FBCIR: DIAGNOSING AND MITIGATING PARTIAL-QUERY RELIANCE IN COMPOSED IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Composed image retrieval (CIR) is inherently conjunctive: a correct target must satisfy both a reference image and a textual modification. Yet a model can often retrieve the correct target without using the full composed query, because conventional easy negatives can be rejected from only one query component. We refer to this behavior as *focus imbalances*. We propose **FBCIR**, a multi-modal focus interpretation method that identifies visual and textual input components sufficient to preserve a model's retrieval decisions. Motivated by this diagnosis, we further develop **FBCIR-Data**, a CIR data augmentation workflow that enriches existing datasets with curated hard negatives requiring discrimination of both visual and textual constraints, together with a controlled benchmark and finetuning data for evaluating and mitigating focus imbalances. Experiments across representative CIR models reveal substantial vulnerabilities under such hard-negative settings. Under matched finetuning configurations, **FBCIR-Data** consistently improves hard-case retrieval for VLM-based retrievers and outperforms alternative training datasets, with additional gains observed on standard benchmarks. These results show that explicitly exposing models to negatives that invalidate single-component shortcuts provides an effective way to diagnose and improve robustness in composed image retrieval.

1 INTRODUCTION

Image retrieval (IR) has long been a fundamental problem in computer vision Isola et al. (2015) and vision-language research Liu et al. (2021). Traditional IR methods with uni-modal paradigms (i.e., queries are specified either by images or by text) are limited in addressing increasingly complex user requirements that involve multi-modal specifications. Composed image retrieval (CIR) Vo et al. (2019) addresses this limitation by enabling multi-modal queries (images & texts in most cases). Given an image and a related modifying instruction, CIR aims to retrieve images that are semantically consistent with both the visual and the textual content. Early CIR methods Baldrati et al. (2022); Saito et al. (2023); Li et al. (2022) extend CLIP-like models Radford et al. (2021) with multi-modal input capabilities, while more recent methods Bai et al. (2023) achieve improved performance by adopting vision-language models (VLMs).

A central challenge in CIR is the requirement of jointly reasoning over semantics from both modalities. In many common cases, there are large semantic gaps between targets and negative candidates. This setting allows models to rely on "shortcut" strategies, producing correct retrieval results without equally processing information from both modalities, but generally fail in more challenging scenarios where negatives closely resemble targets and are semantically consistent with one of the input modalities. An intuitive example is shown in Figure 1. In common-case #1, all negative candidates lack the key visual concept (castle) required by the query image, enabling the model to correctly retrieve the target without referring to the text semantics. Similarly, all negatives in common-case #2 lack the key textual concept (winter) specified by the text, enabling correct retrieval without referring to the image semantics. However, in harder cases where both modalities contribute essential but complementary information, such shortcut behavior often leads to retrieval failures. This difficulty is a key factor that makes CIR more challenging than conventional IR, and contributes to the generally lower performance of CIR models Jiang et al. (2024). Additional analyses of such behavior and its impacts on model performance are detailed in Appendix A.

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.

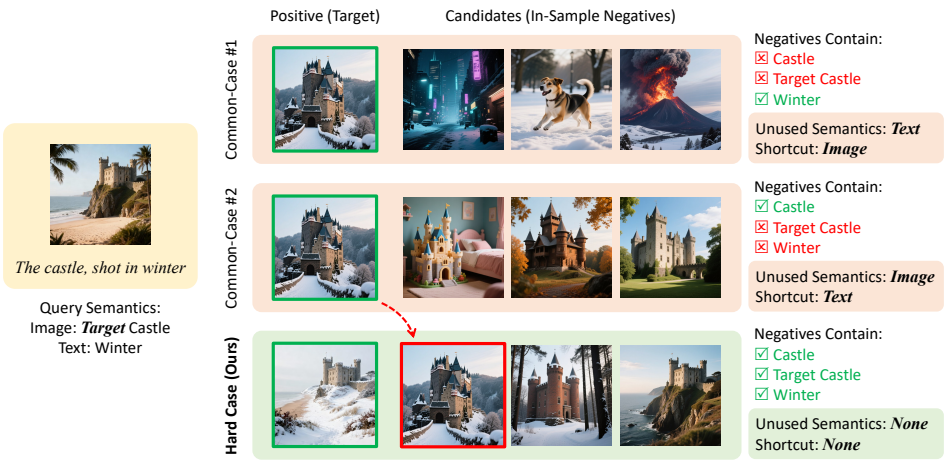


Figure 1: Models trained on common-case data learn "shortcuts" to get correct common-case retrieval results, but tend to fail on hard cases that require balanced cross-modal focuses. In this paper, we try to solve this issue by constructing targeted hard negatives. Original positives that are not highly consistent with the query are also considered negative in the proposed framework and replaced by more consistent synthetic images.

In this paper, we term such shortcuts *focus imbalances*, where a model over-attends to one modality while neglecting the other. To validate the prevalence and severity of this issue, we first introduce a multi-modal focus interpretation method, termed **FBCIR** (**F**ocus-**B**alancing). Using **FBCIR**, we reveal that focus imbalances are common across existing CIR models. We also reveal a diagnostic association between hard-case failures and focus imbalances in Appendix B. Building on the analyses, we further propose a CIR data augmentation workflow designed to mitigate focus imbalances. The **FBCIR-Data** workflow systematically enriches existing CIR triplets with curated hard negatives that require balanced reasoning over both image and text semantics. Using the workflow, we construct a novel benchmark and a finetuning dataset that enable quantitative evaluations and targeted improvements of CIR models from the perspective of focus balancing. Extensive experiments demonstrate that the proposed workflow effectively improves model performance on challenging cases while alleviating focus imbalances. An overview of **FBCIR** is shown in Figure 2.

The contributions of this work are three-fold:

- We identify and formalize the problem of *focus imbalance* in CIR, where models exploit shortcut cues from a single modality instead of jointly reasoning over image and text semantics, and empirically demonstrate its prevalence across representative CIR models.
- We propose **FBCIR**, a multi-modal focus interpretation method that provides fine-grained analyses of multi-modal focus behavior, enabling systematic diagnosis and quantitative assessment of focus imbalances in CIR models.
- We develop a data augmentation workflow based on **FBCIR**, yielding a targeted hard-case benchmark and a finetuning dataset with curated hard negatives, which effectively improve models' hard-case performance and encourage more balanced cross-modal focuses. The proposed workflow is among the first works that connect 1) token-level focus identification for diagnosing cross-modal focuses, and 2) data augmentation that explicitly constructs counterfactual negatives targeting focus imbalance.

