{c}^{\mathrm{p}})$; user image-search click log $\mathcal{L}$ with tuples $(\mathbf{I}^{\mathrm{q}}, \mathbf{b}^{\mathrm{q}}, \mathbf{I}^{\mathrm{p}}, \mathbf{T}^{\mathrm{p}}, \mathbf{c}^{\mathrm{p}})$; GroundingDINO detector $\mathcal{G}_{\mathrm{det}}$; CLIP model $\mathcal{C} = (\mathcal{C}_v, \mathcal{C}_t)$; online representation model $\mathcal{E}$; SPU taxonomy $\mathcal{S}$; VLM/human verifier $\mathcal{V}$; thresholds τ_{clip}, τ_{low}=0.80, τ_{high}=0.97; number of distractors k

Ensure: Training set $\mathcal{D}_{\mathrm{train}}$ and evaluation sets $\mathcal{D}^{\mathrm{N}}_{\mathrm{eval}}$, $\mathcal{D}^{\mathrm{M}}_{\mathrm{eval}}$; all tuples share the schema $(\mathbf{I}^{\mathrm{q}}, \mathbf{b}^{\mathrm{q}}, \mathbf{I}^{\mathrm{p}}, \mathbf{b}^{\mathrm{p}}, \mathbf{T}^{\mathrm{p}}, \mathbf{c}^{\mathrm{p}})$

procedure CANDIDATEMINING($\mathcal{P}, \mathcal{L}$) ▷ query-side mining from two heterogeneous sources
  $\mathcal{Q} \leftarrow \emptyset$
  # Source 1: product main / sub image pairs ($\mathbf{b}^{\mathrm{q}}$ must be inferred)
  for each product $\mathrm{p} \in \mathcal{P}$ **do**
    if PHASH($\mathbf{I}^{\mathrm{p}}_{\mathrm{main}}$) = PHASH($\mathbf{I}^{\mathrm{p}}_{\mathrm{sub}}$) **then continue** ▷ drop visually identical main/sub
    end if
    $\mathbf{I}^{\mathrm{q}} \leftarrow \mathbf{I}^{\mathrm{p}}_{\mathrm{sub}}$, $\mathbf{I}^{\mathrm{p}} \leftarrow \mathbf{I}^{\mathrm{p}}_{\mathrm{main}}$ ▷ treat sub-image as query, main image as item
    $\mathbf{b}^{\mathrm{q}} \leftarrow \mathcal{G}_{\mathrm{det}}(\mathbf{I}^{\mathrm{q}}; \mathbf{T}^{\mathrm{p}})$ ▷ title-grounded box proposal
    if $\cos\big(\mathcal{C}_v(\mathrm{crop}(\mathbf{I}^{\mathrm{q}}, \mathbf{b}^{\mathrm{q}})), \mathcal{C}_t(\mathbf{T}^{\mathrm{p}})\big) < \tau_{\mathrm{clip}}$ **then**
      continue ▷ CLIP text–image alignment filter
    end if
    $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{(\mathbf{I}^{\mathrm{q}}, \mathbf{b}^{\mathrm{q}}, \mathbf{I}^{\mathrm{p}}, \mathbf{T}^{\mathrm{p}}, \mathbf{c}^{\mathrm{p}})\}$
  end for
  # Source 2: click log ($\mathbf{b}^{\mathrm{q}}$ is the user-drawn crop from production traffic)
  for each $(\mathbf{I}^{\mathrm{q}}, \mathbf{b}^{\mathrm{q}}, \mathbf{I}^{\mathrm{p}}, \mathbf{T}^{\mathrm{p}}, \mathbf{c}^{\mathrm{p}}) \in \mathcal{L}$ **do**
    $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{(\mathbf{I}^{\mathrm{q}}, \mathbf{b}^{\mathrm{q}}, \mathbf{I}^{\mathrm{p}}, \mathbf{T}^{\mathrm{p}}, \mathbf{c}^{\mathrm{p}})\}$
  end for
  return $\mathcal{Q}$
end procedure

procedure ITEMBOXPAIRING($\mathcal{Q}$) ▷ lift item images to item crops via embedding similarity
  $\mathcal{D} \leftarrow \emptyset$
  for each $(\mathbf{I}^{\mathrm{q}}, \mathbf{b}^{\mathrm{q}}, \mathbf{I}^{\mathrm{p}}, \mathbf{T}^{\mathrm{p}}, \mathbf{c}^{\mathrm{p}}) \in \mathcal{Q}$ **do**
    $\mathcal{B}^{\mathrm{p}} \leftarrow \mathcal{G}_{\mathrm{det}}(\mathbf{I}^{\mathrm{p}})$ ▷ open-vocabulary proposals on the item image
    $\mathbf{e}^{\mathrm{q}} \leftarrow \mathcal{E}(\mathrm{crop}(\mathbf{I}^{\mathrm{q}}, \mathbf{b}^{\mathrm{q}}))$
    $\mathbf{b}^{\mathrm{p}\star} \leftarrow \arg\max_{\mathbf{b} \in \mathcal{B}^{\mathrm{p}}} \cos\big(\mathbf{e}^{\mathrm{q}},\, \mathcal{E}(\mathrm{crop}(\mathbf{I}^{\mathrm{p}}, \mathbf{b}))\big)$
    $s \leftarrow \cos\big(\mathbf{e}^{\mathrm{q}},\, \mathcal{E}(\mathrm{crop}(\mathbf{I}^{\mathrm{p}}, \mathbf{b}^{\mathrm{p}\star}))\big)$
    if $s > \tau_{\text{high}}$ **then continue** ▷ near-duplicate crop: leakage / trivial match
    else if $s < \tau_{\text{low}}$ **then continue** ▷ likely not the same instance
    end if
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{I}^{\mathrm{q}}, \mathbf{b}^{\mathrm{q}}, \mathbf{I}^{\mathrm{p}}, \mathbf{b}^{\mathrm{p}\star}, \mathbf{T}^{\mathrm{p}}, \mathbf{c}^{\mathrm{p}})\}$
  end for
  return $\mathcal{D}$
end procedure

