DISTRIBUTED LOCAL EVIDENCE ALLOCATION UNDER STRONG GLOBAL ANCHORS FOR COMPOSED IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

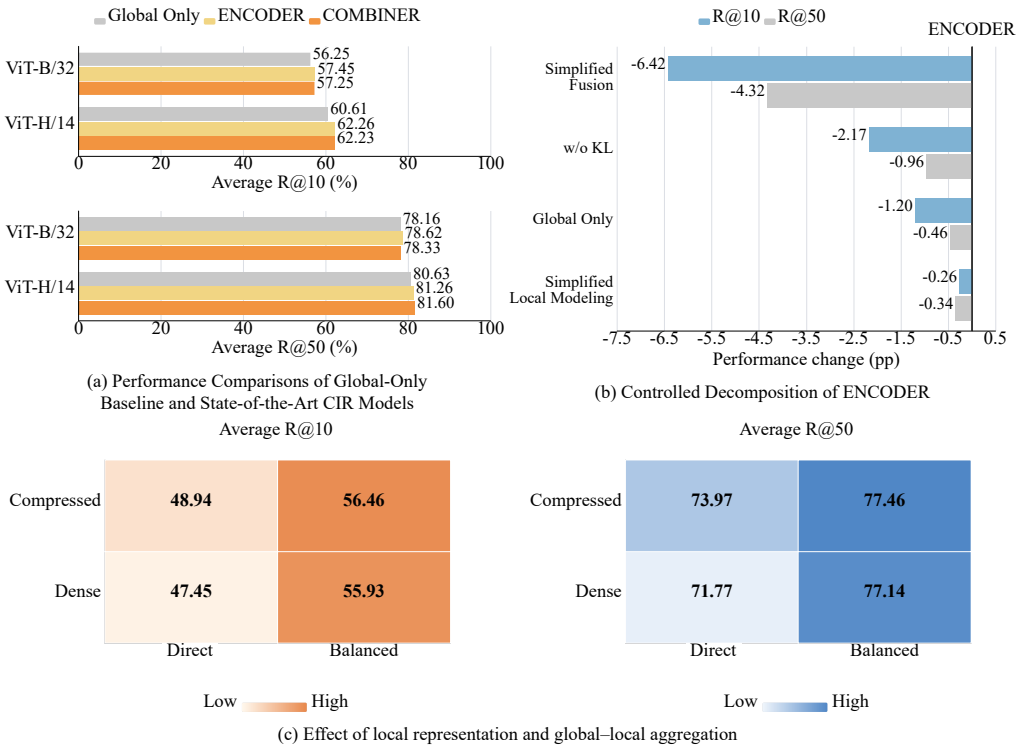
Composed Image Retrieval (CIR) retrieves a target image from a reference image and a modification text. Recent supervised CIR methods complement strong vision–language pretrained (VLP) representations with specialized local modeling, but it remains unclear whether the benefit of local modeling under strong global representations is tied to such specialized local modeling pipelines. Through controlled studies, we find that global VLP features provide a strong retrieval anchor and that, in a controlled ENCODER study, replacing its specialized local branch with lightweight token compression causes only a small degradation. On FashionIQ, dense and compressed local representations also perform similarly once global–local branch weighting is controlled. These observations motivate a narrower design question: how should a compact local branch allocate its limited capacity across reference evidence for each modification? We propose DART, a Distributed Information Bottleneck (DIB)-inspired module that conditions token-wise allocation on modification semantics and regularizes the allocation coefficients around a compression prior. DART is combined with balanced global–local aggregation while preserving the host VLP backbone and multimodal composition mechanism. Experiments on FashionIQ and CIRR support distributed local evidence allocation as a compact local modeling strategy under strong global representations.

1 INTRODUCTION

Composed Image Retrieval (CIR) (Vo et al., 2019) retrieves a target image from a multimodal query consisting of a reference image and a modification text (Song et al., 2025). Recent supervised methods increasingly build on strong vision–language pretrained (VLP) representations such as CLIP (Radford et al., 2021) and BLIP (Li et al., 2022), while introducing specialized local modeling mechanisms, including latent-factor filtering and entity–action binding, focus-based region modeling, and attribute/prototype modeling (Li et al., 2025; Chen et al., 2025; Li et al., 2026). These approaches exploit local evidence in different ways, but all introduce specialized processing into the local branch. Although these methods differ in implementation, they share an emphasis on exploiting local evidence beyond global VLP features. This motivates us to examine whether the benefit of local modeling under strong global representations is tied to specialized local representation construction, or can be retained with a compact local representation.