2 Related Works

2.1 Composed Image Retrieval (CIR) Models

Composed Image Retrieval (CIR) retrieves images by conditioning jointly on a reference image and a modifying textual description. Early methods Vo et al. (2019) focus on explicit feature composition

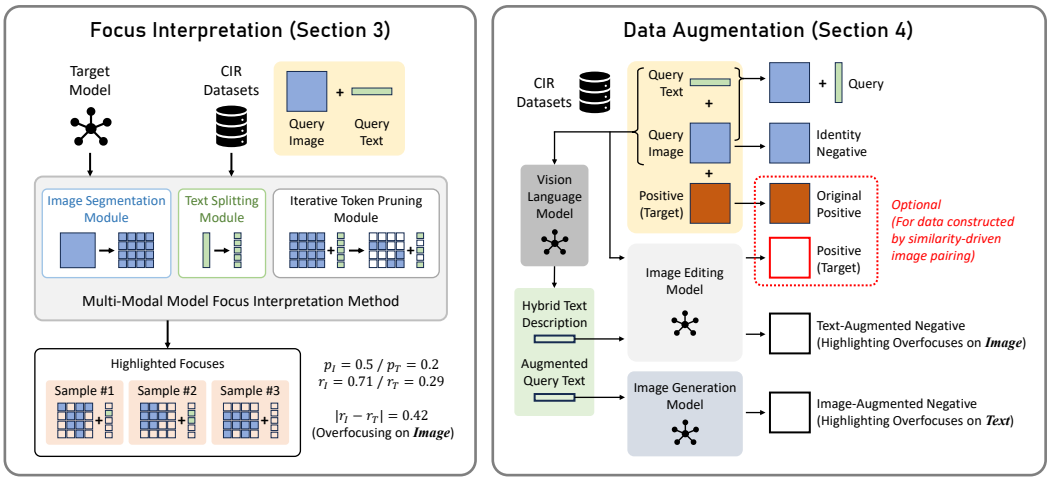


Figure 2: The overall framework of **FBCIR**, including a multi-modal model focus interpretation method and a dataset augmentation workflow. Given CIR triplets, the focus interpretation method highlights specific image segments and instruction keywords as the model's focuses, and reveals possible focus imbalances. The dataset augmentation framework facilitates existing CIR triplets with crafted hard negatives that encourage more balanced focuses. The focus interpretation module serves as the problem indicator and post-hoc validator of the data augmentation module.

over uni-modal or weakly aligned multimodal representations. With the development of vision-language pretraining, recent CLIP-based Radford et al. (2021); Baldrati et al. (2022); Saito et al. (2023) and VLM-based Bai et al. (2023); Jian et al. (2025); Lin et al. (2024) approaches employ more expressive fusion modules or dedicated query encoders, achieving strong performance on standard CIR benchmarks.

However, most CIR models are trained and evaluated under relatively easy task settings, where negative candidates are semantically distant from the target. This allows models to over-rely on a single modality rather than integrating multi-modal information. Such "shortcuts" have received limited explicit analysis, motivating in-depth investigation of model focuses under more challenging retrieval scenarios.

2.2 CIR DATASETS AND BENCHMARKS

Several benchmark datasets drive CIR research. FashionIQ Wu et al. (2021) focuses on fashion images with modification instructions, while CIRR Liu et al. (2021) extends to diverse real-world objects with relative captions. Earlier datasets Isola et al. (2015) also support attribute-based or relative retrieval, but within more constrained domains. Despite their impacts, these benchmarks generally rely on simple negative sampling strategies, making retrieval achievable even when models attend primarily to one modality. In particular, they rarely include candidates that correlate strongly with one specific modality of the composed query, which are crucial for revealing focus imbalances.

Recent works Psomas et al. (2025); Feng et al. (2024) introduce curated harder negatives and refined protocols. Though having similar data augmentation strategies to ours, they remain limited in systematic evaluation of cross-modal focuses of CIR models. In addition, the semantic correlations between negative candidates and queries in these datasets are not sufficiently strong for cross-modal balancing. This work explicitly targets such focus balancing and provides a complementary evaluation perspective beyond conventional retrieval accuracy.

2.3 MODEL FOCUS IDENTIFICATION AND INTERPRETATION

A line of research on model interpretation aims to identify input components most influential to model decisions. Gradient-based saliency methods Simonyan et al. (2013); Selvaraju et al. (2017) and perturbation-based methods Zeiler & Fergus (2014); Fong & Vedaldi (2017); Ribeiro et al.

(2016); Lundberg & Lee (2017) analyze model behavior by attributing predictions to specific inputs or by observing output changes under controlled input modifications. Some works Zhao et al. (2026); Chang et al. (2018) further identify minimal sufficient input regions that preserve model predictions, yielding deterministic and interpretable explanations.

While effective for single-modality vision tasks, existing interpretation are not designed to identify multi-modal model focuses, nor to account for cross-modal focus balances. This limitation motivates focus-aware analysis frameworks that jointly examine visual and textual contributions in CIR.

3 **FBCIR**: A MULTI-MODAL FOCUS INTERPRETATION METHOD

FBCIR is a model behavior interpretation method from the "focus" perspective. Given a model and a validation dataset, it identifies focuses of the model on the query images and texts, enabling quantitative measurement of cross-modal focus imbalances. This allows direct validation of the issue stated in Section 1 and Figure 1.

3.1 MULTI-MODAL ITERATIVE FOCUS REFINEMENT

We define "multi-modal focuses" of a CIR model as the minimal set of image segments (image tokens) and instruction keywords (text tokens) sufficiently supporting retrieval results. Image tokens are obtained through segmentation (Segment Anything Kirillov et al. (2023) in this paper), while text tokens are obtained via word-level splitting. These multi-modal tokens are jointly refined through an iterative pruning process. Given an image-text query and a candidate set, **FBCIR** iteratively removes less-deterministic tokens while preserving those essential to maintaining the original retrieval rankings. Following the practice in Fong & Vedaldi (2017), image tokens are pruned by zero-masking, and text tokens are pruned by replacing them with empty strings.