Stage 1 — Training set (ECom-RF-IMMR-10M)
$\mathcal{Q} \leftarrow$ CANDIDATEMINING($\mathcal{P}, \mathcal{L}$)
$\mathcal{D} \leftarrow$ ITEMBOXPAIRING($\mathcal{Q}$)
$(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{held-out}}) \leftarrow$ PARTITIONBYSPLIT($\mathcal{D}$)
▷ use the existing disjoint training and evaluation split
$\mathcal{D}_{\text{train}} \leftarrow$ SPUCOARSEDEDUP($\mathcal{D}_{\text{train}}, \mathcal{S}$)
▷ remove pairs sharing the same SPU so that distinct pairs serve as valid in-batch negatives
$\mathcal{D}_{\text{train}} \leftarrow$ CATEGORYBALANCEDSAMPLING($\mathcal{D}_{\text{train}}$)
▷ re-balance category distribution to avoid long-tail dominance

Stage 2 — ECom-RF-IMMR-Normal evaluation set
$\mathcal{D}^{\mathrm{N}}_{\text{eval}} \leftarrow \emptyset$
for each $x = (\mathbf{I}^{\mathrm{q}}, \mathbf{b}^{\mathrm{q}}, \mathbf{I}^{\mathrm{p}}, \mathbf{b}^{\mathrm{p}}, \mathbf{T}^{\mathrm{p}}, \mathbf{c}^{\mathrm{p}}) \in \mathcal{D}_{\text{held-out}}$ **do** ▷ held-out pairs, primarily from click logs
  if $\mathcal{V}(x) =$ VALID **then** ▷ VLM pre-filter followed by human audit
    $\mathcal{D}^{\mathrm{N}}_{\text{eval}} \leftarrow \mathcal{D}^{\mathrm{N}}_{\text{eval}} \cup \{x\}$
  end if
end for
$\mathcal{D}^{\mathrm{N}}_{\text{eval}} \leftarrow$ CATEGORYBALANCEDSAMPLING($\mathcal{D}^{\mathrm{N}}_{\text{eval}}$)

Stage 3 — ECom-RF-IMMR-Mosaic evaluation set (derived from Normal)
$\mathcal{I}_{\text{bg}} \leftarrow \{\mathbf{I}^{\mathrm{p}} \mid (\cdot, \cdot, \mathbf{I}^{\mathrm{p}}, \cdot, \cdot, \cdot) \in \mathcal{D}^{\mathrm{N}}_{\text{eval}}\}$ ▷ background pool: full item images from Normal
$\mathcal{I}_{\text{dist}} \leftarrow \{(\mathrm{crop}(\mathbf{I}^{\mathrm{p}}, \mathbf{b}^{\mathrm{p}}),\, \mathbf{c}^{\mathrm{p}}) \mid (\cdot, \cdot, \mathbf{I}^{\mathrm{p}}, \mathbf{b}^{\mathrm{p}}, \cdot, \mathbf{c}^{\mathrm{p}}) \in \mathcal{D}^{\mathrm{N}}_{\text{eval}}\}$
▷ distractor pool: target crops with categories (used for cross-category sampling)
$\mathcal{D}^{\mathrm{M}}_{\text{eval}} \leftarrow \emptyset$
for each $(\mathbf{I}^{\mathrm{q}}, \mathbf{b}^{\mathrm{q}}, \mathbf{I}^{\mathrm{p}}, \mathbf{b}^{\mathrm{p}}, \mathbf{T}^{\mathrm{p}}, \mathbf{c}^{\mathrm{p}}) \in \mathcal{D}^{\mathrm{N}}_{\text{eval}}$ **do**
▷ keep $(\mathbf{I}^{\mathrm{q}}, \mathbf{b}^{\mathrm{q}}, \mathbf{T}^{\mathrm{p}}, \mathbf{c}^{\mathrm{p}})$ unchanged from Normal
  Sample background $\mathbf{I}_{\text{bg}} \sim \mathcal{I}_{\text{bg}}$
  Sample k distractors $\{(\mathbf{I}_{d_i}, \mathbf{c}_{d_i})\}_{i=1}^{k} \sim \mathcal{I}_{\text{dist}}$ s.t. $\mathbf{c}_{d_i} \neq \mathbf{c}^{\mathrm{p}},\ \forall i$ ▷ cross-category distractors only
  $\mathbf{I}_{\text{tgt}} \leftarrow \mathrm{crop}(\mathbf{I}^{\mathrm{p}}, \mathbf{b}^{\mathrm{p}})$ ▷ ground-truth target crop preserved verbatim
  $(\tilde{\mathbf{I}}^{\mathrm{p}}, \tilde{\mathbf{b}}^{\mathrm{p}}) \leftarrow$ COMPOSE$\big(\mathbf{I}_{\text{bg}}, \mathbf{I}_{\text{tgt}}, \{\mathbf{I}_{d_i}\}_{i=1}^{k}\big)$
  ▷ paste target and distractors on the background at random scale/location; $\tilde{\mathbf{b}}^{\mathrm{p}}$ is the target bbox in $\tilde{\mathbf{I}}^{\mathrm{p}}$
  $\mathcal{D}^{\mathrm{M}}_{\text{eval}} \leftarrow \mathcal{D}^{\mathrm{M}}_{\text{eval}} \cup \{(\mathbf{I}^{\mathrm{q}}, \mathbf{b}^{\mathrm{q}}, \tilde{\mathbf{I}}^{\mathrm{p}}, \tilde{\mathbf{b}}^{\mathrm{p}}, \mathbf{T}^{\mathrm{p}}, \mathbf{c}^{\mathrm{p}})\}$
end for

return $\mathcal{D}_{\text{train}},\ \mathcal{D}^{\mathrm{N}}_{\text{eval}},\ \mathcal{D}^{\mathrm{M}}_{\text{eval}}$

D Qualitative Retrieval Comparison

To complement the quantitative results in Section 4, we visualize the top-6 retrieved candidates of three representative methods—our **TIGER-FG**, the strongest VLM-based baseline **BLIP$_{\text{FF}}$**, and the strongest MLLM-based embedder **Qwen3-VL-Embedding**—on one query drawn from the public **eSSPR** benchmark (Figure 6) and one from our **ECom-RF-IMMR-Normal** (Figure 7). For each method we show the retrieved item title and image at ranks 1–6, with the ground-truth candidate marked by a green check and incorrect candidates by a red cross.