We study this question through the controlled analyses summarized in Figure 1. A global-only baseline remains close to ENCODER and COMBINER across two VLP scales, showing that the global representation already provides a strong retrieval anchor. Within ENCODER, replacing its specialized local pipeline with lightweight token compression causes only a small degradation, whereas changing the fusion mechanism or training objective produces larger changes. We therefore keep these host-specific components fixed rather than using the comparison to rank their importance. Finally, the comparison between dense and compressed local representations is sensitive to how the global and local branches are weighted; under balanced aggregation, their performance remains close. These results establish a compact local branch under controlled branch weighting as a viable starting point, but leave a distinct challenge: how should its limited capacity be distributed across reference tokens? This motivates our design hypothesis: **under a strong global anchor, the limited**

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Performance Comparisons of Global-Only Baseline and State-of-the-Art CIR Models

(b) Controlled Decomposition of ENCODER

(c) Effect of local representation and global–local aggregation

Figure 1: Controlled observations motivating local evidence allocation in supervised CIR. (a) Comparison of a global-only baseline with ENCODER and COMBINER under ViT-B/32 and ViT-H/14. (b) Controlled decomposition of ENCODER, with performance changes measured relative to the original model. (c) Controlled comparison of dense and compressed local representations under direct and balanced global–local aggregation.

capacity of a compact local branch should be distributed across reference tokens according to the modification.

Based on this hypothesis, we propose **DART** (**D**IB-inspired **A**llocation of **R**eference **T**okens), a lightweight module that uses modification semantics to adapt the reference-token weights used for compact-token construction. The modification and target branches use the same lightweight local compressor without conditional allocation, and the resulting compact representations are processed by the original framework-specific multimodal composition mechanism. We further use balanced global–local aggregation to prevent branch weighting from being determined by token count. In this way, DART changes how local information is allocated and integrated while leaving the host VLP backbone and multimodal composition mechanism unchanged.

Our main contributions are summarized as follows:

- We show through controlled analyses that compact local modeling can retain most of the retrieval performance under a strong global anchor, motivating a shift toward how limited local capacity should be allocated.
- We propose **DART**, a lightweight local evidence allocation module that adapts reference-token weights within a compact K-token representation according to modification semantics, together with DIB-inspired allocation regularization and balanced global–local integration.
- We evaluate DART across three representative supervised CIR frameworks on FashionIQ and CIRR, demonstrating comparable or improved retrieval performance with a unified lightweight local modeling design.

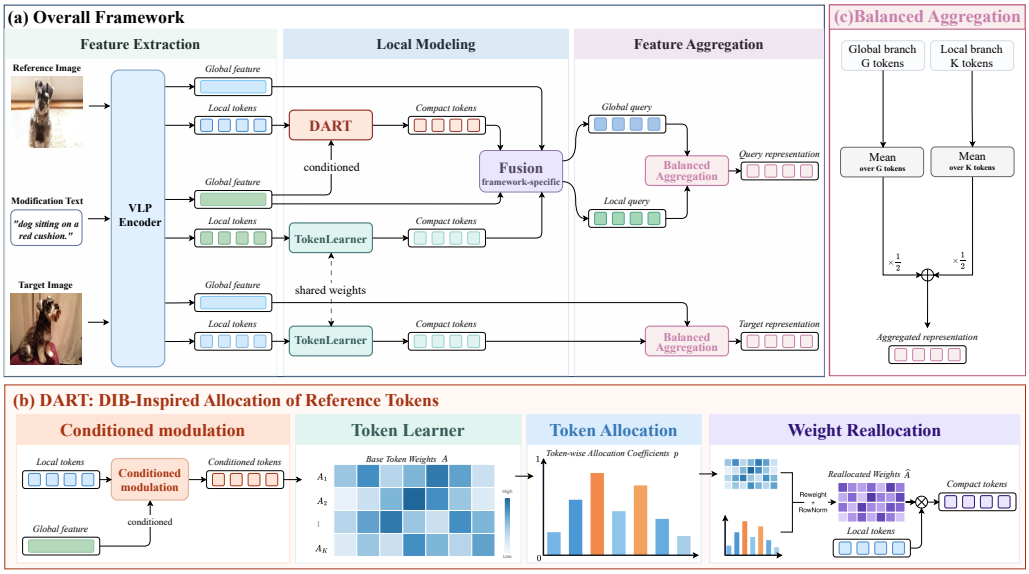


Figure 2: DART performs modification-conditioned local evidence allocation on the reference branch, while the modification and target branches use the shared lightweight compressor. The resulting compact local representations are then processed by the framework-specific multimodal composition module and integrated with the global representation through balanced global–local aggregation.

2 Related Work

2.1 Composed Image Retrieval

Composed Image Retrieval (CIR) retrieves a target image from a reference image and a modification text. Early methods mainly focused on learning image–text composition functions, such as TIRG (Vo et al., 2019), ComposeAE (Anwaar et al., 2021), and ARTEMIS (Delmas et al., 2022). With the emergence of vision-language pre-training (VLP), recent approaches increasingly build upon strong pretrained representations. For example, CLIP4CIR (Baldrati et al., 2023) fine-tunes CLIP for CIR and learns a multimodal Combiner, while DQU-CIR (Wen et al., 2024) unifies the multimodal query at the raw-data level.

More recent supervised CIR methods further exploit local representations to complement global VLP features. ENCODER (Li et al., 2025) filters modification-relevant visual and textual tokens and models entity–action relations before composition. OFFSET (Chen et al., 2025) performs region-level visual modeling and text-guided focus revision, while COMBINER (Li et al., 2026) compresses global and local features into multi-grained attribute prototypes for multimodal composition. These approaches demonstrate the utility of exploiting local evidence beyond global VLP embeddings.