For each input state s with zero or more tokens pruned, we use an index set K_s to store the indices of all preserved tokens of the state. The process starts with an initial state s_0 where all tokens are preserved and $K_{s_0} = \{1, 2, \cdots, n_I + n_T\}$, given that the query image I has n_I tokens and the query text T has n_T tokens. In each iteration, new states are derived by pruning exactly one additional token from existing states. Each new state is validated via model inference: states that preserve the original retrieval result are retained and propagated to the next iteration. The process terminates when no further valid states can be obtained. A demonstration of the process is shown in Figure 3. Following Zhao et al. (2026), to improve efficiency, we adopt a beam-search strategy Bisiani (1992) and set a maximum number of new valid states per iteration (5 in this paper), which significantly accelerates refinement without degrading focus precisions, as analyzed in Appendix I.

The resulting final state set S satisfies:

- **Validity**: All final states yield the same retrieval result as the original input: $\forall s \in S, \text{rankings}(s) = \text{rankings}(s_0)$.
- **Minimality**: No further token removal is possible without altering the retrieval result: $\forall \hat{s} \notin S, \forall s \in S, |K_{\hat{s}}| < |K_s| \rightarrow \text{rankings}(\hat{s}) \neq \text{rankings}(s_0)$.

Tokens preserved in each final state are therefore minimal sufficient supports for model decisions, providing a precise and straightforward input-level interpretation of model focuses.

3.2 QUANTITATIVELY ASSESSING CROSS-MODAL FOCUS BALANCES

Based on the refinement results obtained in Section 3.1, we further propose a novel metric named *focus balance ratio* to quantify cross-modal *focus imbalances*.

For each final state s and modality $M \in \{I, T\}$, **FBCIR** extracts its corresponding index subset $K_{s,M}$ and calculates its *focus token proportion*:

$$p_{s,M} = \sum \alpha_M \cdot 1_{K_{s,M}} \tag{1}$$

in which $1_{K_{s,M}}$ is a binary vector with preserved indices activated, weighted by a token importance vector α_M. α_I is the area proportions of the segments to the whole image, while text tokens are

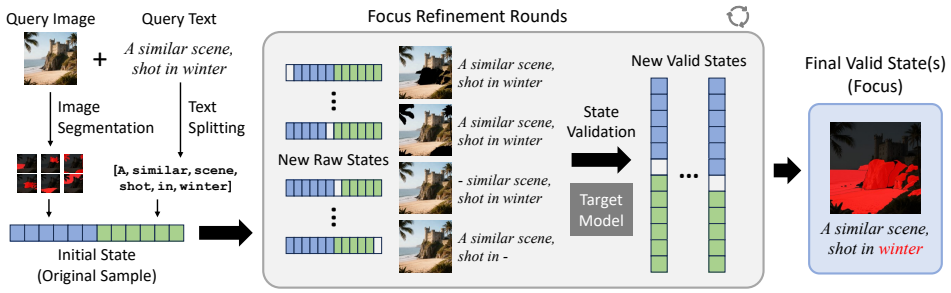


Figure 3: The multi-modal iterative focus refinement process of **FBCIR**.

uniformly weighted (i.e., $\alpha_T = \frac{1}{n_T}$). Note that although initial image and text tokens may possess different semantic weights from such formulation, obtaining reliable ground-truth token weights is inherently challenging. As the weights are only applied to the preserved most crucial tokens, treating them with area-proportional (image) and uniform (text) weights is a valid approximation of the model's "dependency" on the respective modalities.

The *focus balance ratio* of the model is then represented by the mean *relative* focus token proportion:

$$r_M = \frac{1}{|S|} \sum_s \frac{p_{s,M}}{\sum_{M' \in \{I,T\}} p_{s,M'}} \tag{2}$$

The absolute difference between the focus balance ratios $|r_I - r_T|$ over a validation set reflects the global focus imbalances of the model: Large gaps between r_I and r_T possibly indicate the model's over-reliance on the single modality with higher r_M, potentially revealing shortcut behavior. Note that the ratios are not calibrated measures of attention or causal contribution, and effective multi-modal reasoning does not require equal image and text ratios. We therefore interpret it alongside retrieval performance and controlled failure analyses.

4 The **FBCIR-Data** workflow and datasets

To further highlight the severity of focus imbalance and to explore potential solutions, we propose a novel dataset augmentation workflow for CIR. This workflow produces two datasets: a benchmark dataset designed to quantitatively evaluate existing CIR models on focus-challenging cases, and a finetuning dataset intended to encourage more balanced cross-modal focuses during training. Both datasets are constructed under a unified paradigm stated below.

4.1 Data construction

The core idea of the **FBCIR-Data** workflow is to expose focus imbalances through crafted hard negatives. If the model overfocuses on image, its hardest confusions are candidates visually similar to the query image but semantically inconsistent with the query text. Conversely, when a model overfocuses on text, its hard negatives are semantically aligned with the query text while exhibiting visual discrepancies. In both cases, neglecting one modality leads to incorrect retrieval results.

The workflow incorporates a VLM (Qwen3-VL-30B-A3B-Instruct Yang et al. (2025) in this paper), an image editing model (Qwen-Image-Edit Wu et al. (2025) in this paper), and an image generation model (Qwen-Image Wu et al. (2025) in this paper), and operates on existing CIR triplets. Starting from a triplet, the workflow constructs a focus-challenging sample with the following components:

- **Query**: The original query from the triplet.
- **Text-augmented negatives**: Leveraging the VLM, the workflow alters the semantics of the query text to produce multiple modified negative texts. Then, the workflow conducts image editing on the query image with the modified negative texts to yield multiple negative images that remain visually similar to the query image while violating the textual intent, highlighting overfocuses on the image modality.

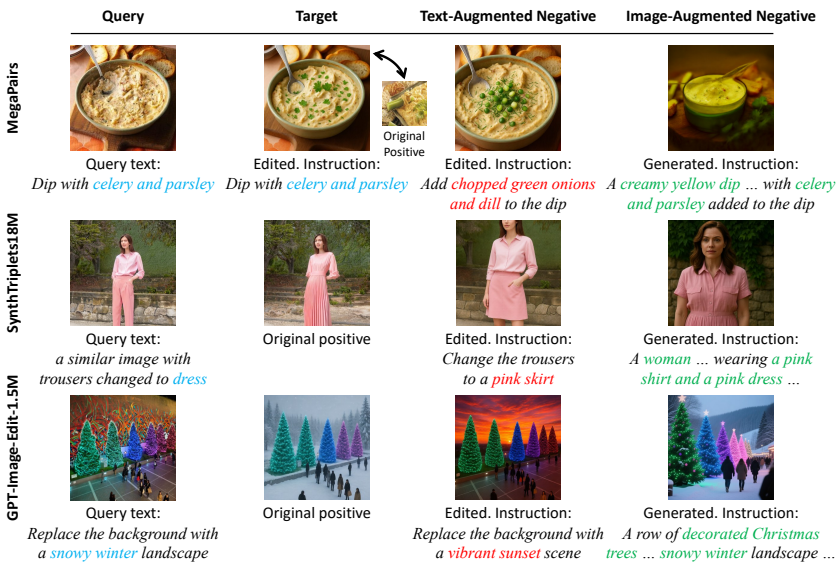


Figure 4: Examples of the data in **FBCIR-Data**. For the real-life dataset MegaPairs, we synthesize a positive for each triplet, and regard the original positive a special candidate.