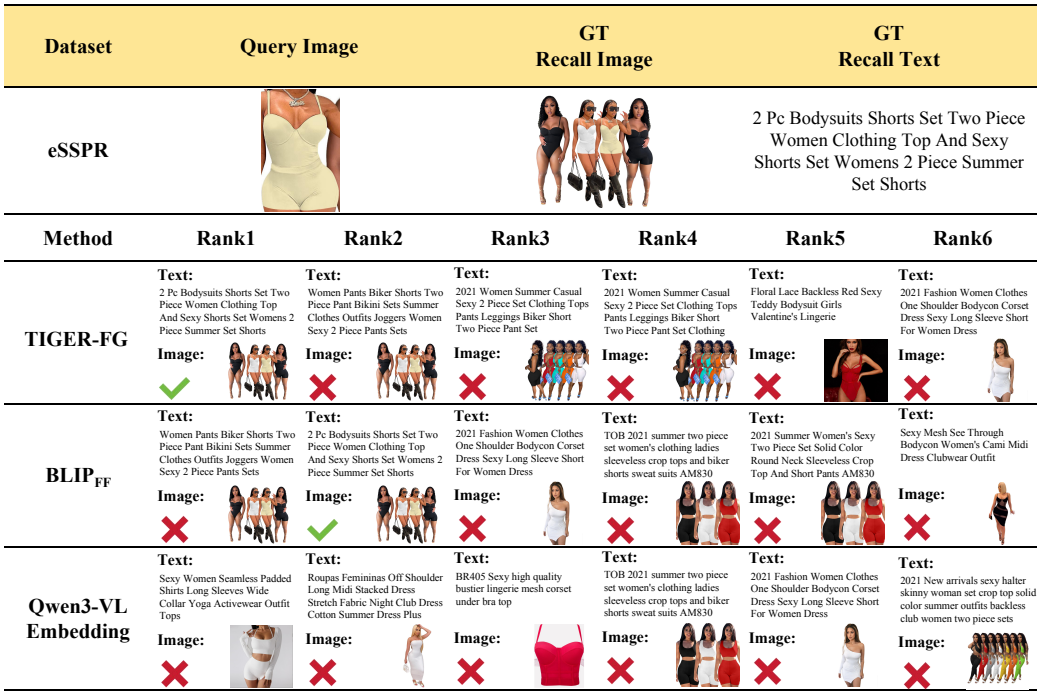


Figure 6: **Top-6 retrieval results on eSSPR.** TIGER-FG retrieves the ground-truth item at rank 1, while BLIP_{FF} retrieves it at rank 2. BLIP_{FF} ranks a visually similar but title-mismatched biker-shorts item first. Qwen3-VL-Embedding misses the ground truth in the top six and retrieves related items such as yoga tops, bodycon dresses, and lingerie.

eSSPR (Figure 6). The query is a cropped image of a beige one-piece bodysuit, and the ground truth is a two-piece bodysuit–shorts set whose item image shows four models jointly wearing dresses, bodysuits, and shorts—a cluttered candidate that global matching tends to mishandle. **TIGER-FG** places the ground truth at rank 1, with several semantically coherent two-piece sets and bodysuits filling out the rest of the top-6. **BLIP_{FF}** retrieves the ground truth at rank 2; its rank-1 candidate is image-matching but title-mismatching (a biker-shorts set), exactly the failure mode that text-guided grounding is meant to fix. **Qwen3-VL-Embedding** misses the ground truth in the top-6 entirely, drifting into adjacent categories such as yoga activewear, bodycon dresses, and lingerie—a sign that a global MLLM embedding is insufficient when the query crop shares low-level visual cues with many garment categories.

ECom-RF-IMMR-Normal (Figure 7). The query shows a woman with a lace fabric draped over her head—without an accompanying title, the user could equally be searching for a scarf or for a piece of clothing/headwear, and there is no way to tell from the image alone. The ground truth is in fact the same scarf product, but its item image shows the scarf laid flat next to a flower vase—a viewpoint that shares almost no low-level cues with the query, so coarse global matching alone has little chance of bridging the two. **TIGER-FG** still ranks the ground truth first, and its remaining top-6 spans both scarves/shawls and clothing items—a reasonable spread given the inherent ambiguity of the query, and a sign that the model is actually using the title to ground the search rather than collapsing to one visual interpretation. **BLIP_{FF}** commits to the headwear reading and returns six hat/beanie/headwrap products in a row, never reaching the target. **Qwen3-VL-Embedding** retrieves the ground truth at rank 2, but its rank-1 result is an unrelated maternity swaddle blanket—the kind of category drift that follows from squeezing a multimodal candidate into a single global embedding.

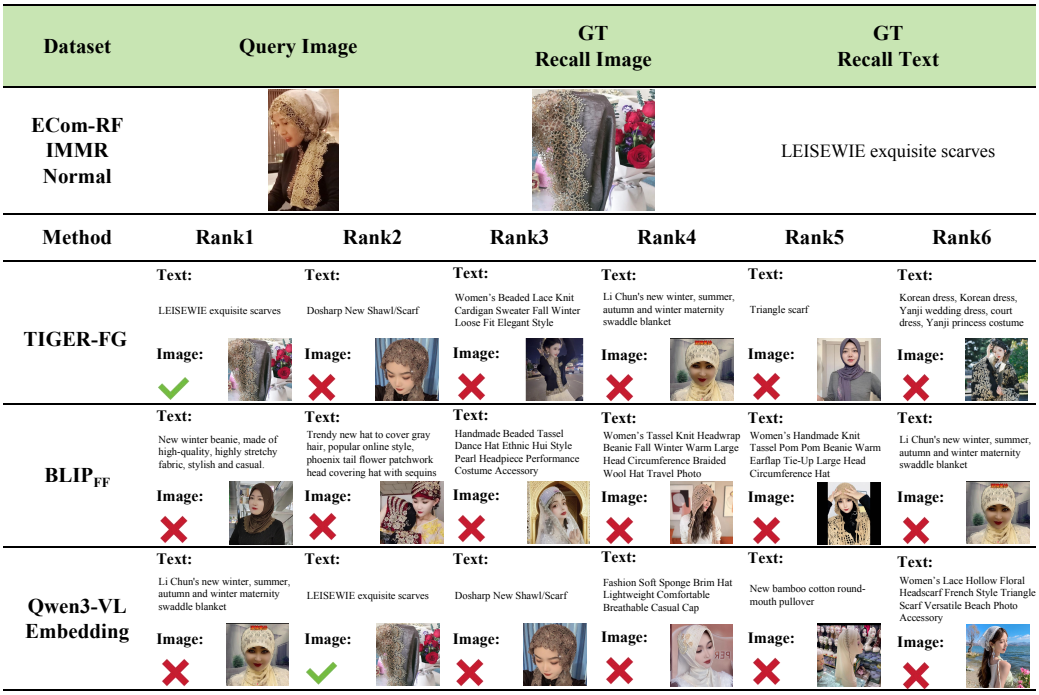


Figure 7: **Top-6 retrieval results on ECom-RF-IMMR-Normal.** TIGER-FG retrieves the ground-truth scarf at rank 1. BLIP$_{FF}$ returns headwear items throughout the top six and misses the target. Qwen3-VL-Embedding ranks the target at rank 2, behind an unrelated maternity blanket.

