arXiv:2611.05193v1 [cs.CV] 3 Nov 2026

Factorized MLLM Embeddings for Composed Image Retrieval

WonJun Moon[1] Seongchan Kim[1] Dahun Kim[2] Seungryong Kim[1,*]
KAIST AI[1] Google DeepMind[2]

https://cvlab-kaist.github.io/FactEmbed/

Abstract

Composed Image Retrieval (CIR) aims to retrieve a target image given a reference image and a modification text. Although recent approaches leveraging Multimodal Large Language Models (MLLMs) have shown significant progress, we argue that the MLLM's compositional understanding remains underutilized. This stems from directly mapping composed queries into a retrieval embedding via contrastive supervision alone, without explicitly resolving the intended visual modifications. In this work, we introduce Factorized MLLM Embeddings (FactEmbed), which decouples compositional query comprehension from building discriminative embeddings. Rather than directly projecting the composed query into a single embedding without internal reasoning, FactEmbed employs a set of factorized latents that simulate the intended visual modifications via an intermediate target imagination objective. This process encourages the latents to ground into target semantics before forming discriminative representations for retrieval. Subsequently, to propagate these factorized representations throughout the model for retrieval, we optimize a factor-level contrastive loss while distilling the decomposed representations into a deeper attention layer of the MLLM. Finally, for efficient inference, we retain the `[EOS]` token to establish a global-level embedding on top of these factorized latents, while we exploit the mean-pooled factorized latents to rerank a short candidate list at negligible cost, as both are produced by the same forward pass. Extensive evaluations demonstrate that FactEmbed outperforms prior works across CIR benchmarks.

1. Introduction

Composed Image Retrieval (CIR) aims to retrieve a target image from a gallery using a composed query, consisting of a reference image and a modification text [1]. CIR allows users to express their search intent through relative descriptions: they may provide an image and specify how the desired result should differ from it. This paradigm is appealing when users can recognize a desired image by its visual relation to a reference, but cannot easily describe the target itself.

Recently, Multimodal Large Language Models (MLLMs) have emerged as a powerful foundation for CIR, owing to their ability to deeply fuse visual and linguistic information within a unified autoregressive representation space [2, 3]. A common strategy is to finetune an MLLM contrastively by compressing the multimodal query and the target image into a single embedding, typically obtained from the final `[EOS]` token, and directly optimizing their similarity [2, 4–6]. Despite its effectiveness, this formulation exposes a key limitation: forcing a direct mapping onto a single embedding provides no explicit signal for the MLLM to exploit its compositional understanding capability to resolve the requested edit before forming a discriminative embedding.

To overcome this, we aim to decouple query comprehension from discriminative embedding learning, allowing the model to first simulate the requested edit, *i.e.*, imagine the target, and then build its retrieval embedding upon the imagined target. We present FactEmbed, which realizes this intermediate step through a compact set of factorized latents that parse the query before producing the final embedding. These factorized latents are prepended to the `[EOS]` token in the input sequence, serving as information bottlenecks that decompose fused multimodal queries into factor-level components. This explicit decomposition enables FactEmbed to isolate specific visual attributes and distinguish the factors the text modifies from those it preserves, in contrast to direct contrastive mapping which only implicitly infers the modified embedding. To encourage each latent to capture different details in the target image, we adopt a target imagination objective parameterized by a lightweight decoder over these latents. Given a coarse layout prior of the target, the latents are trained to imagine the residual fine-grained factors that the prior cannot explain. Constraining reconstruction through these bottlenecks forces the latents to cluster semantically coherent target factors to maximize the coverage. This provides structural supervision without requiring explicit reasoning traces or dense annotations, thereby allowing the model to implicitly simulate the specified edit.

Reconstruction alone, however, shapes the latents generatively without ensuring their utility for retrieval: paired query

* Corresponding author.

and target could be faithfully decomposed yet embedded in incompatible coordinate spaces. We address this by adapting target imagination for retrieval in two ways. First, we apply a factor-level InfoNCE objective directly on the factorized latents, ensuring that the decomposed factors are aligned for retrieval rather than only for reconstruction. Second, we distill the decoder's factorized representation into the attention maps of the latents at a deeper layer, transferring the fine-grained decomposition from the auxiliary head into the backbone layers accessed during retrieval.

Lastly, we assign complementary roles to the two outputs produced in a single forward pass. Positioned prior to the `[EOS]`, the factorized latents allow `[EOS]` to integrate the decomposed factors into a global retrieval embedding, while `[EOS]` is optimized as a single retrieval embedding for efficient retrieval. Yet, as fine-grained information is inevitably diluted when aggregating multiple factors into a single discriminative embedding, we rerank a short candidate list retrieved by `[EOS]` using the mean-pooled latents. This complementary structure allows FactEmbed to achieve fast global recall via single `[EOS]` token while precision is significantly boosted by recovering fine-grained factor evidence during reranking. Because both representations originate from a single forward pass and the mean-pooled latent adds only one vector per image, this design complements the global embedding with factor-level evidence without additional encoding or heavy storage overhead. Combined with our MLLM data augmentation, FactEmbed achieves state-of-the-art results on CIR benchmarks, outperforming both CIR methods and general-purpose embedding models.

2. Related Work

Composed Image Retrieval. CIR aims to retrieve a target image from a gallery given a reference image with a modification text that states how the desired result should differ from it [1, 7–9]. Earlier works [1, 10–14] attempted to resolve representational conflict by performing the composition inside CLIP [15] or BLIP [16, 17] embeddings. Then, since collecting high-precision annotations for CIR incurs significant cost [18, 19], two parallel lines of research have been actively explored: learning with noisy CIR annotations [20] and zero-shot approaches [21]. While the former stream attempted to mitigate the impact of noisy annotations by leveraging loss dynamics [22], similarity curves [23], MLLM-constructed anchor set [24], or unlearning process [25], zero-shot approaches aim to bypass the high cost of collecting annotated triplets [26–31]. Notably, recent zero-shot methods also anticipate the target, either by generating a proxy image [30] or by leveraging a world model [31]; in contrast, our imagination is embedded within the embedding model itself, requiring neither a generative model nor an additional inference pass.

More recently, LLMs and MLLMs have been widely adopted in CIR owing to their strong comprehension capabilities. CIR-LVLM [4] adapted a local MLLM for CIR by introducing intent-aware instructions with learnable soft prompts, while ReCALL [2] proposed MLLM-based triplet augmentation strategy by diagnosing the blind spots of adapted MLLM retriever. To further exploit the comprehension capability of LLMs, CCIN [32] and FIRE-CIR [33] incorporated an external LLM at query time to decompose the query at the attribute level, which, however, incurs substantial inference overhead, with FIRE-CIR further requiring verification linear in the number of candidates. Thus, in this work, we adapt a local MLLM to exploit its inherent comprehension capability; whereas prior MLLM-based methods [2, 4] directly map the composed query onto a single embedding, we encourage the model to decompose the query within a single forward pass and build the retrieval embedding on top of the decomposed factors.

Multimodal Retrieval with MLLMs. MLLM embeddings are being extensively studied in multimodal retrieval where complex comprehension is required [34–36]. VLM2Vec [37] initially used the last hidden state of the `[EOS]` token of a causal MLLM. Yet, as `[EOS]` is not designed to generate embeddings, different methods are being studied, *e.g.*, mean-pooling [38–40] and multi-vector embeddings [41–43]. Also, integrating computation via explicit reasoning [44], latent rollouts [45], or reinforcement-learned generative embeddings [46] is another prominent line of research. Our work resembles the stream of appending a small bank of learned tokens whose hidden states become the representation [47, 48]. None of these works, however, constrains what an individual token should represent. In contrast, we leverage target reconstruction as a structural supervisory signal, forcing each factorized latent to ground and specialize in spatially coherent regions of the target scene.

Compact Latent Representations. Representing an input as a small set of latents has been studied extensively across fields. Retrievers used it for late interaction [49–52], Perceiver [53, 54] exploited latents to decouple compute from input size, DETR [55] used them as object queries to perform object detection, and slot attention [56–58] utilized slots to decompose the scene into object-wise regions by normalizing the attention across latents. FactEmbed adopts the same competitive normalization in its decoder to induce factor-level specialization, but keeps the MLLM's forward pass intact, thereby preserving kernel optimizations such as FlashAttention [59].

3. FactEmbed: Factorized MLLM Embeddings

Task. Given a query $q = (I_r, T)$ consisting of a reference image I_r and a modification text T, the goal of Composed Image Retrieval (CIR) is to retrieve the target image $t = I_t$ that depicts the reference after the described modification. The task is intrinsically compositional: the model must de-

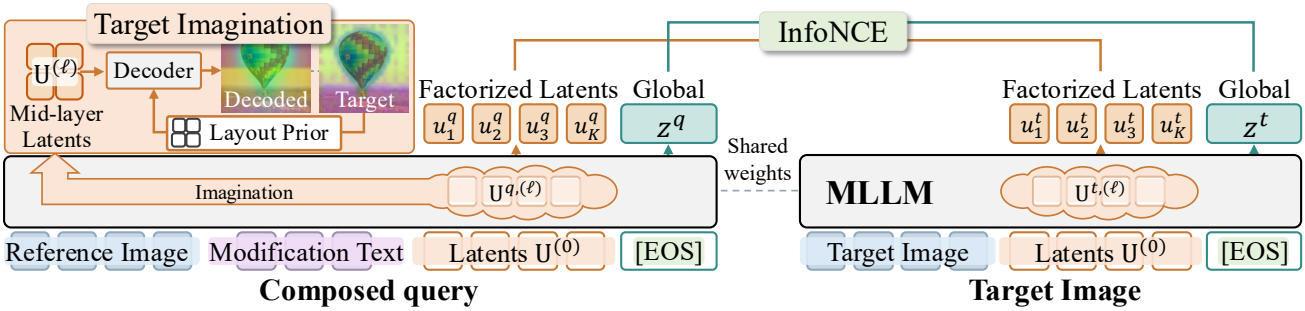


Figure 1. **Overview of FactEmbed.** Given a composed query and a target image, FactEmbed introduces a compact set of learnable factorized latents into the input sequence. To enforce structured factorization without requiring dense annotations, we attach a decoder to an intermediate MLLM layer, which utilizes these latents as an information bottleneck to reconstruct target patch features via a target imagination objective. In addition, we apply a factor-level InfoNCE to the resulting latents $\{u_k\}_{k=1}^{K}$ to leverage the decomposed factors for retrieval, while global-level embedding z is trained with the instance-level InfoNCE objective.

cide which visual factors of I_r the text modifies (*e.g.*, "replace the logo on the shirt") and which it must preserve (*e.g.*, the rest of the shirt and the background), then express the result of that edit as a retrievable representation.

Overview. As shown in Fig. 1, FactEmbed adapts an MLLM to produce CIR embeddings, utilizing a shared backbone to encode both the multimodal query and the target image. Rather than mapping the input directly to a single vector, we first introduce K learnable factorized latents into the multimodal input, on top of which the `[EOS]` embedding is formed. This provides the model with the capacity to represent the complex query as a structured set of compositional components. To ground these latents in the visual content of the target image, we train them to reconstruct the target's patch features beyond a coarse layout prior, a process we refer to as target imagination. We perform this reconstruction by attaching a lightweight decoder at an intermediate layer of the MLLM, which allows the deeper layers to transform the decomposed factors into discriminative representations for retrieval. For retrieval training, we apply InfoNCE [60] at two levels: a factor-level loss encourages the factorized latents to contribute to retrieval by matching the query and target latents via optimal transport, while an instance-level loss optimizes the global embedding. In this way, the imagined decomposition propagates fine-grained evidence into a compact embedding that remains efficient for retrieval.