In contrast, our work studies local modeling from the perspective of selective information allocation under a strong global semantic anchor. Rather than focusing on specialized local processing, we focus on how compact local capacity is allocated across reference evidence according to modification semantics.

2.2 Local Representation Learning and Information Bottleneck

Local representations have been widely used to preserve fine-grained visual or textual cues that may be overlooked by global embeddings. Representative approaches learn to select or aggregate informative local tokens, such as TokenLearner, which adaptively summarizes spatial features into a compact set of tokens (Ryoo et al., 2021).

In parallel, the information bottleneck (IB) principle seeks compressed representations that retain task-relevant information while suppressing nuisance variation (Tishby et al., 2000). Conventional IB formulations typically regularize a representation as a whole, leaving how information should be distributed across its constituent local elements unspecified. Distributed formulations of IB instead impose information constraints across individual components, enabling component-wise characterization and allocation of information (Murphy & Bassett, 2024). Inspired by DIB, we introduce modification-conditioned token allocation for CIR, where token-wise allocation coefficients reweight the base token weights for local compression.

3 Empirical Motivation for Local Evidence Allocation

We conduct controlled studies to identify a compact local modeling setting and use the resulting observations to motivate our design.

Experimental protocol. Unless otherwise specified, all analyses are conducted on FashionIQ using ViT-B/32 with batch size 128. We additionally use ViT-H/14 in Section 3.1 to examine whether the observation persists with a stronger backbone. Each comparison changes only the component under study while keeping the remaining training and architectural settings fixed.

3.1 Global Representations as a Strong Retrieval Foundation

We first examine how much retrieval capability is already provided by global VLP representations. The global-only baseline removes explicit local modeling while retaining the global image and text features extracted by the VLP backbone and the remaining retrieval pipeline. As shown in Figure 1(a), under ViT-B/32 it trails ENCODER and COMBINER by 1.00–1.20 percentage points in average R@10 and 0.17–0.46 percentage points in average R@50. Under ViT-H/14, the corresponding gaps are 1.62–1.65 and 0.63–0.97 points. This suggests that global VLP features already provide a strong retrieval foundation, while local information remains beneficial.

3.2 Local Modeling under Fixed Host Components

To examine how retrieval performance changes when specialized local modeling is simplified, we use ENCODER as a controlled case: **Global Only** removes its local branch; **Simplified Local Modeling** replaces the local pipeline with TokenLearner (Ryoo et al., 2021) while matching the original output token count; **Simplified Fusion** replaces the original feature-wise affine fusion with simple concatenation followed by a linear projection; and **w/o KL** removes the KL-based regularizer.

Figure 1(b) shows that removing the local branch reduces average R@10/R@50 by 1.20/0.46 percentage points, whereas replacing the specialized local pipeline with TokenLearner reduces them by only 0.26/0.34 points. In comparison, simplifying the fusion mechanism and removing the KL regularizer lead to larger drops of 6.42/4.32 and 2.17/0.96 points, respectively. Because the fusion mechanism and training objective are sensitive to intervention, we keep them fixed and focus subsequent analyses on the local branch and its aggregation with the global representation.

3.3 Local Representation and Global–Local Aggregation

To compare dense and compressed local representations without confounding their contribution with token-count-dependent branch weighting, we evaluate both under two aggregation strategies in a controlled 2×2 study.

Dense representations directly use the local tokens produced by the VLP encoder, whereas **compressed** representations summarize them into K learned tokens ($K \ll N$) using TokenLearner (Ryoo et al., 2021). **Direct aggregation** jointly averages the global token and all local tokens. **Balanced aggregation** instead first averages the local tokens and then assigns equal weight to the global and mean-local representations.

As shown in Figure 1(c), switching from direct to balanced aggregation improves average R@10 by 7.52–8.48 points and average R@50 by 3.49–5.37 points for both representation forms. This performance gap may partly arise because directly averaging one global token with many local

tokens may allow the local branch to dominate the aggregated representation, weakening the contribution of the global feature. Balanced aggregation mitigates this imbalance by assigning equal weight to the global and mean-local representations, and we therefore adopt it in our method. Under this balanced setting, dense and compressed representations differ by only 0.53/0.32 points in average R@10/R@50, with the compressed variant slightly higher in both metrics.

These results establish compact local compression under branch-balanced aggregation as a controlled starting point. However, compression alone does not determine the reference-token weights used to construct the K compact tokens for a given modification. This motivates the token-wise allocation introduced in DART.

4 METHOD

Section 3 establishes compact local compression with balanced global–local aggregation as a viable starting point, while leaving open how its limited capacity should be distributed across reference tokens. DART addresses this challenge by using modification semantics to reallocate the reference-token weights used for compact-token construction. As shown in Figure 2, the reference, modification, and target branches share the same lightweight local compressor, while DART is applied only to the reference branch. The resulting compact representations are processed by the original framework-specific multimodal composition mechanism and combined with the global representation through balanced aggregation.