- **Image-augmented negatives**: Leveraging the VLM, the workflow generates a comprehensive text description that integrates the semantics of the query. Then, the workflow conducts image generation with the description to obtain multiple negative images that are semantically aligned with the query text while having visual discrepancies from the query image, highlighting overfocuses on the text modality.
- **Identity negatives**: The original query image itself is included as a hard negative, given that there is no query text for identity retrieval.
- **Positive (Target)**: Triplets in most existing real-world CIR datasets are constructed via similarity-driven image pairing and query text generation Zhou et al. (2025). This results in relatively low data quality, as a large number of targets are only loosely consistent with their queries. To address this issue, the workflow treats the original positive from such datasets as a special candidate rather than the truth target, and synthesizes a strictly-consistent positive image based on the query. For datasets built under editing-driven triplet construction paradigms Wang et al. (2025); Gu et al. (2023), the target is intrinsically consistent with the query, and the original positive is retained.

Since real-world datasets lack such specific counterfactual negatives required to isolate focus imbalances, the synthetic approach is the optimal way to quantitatively measure this specific model behavior.

The above construction strategy is applied to both the benchmark and the finetuning dataset. We offer detailed quality analyses of them in Appendix D. Specifically, for the **FBCIR-Data** benchmark, its statistics and comparisons with existing CIR benchmarks are reported in Table 1. Representative examples of the constructed data are shown in Figure 4. Note that **FBCIR-Data** is not intended to replace global benchmarks such as CIRR, but to complement them. In practice, we advocate using both types of benchmarks together: Global ones evaluate overall retrieval capability in realistic large-scale settings, while **FBCIR-Data** probes robustness under focus-challenging conditions and reveals failure modes that may remain hidden in global evaluation.

4.2 Quantitative metrics in the benchmark

The proposed benchmark is designed to focus only on in-sample subset retrieval performance, as existing benchmarks Liu et al. (2021); Baldrati et al. (2023a); Vaze et al. (2023); Wu et al. (2021) already provide comprehensive evaluations on models' global retrieval capabilities. Specifically, the benchmark employs two quantitative metrics:

Table 1: Statistical comparisons between the **FBCIR-Data** and the standard benchmarks. Note that our proposed benchmark can be easily expanded based on existing triplets.

Benchmark	Queries	Candidates	Candidates (Local)
CIRR	4180	2265	5.97
FashionIQ	12032	6003	N/A
GeneCIS	3920	3030	15.0
FBCIR-Data	1000	3844	5.0

- **Subset recall** (R_s): Recall computed over the candidate set of each sample, consisting of the positive and the multiple types of hard negatives stated in Section 4.1. This metric measures model performance under focus-challenging conditions. Note that we do not adopt global recall for analysis, and the local subset-driven setting is a deliberate design for controlled evaluation of hard cases and focus-related failure modes.
- **Focus imbalance** ($|r_I - r_T|$): This metric directly quantifies models' cross-modal focus imbalances.

Together, these metrics characterize CIR models from two complementary perspectives: Subset recall reflects task performance under challenging cases, while focus imbalance captures underlying focus-related model behavior.

5 Experiments and discussions

5.1 Experimental setup

We construct the **FBCIR-Data** benchmark and finetuning dataset based on three large-scale CIR datasets: MegaPairs Zhou et al. (2025), SynthTriplets18M Gu et al. (2023), and GPT-Image-Edit-1.5M Wang et al. (2025). MegaPairs consists of real-world images, while the other two datasets are synthesized via image editing. Note that the text instructions in GPT-Image-Edit-1.5M are notably complex and frequently involve multiple editing requirements. Therefore, we decompose them into multiple simpler units to better align with the text distribution in standard CIR settings. To reflect realistic retrieval scenarios, we sample data from the three datasets with a ratio of 3:1:1. In each constructed sample, the numbers of positives, text-augmented negatives, and image-augmented negatives are all set to 1. In addition to the proposed hard-case benchmark, we also conduct extensive model performance evaluations on three widely-used standard benchmarks - CIRR Liu et al. (2021), FashionIQ Wu et al. (2021), and GeneCIS Vaze et al. (2023) - for general retrieval analysis.

The tested models include the CLIP-based CLIP4CIR Baldrati et al. (2023b), SEARLE Baldrati et al. (2023a), and BGE Zhou et al. (2025) families, and the VLM-based GME Zhang et al. (2024) family, RzenEmbed Jian et al. (2025), and MM-Embed Lin et al. (2024). All models go through the identical evaluation and finetuning process.

5.2 Pretrained model performance on the **FBCIR-Data** benchmark

Model performance on the **FBCIR-Data** and the standard benchmarks are summarized in Table 2. Results show that the tested models achieve comparable retrieval recalls when retrieving from large global candidate pools (with K-magnitude numbers of samples) in standard benchmarks and from small local candidate pools (similar to CIRR settings with 5 samples in average) in the **FBCIR-Data** benchmark. This observation highlights the difficulty of the proposed benchmark and suggests the presence of significant focus imbalances in existing models. Moreover, models exhibit inconsistent trends between common-case performance and focus imbalance, indicating that current benchmarks are insufficient for comprehensively evaluating model performance and behavior. In Appendix B, we provide focus statistics on correctly- and incorrectly-retrieved samples in **FBCIR-Data**, revealing a diagnostic association between hard-case failures and focus imbalances. Additional experimental results and qualitative focus visualizations are provided in Appendix E.

Table 2: Pretrained performance of the tested CIR models on the standard and the **FBCIR-Data** benchmarks. The "Average" column shows the average value of CIRR R@1, CIRR Rs@1, FashionIQ R@10, and GeneCIS Rs@1. Higher is better for average and **FBCIR-Data** performance, while lower is better for focus imbalance.