3.1. Factorized Latents with Target Imagination

Motivation. CIR requires modifying specific visual concepts given text instructions (*e.g.*, replacing an item). However, directly mapping a composed query to a global embedding in an MLLM provides no explicit mechanism for disentangling the visual attributes to be modified from those that should remain invariant. To overcome this, we introduce a compact set of factorized latents whose explicit job is to parse the scene into structured factors. By decomposing the representation, these latents are designed to give the model room to express the instructed modification on the relevant factors without distorting the others.

Factorized Latents. Factorized latents are defined as a set of K learnable tokens, $\mathbf{U}^{(0)} \in \mathbb{R}^{K\times D}$, which act as information bottlenecks to parse multimodal inputs into factorized representations. These tokens are prepended to the `[EOS]` in both the query and target sequences and interact with the multimodal inputs via the MLLM's causal attention. We denote their hidden states at layer ℓ by $\mathbf{U}^{s,(\ell)} = [u_1^{s,(\ell)}, \dots, u_K^{s,(\ell)}]$ for $s \in \{q, t\}$, where s indicates the query- or target-side input, and drop the layer index for the last layer, *i.e.*, $\mathbf{U}^s$.

Lightweight Decoder for Imagination. For target imagination, we employ a lightweight decoder parameterized by a self-attention layer and 3-layer MLP attached to an intermediate layer ℓ ($\ell = 14$ out of 28 layers by default) of the MLLM to reconstruct the target patch representation from a vision foundation model [61]. Note that we omit the superscript s for brevity since the decoder is applied to each side independently. Each 1D factorized latent $u_k^{(\ell)} \in \mathbb{R}^D$ is spatially broadcast across a 2D grid with learnable positional embeddings and passed through the decoder. Yet, reconstructing the target from the latents alone would force them to encode the global layout of the target, which is dominated by low-frequency signals (often irrelevant to the localized edits in CIR). Thus, we condition the decoding process on a layout prior that provides only the coarse layout of the target; ground-truth (GT) target features $F \in \mathbb{R}^{H\times W\times D_{\text{DINO}}}$ are average-pooled to $h \times h$, expanded back to full resolution, and added to the broadcast latents. As this layout prior already explains the coarse layout, minimizing the reconstruction error drives the latents to carry only the information that is not contained in the coarse layout. Consequently, for each latent k, the decoder predicts spatial logits $\mathcal{M}_k \in \mathbb{R}^{H\times W}$ along with reconstruction features $\hat{F}_k \in \mathbb{R}^{H\times W\times D_{\text{DINO}}}$, where we index the HW spatial positions by p. To ensure that the contributions of all latents sum to 1 at each position p, we obtain the normalized alpha mask $\bar{\mathcal{M}}_k$ via a softmax over the latent dimension:

$$\bar{\mathcal{M}}_{k,p} = \frac{\exp(\mathcal{M}_{k,p})}{\sum_{k'=1}^{K} \exp(\mathcal{M}_{k',p})}. \tag{1}$$

Then, the reconstructed feature is formed by the mask-weighted sum:

$$\hat{F} = \sum_{k=1}^{K} \bar{\mathcal{M}}_k \odot \hat{F}_k, \tag{2}$$

where $\odot$ is element-wise multiplication with broadcasting along the channel dimension. Let $\hat{F}^q$ and $\hat{F}^t$ denote the features predicted from the query- and target-side latents, respectively. Both are supervised against the same GT target features F, and the imagination loss is the mean squared error averaged over both sides:

$$\mathcal{L}_{\text{imag}} = \frac{1}{2} \sum_{s \in \{q,t\}} \text{MSE}(\hat{F}^s, F). \tag{3}$$

As the coarse target layout prior is shared by both sides, $\mathcal{L}_{\text{imag}}$ is minimized only by improving the residual details that are not covered by the layout prior. On the query side, the latents must infer this residual from (I_r, T) without observing the target, which is the component we refer to as imagination. On the target side, the same objective anchors the target latents to the same residual factors, so that query and target latents occupy comparable factor slots for the latent matching in Sec. 3.2.

3.2. From Imagination to Retrieval

Factor-level Matching. To explicitly involve the factorized latents in retrieval, we apply an InfoNCE [60] on top of them. Building upon the factor alignment established by $\mathcal{L}_{\text{imag}}$, we define the matching score $s_{\text{OT}}(\mathbf{U}^q, \mathbf{U}^t)$ using Sinkhorn optimal transport (OT) over pairwise cosine similarities to allow flexible factor-to-factor correspondences (see Appendix E). For the i-th query and j-th target in a batch of size B, the factor-level similarity is $\mathbf{S}^f_{ij} = s_{\text{OT}}(\mathbf{U}^q_i, \mathbf{U}^t_j)$. With the batch-wise contrastive loss over a similarity matrix $\mathbf{S} \in \mathbb{R}^{B \times B}$ defined as:

$$\mathcal{C}(\mathbf{S}) = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(\mathbf{S}_{ii}/\tau)}{\sum_{j=1}^{B} \exp(\mathbf{S}_{ij}/\tau)}, \tag{4}$$

applying $\mathcal{C}(\mathbf{S}^f)$ encourages each query factor to align with its counterpart in the target, which a single pooled vector cannot enforce. τ is a temperature parameter.

Instance-level Matching. The `[EOS]` token, which attends to the factorized latents, aggregates the comprehended factors into a single embedding. We read out its normalized hidden state as the global embedding z and apply the same contrastive loss $\mathcal{C}(\mathbf{S}^g)$ with the instance-level similarity $\mathbf{S}^g_{ij} = z^{q\top}_i z^t_j$ between the i-th query and the j-th target. Unlike the factorized set, which requires storing K vectors per target image, z remains a single vector and thus serves as the primary retrieval embedding at scale. Our retrieval objective thus combines both levels:

$$\mathcal{L}_{\text{NCE}} = \mathcal{C}(\mathbf{S}^f) + \mathcal{C}(\mathbf{S}^g), \tag{5}$$

where $\mathcal{C}(\mathbf{S}^f)$ aligns individual factors through OT and $\mathcal{C}(\mathbf{S}^g)$ directly optimizes the embedding used at test time.

Propagating Factorization into the MLLM Backbone. While intermediate-layer decoding effectively encourages latent specialization, deeper layers near the retrieval output tend to lose spatial decomposition as representations abstract into global embeddings. To preserve factorized knowledge within the MLLM backbone, we distill the spatial alpha masks $\bar{\mathcal{M}}$ generated by the decoder into the latent-to-patch attention of the factorized latents at a deeper layer. Note that we formulate the spatial distillation loss exclusively on the target-side forward pass, as the query-side attention maps are defined over the reference image space, which does not share a spatial grid with the target alpha masks. Specifically, let $\mathcal{A}_k \in \mathbb{R}^{H \times W}$ denote the head-averaged attention of the k-th factorized latent toward the target patch tokens at layer ℓ_{top}, with the softmax taken over the patch keys only, bilinearly resampled to the decoder grid. Then, we normalize $\mathcal{A}_k$ and $\bar{\mathcal{M}}_k$ over the spatial positions into distributions $\mathbf{a}_k$ and $\mathbf{m}_k$, each describing the region the k-th latent occupies:

$$\mathbf{a}_{k,p} = \frac{\mathcal{A}_{k,p}}{\sum_{p'} \mathcal{A}_{k,p'}}, \quad \mathbf{m}_{k,p} = \frac{\bar{\mathcal{M}}_{k,p}}{\sum_{p'} \bar{\mathcal{M}}_{k,p'}}. \tag{6}$$

Consequently, we formulate a spatial distillation loss using Kullback-Leibler (KL) divergence to align the deeper layer attention distributions with the decoder's spatial alpha masks:

$$\mathcal{L}_{\text{distill}} = \frac{1}{K} \sum_{k=1}^{K} \text{KL}\big(\mathbf{m}_k \,\|\, \mathbf{a}_k\big). \tag{7}$$

This explicitly transfers spatial decomposition from the decoder to the MLLM. Although the supervision is applied only on the target side, it shapes the backbone and the latents shared by both sides, thereby encouraging the query-side latents to also retain localized semantics at inference, where the target layout prior is unavailable.

3.3. Training Objective and Inference

Training. Our joint objective reconciles two competing dynamics: $\mathcal{L}_{\text{imag}}$ forces factorized latents to partition target features into disjoint factors, whereas $\mathcal{L}_{\text{NCE}}$ encourages latents within an instance to collapse into a single discriminative embedding [62]. To address this, we employ a staged learning strategy. In Stage 1, we optimize only the reconstruction loss $\mathcal{L}_{\text{imag}}$ to warm up the factorized latents and the decoder, establishing a stable factorized prior by mapping latents to a spatially coherent region of the target scene. In Stage 2, we further train the MLLM backbone with DoRA [63] adapters, jointly with the latents and the decoder, using full objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{NCE}} + \lambda_{\text{imag}} \mathcal{L}_{\text{imag}} + \lambda_{\text{distill}} \mathcal{L}_{\text{distill}}, \tag{8}$$

where λ_{imag} and λ_{distill} are coefficients. Finally, we follow ReCALL [2] to add a post-training stage which in-

Table 1. **Results on Refined-CIRR and Refined-FashionIQ.** † are the results from CoLLM [19]. All other baselines are reproduced using official checkpoints, except for QWEN2.5-VL and ReCALL, which were re-run from scratch due to the lack of available checkpoints. $*$ denotes the use of additional gallery image captions. "Avg." for Refined-CIRR is computed as $(R@1 + R_{\text{subset}}@1)/2$, while on Refined-FashionIQ, it represents the mean across all three categories. Best and second-best results are shown in **bold** and underlined, respectively.