4.1 OVERVIEW

Given a reference image I_r, a modification text T_m, and a target image I_t, a pretrained vision–language backbone extracts global and local representations:

$$(\mathbf{g}_r, \mathbf{X}_r) = E_v(I_r), \qquad (\mathbf{g}_m, \mathbf{X}_m) = E_t(T_m), \qquad (\mathbf{g}_t, \mathbf{X}_t) = E_v(I_t), \tag{1}$$

where $\mathbf{g}_{(\cdot)} \in \mathbb{R}^d$ denotes a global representation and $\mathbf{X}_{(\cdot)} \in \mathbb{R}^{N_{(\cdot)} \times d}$ denotes the corresponding local tokens.

We use a shared TokenLearner-style compressor to summarize each local sequence into K compact tokens, where the compact-token count K defines the local capacity considered in this work.

For the modification and target branches, the learned base token weights are used directly to construct the compact tokens. For the reference branch, DART uses the modification semantics to reallocate the base token weights used for compact-token construction. In this way, the limited capacity of the K compact tokens is distributed across reference evidence according to the current modification, while preserving the VLP architecture and the host framework's multimodal composition mechanism.

4.2 MODIFICATION-CONDITIONED LOCAL EVIDENCE ALLOCATION

Modification-conditioned reference features. Let $\mathbf{X}_r = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$. We first normalize the reference tokens and modification feature as $\mathbf{x}_i^n = \mathrm{LN}_x(\mathbf{x}_i)$ and $\mathbf{g}_m^n = \mathrm{LN}_m(\mathbf{g}_m)$, where LN_x and LN_m are separate learnable LayerNorms.

A lightweight MLP predicts token-specific modulation parameters:

$$\begin{aligned} [\boldsymbol{\gamma}_i, \boldsymbol{\delta}_i] &= f_\phi([\mathbf{x}_i^n; \mathbf{g}_m^n]), \\ \tilde{\mathbf{x}}_i &= (1 + \boldsymbol{\gamma}_i) \odot \mathbf{x}_i^n + \boldsymbol{\delta}_i \odot \mathbf{g}_m^n. \end{aligned} \tag{2}$$

The modulation layer is zero-initialized, such that $\tilde{\mathbf{x}}_i = \mathbf{x}_i^n$ at initialization and modification conditioning is introduced progressively during training.

A shared TokenLearner-style scorer F_θ predicts K scores for each local token, corresponding to the K compact output slots. Softmax over the N input tokens converts these scores into K normalized base token weights:

$$\mathbf{A} = \mathrm{Softmax}_N\left(F_\theta(\tilde{\mathbf{X}}_r)\right) \in \mathbb{R}^{K \times N}. \tag{3}$$

Each row of $\mathbf{A}$ defines the normalized base weights over the N local tokens to one compact token. F_θ is shared across the reference, modification, and target branches.

Token-wise allocation. Given the compression from N reference tokens to K compact tokens, we define

$$\rho = \frac{K}{N} \tag{4}$$

as a compression prior.

For each conditioned reference token $\tilde{\mathbf{x}}_i$, we compute its compatibility with the modification feature:

$$s_i = \cos(\tilde{\mathbf{x}}_i, \mathbf{g}_m), \qquad \bar{s} = \frac{1}{N}\sum_{j=1}^{N} s_j. \tag{5}$$

The centered compatibility score then adjusts the compression prior to obtain a token-wise allocation coefficient:

$$p_i = \sigma(\mathrm{logit}(\rho) + \alpha(s_i - \bar{s})), \qquad p_i \in (0, 1), \tag{6}$$

where α is a learnable scaling factor initialized to 2.0. When $s_i = \bar{s}$, $p_i = \rho$; the centered compatibility score introduces modification-dependent variation across reference tokens.

Modification-conditioned weight reallocation. The token-wise allocation coefficients reweight the base token weights A before compact-token construction. For the k-th compact output, we define

$$\hat{a}_{ki} = \frac{a_{ki} p_i}{\sum_{j=1}^{N} a_{kj} p_j}, \qquad \mathbf{z}_{r,k} = \sum_{i=1}^{N} \hat{a}_{ki} \mathbf{x}_i. \tag{7}$$

Thus, DART reallocates the base token weights across reference tokens according to the modification-conditioned allocation coefficients, rather than directly rescaling their feature magnitudes. In this way, compact-token weighting combines the slot-specific base weights $\mathbf{A}$ with separate modification-conditioned token-wise allocation coefficients p_i, yielding the reallocated weights $\hat{\mathbf{A}}$. The resulting reallocated weights are used to construct the compact representation from the original VLP tokens.

DIB-inspired allocation regularization. Inspired by DIB, which imposes information constraints on individual components (Murphy & Bassett, 2024), we regularize each token-wise allocation coefficient toward the compression prior ρ:

$$\mathcal{L}_{\mathrm{alloc}} = \frac{1}{K}\sum_{i=1}^{N} D_{\mathrm{KL}}\left[\mathrm{Ber}(p_i) \,\|\, \mathrm{Ber}(\rho)\right]. \tag{8}$$

This regularizer penalizes excessive deviation from the compression prior while allowing the allocation coefficients to vary across reference tokens according to the current modification.