E Extended Ablation

E.1 Additive Ablation

To complement the subtractive ablation in the main paper, we conduct an additive study that builds the model progressively from a plain dual-encoder baseline. The study serves two purposes. First, it quantifies the contribution of each component under the order in which it is introduced into TIGER-FG. Second, it examines how progressively adding fine-grained region supervision and Mosaic-augmented training data changes both the quantitative retrieval results and the qualitative localization patterns. All experiments are conducted on **ECom-RF-IMMR-Mosaic** under the same training setup as in Section 4.

Table 5 is organized into three blocks that follow the additive construction of TIGER-FG. Block *(a)* starts from the plain dual-encoder and updates the item-side encoder by replacing the visual backbone and then introducing the text-guided fusion design shown in Figure 2, corresponding to **S**. Block *(b)* keeps the training data fixed to raw samples only and progressively adds the main item-side constraints, including target-anchored fused–region alignment **B**, spatial-relational distillation **R**, and mismatched-text hard negatives **H**. Block *(c)* then switches the training data from 1 to the raw+Mosaic setting $1{+}4$, and further introduces similarity-distribution distillation **D** and the image–text contrastive objective **T**. This organization makes it easier to track how the retrieval metrics evolve as the model, supervision, and training data are introduced step by step.

Early item-side changes bring modest gains. In block *(a)*, replacing the CLIP visual backbone with DINOv3 improves Recall@1 from 46.80 to 49.74, with consistent gains across Recall, MRR, and NDCG. Further introducing the fusion design **S** yields only a marginal gain in the metrics, increasing Recall@1 from 49.74 to 49.94. The qualitative cases in Figures 8–10 nevertheless show a clearer change from panels (a) to (c). The highlighted regions become more complete and more concentrated on product entities in the candidate image, suggesting improved entity-level localization. At the same time, the responses remain broad and are not yet reliably routed to the text-specified target, which is consistent with the limited metric gain at this stage.

Table 5: **Additive ablation on ECom-RF-IMMR-Mosaic.** Starting from a plain dual-encoder, we progressively enable components of TIGER-FG. *Data*: "1" uses raw samples only; "1+4" mixes Mosaic-augmented samples at a 1:1 ratio. *Config* is the cumulative set of components, abbreviated as: **S**=slot- and CLS-guided cross-attention; **B**=target-anchored fused–region alignment; **R**=spatial-relational distillation; **H**=mismatched-text hard negatives; **D**=similarity-distribution distillation; **T**=image–text contrastive regularizer. All values are in %.

Model	Data	Config	Recall			MRR		NDCG	
			@1	@4	@10	@4	@10	@4	@10
(a) Backbone only									
CLIP	1	–	46.80	65.24	73.35	54.40	55.64	57.15	59.92
DINOv3	1	–	49.74	67.83	75.42	57.21	58.38	59.91	62.50
DINOv3-Slot	1	S	49.94	68.01	75.77	57.39	58.58	60.08	62.74
(b) + region supervision on raw data									
DINOv3-Slot	1	S+B	50.11	68.09	75.87	57.53	58.72	60.21	62.87
DINOv3-Slot	1	S+B+R	46.54	64.86	73.16	54.07	55.34	56.81	59.65
DINOv3-Slot	1	S+B+R+H	46.70	64.62	72.66	54.06	55.30	56.74	59.50
(c) + Mosaic augmentation and alignment objectives									
DINOv3-Slot	1+4	S+B+R+H	71.30	91.55	97.04	79.91	80.78	82.87	84.78
DINOv3-Slot	1+4	S+B+R+H+D	74.93	93.20	97.72	82.70	83.42	85.38	86.95
DINOv3-Slot	1+4	S+B+R+H+D+T	**75.22**	**93.47**	**97.85**	**83.02**	**83.71**	**85.68**	**87.21**

Region supervision improves localization but hurts retrieval under original-only training. Block *(b)* shows a clear mismatch between qualitative and quantitative behavior. Adding box supervision **B** on top of **S** slightly improves the retrieval metrics, with Recall@1 increasing from 49.94 to 50.11. In the visualizations, panels (d) in Figures 8–10 show that **B** further strengthens entity localization in the candidate image. Adding spatial-relational distillation **R** then makes the responses cleaner and more compact, with substantially reduced background activation, while adding hard negatives **H** begins to induce query-dependent behavior, so that different titles lead the model to attend to different objects. Despite these qualitative improvements, the retrieval metrics drop after **R** and remain low after **H**.

The metric drop reflects a distribution mismatch. This behavior is best understood as a mismatch between the raw-data training distribution and the Mosaic evaluation distribution. Under raw-data training, the item images are relatively clean, so stronger localization constraints encourage the model to rely on highly concentrated local evidence. This makes the responses visually sharper, but it also reduces reliance on broader cues that can still be useful for retrieval. Moreover, after **R** and especially **H**, the model begins to show weak query-dependent behavior, but this text-guided ability is still not reliable. When the guidance is correct, it can help the model focus on the relevant product entity. When it is incorrect, it can instead drive the model toward a mismatched entity that is visually salient but inconsistent with the title. In ECom-RF-IMMR-Mosaic, where candidate images contain more distractors and cross-object interference, such over-concentrated and occasionally misdirected matching can become brittle. The qualitative gains in block *(b)* therefore indicate improved localization and emerging text guidance under the training distribution, but they do not yet translate into better retrieval under the Mosaic distribution.