Method	Refined-CIRR								Refined-FashionIQ							
	Recall@N				$\text{Recall}_{\text{subset}}$@N			Avg.	Dress		Shirt		Top&Tee		Avg.	
	R@1	R@5	R@10	R@50	R@1	R@2	R@3		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
CoVR-2†	56.1	–	91.1	97.9	–	–	–	–	–	–	–	–	–	–	54.2	72.9
CoLLM†	60.4	–	92.6	98.2	–	–	–	–	–	–	–	–	–	–	57.2	76.4
VLM2Vec 7B	53.39	83.50	90.96	97.89	78.16	91.65	96.72	65.78	18.59	37.41	30.57	50.21	26.98	47.61	25.38	45.08
VLM2Vec-v2 2B	59.93	86.79	92.57	98.32	81.47	93.30	97.73	70.70	35.57	60.19	53.83	71.53	49.13	70.09	46.18	67.27
LIMN	52.32	82.74	90.40	98.04	74.41	90.55	96.28	63.37	53.27	74.43	60.03	78.73	63.84	82.36	59.05	78.50
ENCODER	57.76	86.15	92.80	98.75	82.76	94.53	97.94	70.26	56.62	78.50	62.76	80.78	66.07	84.64	61.82	81.31
TME	65.05	90.53	94.76	98.91	86.86	95.93	98.50	75.96	63.15	82.19	70.48	86.03	72.58	88.38	68.74	85.53
MM-Embed 8B	60.64	87.73	93.66	98.57	81.54	92.64	97.35	71.09	52.21	74.37	63.71	81.67	64.98	83.88	60.30	79.97
Qwen3-VL-Emb 8B	67.03	90.45	95.37	98.93	84.60	94.76	98.12	75.82	63.71	80.96	69.75	85.29	72.37	86.48	68.61	84.25
Core-Emb 8B	64.00	88.85	93.30	98.73	81.11	93.05	97.53	72.56	59.13	78.11	66.54	81.93	66.34	83.06	64.00	81.04
ReMatch 3B	62.40	89.36	94.04	98.98	83.17	94.35	98.17	72.79	33.50	57.40	50.53	69.43	46.85	66.18	43.63	64.34
CSMCIR*	67.16	90.96	**95.44**	**99.06**	86.15	95.75	98.55	76.66	63.65	82.08	71.48	86.61	73.51	**88.82**	69.55	85.83
Qwen2.5-VL (Base)	64.64	89.51	94.53	98.91	84.65	94.68	97.86	74.65	61.14	80.74	69.01	84.40	70.30	85.50	66.82	83.55
ReCALL	67.41	90.99	95.24	**99.06**	87.12	96.00	98.55	77.27	64.21	82.47	71.17	86.08	73.07	88.55	69.48	85.70
FactEmbed (Ours)	**68.99**	**91.29**	95.42	**99.06**	**87.45**	**96.16**	**98.63**	**78.22**	**66.61**	**83.08**	**72.79**	**87.66**	**74.00**	88.49	**71.13**	**86.41**

tegrates MLLM-augmented triplets to reduce false negatives [23, 33] and mitigate the influence of mislabeled triplets [20, 22, 25], both of which stem from the sparse supervision of CIR datasets. In brief, we adopt the existing augmentation pipeline to generate modification texts between each reference image and its mined hard negatives; the difference from [2] is that we condition the MLLM on the original target image and prompt it to describe what distinguishes the hard negative, which yields captions non-redundant with the original modification text in a single generation pass. Details on the augmentation are in Appendix F.

Inference. During inference, a single forward pass yields both global embedding z and factorized latents $\{u_k\}_{k=1}^{K}$. To minimize the memory footprint and computational cost, we aggregate latents into a single mean vector $\tilde{u} = \frac{1}{K}\sum_{k=1}^{K} u_k$, and utilize it exclusively for efficient reranking. Note that $\tilde{u}$ is not explicitly optimized during training; rather, it serves as a compact proxy for the factorized set, whose members are aligned via OT, and thus requires no additional training. Specifically, candidates in the gallery are initially searched with z alone, and the top-100 candidate shortlist is reranked by integrating the cosine similarity of the mean factor as:

$$r(q,t) = \cos\left(z^q, z^t\right) + \cos\left(\tilde{u}^q, \tilde{u}^t\right). \tag{9}$$

Since only the top candidates are rescored with a single vector, the additional cost is negligible and independent of the gallery size, while still leveraging the factor-level evidence.

4. Experiments

Datasets & Metrics. We test FactEmbed on CIRR [7] and FashionIQ [72], two standard CIR benchmarks. However, as original benchmarks contain ambiguous samples [19] and fail to provide robust evaluation [73], we conduct our primary evaluations on Refined-CIRR and Refined-FashionIQ [19]. Results on original datasets are also reported. For evaluation metrics, we use Recall@N (R@N), the fraction of queries whose target appears in the top-N results. Along with standard Recall, we also report $\text{Recall}_{\text{subset}}@\{1,2,3\}$ on CIRR, which are computed within each query's curated six-candidate subset.

Implementation Details. Adopting the base setup from ReCALL's repository [2], we use DoRA [63] to finetune Qwen2.5-VL-7B-Instruct [67] with rank 16, scaling factor 64. We use AdamW optimizer, a cosine schedule with 10% warmup, a batch size of 512, and bfloat16 mixed precision. The last checkpoint is used for evaluation. For hyperparameters, we set the number of latents $K = 8$, the resolution of target layout prior to 2×2 ($h = 2$), and the decoding layer to $\ell = 14$. Additionally, we distill the factorized map into the attention map from the 3rd layer from the top ($\ell_{\text{top}} = -3$). Loss coefficients are set to $\lambda_{\text{imag}} = 1.0$ and $\lambda_{\text{distill}} = 0.5$ for all experiments.

4.1. Main Results

In Tab. 1, we benchmark FactEmbed on refined CIR benchmarks against recent CIR methods and general-purpose embedding models [34, 36, 37, 74–76] using their official checkpoints. While prior works primarily relied on broader metrics like Recall@5 on CIRR due to sample ambiguities in original benchmarks, we emphasize that Recall@1 and Subset Recall@1 on CIRR serve as more precise ranking metrics under these refined evaluation settings, where broader metrics saturate near top performance. Specifically, FactEmbed outperforms two strong baselines, CSMCIR [66] (which additionally relies on synthetic text captions for every gallery image) and ReCALL [2]: it achieves improvements of 2.7% and 2.3% in Recall@1 on Refined-CIRR, and 2.3% and 2.4% in average Recall@10 on Refined-FashionIQ, respectively. Notably, even when post-trained with ReCALL's augmentation

Table 2. **Results on CIRR test set.** "Avg." for CIRR test set is computed as $(R@5 + R_{\text{subset}}@1)/2$.

Method	Recall@N				Recall$_{\text{subset}}$@N			Avg.
	R@1	R@5	R@10	R@50	R@1	R@2	R@3	
ARTEMIS [11]	16.96	46.10	61.31	87.73	39.99	62.20	75.67	43.05
TG-CIR [12]	45.25	78.29	87.16	97.30	72.84	89.25	95.13	75.57
SPRC [10]	51.96	82.12	89.74	97.69	80.65	92.31	96.60	81.39
LIMN [13]	43.64	75.37	85.42	97.04	69.01	86.22	94.19	72.19
CoVR-2 [64]	50.43	81.08	88.89	98.05	76.75	90.34	95.78	78.92
CaLa [14]	49.11	81.21	89.59	98.00	76.27	91.04	96.46	78.74
ENCODER [65]	46.10	77.98	87.16	97.64	76.92	90.41	95.95	77.45
CIR-LVLM [4]	53.64	83.76	90.60	97.93	79.12	92.33	96.67	81.44
QuRe [23]	52.22	82.53	90.31	98.17	78.51	91.28	96.48	80.52
CCIN [32]	53.41	84.05	91.17	98.00	–	–	–	–
TME [22]	53.42	82.99	90.24	98.15	81.04	92.58	96.94	82.01
Air-Know [24]	51.97	82.25	90.42	98.25	79.21	91.94	96.87	80.73
ConeSep [25]	53.06	83.45	90.60	98.34	81.23	92.62	97.24	82.34
Q3-VL-Emb 8B [36]‡	51.95	80.92	88.58	97.49	75.23	90.29	95.66	78.08
CSMCIR* [66]	53.88	**85.06**	91.69	**98.58**	80.10	92.24	96.92	82.58
Q2.5-VL (Base) [67]‡	51.98	81.90	90.27	98.19	77.69	91.85	96.92	79.80
ReCALL [2] ‡	53.76	83.76	91.54	98.34	81.11	93.35	97.52	82.44
FactEmbed (Ours)	**56.60**	84.87	**91.78**	98.41	**81.86**	**93.78**	**97.61**	**83.37**

Table 3. **Results on FashionIQ validation set.** "Avg." represents the mean performance across all three categories.

Method	Dress		Shirt		Top&Tee		Avg.	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
FashionSAP [68]	33.71	60.43	41.91	70.93	33.17	61.33	36.26	64.23
FAME-ViL [69]	42.19	67.38	47.64	68.79	50.69	73.07	46.84	69.75
SyncMask [70]	33.76	61.23	35.82	62.12	44.82	72.06	38.13	65.14
SADN [71]	40.01	65.10	43.67	66.05	48.04	70.93	43.91	67.36
CaLa [14]	42.38	66.08	46.76	68.16	50.93	73.42	46.69	69.22
CoVR-2 [64]	46.53	69.60	51.23	70.64	52.14	73.27	49.96	71.17
SPRC [10]	49.18	72.43	55.64	73.89	59.35	78.58	54.72	74.97
CIR-LVLM [4]	50.42	73.60	**58.59**	75.86	59.61	78.99	56.21	76.14
CCIN [32]	49.38	72.58	55.93	74.14	57.93	77.56	54.41	74.76
TME [22]	49.73	71.69	56.43	74.44	59.31	78.94	55.15	75.02
QuRe [23]	46.80	69.81	53.53	72.87	57.47	77.77	52.60	73.48
Air-Know [24]	51.03	73.08	56.76	75.05	59.47	79.02	55.75	75.72
ConeSep [25]	50.96	73.02	56.98	75.22	58.80	79.40	55.58	75.88
Q3-VL-Emb 8B [36]‡	47.84	69.46	52.36	71.34	55.58	74.40	51.93	71.73
CSMCIR* [66]	52.45	**74.81**	57.70	75.76	61.14	**80.98**	57.10	**77.18**
Q2.5-VL (Base) [67]‡	45.81	70.65	55.50	73.45	57.62	76.95	52.98	73.68
ReCALL [2] ‡	48.93	72.63	54.76	73.60	57.98	78.74	53.89	74.99
FactEmbed (Ours)	**53.25**	74.67	58.34	**76.30**	**61.55**	79.65	**57.71**	76.87

Bold and underline denote the best and second-best results, respectively. ‡ denotes results reproduced by us with official implementation. * denotes the additional use of image captions for gallery images, and Q denotes QWEN.

strategy, FactEmbed still surpasses prior methods (Fig. 6), indicating that the gains are not attributed to the augmentation alone. These gains demonstrate the advantage of decoupling complex query understanding from direct visual alignment: our decomposed factorized latents yield fine-grained, disambiguated semantic representations that standard contrastive architectures do not explicitly encourage.

In Tab. 2 and Tab. 3, we further report performance on the original CIRR and FashionIQ datasets. Notably, all methods score substantially lower on the original benchmarks than on their refined counterparts. Excluding Qwen3-VL-Embedding 8B, this gap reaches at most 13.7 points in R@1 on CIRR and 15.6 points in average R@10 on FashionIQ, whereas Qwen3-VL-Embedding 8B, the state-of-the-art generalist model, exhibits the largest gap (15.1 and 16.7 points, respectively). This is consistent with prior findings [19, 73] that the original benchmarks contain ambiguous annotations and evaluation artifacts, which may underestimate the compositional reasoning capabilities of high-capacity models. Nevertheless, FactEmbed achieves the best overall result.

4.2. Forming Embeddings with Factorized Latents

As factorized latents can be turned into a retrieval score in several ways, we compare the variants in Tab. 4. Matching the full latent set via OT ("Full Factor") achieves the highest R@1 on both benchmarks, but requires storing K vectors per gallery image, whose storage and search costs grow with the gallery size [42, 43]. Among single-vector alternatives, the mean-pooled latents ("Mean Factor") [76] perform the worst, as $\tilde{u}$ is never directly optimized for retrieval, leaving $\tilde{u}$ less discriminative as a standalone global embedding. Yet, directly optimizing $\tilde{u}$ with the instance-level objective in place of `[EOS]` ("Mean Factor$^{\#}$") still falls behind `[EOS]` in R@1 on both benchmarks, suggesting

Table 4. **Inference strategy for balancing performance and efficiency.** While matching the full set of factorized latents via optimal transport yields the best performance, its memory footprint scales with the number of latents. Efficient reranking achieves comparable results to full factor matching while maintaining low overhead. "Mean Factor$^{\#}$" is a variant where the instance-level contrastive loss is applied to the mean-pooled latent $\tilde{u}$ instead of `[EOS]`, together with the factor-level loss. Red highlights our default configuration.