Model	Average	**FBCIR-Data** Rs@1	Focus Imbalance
CLIP-based Models			
CLIP4CIR-RN50	24.9	33.4	0.53
CLIP4CIR-RN50x4	26.0	31.2	0.65
SEARLE-base	23.6	29.3	0.27
SEARLE-large	24.7	30.2	0.70
BGE-base	36.5	41.6	0.10
BGE-large	38.3	38.8	0.02
VLM-based Models			
GME-2B	33.8	29.7	0.54
GME-7B	36.1	32.9	0.55
RzenEmbed-7B	37.6	39.3	0.43
MM-Embed-7B	27.2	26.7	0.60

Table 3: Finetuned performance of the VLM-based CIR models trained on different datasets with different data scales and LoRA ranks. Higher is better for average and **FBCIR-Data** performance, while lower is better for focus imbalance.

Metrics: Average (Gain) || **FBCIR-Data** Rs@1 (Gain) || Focus Imbalance (Reduction)						
Data / Rank	GME-2B			GME-7B		
Pretrained	33.8	29.7	0.54	36.1	32.9	0.55
FBCIR-Data 5K / 2	**36.5 (+2.7)**	**37.6 (+7.9)**	**0.51 (-0.03)**	39.3 (+3.1)	**42.2 (+9.3)**	**0.54 (-0.01)**
iCIR 5K / 2	36.1 (+2.3)	34.4 (+4.7)	0.51 (-0.03)	**39.4 (+3.2)**	39.0 (+6.1)	0.55 (-0.00)
SPN4CIR 5K / 2	36.0 (+2.2)	34.6 (+4.9)	0.54 (-0.00)	39.0 (+2.9)	40.1 (+7.2)	0.54 (-0.01)
FBCIR-Data 10K / 4	35.4 (+1.6)	**36.1 (+6.4)**	**0.40 (-0.14)**	**40.3 (+4.2)**	**43.4 (+10.5)**	**0.54 (-0.02)**
iCIR 10K / 4	35.3 (+1.5)	34.1 (+4.4)	0.51 (-0.03)	39.9 (+3.8)	39.7 (+6.8)	0.54 (-0.02)
SPN4CIR 10K / 4	**35.9 (+2.1)**	35.7 (+6.0)	0.49 (-0.05)	40.1 (+4.0)	39.4 (+6.5)	0.54 (-0.01)
Data / Rank	RzenEmbed-7B			MM-Embed-7B		
Pretrained	37.6	39.3	0.43	27.2	26.7	0.60
FBCIR-Data 5K / 2	38.6 (+1.1)	**42.2 (+2.9)**	0.44 (+0.01)	**28.5 (+1.3)**	**30.9 (+4.2)**	**0.56 (-0.04)**
iCIR 5K / 2	37.8 (+0.2)	39.7 (+0.4)	**0.43 (-0.00)**	24.7 (-2.5)	25.2 (-1.5)	0.64 (+0.04)
SPN4CIR 5K / 2	**38.9 (+1.3)**	39.2 (-0.1)	0.45 (+0.02)	28.0 (+0.8)	28.8 (+2.1)	0.57 (-0.03)
FBCIR-Data 10K / 4	38.9 (+1.3)	**42.5 (+3.2)**	**0.08 (-0.35)**	**29.3 (+2.1)**	**32.8 (+6.1)**	**0.17 (-0.43)**
iCIR 10K / 4	**39.2 (+1.6)**	40.3 (+1.0)	0.40 (-0.03)	24.5 (-2.7)	25.6 (-1.1)	0.62 (+0.02)
SPN4CIR 10K / 4	39.2 (+1.6)	40.6 (+1.3)	0.43 (-0.00)	29.0 (+1.8)	31.2 (+4.5)	0.58 (-0.02)

5.3 Finetuning performance comparisons with other existing datasets

We finetune the models for 1 epoch using the constructed dataset, evaluate their zero-shot (as the constructed data has no overlap with existing benchmarks) performance, and compare it with two other existing datasets: iCIR Psomas et al. (2025) and SPN4CIR Feng et al. (2024), as discussed in Section. 2.2. We finetune the last few layers of the CLIP-based models, and apply LoRA Hu et al. (2021) to the VLM-based models. Detailed finetuning configurations are provided in Appendix F. Unless otherwise specified, augmentation is applied to 50% of the finetuning samples, with the remaining samples consisting of standard CIR triplets. Results under more experimental settings are reported in Section 5.4 and Appendix G.

Table 3 shows VLM-based finetuning results; CLIP-based results are in Appendix G.2. Across settings, **FBCIR-Data** consistently outperforms baseline datasets, especially on hard-case recall (**FBCIR-Data** Rs@1), indicating improved robustness under focus-challenging conditions. More related analyses are in Appendix G.3. Improvements on CIRR Rs@1 are generally larger than

Table 4: Finetuned (with LoRA rank 2) performance of the VLM-based CIR models trained with different data scales and negative ratios on the **FBCIR-Data** benchmark. Higher is better.

Metrics: Average (Gain) / **FBCIR-Data** Rs@1 (Gain)								
Neg Ratio	5K	10K	20K	50K	5K	10K	20K	50K
	GME-2B (Pretrained: 33.8 / 29.7)				GME-7B (Pretrained: 36.1 / 32.9)			
0%	35.7 / 33.6	36.0 / 33.1	35.3 / 34.5	35.6 / 33.3	38.8 / 38.6	39.6 / 38.6	39.2 / 39.3	39.4 / 39.1
50%	36.5 (+0.8) 37.6 (+4.0)	36.6 (+0.6) 37.7 (+4.6)	35.3 (+0.0) 36.6 (+2.1)	36.4 (+0.7) 37.9 (+4.6)	39.3 (+0.5) 42.2 (+3.6)	41.1 (+1.6) 42.7 (+4.1)	40.7 (+1.6) 44.2 (+4.9)	41.3 (+1.8) 46.6 (+7.5)
100%	36.6 (+0.9) 40.7 (+7.1)	36.4 (+0.4) 42.6 (+9.5)	34.4 (-0.9) 41.1 (+6.6)	36.5 (+0.9) 45.5 (+12.2)	40.2 (+1.4) 45.0 (+6.4)	41.7 (+2.2) 49.3 (+10.7)	39.1 (+0.0) 48.7 (+9.4)	39.8 (+0.4) 50.4 (+11.3)
	RzenEmbed-7B (Pretrained: 37.6 / 39.3)				MM-Embed-7B (Pretrained: 27.2 / 26.7)			
0%	37.3 / 38.9	36.5 / 37.7	36.8 / 37.8	36.7 / 38.3	24.5 / 25.4	24.6 / 25.7	26.1 / 26.5	21.9 / 24.0
50%	38.6 (+1.3) 42.2 (+3.3)	40.4 (+3.9) 44.7 (+7.0)	41.3 (+4.4) 44.1 (+6.3)	40.7 (+4.1) 45.4 (+7.1)	28.5 (+4.0) 30.9 (+5.5)	29.6 (+5.0) 34.2 (+8.5)	31.2 (+5.1) 35.0 (+8.5)	31.9 (+10.0) 38.5 (+14.5)
100%	38.9 (+1.6) 45.7 (+6.8)	40.1 (+3.6) 47.8 (+10.1)	39.2 (+2.3) 48.2 (+10.4)	41.4 (+4.7) 50.1 (+11.8)	29.4 (+4.8) 34.7 (+9.3)	33.2 (+8.6) 40.3 (+14.6)	33.0 (+6.9) 43.0 (+16.5)	32.8 (+10.9) 45.4 (+21.4)