4.3 Balanced Global–Local Integration and Optimization

Following the controlled analysis in Section 3.3, we use balanced aggregation for global–local integration. We first average each branch to obtain $\bar{\mathbf{h}}_g$ and $\bar{\mathbf{h}}_l$, and then combine them with equal weight:

$$\mathbf{h} = \frac{1}{2}\left(\bar{\mathbf{h}}_g + \bar{\mathbf{h}}_l\right). \tag{9}$$

The same rule is applied to the composed query and target representations.

We retain the original retrieval and auxiliary losses unrelated to the replaced local modeling module of each host framework and augment them with the allocation regularizer:

$$\mathcal{L} = \mathcal{L}_{\mathrm{host}} + \beta \mathcal{L}_{\mathrm{alloc}}, \tag{10}$$

where $\mathcal{L}_{\mathrm{host}}$ denotes the original host objective and β controls the strength of the allocation regularization.

Table 1: Main comparison across representative CIR frameworks on FashionIQ and CIRR. All results are reported as mean ± standard deviation over three independent runs. FashionIQ reports category-averaged R@10 and R@50. CIRR Average is computed as $(\mathrm{R@5} + \mathrm{R}_{\mathrm{subset}}\mathrm{@1})/2$. Ours uses the shared lightweight compressor, applies DART to the reference branch, and adopts balanced global–local aggregation while preserving the framework-specific multimodal composition mechanism. † denotes our reproduction. Full results are provided in Appendix A.1.

Framework	Variant	FashionIQ		CIRR
		Average R@10	Average R@50	Average
COMBINER (Li et al., 2026)	Original†	62.23 ± 0.07	81.60 ± 0.17	74.40 ± 0.09
	Ours	$\mathbf{63.58 \pm 0.07}$	$\mathbf{82.31 \pm 0.18}$	$\mathbf{74.60 \pm 0.09}$
ENCODER (Li et al., 2025)	Original†	62.26 ± 0.59	81.26 ± 0.48	74.29 ± 0.10
	Ours	$\mathbf{63.52 \pm 0.07}$	$\mathbf{81.77 \pm 0.10}$	$\mathbf{74.94 \pm 0.28}$
OFFSET (Chen et al., 2025)	Original†	62.18 ± 0.06	81.24 ± 0.14	73.19 ± 0.18
	Ours	$\mathbf{63.68 \pm 0.07}$	$\mathbf{82.28 \pm 0.05}$	$\mathbf{75.21 \pm 0.03}$

5 EXPERIMENT

Implementation Details. We integrate DART into three representative supervised CIR frameworks. To facilitate comparison, we use CLIP ViT-H/14, a batch size of 32, and 10 training epochs for all frameworks. All other original host objectives unrelated to the replaced local modeling module, optimization settings, and framework-specific hyperparameters, including the compact-token count K, follow the corresponding official implementations and are kept fixed between the original and DART variants. Accordingly, we use $K = 4$, 4, and 4 for COMBINER, ENCODER, and OFFSET, respectively. DART augments the host objective with the allocation regularizer using $\beta = 0.5$, selected according to the sensitivity study in Appendix A.2. All experiments are conducted on a single NVIDIA A100 80GB PCIe GPU.

Datasets and Evaluation. Following previous supervised CIR works, we evaluate DART on two widely used benchmarks, FashionIQ (Wu et al., 2021) and CIRR (Liu et al., 2021), using their standard dataset splits and evaluation protocols. FashionIQ is a fashion-domain dataset, whereas CIRR is an open-domain benchmark. For FashionIQ, we report R@10 and R@50 for each category, together with their category-wise averages. For CIRR, we report R@1, R@5, R@10, and R@50 on the full retrieval set, as well as $R_{\mathrm{subset}}@K$ ($K = 1, 2, 3$) under the subset setting.

5.1 MAIN RESULTS

As shown in Table 1, Ours consistently improves the three evaluated host frameworks on FashionIQ. Compared with the reproduced baselines, Ours improves average R@10/R@50 by 1.35/0.71 points for COMBINER, 1.26/0.51 points for ENCODER, and 1.50/1.04 points for OFFSET. The same improvement trend is observed on CIRR, although the magnitude varies across host frameworks. The summary average increases by 0.20 points for COMBINER, 0.65 points for ENCODER, and 2.02 points for OFFSET. The complete results in Appendix A.1 further show that Ours improves all reported CIRR recall metrics for all three frameworks, while the FashionIQ gains are observed across nearly all category-level metrics.

These results support distributed local evidence allocation together with balanced global-local aggregation as an effective alternative to framework-specific local modeling in the three evaluated CIR frameworks, without changing their multimodal composition mechanisms.

5.2 ABLATION STUDIES

Component analysis. As shown in Table 2, replacing ENCODER's specialized local modeling with lightweight TokenLearner (Ryoo et al., 2021) alone reduces performance, while balanced global–local aggregation largely recovers the loss. Modification conditioning brings a small R@10 gain, and token-wise allocation further improves R@10 from 62.53 to 63.35. The DIB-inspired

Table 2: Progressive component analysis of our method on FashionIQ within the ENCODER framework. TL denotes lightweight local modeling with TokenLearner; BA denotes balanced global–local aggregation; CM denotes modification-conditioned token modulation; Alloc. denotes token-wise allocation; and Reg. denotes the DIB-inspired allocation regularizer. Each row cumulatively adds one component to the preceding variant.