	CIRR			Refined-CIRR		
	R@1	R@5	R@10	R@1	R@5	R@10
Full Factor	56.84	84.87	91.81	69.20	91.47	95.24
Mean Factor	54.87	83.21	90.53	67.21	90.30	94.68
Mean Factor$^{\#}$	55.69	84.36	91.37	68.25	91.60	95.42
`[EOS]`	56.12	84.58	91.61	68.84	90.96	94.93
Efficient Rerank	56.60	84.87	91.78	68.99	91.29	95.42

that imposing instance-level objective on the latents pulls them toward a single discriminative embedding at the cost of fine-grained evidence. We therefore keep the latents factor-aligned and use `[EOS]` as the global embedding for initial retrieval, while exploiting $\tilde{u}$ as complementary factor-level evidence for reranking. This efficient reranking narrows the R@1 gap to Full Factor to 0.24 and 0.21, with only two vectors per gallery image, a memory footprint on par with standard multi-vector CIR methods [10, 22, 23, 25, 32].

4.3. Ablation Study

In Tab. 5, we conduct component ablation study. Introducing factorized latents with the target imagination alone (b) does not consistently improve retrieval: while it improves R@10 on FashionIQ and Refined-FashionIQ by 1.00 and 1.16, and R@5 on CIRR by 0.99, it slightly lowers R@1 on CIRR. We attribute this to the gap between generation and discrimination: target reconstruction structures the latents into semantically factorized components, which already

Table 5. **Ablation study.** Imag., NCE, and Distill represent $\mathcal{L}_{\text{imag}}$ (Eq. 3), factor-level InfoNCE, and attention distillation (Eq. 7), respectively. ERR represents efficient reranking (Eq. 9) and Post denotes the post-training stage with augmented triplets. R_s@N denotes $\text{Recall}_{\text{subset}}$@N.

	Imag	NCE	Distill	ERR	Post	CIRR test					FIQ		Refined-CIRR					Refined-FIQ	
						R@1	R@5	R@10	R_s@1	R_s@2	R@10	R@50	R@1	R@5	R@10	R_s@1	R_s@2	R@10	R@50
a	-	-	-	-	-	51.98	81.90	90.27	77.69	91.85	52.98	73.68	64.64	89.51	94.53	84.65	94.68	66.82	83.55
b	✓	-	-	-	-	51.45	82.89	90.36	77.64	91.76	53.98	74.26	63.72	89.56	94.27	83.73	94.76	67.98	84.47
c	✓	✓	-	-	-	52.72	83.28	90.96	78.68	92.22	55.25	75.01	65.17	89.94	94.65	84.83	94.81	68.94	84.97
d	✓	-	✓	-	-	52.53	82.84	90.82	78.15	92.02	54.69	75.36	64.92	89.74	94.53	84.14	94.78	68.40	84.88
e	✓	✓	✓	-	-	53.28	83.42	91.18	78.94	92.51	55.98	75.94	65.43	90.17	94.83	84.73	94.86	69.19	85.50
f	✓	✓	✓	✓	-	54.75	83.86	91.61	80.84	93.18	56.67	76.19	66.55	90.33	95.11	85.54	95.47	69.78	85.67
g	-	-	-	-	✓	53.69	83.21	91.01	80.63	92.99	54.84	75.12	66.42	90.19	95.21	86.33	95.67	68.76	84.60
h	✓	✓	✓	✓	✓	56.60	84.87	91.78	81.86	93.78	57.71	76.87	68.99	91.29	95.42	87.45	96.16	71.13	86.41

benefits FashionIQ where modifications are confined to localized attributes, but it provides no explicit alignment metric for ranking, so the decomposition alone is insufficient for precise top-1 retrieval on open-domain CIRR. Yet, as it lays a robust foundation for fine-grained query decomposition, we find that bridging the gap with factor-level InfoNCE (c) and incorporating the attention distillation loss (e) effectively enhances the retrieval performance by transferring localized decomposition into the MLLM's deeper attention layer. Subsequently, incorporating our ERR in (f) delivers a prominent gain of 1.17 and 0.97 in Avg. on CIRR and Refined-CIRR, respectively. In contrast, applying the same reranking to latents trained without any auxiliary objective changes Avg. by only -0.37 and $+0.24$, indicating that the gain does not stem from merely fusing an additional vector but from the structure imposed by our objectives. We note that this gain comes at a marginal computational overhead, as the factorized latents are generated within the same forward pass (see Appendix C). Lastly, integrating the post-training stage (h) achieves additional gains, demonstrating the effectiveness of mitigating the noise arising from the sparse supervision of CIR datasets. To verify that these gains are not attributed to the augmentation alone, we also apply the post-training stage directly to the baseline (g). While post-training alone improves the baseline, our full model (h) further surpasses it by 2.91 and 2.57 in R@1 on CIRR and Refined-CIRR, and by 2.87 and 2.37 in R@10 on FashionIQ and Refined-FashionIQ, respectively. Moreover, the gains of our components remain comparable with and without post-training ((a)→(f) vs. (g)→(h)), indicating that our framework and the augmentation are complementary.

4.4. Analysis and Discussion

Imagination Layer Selection. Determining the MLLM layer ℓ for target imagination governs a trade-off; when the decoder is attached too early, the factorized latents may lack sufficient contextual fusion of multimodal inputs to imagine the target, while attaching it too late leaves insufficient subsequent layers to transform the decomposed factors into a discriminative retrieval embedding. In Fig. 2a, we evaluate decoder placements at layers $\ell \in \{7, 14, 21\}$ of the 28-layer MLLM, and find that $\ell = 14$ yields the best result, which

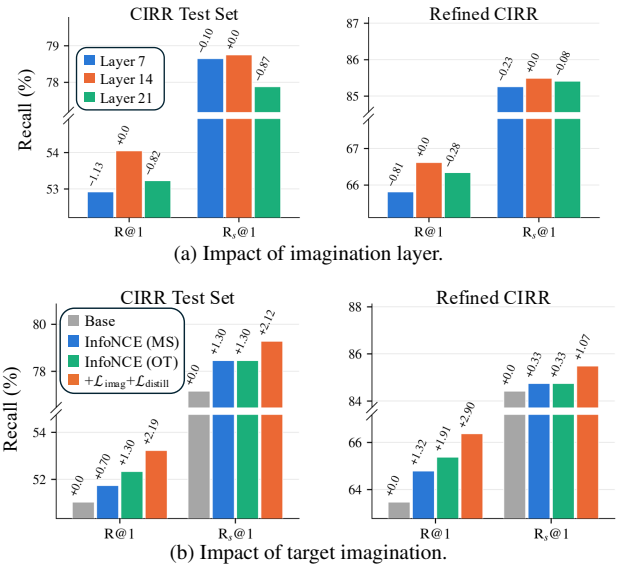


Figure 2. **Studies on imagination layer and comparison against NCE-only baselines with different matching metrics.** MS and OT denote maximum similarity and optimal transport applied to latents, respectively, without the imagination objective for factorization.

aligns with recent findings [77, 78].

Impact of Target Imagination. To isolate the effect of target imagination, we compare against variants that train the latents solely with InfoNCE, matched by either maximum similarity (MS) [47] or OT, in Fig. 2b. While these variants outperform the naïve baseline, we observe that target imagination brings a further gain by enforcing the latents to decompose and structure the target scene rather than relying on the contrastive signal alone. Also, OT matching consistently outperforms MS matching strategy on both benchmarks, supporting our choice of OT for factor-level matching.

What Does the Decoder Imagine? Fig. 3 compares the imagined features $\hat{F}^q$ against the GT target features F, using joint PCA basis. On the query side, the imagined features reconstruct the overall structure of the target: while the layout prior supplies only the coarse layout, the factorized latents recover fine-grained details such as object silhouettes. The rightmost column visualizes $\arg\max_k \bar{\mathcal{M}}_{k,p}$, demonstrating that each factorized latent contributes to reconstructing



Figure 3. **Qualitative analysis of target imagination.** For each composed query, we visualize target DINOv2 features (DINOv2 PCA) and features imagined by our decoder from the query (Recon PCA), projected onto a joint PCA basis. Alpha Mask Argmax shows, for each spatial position p, the factorized latent with the largest decoder alpha weight, *i.e.*, $\arg\max_k \bar{\mathcal{M}}_{k,p}$, where each color denotes a different latent. On top of the coarse layout from the target prior, fine details are recovered by the latents, jointly reconstructing the target structure.



Figure 4. **Impact of layout prior in target imagination.** Three rows compare factor decomposition across different layout prior configurations: (top) without a layout prior, (middle) with a 1×1 layout prior, and (bottom) with our default 2×2 layout prior. While the unconditioned model collapses into coarse global semantics, introducing the 2×2 layout prior guides the latents to isolate semantically editable parts.

semantically coherent regions. This suggests that each latent specializes in a coherent semantic factor of the target. Additional visualizations are in Appendix A.

Impact of Target Layout Prior. As described in Sec. 3.1, the decoder is conditioned on a coarse target layout prior. Fig. 4 visualizes the decomposition under different prior resolutions. Without the layout prior, the latents must also encode the global layout, which dominates the reconstruction objective and causes them to collapse into a few coarse regions. In contrast, when the prior is injected, we found that the factorized latents are encouraged to focus on fine-grained visual structures, decomposing the target into localized parts such as arms, legs, and necklines. We set $h = 2$ so that the layout prior provides only the coarse layout while leaving the localized factors to the latents.

Generalization across Backbones. We further apply FactEmbed to stronger MLLM backbones in Tab. 6. Specifically, we observe that upgrading the backbone from Qwen2.5-VL-7B to Qwen3-VL-8B improves R@1 by 2.65 points on the CIRR test set, exceeding the 1.60 points gain reported by ReCALL [2] for the same backbone transition. This demonstrates that FactEmbed scales more effectively with improved backbone capabilities. The performance gains further extend to Qwen3-VL-Embedding-8B, a backbone already specialized for retrieval, where FactEmbed achieves top performance across most metrics.

Table 6. **FactEmbed with different MLLM backbones.** "Avg." follows Tab. 2 and Tab. 1, *i.e.*, (R@5 + R_s@1)/2 on CIRR test and (R@1 + R_s@1)/2 on Refined-CIRR.

	Backbone	R@1	R@5	R@10	R@50	R_s@1	R_s@2	R_s@3	Avg.
CIRR test	Qwen2.5-VL-7B	56.60	84.87	91.78	98.41	81.86	93.78	97.61	83.37
	Qwen3-VL-8B	59.25	86.10	91.93	98.36	83.54	94.53	97.74	84.82
	Qwen3-VL-Emb-8B	**60.43**	**86.99**	**93.35**	**98.70**	**84.15**	**94.82**	**97.81**	**85.57**
Refined CIRR	Qwen2.5-VL-7B	68.99	91.29	95.42	99.06	87.45	96.16	98.63	78.22
	Qwen3-VL-8B	71.31	92.29	95.72	**99.16**	**89.33**	96.59	**98.88**	80.32
	Qwen3-VL-Emb-8B	**73.60**	**93.02**	**96.41**	99.13	88.95	**97.12**	**98.88**	**81.28**

5. Conclusion

In this work, we present FactEmbed, a framework that better exploits the MLLM's compositional understanding capability to simulate visual edits by decoupling compositional query comprehension from learning discriminative embedding. To structure query comprehension, we prepend factorized latents that act as an information bottleneck, trained to imagine the residual target factors beyond a coarse layout prior that groups semantically coherent visual factors of the target without dense annotations. We then bridge generative imagination and discriminative retrieval by optimizing a factor-level InfoNCE loss via optimal transport and distilling the decoder's alpha masks into the MLLM's deeper attention maps. At inference, the `[EOS]` token serves as a global-level embedding to perform initial retrieval, while the mean-pooled factorized latents, produced in the same forward pass, rerank a short candidate list at negligible cost. Extensive evaluations show that FactEmbed achieves state-of-the-art performance, suggesting that factor-level supervision on top of instance-level matching improves MLLM-based CIR.