other gains, further showing improved zero-shot hard-case performance. To further validate its generalization capabilities, we apply the identical augmentation strategy to CIRR to construct a distribution-shifted zero-shot hard-case benchmark (**FBCIR**-CIRR), with details presented in Appendix H. Results show high stability and generalizability of model improvements.

5.4 Ablation Studies

We further analyze the impacts of data scales and negative ratios on finetuned model performance.

Table 4 reports results under varying data scales and negative ratios. The data with negative ratio 0% contains no **FBCIR** augmentation and serves as a baseline. Compared with this baseline, **FBCIR**-augmented data consistently improves both standard and hard-case benchmarks. Specifically, **FBCIR-Data** performance increases with negative ratios and data scales, showing the effectiveness of more difficult and diverse hard negatives. Additional focus analyses are in Appendix G.4.

6 Discussions and Limitations

Despite the effectiveness of **FBCIR**, this work has several limitations worth discussion. First, the metrics in the **FBCIR-Data** benchmark rely on heuristic token weighting schemes. Importantly, since our analysis relies on relative comparisons instead of absolute values, potential biases are effectively mitigated. In Appendix C, we validate the stability of our analyses across metric designs. Second, focus refinement in **FBCIR** requires multiple model inferences and has high time complexity. As an offline diagnostic tool applied only to a relatively small subset of samples in practice, it remains tractable for benchmark construction and model analyses. Empirical runtime statistics are provided in Appendix I. Third, according to Table 5 in Appendix A, the relationship between focus diagnostics and retrieval performance is not always straightforward and may vary across models and task settings. Future work could further identify when changes in model focuses are most informative about retrieval improvements. Overall, these limitations point to promising directions for future work, while not affecting method performance and general conclusions presented in this work.

7 Conclusions

We investigate *focus imbalances* in composed image retrieval, where models over-rely on one modality and may possibly fail on hard negatives. We propose **FBCIR** for multi-modal focus interpretation and **FBCIR-Data**, which constructs a hard-case benchmark and finetuning dataset with curated negatives. Experiments show that **FBCIR-Data** improves hard-case performance and focus balance while preserving standard benchmark capability. This work provides a complementary evaluation perspective for analyzing and improving CIR robustness.

REFERENCES

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21466–21474, 2022.

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*, pp. 15338–15347, 2023a.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Composed image retrieval using contrastive learning and task-oriented clip-based features. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–24, 2023b.

Roberto Bisiani. Beam search. *Encyclopedia of artificial intelligence*, 1992.

Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. *arXiv preprint arXiv:1807.08024*, 2018.

Muxi Chen, Yu Li, and Qiang Xu. Hibug: on human-interpretable model debug. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhangchi Feng, Richong Zhang, and Zhijie Nie. Improving composed image retrieval via contrastive learning with scaling positives and negatives. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1632–1641, 2024.

Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pp. 3429–3437, 2017.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1383–1391, 2015.

Weijian Jian, Yajun Zhang, Dawei Liang, Chunyu Xie, Yixiao He, Dawei Leng, and Yuhui Yin. Rzenembed: Towards comprehensive multimodal retrieval. *arXiv preprint arXiv:2510.27350*, 2025.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

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

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*, pp. 2125–2134, 2021.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Bill Psomas, George Retsinas, Nikos Efthymiadis, Panagiotis Filntisis, Yannis Avrithis, Petros Maragos, Ondrej Chum, and Giorgos Tolias. Instance-level composed image retrieval. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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*, pp. 8748–8763. PmLR, 2021.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

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*, pp. 19305–19314, 2023.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Sagar Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6862–6872, 2023.

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*, pp. 6439–6448, 2019.

Yuhan Wang, Siwei Yang, Bingchen Zhao, Letian Zhang, Qing Liu, Yuyin Zhou, and Cihang Xie. Gpt-image-edit-1.5 m: A million-scale, gpt-generated image dataset. *arXiv preprint arXiv:2507.21033*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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 *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11307–11317, 2021.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

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.

Chenchen Zhao, Muxi Chen, and Qiang Xu. Focalogic: Logic-based interpretation of visual model decisions. *arXiv preprint arXiv:2601.12049*, 2026.

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)*, pp. 19076–19095, 2025.

A Significance of focus balancing in improving model performance

In this section, we provide a more detailed discussion of focus balancing and its impact on CIR performance by addressing the following three questions:

Question #1: From which perspective does focus balancing diagnostically associated with more stable CIR performance?

As illustrated in Figure 1 in the main paper, models with imbalanced focuses tend to fail on hard cases where the negative candidate pool contains the semantics from both input modalities. In such cases, correct retrieval requires the model to jointly reason over visual and textual information, making balanced focus a necessary condition for stable CIR performance.

In contrast, common-case scenarios are often dominated by large semantic gaps between positives and negatives. Under these conditions, retrieval success depends more on a model's general representational capacity than on balanced cross-modal reasoning. As a result, models may still perform well even when over-relying on a single modality, as also demonstrated in Figure 1. This explains why focus imbalances are less visible in standard benchmarks but become critical in the **FBCIR-Data** benchmark with more difficult settings. Appendix G.3 offers a quantitative analysis on the specific types of capabilities that are improved through finetuning with **FBCIR-Data**.