Variant	TL	BA	CM	Alloc.	Reg.	Average R@10	Average R@50
ENCODER						62.26	81.26
+TL	✓					61.43	80.67
+BA	✓	✓				62.30	81.73
+CM	✓	✓	✓			62.53	81.66
+Alloc.	✓	✓	✓	✓		63.35	81.38
+Reg.	✓	✓	✓	✓	✓	**63.52**	**81.77**

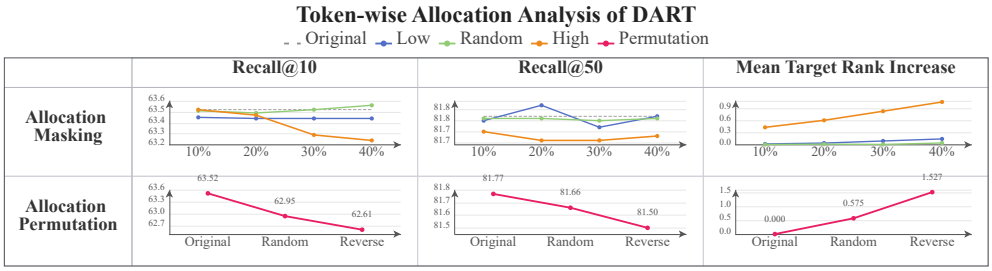


Figure 3: **Token-wise allocation analysis of DART on FashionIQ.** Top: masking high-allocation reference tokens causes larger target-rank degradation than masking low-allocation or random tokens. Bottom: preserving the allocation values while permuting or reversing their token positions progressively degrades retrieval, indicating that DART learns meaningful token-wise allocation.

allocation regularizer then improves R@50 while retaining the R@10 gain, yielding the final 63.52/81.77. These results show that token-wise allocation provides the main R@10 gain, while the allocation regularizer improves R@50 while retaining the R@10 improvement.

5.3 Further Analysis

Token-wise allocation analysis. Figure 3 examines whether the learned allocation is tied to specific reference tokens. Mean target rank increase measures the change in the ground-truth target rank after perturbation, where a larger value indicates stronger degradation. Under allocation masking, removing high-allocation tokens causes substantially larger rank degradation than removing low-allocation or random tokens. To further test whether retrieval depends on the correspondence between allocation values and token positions, we permute the learned allocation values while preserving their distribution. Random permutation degrades retrieval, while reversing the allocation causes a larger drop. This shows that performance depends not only on the allocation values, but also on their correspondence to reference-token positions.

Efficiency and complexity. Table 3 compares the local-modeling cost and retrieval performance of DART with the original frameworks. For ENCODER and OFFSET, the DART local branch reduces local-specific parameters by 90.5–97.2%, query-side local GFLOPs by 69.1–87.2%, and latency by 72.2–73.2%, while improving Average R@10/R@50 by 1.26/0.51 and 1.50/1.04 points, respectively. For COMBINER, whose original local summarizer is already lightweight, DART introduces additional local cost while improving Average R@10/R@50 by 1.35/0.71 points. Thus, DART substantially simplifies the heavier local pipelines of ENCODER and OFFSET, while improving retrieval performance in all three evaluated frameworks.

Table 3: Efficiency comparison of the original local modeling and DART on FashionIQ. Params denotes trainable parameters specific to local modeling, excluding shared projection layers. GFLOPs and latency are measured on the actual post-VLP query-side local modeling path, including the required feature projections but excluding VLP encoding, global–local aggregation, multimodal composition, and retrieval scoring. Latency is measured with batch size 1. Retrieval results are reported as mean ± standard deviation over three runs. Bold indicates the better result within each framework.

Framework	Variant	Params (M)↓	GFLOPs↓	Latency (ms)↓	Average R@10↑	Average R@50↑
COMBINER	Original	**0.526**	**1.190**	**1.03**	62.23 ± 0.07	81.60 ± 0.17
	DART	0.793	1.320	1.86	$\mathbf{63.58 \pm 0.07}$	$\mathbf{82.31 \pm 0.18}$
ENCODER	Original	8.315	4.274	7.42	62.26 ± 0.59	81.26 ± 0.48
	DART	**0.793**	**1.323**	**1.99**	$\mathbf{63.52 \pm 0.07}$	$\mathbf{81.77 \pm 0.10}$
OFFSET	Original	27.939	10.275	4.38	62.18 ± 0.06	81.24 ± 0.14
	DART	**0.793**	**1.320**	**1.22**	$\mathbf{63.68 \pm 0.07}$	$\mathbf{82.28 \pm 0.05}$

6 Conclusion

In this work, we study local modeling in supervised composed image retrieval under strong global VLP representations. Controlled analyses show that compact local compression with balanced global–local aggregation provides a viable starting point, leaving the question of how limited local capacity should be distributed across reference tokens. We introduce DART, a DIB-inspired local evidence allocation module that reallocates reference-token weights according to modification semantics within a limited K-token local representation, together with DIB-inspired regularization around a compression prior. Across three evaluated CIR frameworks on FashionIQ and CIRR, our method improves the reported summary retrieval metrics; for ENCODER and OFFSET, it also substantially reduces local-modeling cost. These results support distributed local evidence allocation as an effective compact local modeling strategy under strong global representations.