Limitations. Despite its effectiveness, FactEmbed has several limitations. First, using a fixed K ignores the input complexity since no explicit supervision specifies the number of semantic factors. Exploring query-adaptive latents and matching schemes across varying set sizes remains for future work. Second, while full-factor retrieval achieves higher accuracy, we mean-pool the latents for efficient reranking, unexploiting part of the fine-grained information. Incorporating Matryoshka representation learning [79] into factor representations may preserve accuracy of full-factor matching near single-vector storage costs. Lastly, our evaluation is limited to CIRR and FashionIQ; thus, generalizability as a unified multimodal retriever requires further verification.

A. Additional Visualizations for Imagination

In this section, we present additional results for target imagination in Fig. 5. Following the protocol in Sec. 4.4, we jointly apply PCA to the ground-truth DINOv2 [61] feature maps and our imagined representations for visualization. Although the shared color mapping between both visualized images is partly facilitated by the target layout prior, we remark that this prior provides only a highly coarse, average-pooled 2×2 spatial layout. Consequently, the reconstruction of details (*e.g.,* geometry and position) demonstrates that our factorized latents successfully capture rich spatial positioning and object geometry, allowing the decoder to synthesize the spatial maps across local regions.

B. Diversity of Factorized Latents

To analyze whether each factorized latent specializes in distinct concepts, we examine the cosine similarity between factorized latents both within the same instance and across different instances (Tab. 7). Specifically, we compare our approach against two baseline variants: one where learnable latents are prepended to the `[EOS]` token without additional constraints, and another where the factor-level InfoNCE objective is introduced.

As discussed in Sec. 3.3, applying the factor-level InfoNCE alone without the imagination objective $\mathcal{L}_{\text{imag}}$ (b) increases the similarity among latents within each instance, causing them to collapse into a single instance-level representation. This is because the retrieval objective only requires the latents to be collectively discriminative across instances, providing no incentive for individual latents to take on distinct roles. Furthermore, this representation collapse becomes even more pronounced when the per-latent mean vector (computed across the dataset) is subtracted from each latent, indicating that the latents share instance-specific rather than factor-specific information. In contrast, incorporating our imagination and distillation objectives (c) yields the lowest within-instance similarity, with a clear margin after removing the per-latent mean (0.548 vs. 0.599 and 0.841). This indicates that our framework encourages the latents to take on distinct roles, whereas naïvely employing multi-vector latents with a retrieval objective drives them toward a shared representation.

Note that the latents without any auxiliary objective (a) also exhibit low within-instance similarity. We attribute this phenomenon to the fact that the latents are orthogonally initialized and largely retain this initial bias due to the lack of direct supervision. Indeed, these latents do not benefit retrieval: even with efficient reranking, they achieve an average of 73.21 on Refined-CIRR and 79.60 on the CIRR test set, falling below the baseline without latents (74.65 and 79.80), where the averages follow Tab. 1 and Tab. 2, *i.e.*, (R@1 + R_{subset}@1)/2 and (R@5 + R_{subset}@1)/2, respectively. In contrast, under the same setting without post-training, our latents raise these scores to 76.05 and 82.35 (Tab. 5 (f)), confirming that they are not only distinct but also informative for retrieval.

C. Inference Cost

A natural concern with multi-vector representations is that they trade retrieval quality for memory footprint and query latency [42, 47]. To address this, Tab. 8 quantifies the inference overhead on the CIRR validation set (2,297 gallery images) using a single NVIDIA RTX PRO 5000 Blackwell GPU. Crucially, the forward pass of the query encoder dominates the end-to-end latency pipeline, requiring 353ms per query at batch size 1 and an average of 164ms per query at batch size 32. In contrast, scoring the gallery using only the `[EOS]` embedding requires only 0.06ms per query. Our efficient reranking procedure, which re-scores the top-100 shortlist with the sum of the cosine similarities from each `[EOS]` and mean factorized vector, adds one extra vector per image to the index and 0.17ms per query at batch size 1 (0.003ms amortized), which only induces less than 0.05% overhead on the total latency. We adopt this design as our default inference configuration, as its storage footprint remains comparable to, or slightly smaller than that of recent CIR methods that maintain multi-vector galleries [10, 14, 22, 23, 25, 32]. When storage and compute budgets permit, however, the full latent variant can be preferred, as full OT matching offers slightly higher R@1 (Tab. 4). Note that its time overhead remains marginal at the scale of the CIRR validation gallery, but both its storage (K vectors per image) and its scoring cost grow linearly with the gallery size, whereas efficient reranking only rescores a fixed top-100 shortlist.

D. Additional Implementation Details

In this section, we elaborate additional implementation details. While we use base learning rates of 2×10^{-5} and 4×10^{-5} to train DoRA modules attached to the pretrained backbone parameters for CIRR and Fashion-IQ, respectively, we apply 4×10^{-4} as the learning rate to modules such as the factorized latent and imagination decoder since our newly introduced components are trained from scratch. For the training schedule, we use 3,000, 500, and 350 steps across the three stages for CIRR, and 3,000, 350, and 350 steps for FashionIQ. While Stage 1 could technically be trained jointly with the later stages, the competing dynamics between $\mathcal{L}_{\text{imag}}$ and $\mathcal{L}_{\text{NCE}}$ slow down the formation of factorized latents, requiring substantially longer training to converge. Isolating Stage 1 to train only the decoder and the latents with $\mathcal{L}_{\text{imag}}$ removes this competition and further allows bypassing the upper layers of the MLLM backbone, yielding a substantially more compute-efficient setup: Stage 1 completes in approximately 13 hours on 4$\times$ RTX PRO 5000 Blackwell

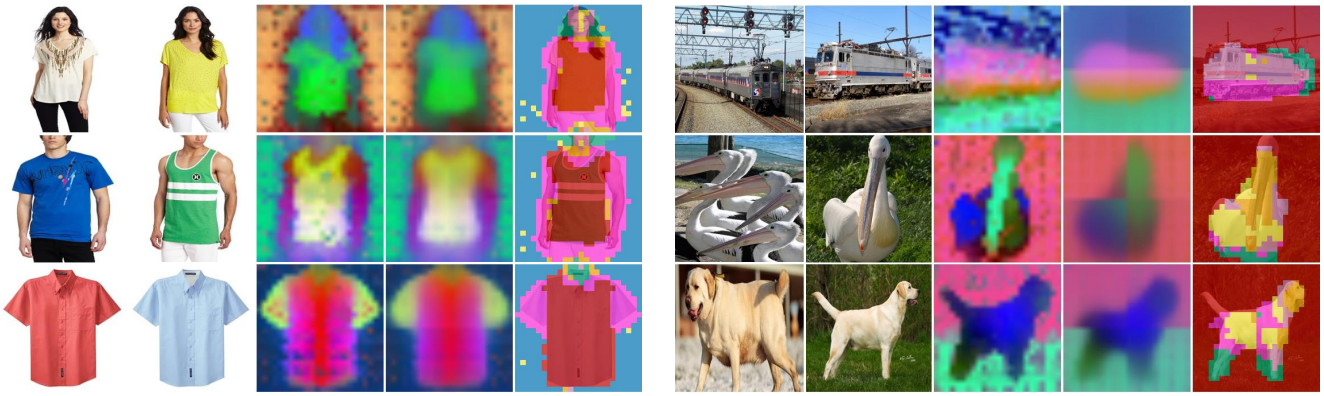


Figure 5. **Additional Qualitative Analysis of Target Imagination.**

Table 7. **Diversity analysis on factorized latents.** "Within instance" denotes the mean cosine similarity across all latent pairs within the same image, whereas "Across instances" denotes the mean cosine similarity between latents of different images. Further, we report prototype-removed diversity which is computed by subtracting the dataset-wide per-latent mean vector (computed over the CIRR dataset) from each latent to isolate diversity from MLLM residual biases.

	Methods	Raw Similarity		Prototype-removed
		within instance	across instances	within instance
a	Learnable Latents only (no auxiliary loss)	0.606	0.408	0.599
b	Latents with InfoNCE	0.764	0.548	0.841
c	Ours	0.586	0.426	0.548

GPUs. Then, subsequent training stages were conducted using 2× RTX PRO 6000 Blackwell GPUs using PyTorch FSDP. Images are resized within the processor to between 3,136 and 147,456 pixels, with input sequences truncated at 512 tokens.

E. Optimal Transport for Matching Score between Factorized Sets

To measure the similarity between two unordered factorized sets, we used Optimal Transport (OT) to find the best possible alignment between them. Given the query and target factorized latents $\mathbf{U}^q = [u_1^q, \ldots, u_K^q]$ and $\mathbf{U}^t = [u_1^t, \ldots, u_K^t]$, we first compute a pairwise cosine similarity matrix $\mathbf{R} \in \mathbb{R}^{K\times K}$, where $\mathbf{R}_{k,m} = \cos(u_k^q, u_m^t)$ is the cosine similarity between the k-th query factor and the m-the target factor. Based on this similarity matrix $\mathbf{R}$, the OT matching score is defined as:

$$s_{\mathrm{OT}}(\mathbf{U}^q, \mathbf{U}^t) = \langle \mathbf{P}^\star, \mathbf{R} \rangle,$$
$$\mathbf{P}^\star = \arg\max_{\mathbf{P}\in\Pi(\frac{1}{K}\mathbf{1}, \frac{1}{K}\mathbf{1})} \langle \mathbf{P}, \mathbf{R} \rangle + \varepsilon \mathcal{H}(\mathbf{P}), \tag{10}$$

where $\mathbf{P}^\star$ is the optimal transport plan that maximizes the overall alignment score and $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product. $\Pi(\frac{1}{K}\mathbf{1}, \frac{1}{K}\mathbf{1})$ denotes the set of transport plans whose rows and columns each sum to $1/K$, so that every query and target factor contributes equally. The entropy regularization $\mathcal{H}(\mathbf{P}) = -\sum_{k,m} \mathbf{P}_{k,m} \log \mathbf{P}_{k,m}$ allows $\mathbf{P}^\star$ to be computed efficiently with a few Sinkhorn iterations, serving as a differentiable surrogate for one-to-one matching. We set the $\varepsilon = 0.1$ and run 30 Sinkhorn iterations. Since the elements of $\mathbf{P}^\star$ sum to 1, $\langle \mathbf{P}^\star, \mathbf{R} \rangle$ represents the average matching score in $[-1, 1]$, which is subsequently used as the factor-level similarity $\mathbf{S}^f$ in the first term of Eq. 5. This encourages the factorized set of a query to align with that of its target while repulsing the sets of non-matching targets.

F. Post-training with MLLM Augmentation

A typical challenge in CIR is that the modification text often underdetermines the target image. In other words, instructions are often rarely exhaustive; they specify what should change while leaving everything else implicit. Consequently, a single composed query can often be satisfied by multiple candidate images in the gallery. To address this, prior works have leveraged LLM/MLLMs to sharpen the query itself to disambiguate the instruction [32, 33], enlarged the clean paired triplets [2, 18, 19, 35]. Similarly, we follow ReCALL [2] to augment training triplets by pairing each reference image with its mined hard negative instances, and

Table 8. **Per-query inference cost on CIRR validation set (2,297 gallery images).** We report the number of vectors indexed per gallery image, the scoring cost (similarity matching), and the total inference latency for both single and batched queries. Total inference latency involves the scoring cost and the encoding cost of the query-side, while the cost of encoding offline gallery (target-side) is excluded. Δ indicates the latency overhead relative to the `[EOS]`-only baseline.