Mosaic augmentation aligns training with the retrieval setting. Switching the training data from raw samples only to the raw+Mosaic setting yields the largest gain in Table 5. Recall@1 jumps from 46.70 to 71.30, with similarly large improvements across Recall, MRR, and NDCG. The main reason is that Mosaic augmentation better matches the retrieval setting, where candidate images often contain multiple products and stronger distractors. Under this training mixture, the model is no longer optimized only for clean single-object item images, but is instead exposed to the kind of multi-object interference that appears at retrieval time. This substantially reduces the train–test distribution disparity. More importantly, Mosaic augmentation helps the model learn which

product entity in a multi-object item image should be matched to the title, rather than only sharpening attention on visually salient regions. As a result, the localization ability introduced by **R** and **H** becomes effective under realistic retrieval conditions, which leads to the large improvement in block *(c)*.

Similarity-distribution distillation and contrastive learning. Building on the raw+Mosaic setting, adding similarity-distribution distillation **D** further raises Recall@1 from 71.30 to 74.93. SDD aligns the batch-level query–item similarity distributions of TIGER-FG with those produced by a frozen image-to-image retrieval teacher. By transferring relative similarity relationships across candidate items, it helps preserve the discriminative structure of multimodal retrieval representations. Adding the image–text contrastive objective **T** then further improves Recall@1 to 75.22. Although the metric gain is smaller, the qualitative cases in Figures 8–10 show a clearer benefit from panels (h) to (i), where the response becomes more consistently routed to the title-matched object, especially in the challenging cases with competing products. This suggests that **T** helps preserve the text-guided role of the item-side text encoder and prevents the item representation from drifting toward a purely visual matching space. Together, **D** and **T** make title-guided retrieval substantially more reliable in the final TIGER-FG model.

Qualitative visualization. Figures 8–10 visualize how the item-side response evolves as components are added for three representative query–candidate pairs with naturally multiple product entities. For each case, we show two item-text conditions (*Text 1* and *Text 2*), together with the similarity map (*Sim.*) and its overlay on the candidate image (*Overlay*). For models with slot fusion, we further visualize the *Fused*, *Slot*, and *Token* responses. The three cases cover different retrieval difficulties, namely a cluttered fashion scene with a black dress (Case 1), a home-goods scene with small-object queries (storage basket and cosmetic brushes, Case 2), and an apparel scene with two co-present garments (knitwear and dress, Case 3). The visualizations complement the quantitative results in Table 5, showing that more localized responses do not necessarily translate into better retrieval performance. Block *(a)* mainly improves entity localization. Block *(b)* yields sharper and more localized responses, but title-guided routing remains unstable under raw-data training. After introducing Mosaic-augmented training in block *(c)*, the model better identifies which entity in the item image corresponds to the title. The final TIGER-FG configuration further shows more consistent responses to the text-specified target across these examples.

E.2 Joint SRD/SDD Distillation Analysis

SRD preserves pairwise patch relations within the target ROI using a frozen DINOv3 teacher, while SDD aligns batch-level query–item similarity distributions using a frozen image-to-image retrieval teacher. We independently enable or disable the two objectives under the final training configuration while keeping all other settings fixed.

Table 6: **Joint SRD/SDD ablation on ECom-RF-IMMR-Mosaic (%).**

SRD	SDD	R@1	R@4	R@10	MRR@4	MRR@10	NDCG@4	NDCG@10
✓	✓	75.2	93.5	97.8	83.0	83.7	85.7	87.2
–	✓	75.3	93.5	97.8	83.1	83.8	85.7	87.2
✓	–	69.9	91.1	96.8	78.8	79.7	81.9	83.9
–	–	68.8	90.5	96.5	77.9	78.9	81.0	83.1

As shown in Table 6, removing SDD reduces Recall@1 from 75.2 to 69.9, with consistent reductions in MRR and NDCG. Removing SRD alone leaves Recall@1 nearly unchanged (75.3 vs. 75.2), while the box-level grounding analysis in Appendix F.2 (Table 9) shows less concentrated spatial responses without SRD. When SDD is absent, further removing SRD reduces Recall@1 from 69.9 to 68.8. These results support complementary roles for the two objectives. SRD primarily preserves target-region spatial structure, whereas SDD constrains global query–item similarity structure.

F Additional Experimental Analyses

F.1 Controlled Retrieval and Grounding Evaluation

Matched augmentation. To assess whether Mosaic augmentation alone explains TIGER-FG's gains, we retrained CLIP-SF, OClear, and Qwen3-VL-Emb on the same 1:1 original/Mosaic mixture used by TIGER-FG (Table 7). CLIP-SF and OClear received full fine-tuning, while the larger Qwen3-VL-Emb used LoRA adaptation. All three baselines improve markedly over their original-data counterparts in Table 1, demonstrating the benefit of Mosaic augmentation on ECom-RF-IMMR-Mosaic. Nevertheless, TIGER-FG achieves 75.2 Recall@1, outperforming the strongest matched baseline at 63.4 by 11.8 percentage points. These results suggest that Mosaic augmentation alone does not fully explain TIGER-FG's advantage.

Box-level grounding. While the matched-augmentation results demonstrate retrieval improvements under controlled training data, retrieval metrics alone do not directly verify whether the model focuses on the target product in cluttered item images. To quantitatively evaluate the implicit grounding behavior illustrated in Figure 3, we compare spatial response maps against item-side ground-truth boxes on all 100K ECom-RF-IMMR-Mosaic samples (Table 8). We evaluate target localization using pointing accuracy, region recall, region-mass ratio, and attention-box IoU. TIGER-FG achieves 0.995 pointing accuracy and a region-mass ratio of 0.871, substantially outperforming the strongest compared baseline at 0.591 and 0.389, respectively. These results provide direct quantitative evidence that TIGER-FG concentrates spatial responses on target regions despite background clutter and distractor objects, consistent with the qualitative observations in Figure 3.

Table 7: **Matched-Mosaic Retrieval.** All methods use a 1:1 original/Mosaic mixture.

Method	Tuning	ECom-RF-IMMR-Mosaic (%)						
		Recall			MRR		NDCG	
		@1	@4	@10	@4	@10	@4	@10
CLIP-SF	Full	59.3	81.5	89.6	68.7	69.9	72.5	75.5
OClear	Full	43.1	65.8	76.4	52.6	54.2	56.9	60.8
Qwen3-VL-Emb	LoRA	63.4	84.7	92.2	72.4	73.5	76.1	79.0
TIGER-FG	Full	**75.2**	**93.5**	**97.8**	**83.0**	**83.7**	**85.7**	**87.2**

Full: full fine-tuning; LoRA: low-rank adaptation.