AI use statement

In this work, we used generative AI tools, including ChatGPT and Gemini, to assist with research ideation and execution, including refining methodological hypotheses, discussing model and loss-function design, suggesting experimental ablations, assisting with code implementation and debugging, and interpreting experimental results. Generative AI was also used to assist with drafting and revising portions of the manuscript, including improving clarity, organization, and academic presentation.

We did not use generative AI to generate synthetic datasets or to prove mathematical claims. All AI-assisted methodological suggestions, code changes, experimental interpretations, and manuscript text were reviewed by the authors before inclusion. Suggested code was manually inspected and empirically validated through implementation checks and controlled experiments. The authors take full responsibility for the final methodology, experiments, results, claims, and text presented in this paper.

Reproducibility Statement

We provide the formulation and implementation details of DART in Section 4, including the modification-conditioned allocation mechanism, the DIB-inspired regularization objective, and the balanced global–local aggregation strategy. Section 5 specifies the datasets, evaluation protocols, backbone, training configuration, and controlled integration settings used for all compared frameworks. To account for training variability, the main results are reported as the mean and standard deviation over three independent runs. The appendix provides complete category-level FashionIQ results, full CIRR evaluation metrics, and the sensitivity analysis of the DIB-inspired regularization weight. For the efficiency study, we explicitly define the measured local-modeling scope and report local-specific parameter counts, GFLOPs, and inference latency under a common hardware and evaluation setting.

REFERENCES

Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. Compositional learning of image-text query for image retrieval. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1139–1148, 2021. doi: 10.1109/WACV48630.2021.00118.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Composed image retrieval using contrastive learning and task-oriented clip-based features. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(3), October 2023. ISSN 1551-6857. doi: 10.1145/3617597. URL https://doi.org/10.1145/3617597.

Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, Xuemeng Song, and Liqiang Nie. Offset: Segmentation-based focus shift revision for composed image retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 6113–6122, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755366. URL https://doi.org/10.1145/3746027.3755366.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL https://arxiv.org/abs/2201.12086.

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 Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 5101–5109. AAAI Press, 2025. doi: 10.1609/AAAI.V39I5.32541. URL https://doi.org/10.1609/aaai.v39i5.32541.

Zixu Li, Yupeng Hu, Zhiwei Chen, Haokun Wen, Xuemeng Song, and Liqiang Nie. Combiner: Composed image retrieval guided by attribute-based neighbor relations. *IEEE Transactions on Image Processing*, 35:5377–5392, 2026. doi: 10.1109/TIP.2026.3694139.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2105–2114, 2021. doi: 10.1109/ICCV48922.2021.00213.

Kieran A. Murphy and Dani S. Bassett. Information decomposition in complex systems via machine learning. *Proceedings of the National Academy of Sciences*, 121(13):e2312988121, 2024. doi: 10.1073/pnas.2312988121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2312988121.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12786–12797. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/6a30e32e56fce5cf381895dfe6ca7b6f-Paper.pdf.

Xuemeng Song, Haoqiang Lin, Haokun Wen, Bohan Hou, Mingzhu Xu, and Liqiang Nie. A comprehensive survey on composed image retrieval, 2025. URL https://arxiv.org/abs/2502.18495.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, 2000. URL https://arxiv.org/abs/physics/0004057.

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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6432–6441, 2019. doi: 10.1109/CVPR.2019.00660.

Haokun Wen, Xuemeng Song, Xiaolin Chen, Yinwei Wei, Liqiang Nie, and Tat-Seng Chua. Simple but effective raw-data level multimodal fusion for composed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 229–239, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657727. URL https://doi.org/10.1145/3626772.3657727.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11302–11312, 2021. doi: 10.1109/CVPR46437.2021.01115.

A Appendix

A.1 Full Experimental Results

Table 4 provides the category-level breakdown of the FashionIQ results reported in the main paper. Ours improves all six category-level metrics for both COMBINER and ENCODER. For OFFSET, it improves five of the six metrics, including Dress R@10/R@50 by 1.46/0.95 points and Top&Tee R@10/R@50 by 1.66/2.17 points, while Shirt R@50 remains essentially unchanged (81.19 vs. 81.22). Overall, these gains are distributed across categories rather than being driven by a single subset.

Table 5 reports the complete CIRR results. Ours improves all full-set and subset recall metrics across the three host frameworks, with modest gains for COMBINER, larger gains for ENCODER, and the largest improvements for OFFSET. Together with the FashionIQ results, the improvements are consistently observed across the two evaluated benchmarks and three host frameworks.

Table 4: Full category-level results on FashionIQ. We report R@10 and R@50 for Dress, Shirt, and Top&Tee. † denotes our reproduction. Results from repeated runs are reported as mean±std. Bold numbers indicate the better mean within each framework.