Retrieval Unit	Num Vectors	Scoring		Total			
		Single (bs=1)	Batched (bs=32)	Single (bs=1)		Batched (bs=32)	
		ms/q	ms/q	ms/q	Δ	ms/q	Δ
`[EOS]` only	1	0.06	0.004	353.1	–	163.8	–
+ Efficient Rerank	2	0.23	0.007	353.3	+0.17ms	163.8	+0.003ms
Full Factor	8	3.99	0.31	357.1	+3.93ms	164.1	+0.31ms

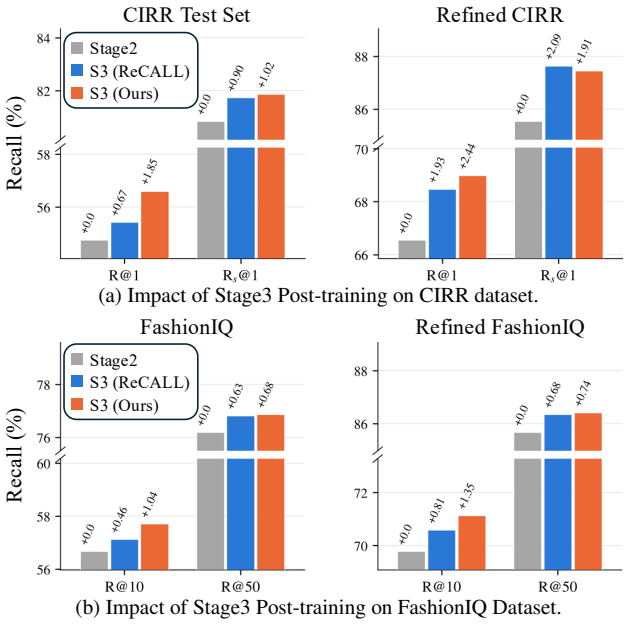
then finetune the model on the resulting data (comprising the reference image, its hard negative, and a generated modification text). Operating on FactEmbed after Stage 2, our augmentation pipeline consists of two phases: hard negative mining and contrastive captioning. During the mining phase, we follow ReCALL to extract hard negative instances I_h for each reference image I_r that ranks above the ground truth I_t, which are then passed to the contrastive captioning phase.

F.1. Contrastive Captioning

Following ReCALL [2], we convert each mined hard negative I_h into a new positive triplet $(I_r, \acute{T}, I_h)$, where an MLLM writes a minimal modification text $\acute{T}$ that transforms I_r into I_h. In practice, however, we found this pipeline to be both noisy and costly. Since mined hard negatives frequently serve as valid targets for (I_r, T) themselves and the captioner observes only I_r and I_h, we observed that the captioner has no cue to separate I_h from I_t. As a result, it often returned T unchanged, producing duplicate queries with conflicting positives (9% and 11% of the samples that survived ReCALL's VQA filter on CIRR and FashionIQ). This pipeline was also inefficient: the local MLLM captioner frequently violated the required output format, with 37% (CIRR) and 56% (FashionIQ) of instances failing on the first pass and 9% and 12% remaining unresolved even after six retry rounds, and the subsequent VQA filter added another full inference pass.

We thus streamline the pipeline with two modifications, while keeping the remaining instructions of ReCALL's captioning prompt unchanged. First, we employ contrastive captioning strategy, which additionally presents the original target I_t alongside I_r and I_h, and instruct the model to describe I_h such that the text no longer applies to I_t. Note that I_t serves purely as a conditioning signal and is omitted from the augmented triplet. Second, we replace the local captioner with GPT-5.6-luna [80] via the batch API, which can follow both the contrast instruction and the output format in a single pass, removing the retry loop and separate verification stage.

Compared with the ReCALL's two-image prompting strategy on 100 sampled mined queries per dataset, we found that our contrastive captioning successfully yields an edited modification text while the two-image prompting strategy produced 27.0% (CIRR) and 33.0% (FashionIQ) of modification texts unchanged. Furthermore, among the edited queries, we found that 30.1% (CIRR) and 10.5% (FashionIQ) of the modification texts rewritten by the two-image prompting strategy were judged by GPT-5.6-luna to describe the original target I_t better than the mined target I_h, while our contrastive captioning reduced this reversal rate to 7.0% and 0.0%, respectively. Consequently, as shown in Fig. 6, our contrastive captioning yields larger gains than the augmentation strategy of ReCALL on most metrics, with comparable results on R_s@1 of Refined-CIRR. This demonstrates that our streamlined pipeline is not only more efficient, as it requires neither a caption retry loop nor a separate verification stage, but



(a) Impact of Stage3 Post-training on CIRR dataset.

(b) Impact of Stage3 Post-training on FashionIQ Dataset.

Figure 6. **Impact of augmentation strategies for post-training.** Starting from our model before post-training (Stage 2), we compare post-training with triplets augmented by the two-image prompting of ReCALL [2] and by our contrastive captioning. Numbers above the bars denote the absolute gains over stage 2 baseline.

also yields higher-quality augmentation. We remark that other than these modifications, we followed ReCALL [2] to mine five hard negatives per composed query, resulting in 57,476 (CIRR) and 56,499 (FashionIQ) captioning requests. This incurs a one-time cost of only $22.82 (CIRR) and $26.17 (FashionIQ), which is highly economical compared to renting GPU servers to deploy a local MLLM, which often struggles with strict output formatting and thus requires repeated retries.

F.2. Hard Negative Filtering

Hard negative mining occasionally yields near-duplicate instances that are virtually identical to the annotated target despite holding distinct identifiers in the same dataset. Rewriting such a pair forces the model to generate text that separates two near-identical images, teaching the retriever to incorrectly push them apart. To prevent this, we filter out these pairs prior to contrastive captioning based on visual feature similarity via DINOv2 [61] before we query the MLLM to write the captions. Specifically, we encoded each image using a frozen DINOv2 ViT-B/14 backbone and mean-pooled the patch tokens into a single descriptor, discarding hard negative candidates (I_h) that exhibit a cosine similarity exceeding 0.99 with the original target (I_t). This filtering step removes 743 CIRR instances and 277 FashionIQ instances out of $58,219$ and $56,776$ mined pairs, respectively. For Stage 3, we additionally filter mislabeled original triplets: when mining with the Stage 2 model, we also record the rank of the ground-truth target, and drop the 3% of original triplets whose ground truth is ranked lowest.

F.3. Post-training

Finally, for the post-training stage, we optimize our model using the total objective, described in Eq. 8, similarly to stage 2. Note that we construct each batch for CIRR by grouping an original triplet with its MLLM-augmented counterparts following [2]. This strategy supplies high-quality in-batch hard negatives, explicitly encouraging the model to discern subtle visual differences.

F.4. Limitation of Contrastive Captioning

While our contrastive captioning mitigates noisy supervision in both the original and augmented triplets, a limitation remains when scaling up the augmentation. Contrastive captioning conditions each rewrite only on the original target I_t, so it reduces ambiguity between I_t and a hard negative but not among the multiple hard negatives mined for the same composed query. As more hard negatives are mined per query, *i.e.,* each pair of reference image I_r and modification text T, we expect these sibling hard negatives to increasingly overlap in semantics, so that a modification text written for one may equally describe another, introducing false negatives within the augmented set. An effective augmentation strategy for CIR therefore requires conditioning across both the original and augmented sets, so that each generated instruction captures features that distinguish its target from all other candidates of the same query. Nevertheless, since the augmented triplets consistently improve CIR performance in our experiments, we will publicly release them to facilitate future research.

G. Qualitative Results

In Fig. 7, Fig. 8, Fig. 9, and Fig. 10, we present qualitative comparisons between FactEmbed and our baseline (Qwen2.5-VL-7B-Instruct [67] fine-tuned via DoRA [63]). In CIR, modifying a reference image with text often projects the query into a broad semantic space where multiple target images can be considered valid matches. Therefore, rather than restricting our analysis to Top-1 prediction, we examine the Top-8 retrieved results to better assess the overall alignment between the composed query and the retrieved distribution.

In the first example of Fig. 7, which belongs to the dress category of FashionIQ dataset, the query specifies an item that is "more fashionable and party wear and is cream and brown with animal print". While both models successfully capture the desired color scheme, FactEmbed ranks candidates that accurately reflect the fine-grained "animal print" pattern higher than the baseline. In the second example, the baseline struggles with structural modifications, incorrectly retrieving an item with short sleeves at Rank-2 and entirely ignoring the requested "strap" detail at Rank-3. Conversely, FactEmbed strictly adheres to these fine-grained structural constraints across the top ranked candidates. Fig. 8 further highlights a common failure mode in standard contrastive learning, which fails to handle negative constraints. For the query containing the phrase "instead of superman logo", the baseline fails to disentangle the negated element, retrieving images containing the superman logo. In contrast, FactEmbed successfully suppresses the negated visual element while maintaining the core semantics of the reference image.

This gap in multimodal understanding extends to the open-domain CIRR dataset. As shown in the first example of Fig. 10, the query requires the model to identify "map displayed on the screen" of a smartphone. While the baseline is easily distracted by generic smartphone images, FactEmbed accurately grounds the text modification into the precise visual context. Also, for queries involving relative counting or relational reasoning (e.g., second example), the baseline exhibits a severe lack of compositional understanding. Overall, we find that the baseline trained with standard contrastive learning often ignores the comparative conditions, while FactEmbed captures these relationships more faithfully.