Table 8: **Box-Level Grounding.** Evaluation on ECom-RF-IMMR-Mosaic.

Method	ECom-RF-IMMR-Mosaic			
	Point	Rec	Mass	IoU
OClear	0.362	0.146	0.346	0.254
CLIP-FF	0.355	0.204	0.380	0.253
Qwen3-VL-Emb	0.591	0.208	0.389	0.440
TIGER-FG	**0.995**	**0.995**	**0.871**	**0.689**

Point: pointing accuracy; Rec: region recall@0.5; Mass: region-mass ratio; IoU: attention-box IoU@q0.8.

F.2 Grounding Metrics and the Role of Spatial Distillation

The box-level grounding evaluation in Table 8 uses all 100,000 ECom-RF-IMMR-Mosaic examples. For this evaluation, the item-side ground-truth boxes serve as references for measuring spatial responses. These boxes are not predicted or required during retrieval.

Formally, for a nonnegative spatial response map A over spatial locations p and target box B, let $m_B = \sum_{p \in B} A(p) / \sum_p A(p)$. Pointing accuracy is the proportion of samples for which $\arg\max_p A(p) \in B$, region recall at 0.5 is the proportion for which $m_B \geq 0.5$, and the region-mass ratio averages m_B. Attention-box IoU is computed between the target box B and the spatial region containing the top 20% of response values.

To isolate the effect of spatial-relational distillation (SRD), all training settings are held fixed and only SRD is disabled. Table 9 reports the effect of SRD on box-level grounding metrics across the evaluation set.

Table 9: **Grounding with and without SRD** on all 100K ECom-RF-IMMR-Mosaic examples. IoU is computed with the top-20% response region.

Variant	Pointing Acc.	Region Recall@0.5	Region-Mass Ratio	Attn.-Box IoU
TIGER-FG	**0.995**	**0.995**	**0.871**	**0.689**
w/o SRD	0.937	0.792	0.564	0.597

Removing SRD decreases the region-mass ratio from 0.871 to 0.564 and region recall from 0.995 to 0.792, although the maximum-response point often remains in the target box. SRD thus improves the localization *concentration* of the target response rather than primarily optimizing the final retrieval ranking.

F.3 Leave-One-Out Component Ablations

The additive ablation in Table 5 evaluates components as they are progressively introduced, but their measured contributions may depend on the order of introduction. To complement the final-setting ablations reported in Table 3, we evaluate four additional components by removing each from the same *final* TIGER-FG configuration and retraining with all other settings unchanged. In the slot ablation, the two-stage text-to-slot-to-vision interaction is replaced with direct text-to-vision cross-attention. The hard-negative ablation removes only mismatched-text terms, and the target-anchor ablation leaves hard-negative supervision intact.

Table 10: **Additional final-setting component ablations on ECom-RF-IMMR-Mosaic.**

Method	R@1	R@4	R@10	MRR@4	MRR@10	NDCG@4	NDCG@10
TIGER-FG	**75.2**	**93.5**	**97.8**	**83.0**	**83.7**	**85.7**	**87.2**
w/o region–text alignment	74.9	93.2	97.7	82.7	83.4	85.4	87.0
w/o target-anchor alignment	74.0	93.0	97.6	82.1	82.8	84.8	86.5
w/o mismatched-text hard negatives	73.5	92.7	97.5	81.7	82.4	84.5	86.2
w/o learnable slots	72.9	92.2	97.1	81.2	82.0	83.9	85.9

Among these four additional component ablations, the learnable slots yield the largest independent improvement of 2.3 Recall@1 points over direct text-to-vision cross-attention, suggesting a benefit from the two-stage aggregation of text-relevant visual evidence. Mismatched-text hard negatives and target-anchor alignment contribute 1.7 and 1.2 points, respectively, with a more modest effect from region–text alignment. These results complement the additive ablation by evaluating the additional components under a fixed final training configuration.

G Additional Experimental Details

G.1 Implementation Details

The query and item visual branches are both initialized from the original DINOv3 ViT-B/16 weights [Siméoni et al., 2025] at 224×224 resolution and trained as two independent copies, without additional vision pretraining on our data. The text encoder is initialized from the text branch of Chinese-CLIP ViT-B/16 [Yang et al., 2022], which adopts a 12-layer RoBERTa-base architecture. Before dual-encoder training, we further pretrain the text encoder with CLIP-style image–text contrastive learning on **ECom-RF-IMMR-10M**.

The unified embedding dimension is $C_u = 256$. The item branch uses $N_q = 8$ learnable query tokens and $K = 1$ mismatched-text hard negative per sample. In the complementary visual branch, we set $\lambda_m = \lambda_c = 0.5$. All contrastive and distillation temperatures are fixed to 0.07. For the item-side objective, we set $(\lambda_{v2t}, \lambda_{i2v}, \lambda_{\mathrm{SRD}}) = (0.5, 0.1, 1.0)$ and $\lambda_h = 0.1$. For the dual-encoder objective, we set $(\lambda_{q2i}, \lambda_{\mathrm{SDD}}) = (1.0, 1.0)$. The final objective uses $\lambda_{\mathrm{dual}} = \lambda_{\mathrm{item}} = 1.0$.

The $\mathcal{L}_{\mathrm{SRD}}$ teacher is a frozen original DINOv3 ViT-B/16. The $\mathcal{L}_{\mathrm{SDD}}$ teacher is our in-house MoCo-pretrained image-to-image retrieval ViT-B/16, which also appears as a baseline in Table 4. Both teachers remain frozen throughout training.

We train for 10 epochs on 8 NVIDIA H800 GPUs with batch size 256 and bf16 mixed precision. Optimization uses AdamW with learning rate $2{\times}10^{-5}$, weight decay 0.01, and a cosine schedule with 5% warmup. A full run takes approximately 14 hours of wall-clock time on these 8 NVIDIA H800 GPUs, corresponding to about 112 GPU-hours of aggregate compute. Unless otherwise specified, TIGER-FG is trained with a $1:1$ mixture of original and Mosaic-augmented samples. We also report TIGER-FG-RAW, which uses the same architecture and objectives but is trained only on the original samples, to isolate the effect of clutter-aware training. Matched-Mosaic retraining is reported separately in Appendix F.1.