Framework	Variant	Dress		Shirt		Top&Tee	
		R@10	R@50	R@10	R@50	R@10	R@50
COMBINER	Original†	55.72 ± 0.34	78.28 ± 0.05	63.62 ± 0.06	80.93 ± 0.33	67.33 ± 0.24	85.58 ± 0.21
	Ours	$\mathbf{58.44 \pm 0.40}$	$\mathbf{79.03 \pm 0.26}$	$\mathbf{64.31 \pm 0.03}$	$\mathbf{81.60 \pm 0.13}$	$\mathbf{67.99 \pm 0.28}$	$\mathbf{86.30 \pm 0.17}$
ENCODER	Original†	57.33 ± 0.80	78.20 ± 0.84	62.63 ± 0.42	80.91 ± 0.00	66.82 ± 0.56	84.67 ± 0.61
	Ours	$\mathbf{58.70 \pm 0.17}$	$\mathbf{78.90 \pm 0.10}$	$\mathbf{64.02 \pm 0.16}$	$\mathbf{81.21 \pm 0.17}$	$\mathbf{67.86 \pm 0.21}$	$\mathbf{85.21 \pm 0.10}$
OFFSET	Original†	57.44 ± 0.13	78.47 ± 0.33	62.37 ± 0.08	$\mathbf{81.22 \pm 0.20}$	66.74 ± 0.20	84.04 ± 0.15
	Ours	$\mathbf{58.90 \pm 0.10}$	$\mathbf{79.42 \pm 0.10}$	$\mathbf{63.74 \pm 0.20}$	81.19 ± 0.10	$\mathbf{68.40 \pm 0.26}$	$\mathbf{86.21 \pm 0.03}$

A.2 Sensitivity to the DIB-inspired allocation-regularization weight β

We study the sensitivity of DART to the coefficient β of the DIB-inspired allocation regularizer on FashionIQ within the ENCODER framework. As shown in Figure 4, introducing moderate allocation regularization improves the combined recall score. We therefore use $\beta = 0.5$, which performs best among the tested values, in all experiments.

Table 5: Full results on CIRR. We report R@K on the full retrieval set and $\mathrm{R}_{\mathrm{subset}}@K$ under the subset evaluation protocol. † denotes our reproduction. Results from repeated runs are reported as mean±std. Bold numbers indicate the better mean within each framework.

Framework	Variant	R@K				$\mathrm{R}_{\mathrm{subset}}@K$		
		1	5	10	50	1	2	3
COMBINER	Original†	46.89 ± 0.11	79.40 ± 0.13	88.19 ± 0.10	97.90 ± 0.04	69.40 ± 0.07	86.81 ± 0.04	93.83 ± 0.06
	Ours	$\mathbf{47.19 \pm 0.22}$	$\mathbf{79.52 \pm 0.04}$	$\mathbf{88.34 \pm 0.06}$	$\mathbf{97.94 \pm 0.09}$	$\mathbf{69.68 \pm 0.14}$	$\mathbf{86.91 \pm 0.07}$	$\mathbf{93.95 \pm 0.06}$
ENCODER	Original†	47.09 ± 0.42	79.09 ± 0.19	87.59 ± 0.14	97.88 ± 0.03	69.50 ± 0.36	86.38 ± 0.26	93.89 ± 0.08
	Ours	$\mathbf{47.60 \pm 0.14}$	$\mathbf{79.54 \pm 0.16}$	$\mathbf{88.28 \pm 0.14}$	$\mathbf{98.01 \pm 0.02}$	$\mathbf{70.33 \pm 0.40}$	$\mathbf{87.24 \pm 0.08}$	$\mathbf{94.28 \pm 0.02}$
OFFSET	Original†	44.55 ± 0.19	77.76 ± 0.42	86.81 ± 0.26	97.27 ± 0.05	68.63 ± 0.16	86.01 ± 0.19	93.53 ± 0.09
	Ours	$\mathbf{48.60 \pm 0.06}$	$\mathbf{80.55 \pm 0.05}$	$\mathbf{88.80 \pm 0.05}$	$\mathbf{98.05 \pm 0.06}$	$\mathbf{69.86 \pm 0.04}$	$\mathbf{87.41 \pm 0.09}$	$\mathbf{94.12 \pm 0.06}$

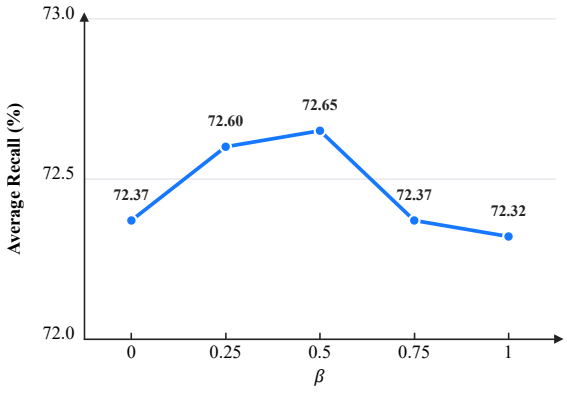


Figure 4: Sensitivity to the DIB-inspired allocation-regularization weight β on FashionIQ within ENCODER. The mean recall score is computed as $(\mathrm{Avg.\ R@10} + \mathrm{Avg.\ R@50})/2$. Among the tested values, $\beta = 0.5$ achieves the best performance.