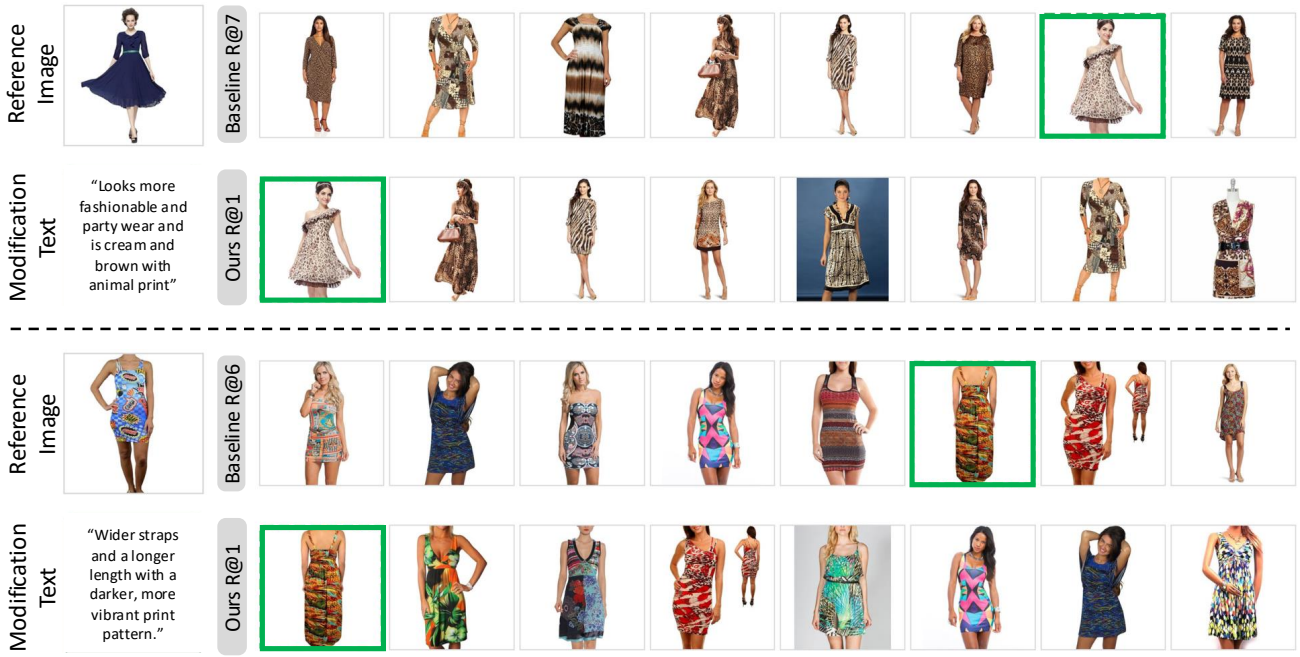


Figure 7. **Results on dress category.**

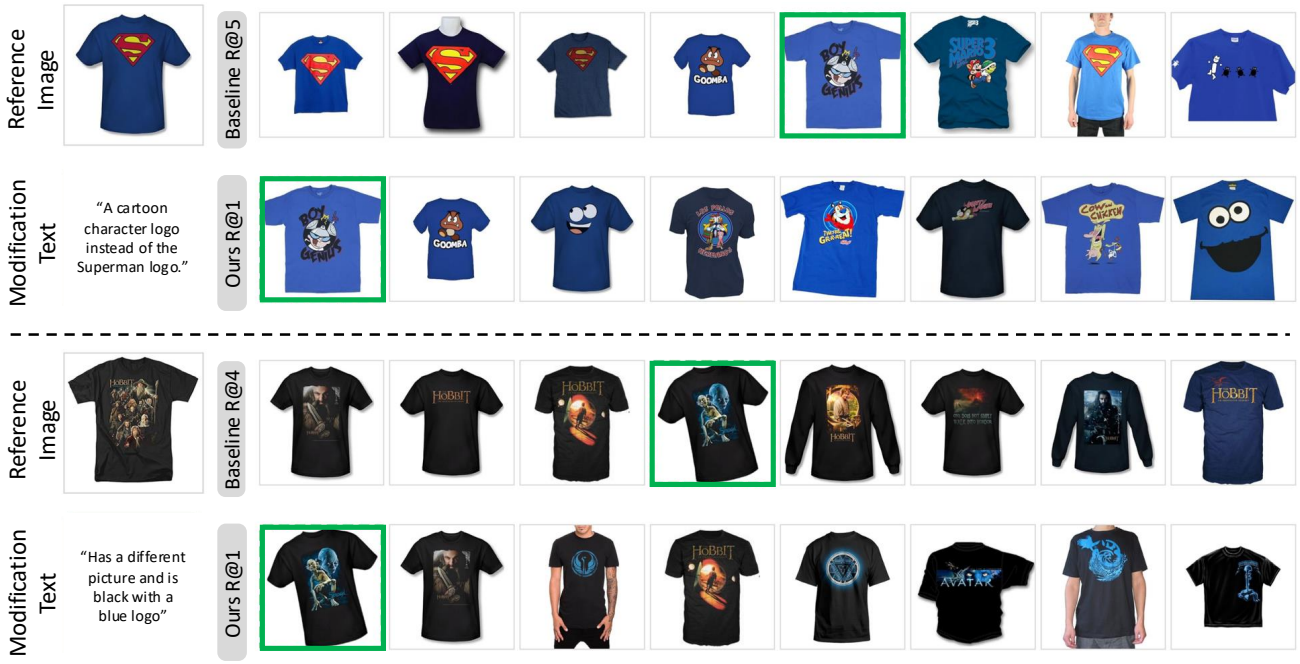


Figure 8. **Results on Shirt category.**

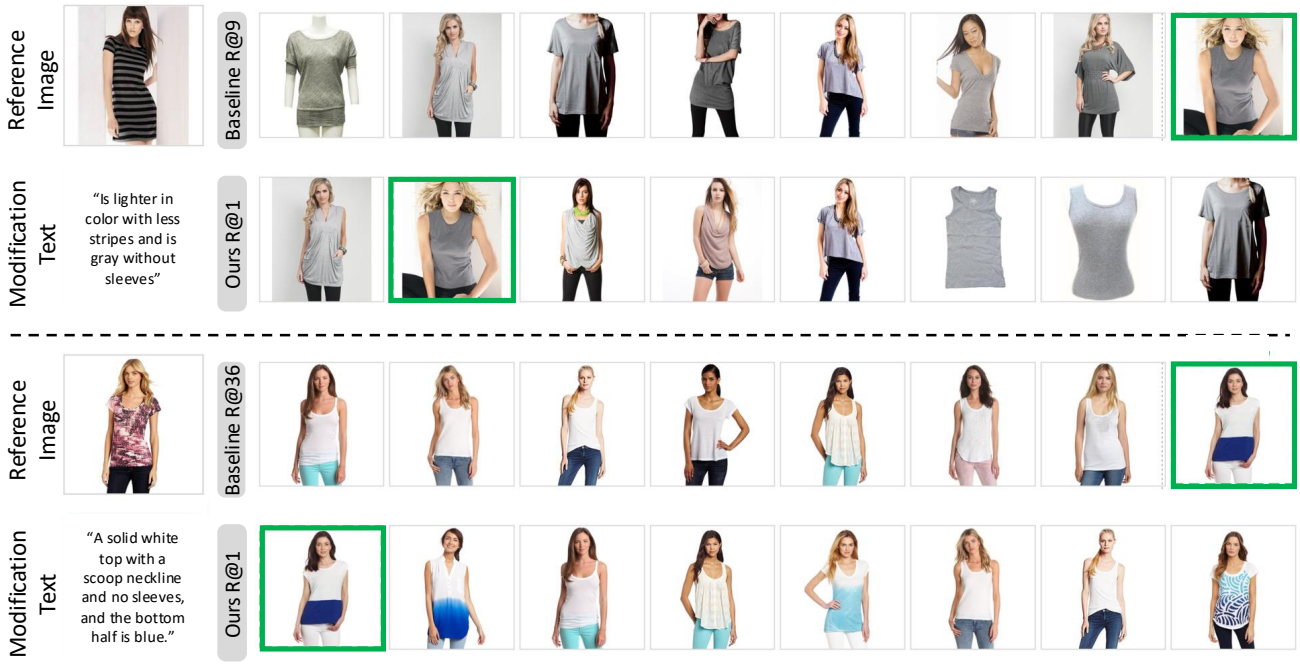


Figure 9. **Results on Top&Tee category.**

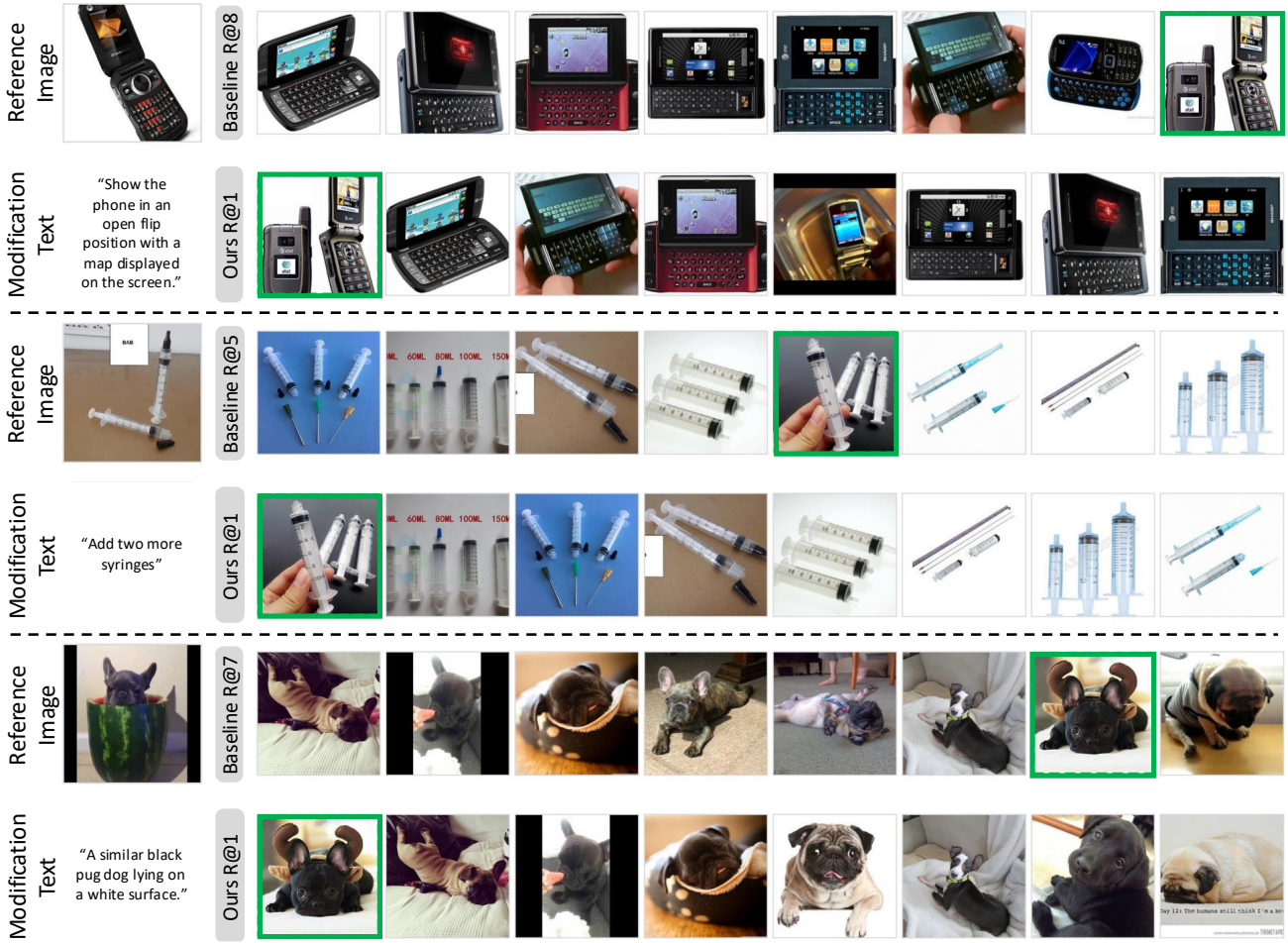


Figure 10. **Results on CIRR.**

References

[1] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval - an empirical odyssey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6439–6448, 2019. 1, 2

[2] Tianyu Yang, ChenWei He, Xiangzhao Hao, Tianyue Wang, Jiarui Guo, Haiyun Guo, Leigang Qu, Jinqiao Wang, and Tat-Seng Chua. Recall: Recalibrating capability degradation for mllm-based composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 38763–38773, June 2026. 1, 2, 4, 5, 6, 8, 10, 11, 12

[3] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024. 1

[4] Zelong Sun, Dong Jing, Guoxing Yang, Nanyi Fei, and Zhiwu Lu. Leveraging large vision-language model as user intent-aware encoder for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7149–7157, 2025. 1, 2, 6

[5] Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4015–4025. IEEE, 2025.