G.2 Public Benchmark Adaptation

We additionally evaluate on two public e-commerce benchmarks, **eSSPR** and **LookBench**, to assess cross-dataset generalization. We adapt their query and candidate formats to the IMMR setting, where a cropped visual query retrieves image–text item candidates.

eSSPR. eSSPR mainly contains clean item images, where each image usually depicts a single foreground item. To adapt eSSPR to IMMR, we convert each query into a cropped visual region and keep the gallery as image–text item candidates. Since ground-truth query boxes are not directly provided in the original benchmark, we first extract region proposals with an off-the-shelf object detector and then use a visual matching module to select the region that best matches the item target. Samples that cannot be reliably parsed are removed. The remaining samples are evaluated with the same retrieval protocol as our in-domain benchmarks.

LookBench. LookBench provides localized query regions and a noisy candidate pool, where a query may correspond to multiple relevant items. We use the provided bounding boxes as cropped visual queries. For the candidate side, we construct multimodal item entries by pairing each candidate image with structured item text. Because item titles are unavailable, Qwen3-VL-32B organizes only the existing category and attribute annotations into candidate-side structured text, without adding image-derived descriptions or semantic expansions. Query boxes, gallery images, and relevance labels remain unchanged, and all image–text-capable baselines receive the same candidate text. We aggregate results across all LookBench subsets. In addition to Recall@K, MRR@K, and NDCG@K, we report HitRate@K on LookBench because each query may have multiple valid matches.

G.3 LookBench Text-Field Sensitivity

To assess the effect of candidate-side text formulation on TIGER-FG, we compare the unpolished category/attribute text used for the LookBench results in the main paper with a polished formulation based on the *same* fields.

Table 11: **LookBench text-field sensitivity for TIGER-FG.** Same categories, attributes, query boxes, item images, and relevance labels in both settings. Values in %.

Text	HR@1	HR@4	HR@10	R@1	R@4	R@10	MRR@10	NDCG@10
Unpolished (main paper)	39.8	61.3	71.0	18.5	38.8	52.2	50.0	45.0
Polished	42.8	62.7	72.1	19.6	39.9	54.1	52.4	47.1

Compared with the unpolished formulation, the polished version improves HitRate@1 from 39.8 to 42.8 and Recall@1 from 18.5 to 19.6, with consistent improvements across the remaining metrics. These results suggest that the organization of candidate-side structured text affects retrieval performance even when the underlying category and attribute fields remain unchanged.

G.4 Licenses and Terms of Use

Table 12 summarizes the license information stated by the upstream repositories and model or dataset cards. Repository-code licenses are distinguished from licenses for model weights and datasets.

The eSSPR README describes research use but does not state a standard license identifier or explicit terms for redistribution of an adapted dataset. We therefore report its published use statement rather than assign an unverified license.

Table 12: **Upstream license information.** Links point to the corresponding license files or release pages. Code-repository entries describe the repository license, not a separate license grant for every associated checkpoint or dataset.

Asset	Scope of statement	License or published terms
DINOv3 ViT-B/16	Code and weights	DINOv3 License
Chinese-CLIP	Repository code	MIT
GroundingDINO	Repository code	Apache-2.0
CLIP	Repository code	MIT
BLIP	Repository code	BSD-3-Clause
UniIR	Repository code	MIT
UniECS	Repository code	MIT
BGE-VL-Large	Model-card declaration	MIT
GME-2B	Model-card declaration	Apache-2.0; additional redistribution and attribution requirements in the model card
Qwen3-VL-Embedding-2B	Model-card declaration	Apache-2.0
Ops-MM-embedding-v1-2B	Model-card declaration	Apache-2.0
Qwen3-VL-32B (Instruct / Thinking releases)	Model-card declarations	Apache-2.0 for both releases
LookBench	Dataset-card declaration	Apache-2.0
eSSPR	Author-provided README	Released for public research; citation requested