Question #2: Are there hard cases that models with balanced focuses may still fail to handle?

In existing CIR literature, the difficulty of a sample is typically determined by how challenging it is for a model to distinguish the positive from its corresponding negatives. Accordingly, semantic discrepancies among candidates play a central role.

Negatives with small semantic discrepancies from the positive can be categorized into three types:

- Negatives that are consistent with the query image but inconsistent with the query text. Models that overfocus on visual information are prone to fail in such cases, and focus balancing directly mitigates this issue.
- Negatives that are consistent with the query text but inconsistent with the query image. Similarly, models that overfocus on textual information tend to fail, and focus balancing is again beneficial.
- Negatives that are semantically consistent with both the query image and the query text, but with lower overall consistency than the positive. Even models with balanced focuses may struggle in these cases.

Therefore, balanced focus does not guarantee success on all hard cases, particularly those involving "balanced but relatively weaker semantic consistency". However, the proposed **FBCIR-Data** workflow explicitly addresses such cases through two mechanisms:

- **Image-augmented negative generation.** As stated in Section 4.1 in the main paper, a specific type of negatives are generated using instructions that integrate both image and text semantics. As typical examples with "balanced but relatively weaker consistency", such negatives help models learn finer-grained distinctions. Note that since the integrated

Table 5: Focus imbalance values of the tested CIR models on correctly- and incorrectly-retrieved samples in **FBCIR-Data**, with their Pearson and Spearman correlations. Lower focus imbalances are better. Note that the association varies across models and is not observed for BGE-large with near-zero imbalance under this metric.

Model	Imbalance (Correct)	Imbalance (Incorrect)	Pearson	Spearman
CLIP-based Models				
CLIP4CIR-RN50	0.40	0.60	-0.39	-0.37
CLIP4CIR-RN50x4	0.60	0.67	-0.18	-0.16
SEARLE-base	0.19	0.30	-0.31	-0.30
SEARLE-large	0.61	0.74	-0.19	-0.18
BGE-base	0.06	0.13	-0.48	-0.45
BGE-large	0.02	0.02	-0.00	-0.00
VLM-based Models				
GME-2B	0.50	0.56	-0.14	-0.13
GME-7B	0.54	0.56	-0.08	-0.07
RzenEmbed-7B	0.39	0.46	-0.27	-0.26
MM-Embed-7B	0.50	0.64	-0.34	-0.32

instructions naturally align more closely with the textual modality, such negatives are especially effective in mitigating overfocuses on text.

- **Filtering of original positives.** As stated in Section 4.1 in the main paper, original positives that are not highly consistent with the query are reclassified as negatives. Such negatives are another type of examples with "balanced but relatively weaker consistency", and further improve model robustness in the above hard scenarios.

Question #3: Are there counter-examples where focus balancing negatively impacts performance?

As stated in Section 1 in the main paper, models with imbalanced focuses may still produce correct retrieval results in simple settings, primarily due to large semantic gaps among candidates. Importantly, encouraging balanced focuses in such cases does not degrade performance, as balanced reasoning fundamentally outperforms simpler retrieval behavior.

Another specific scenario arises when test data contain high-amplitude noise that affects one modality. In such cases, models with balanced focuses may be more susceptible to misleading signals. In this work, we attribute such failures to data quality issues rather than an inherent drawback of focus balancing. Under realistic and well-constructed CIR settings, balanced focuses remain desirable properties for robust composed retrieval.

B Association between retrieval failures and focus imbalances

In Section 5 in the main paper, we validate that models' hard-case performance improves through finetuning with **FBCIR-Data**. To further validate the correlation between retrieval failures and focus imbalances, we compare focus imbalance values between correctly- and incorrectly-retrieved samples in **FBCIR-Data**. Results in Table 5 show that incorrect samples consistently exhibit higher focus imbalances across different CIR models, showing relatively high correlation values. These results indicate that focus imbalance is systematically associated with hard-case failures and is a key contributing factor.

C Stability of focus-related analyses across different token weighting schemes

While our token weighting scheme involves heuristic components, it is designed to serve as a robust and decision-relevant proxy for modality reliance. As described in Section 3.1 in the main paper, the

Table 6: Focus imbalance values of the tested CIR models according to different token weighting schemes.

Model	Focus Imbalance (Heuristic)	Focus Imbalance (VLM-Generated)
CLIP-based Models		
CLIP4CIR-RN50	0.53	0.60
CLIP4CIR-RN50x4	0.65	0.70
SEARLE-base	0.27	0.28
SEARLE-large	0.70	0.74
BGE-base	0.10	0.06
BGE-large	0.02	0.04
VLM-based Models		
GME-2B	0.54	0.56
GME-7B	0.55	0.54
RzenEmbed-7B	0.43	0.46
MM-Embed-7B	0.60	0.55

metrics are computed on minimal sufficient tokens identified via iterative pruning, which reduces the influence of irrelevant input segments and focuses the analysis on the evidence most critical to the retrieval outcome.

To further examine potential biases caused by the scheme, we replace the heuristic weights with VLM-generated token importance. We adopt the powerful Qwen3-VL-30B Yang et al. (2025) to generate comprehensive token weight representations. Concretely, given the query image, query text, and tokenization, we prompt the VLM to assign α_I and α_T based on its understanding of token-level semantic importance and cross-modal relevance. All other components of focus identification remain unchanged from Section 3.2. Results are presented in Table 6. From the results, the focus imbalance values remain highly consistent across models. This consistency indicates that our conclusions are robust to different weighting schemes. While absolute values may vary slightly, the metrics reliably capture relative focus imbalance trends, which is the key objective of our analysis.

D Quality analyses of the **FBCIR-Data** benchmark and finetuning dataset

The high quality of the **FBCIR-Data** benchmark and finetuning dataset can be reflected from the following perspectives:

- **Query quality:** The proposed workflow preserves the original queries from existing CIR triplets. Since these triplets are sampled from established CIR datasets, the resulting queries naturally satisfy the requirements of composed image retrieval. In this work, queries are collected from multiple widely used open-source CIR datasets, providing a reliable foundation for query quality.
- **Candidate quality:** High query quality alone does not guarantee high candidate quality, particularly due to the query-target consistency issue stated in Section 4.1 in the main paper. For datasets constructed via similarity-driven image pairing, we synthesize a more consistent positive image for each triplet, yielding substantially improved alignment between queries and targets compared to the original data. In addition, the image editing and generation models employed in the workflow produce visually coherent images that closely follow the instruction prompts, ensuring both semantic consistency and visual fidelity among candidates.
- **Query coverage (diversity):** Similarly, the query coverage of **FBCIR-Data** is inherited from the source CIR triplets. By aggregating and randomly sampling triplets from three large-scale datasets, the resulting data exhibit broad coverage over query types and semantic variations. In addition, query coverage naturally increases with data scale, and can be adjusted to suit specific experimental or application settings.