[6] 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. 1

[7] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2125–2134, October 2021. 2, 5

[8] Rohan Mahadev, Joyce Yuan, Patrick Poirson, David Xue, Hao-Yu Wu, and Dmitry Kislyuk. Pinpoint: Evaluation of composed image retrieval with explicit negatives, multi-image queries, and paraphrase testing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9742–9751, June 2026.

[9] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*, 2023. 2

[10] Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. Sentence-level prompts benefit composed image retrieval. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 6, 9

[11] Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In *International Conference on Learning Representations*, 2022. 6

[12] Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. Target-guided composed image retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. 6

[13] Haokun Wen, Xuemeng Song, Jianhua Yin, Jianlong Wu, Weili Guan, and Liqiang Nie. Self-training boosted multi-factor matching network for composed image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:3665–3678, 2024. 6

[14] Xintong Jiang, Yaxiong Wang, Mengjian Li, Yujiao Wu, Bingwen Hu, and Xueming Qian. Cala: Complementary association learning for augmenting composed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2177–2187, 2024. 2, 6, 9

[15] 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*, pages 8748–8763. PMLR, 2021. 2

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2

[17] 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*, pages 19730–19742. PMLR, 2023. 2

[18] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. MagicLens: Self-supervised image retrieval with open-ended instructions. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 59403–59420. PMLR, 21–27 Jul 2024. 2, 10

[19] Chuong Huynh, Jinyu Yang, Ashish Tawari, Mubarak Shah, Son Tran, Raffay Hamid, Trishul Chilimbi, and

Abhinav Shrivastava. Collm: A large language model for composed image retrieval. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3994–4004. IEEE, 2025. 2, 5, 6, 10

[20] Zhiwei Chen, Yupeng Hu, Zhiheng Fu, Zixu Li, Jiale Huang, Qinlei Huang, and Yinwei Wei. Intent: Invariance and discrimination-aware noise mitigation for robust composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. 2, 5

[21] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[22] Shuxian Li, Changhao He, Xiting Liu, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Learning with noisy triplet correspondence for composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19628–19637, 2025. 2, 5, 6, 9

[23] Jaehyun Kwak, Ramahdani Muhammad Izaaz Inhar, Se-Young Yun, and Sung-Ju Lee. Qure: Query-relevant retrieval through hard negative sampling in composed image retrieval. In *Forty-second International Conference on Machine Learning*, 2025. 2, 5, 6, 9

[24] Zhiheng Fu, Yupeng Hu, Qianyun Yang, Shiqi Zhang, Zhiwei Chen, and Zixu Li. Air-know: Arbiter-calibrated knowledge-internalizing robust network for composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2026. 2, 6

[25] Zixu Li, Yupeng Hu, Zhiwei Chen, Mingyu Zhang, Zhiheng Fu, and Liqiang Nie. Conesep: Cone-based robust noise-unlearning compositional network for composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2026. 2, 5, 6, 9

[26] Young Kyun Jang, Dat Huynh, Ashish Shah, Wen-Kai Chen, and Ser-Nam Lim. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. In *European conference on computer vision*, pages 239–254. Springer, 2024. 2

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

[28] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only efficient training of zero-shot composed image retrieval. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13225–13234. IEEE, 2024.

[29] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. *International Conference on Learning Representations (ICLR)*, 2024.

[30] You Li, Fan Ma, and Yi Yang. Imagine and seek: Improving composed image retrieval with an imagined proxy. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 3984–3993, June 2025. 2

[31] Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Gaopeng Gou, and Qi Wu. Missing target-relevant information prediction with world model for accurate zero-shot composed image retrieval. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24785–24795. IEEE, 2025. 2

[32] Likai Tian, Jian Zhao, Zechao Hu, Zhengwei Yang, Hao Li, Lei Jin, Zheng Wang, and Xuelong Li. Ccin: Compositional conflict identification and neutralization for composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2025. 2, 6, 9, 10

[33] François Gardères, Camille-Sovanneary Gauthier, Jean Ponce, and Shizhe Chen. Fire-cir: Fine-grained reasoning for composed fashion image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 5694–5703, June 2026. 2, 5, 10

[34] Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025. 2, 5

[35] Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. Megapairs: Massive data synthesis for universal multimodal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19076–19095, 2025. 10

[36] Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv*, 2026. 2, 5, 6

[37] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal em-

bedding tasks. *arXiv preprint arXiv:2410.05160*, 2024. 2, 5

[38] Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, et al. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*, 2025. 2

[39] Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025.

[40] Xiaojie Li, Chu Li, Shi-Zhe Chen, and Xi Chen. Umarvel: Unveiling key factors for universal multimodal retrieval via embedding learning with mllms. In *International Conference on Learning Representations*, volume 2026, pages 121298–121319, 2026. 2

[41] Lihao Liu, Biao Yang, Yan Wang, Da Li, Jiangxia Cao, Yuxiao Luo, Xiang Chen, Xiangyu Wu, Wei Yuan, Fan Yang, Guiguang Ding, Tingting Gao, and Guorui Zhou. Crem: Compression-driven representation enhancement for multimodal retrieval and comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 5714–5724, June 2026. 2

[42] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, 2020. 6, 9

[43] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, volume 2025, pages 61424–61449, 2025. 2, 6

[44] Xiangzhao Hao, Shijie Wang, Tianyu Yang, Tianyue Wang, Haiyun Guo, and Jinqiao Wang. Trace: Task-adaptive reasoning and representation learning for universal multimodal retrieval. *arXiv preprint arXiv:2603.02929*, 2026. 2

[45] Chenwei He, Xiangzhao Hao, Tianyu Yang, Yuxiang Ma, Yuheng Jia, Lingxiang Wu, Chaoyang Zhao, Haiyun Guo, and Jinqiao Wang. Plume: Latent reasoning based universal multimodal embedding. *arXiv preprint arXiv:2604.02073*, 2026. 2

[46] Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. Ume-r1: Exploring reasoning-driven generative multimodal embeddings. In *International Conference on Learning Representations*, volume 2026, pages 115779–115806, 2026. 2

[47] Zilin Xiao, Qi Ma, Mengting Gu, Chun-cheng Chen, Xintao Chen, Vicente Ordonez, and Vijai Mohan. Metaembed: Scaling multimodal retrieval at test-time with flexible late interaction. In *International Conference on Learning Representations*, volume 2026, pages 23619–23638, 2026. 2, 7, 9

[48] Siyu Sun, Jing Ren, Zhaohe Liao, Dongxiao Mao, Xiangyuan Ren, Yiyi Zhang, Haohua Zhao, Weixiong Lin, Jiang Shaohua, Liqing Zhang, et al. Bottleneck tokens for unified multimodal retrieval. *arXiv preprint arXiv:2604.11095*, 2026. 2

[49] Hao Li, Jingkuan Song, Lianli Gao, Xiaosu Zhu, and Hengtao Shen. Prototype-based aleatoric uncertainty quantification for cross-modal retrieval. *Advances in Neural Information Processing Systems*, 36, 2023. 2

[50] WonJun Moon, Cheol-Ho Cho, Woojin Jun, Taeoh Kim, Inwoong Lee, Dongyoon Wee, Minho Shim, and Jae-Pil Heo. Prototypes are balanced units for efficient and effective partially relevant video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21789–21799, October 2025.

[51] Chengzhi Lin, Ancong Wu, Junwei Liang, Jun Zhang, Wenhang Ge, Wei-Shi Zheng, and Chunhua Shen. Text-adaptive multiple visual prototype matching for video-text retrieval. *Advances in neural information processing systems*, 35:38655–38666, 2022.

[52] Pandeng Li, Chen-Wei Xie, Liming Zhao, Hongtao Xie, Jiannan Ge, Yun Zheng, Deli Zhao, and Yongdong Zhang. Progressive spatio-temporal prototype matching for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4100–4110, 2023. 2

[53] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 2

[54] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2022. 2

[55] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[56] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020. 2

[57] Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. *arXiv preprint arXiv:2209.14860*, 2022.

[58] Hyun Seok Seong, WonJun Moon, and Jae-Pil Heo. From vicious to virtuous cycles: Synergistic representation learning for unsupervised video object-centric learning. In *The Fourteenth International Conference on Learning Representations*, 2026. 2

[59] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, volume 2024, pages 35549–35562, 2024. 2

[60] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3, 4

[61] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 3, 9, 12

[62] WonJun Moon, MinSeok Jung, Gilhan Park, Tae-Young Kim, Cheol-Ho Cho, Woojin Jun, and Jae-Pil Heo. Mitigating semantic collapse in partially relevant video retrieval. *Advances in Neural Information Processing Systems*, 38, 2025. 4

[63] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024. 4, 5, 12

[64] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gul Varol. Covr-2: Automatic data construction for composed video retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:11409–11421, 2024. 6

[65] Zixu Li, Zhiwei Chen, Haokun Wen, Zhiheng Fu, Yupeng Hu, and Weili Guan. Encoder: Entity mining and modification relation binding for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 6

[66] Zhipeng Qian, Zihan Liang, Yufei Ma, Ben Chen, Huangyu Dai, Yiwei Ma, Jiayi Ji, Chenyi Lei, Han Li, and Xiaoshuai Sun. CSMCIR: CoT-enhanced symmetric alignment with memory bank for composed image retrieval. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Findings of the Association for Computational Linguistics: ACL 2026*, pages 4411–4424, San Diego, California, United States, July 2026. Association for Computational Linguistics. 5, 6

[67] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5, 6, 12

[68] Yunpeng Han, Lisai Zhang, Qingcai Chen, Zhijian Chen, Zhonghua Li, Jianxin Yang, and Zhao Cao. Fashionsap: Symbols and attributes prompt for fine-grained fashion vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15028–15038, 2023. 6

[69] Xiao Han, Xiatian Zhu, Licheng Yu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2669–2680, 2023. 6

[70] Chull Hwan Song, Taebaek Hwang, Jooyoung Yoon, Shunghyun Choi, and Yeong Hyeon Gu. Syncmask: Synchronized attentional masking for fashion-centric vision-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13948–13957, 2024. 6

[71] Yifan Wang, Wuliang Huang, Lei Li, and Chun Yuan. Semantic distillation from neighborhood for composed image retrieval. In *ACM Multimedia 2024*, 2024. 6

[72] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *CVPR*, 2021. 5

[73] Yikun Liu, Jiangchao Yao, Weidi Xie, and Yanfeng Wang. A sanity check on composed image retrieval, 2026. 5, 6

[74] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multi-

modal llms. In *International Conference on Learning Representations*, volume 2025, pages 44215–44234, 2025. 5

[75] Tingyu Song, Mingxin Li, Yanzhao Zhang, Dingkun Long, Chu Liu, Pengjun Xie, Yilun Zhao, and Shu Wu. Core: Improving compositional reasoning in mllm embedding via reranker distillation. *arXiv preprint arXiv:2609.04083*, 2026.

[76] Qianying Liu, Xiao Liang, Zhiqiang Zhang, Yibo Chen, Xu Tang, Zhongfei Qing, Fengfan Zhou, Yao Hu, and Paul Henderson. Rematch: Boosting representation through matching for multimodal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16823–16833, 2026. 5, 6

[77] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25004–25014, 2025. 7

[78] Heeji Yoon, Jaewoo Jung, Junwan Kim, Hyungyu Choi, Heeseong Shin, Sangbeom Lim, Honggyu An, Chaehyun Kim, Jisang Han, Donghyun Kim, Chanho Eom, Sunghwan Hong, and Seungryong Kim. Visual representation alignment for multimodal large language models, 2025. 7

[79] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022. 8

[80] OpenAI. Gpt-5.6: Frontier intelligence that scales with your ambition, July 2026. 11