(a) CLIP

(b) DINOv3

(c) + S

(d) + B

(e) + R

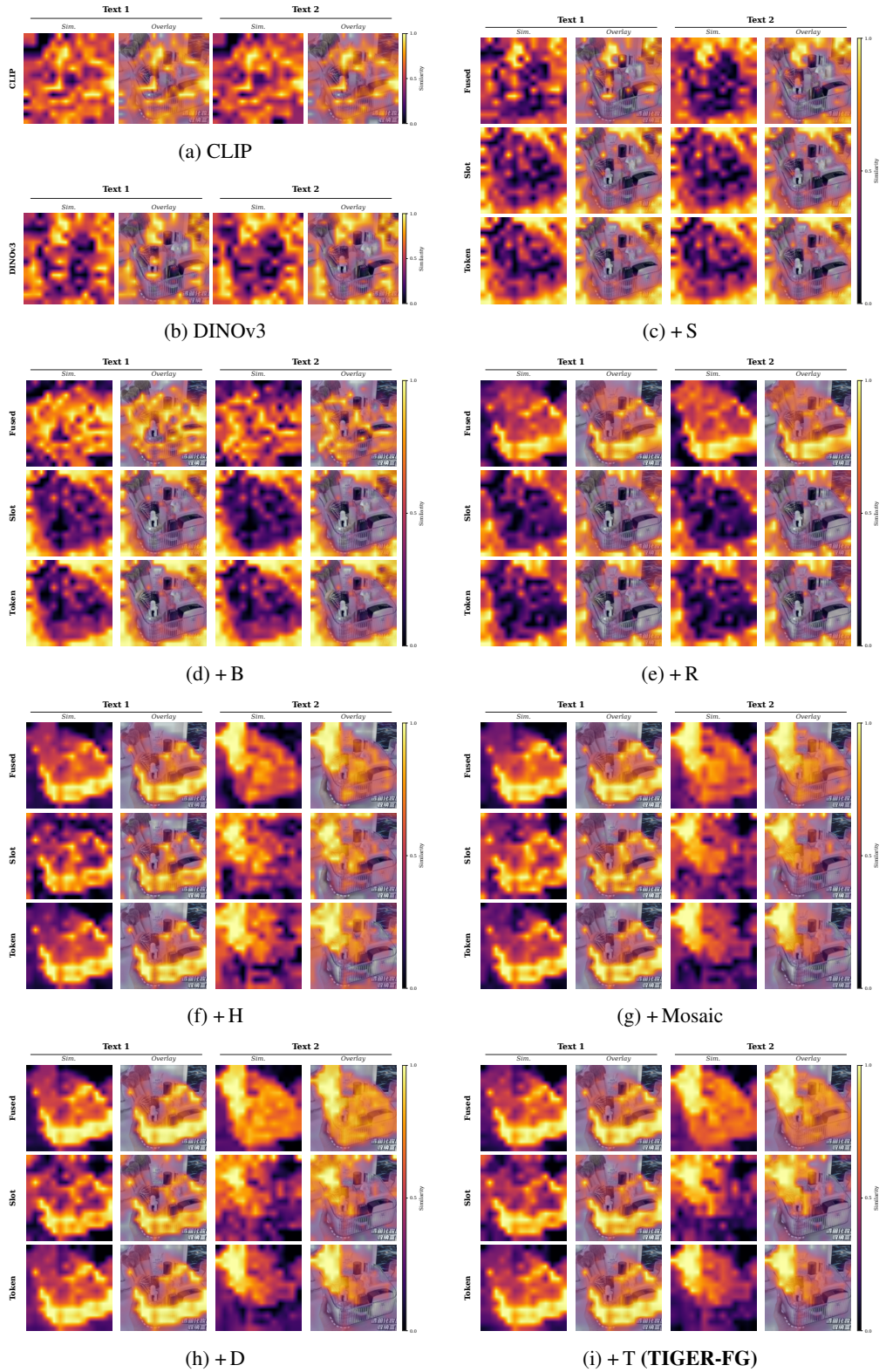
(f) + H

(g) + Mosaic

(h) + D

(i) + T **(TIGER-FG)**

Figure 8: **Qualitative visualization of the additive ablation (Case 1/3, *dress* and *shoes* queries).** Each panel corresponds one-to-one to a row in Table 5 on the same query–candidate pair. From (c) onward, "+ X" denotes adding component **X** on top of the previous panel. Component abbreviations: **S**=slot- and CLS-guided cross-attention; **B**=Target-anchored fused–region alignment; **R**=Spatial-relational distillation; **H**=Mismatched-title hard negatives; **D**=Similarity-distribution distillation; **T**=image–text contrastive regularizer. "Mosaic" denotes switching training data from 1 to 1+4.



(a) CLIP (b) DINOv3 (c) + S

(d) + B (e) + R

(f) + H (g) + Mosaic

(h) + D (i) + T (**TIGER-FG**)

Figure 9: **Qualitative visualization of the additive ablation (Case 2/3: *storage basket* and *cosmetic brushes* queries).** Panels follow the same additive layout as Figure 8; see its caption for component definitions. This case emphasizes small-object queries in a cluttered home-goods scene, where the over-sharpening effect of **R/H** under raw-data training is particularly pronounced.



(a) CLIP

(b) DINOv3

(c) + S

(d) + B

(e) + R

(f) + H

(g) + Mosaic

(h) + D

(i) + T **(TIGER-FG)**

Figure 10: **Qualitative visualization of the additive ablation (Case 3/3: *knitwear* and *dress* queries).** Panels follow the same additive layout as Figure 8; see its caption for component definitions. This case contains two co-present garments in the candidate, and thus directly tests whether the model can route each text query to the correct garment—a capability that only fully emerges after Mosaic augmentation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. The claims in the abstract and introduction accurately reflect the contributions and scope of the paper. The proposed method and its motivation are described in Section 1, with contributions summarized at the end of the introduction, and are supported by the experimental results in Section 4. We restrict our claims to the evaluated datasets and settings.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes. We provide a concise discussion in Appendix A, covering the current e-commerce instantiation of TIGER-FG, the controlled scope of our retrieval benchmarks, and the use of item boxes only for auxiliary supervision and dataset construction.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: Not applicable. The paper does not present formal theoretical results, theorems, or proofs. Section 3 provides the methodological formulation and training objectives, while Section 4 and the appendix empirically validate the design choices through experiments and ablation studies.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes. We provide sufficient details to reproduce the experiments, including dataset descriptions, model components, training objectives, and evaluation protocols in Sections 3 and 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The evaluation datasets are publicly available. The training dataset and code are being prepared for release.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes. We specify key experimental details, including dataset construction, preprocessing, model architecture, training strategies, hyperparameters, and evaluation metrics in Section 4 and the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: No. The reported results are based on single runs without multiple random seeds or confidence intervals, due to the computational cost of large-scale multimodal training.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes. We report the compute resources used in the experiments in Section 4 (Implementation Details), including training on 8× H800 GPUs. Additional details, such as the training setup and runtime, are provided in the Experimental section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes. The research adheres to the NeurIPS Code of Ethics and does not involve any unethical data collection or usage.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: No. The paper focuses on the methodology and evaluation of e-commerce image-to-multimodal retrieval and does not explicitly discuss both its positive and potential negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Not applicable. The paper does not release high-risk generative models, pretrained language models, or safety-critical systems. The proposed benchmark and model are designed for product-level retrieval, and the paper describes data auditing for quality control.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Yes. The paper properly credits the existing models and public benchmarks used in the experiments through standard citations. Their licenses and terms of use are summarized in Appendix G.4, and our use of these assets complies with the applicable terms.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Yes. The paper introduces ECom-RF-IMMR-10M, ECom-RF-IMMR-Normal, and ECom-RF-IMMR-Mosaic, and documents their sources, construction pipeline, schema, quality-control process, and evaluation usage in Section 3.4. Additional construction details and pseudocode are referred to in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The evaluation dataset was quality-checked through VLM-assisted filtering followed by human auditing conducted by the authors, as described in Section 3.4 and Appendix C. No external annotators or crowdsourcing workers were recruited, and the study did not involve research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Not applicable. The paper does not report experiments involving human subjects or interventions with study participants. The human auditing is used for dataset quality control rather than as a human-subject study.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Qwen3-VL-32B is used to organize existing category and attribute annotations into candidate-side structured text for LookBench adaptation, without introducing image-derived descriptions or additional semantic information. This preprocessing procedure is described in Appendix G.2.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.