Table 7: User feedback on the data quality and the query-target consistency of **FBCIR-Data** and other existing CIR datasets. Higher is better.

Data	Quality	Consistency
FBCIR-Data	**1.00**	**0.93**
MegaPairs	0.96	0.75
SynthTriplets18M	0.98	0.90
CIRCO Baldrati et al. (2023a)	0.90	0.74

Table 8: Finetuning performance of the **FBCIR** workflow w.r.t. different image synthesis models. Higher is better for average and **FBCIR-Data** performance, while lower is better for focus imbalance.

Model / Synthesis	Average (Gain)	**FBCIR-Data** Rs@1 (Gain)	Focus Imbalance (Reduction)
GME-2B / Qwen	36.5 (+2.7)	37.6 (+7.9)	0.51 (-0.03)
GME-2B / FLUX	38.2 (+4.4)	37.3 (+7.6)	0.51 (-0.03)
GME-7B / Qwen	40.9 (+4.8)	43.1 (+10.2)	0.54 (-0.01)
GME-7B / FLUX	41.4 (+5.3)	41.3 (+8.4)	0.53 (-0.02)

- **Candidate coverage (diversity):** As presented in Table 1 in the main paper, the proposed benchmark achieves a substantially higher candidate-to-query ratio than existing benchmarks. This indicates low candidate reuse and semantic redundancy, leading to greater diversity among candidates. Such diversity is particularly important for evaluating and diagnosing focus imbalances under challenging retrieval scenarios.

To further validate the superior quality of **FBCIR-Data** over existing CIR data, we conduct user studies on both image quality and query-target consistency. The user studies involve 10 researchers with experience in multimodal models and data analysis.

- Dataset-level absolute evaluation. Each participant was shown 4 groups, each containing 100 samples drawn from the respective dataset. The dataset source was hidden from the participant (single-blind setting). For each sample, the participant assigned two binary judgments:
  - Image quality: whether the target image is visually plausible and of sufficiently high perceptual quality.
  - Query-target consistency: whether the target image is semantically consistent with both the query image and the query text, i.e., whether it integrates information from both modalities rather than matching only one of them.

  The final scores were obtained by averaging these binary judgments across samples and participants.

  The feedback from the users is presented in Table 7, showing that **FBCIR-Data** achieves higher quality and consistency than existing CIR datasets.
- Pairwise comparison between original and constructed positives. Each participant was shown 100 samples, each with two candidate positives: the original positive and our constructed positive. Their identities were hidden. For each pair, the participant selected which candidate was better in terms of image quality and query-target consistency, with detailed requirements stated above.

  According to the results, the users prefer 93% of our constructed positives, 3% of the original positives, and 4% of neither, reflecting the superior positive quality of **FBCIR-Data**.

To further investigate the stability of our data synthesis pipeline, we replace Qwen-Image-Edit in the pipeline with FLUX.2-Klein-4B Labs (2025), and test its finetuning performance. In this experiment, GME-2B and GME-7B are finetuned with 5K data with 50% augmentation. LoRAs with rank 2 and 4 are added to the respective models. Results in Table 8 demonstrate the highly-stable

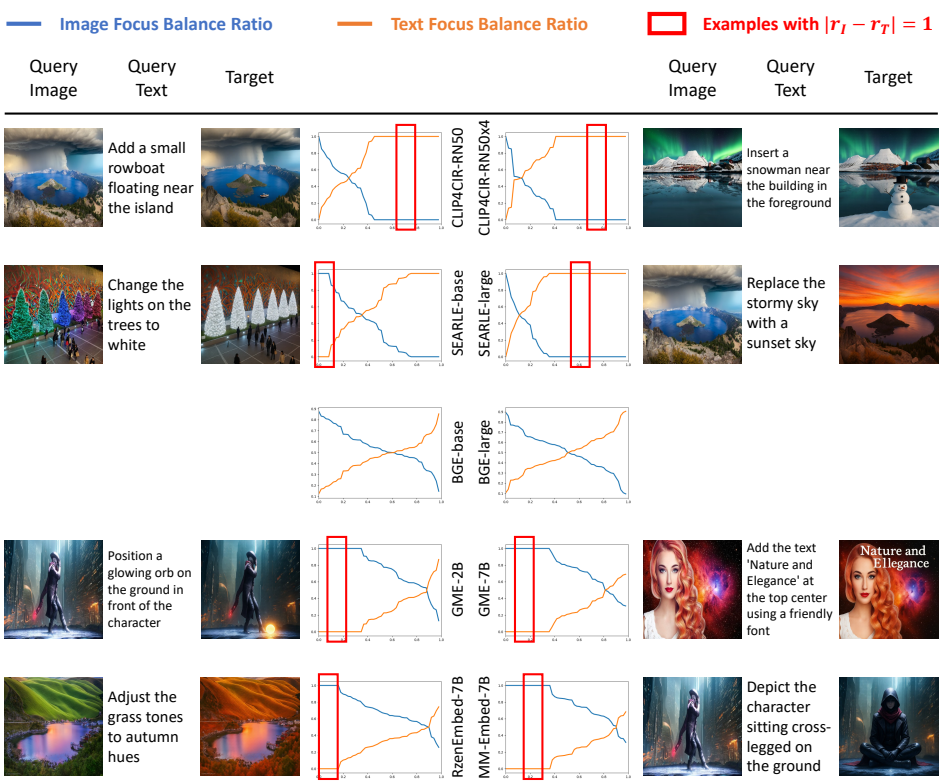


Figure 5: Visualizations of image and text focus balance ratios, and examples with semantics of a specific modality completely ignored by the models.

performance improvements by the **FBCIR** workflow against synthetic artifacts caused by specific data synthesis models.

In summary, beyond increasing task difficulty and exposing focus imbalances, the proposed **FBCIR-Data** workflow also improves the overall quality of existing CIR triplets by enhancing query-target consistency and candidate diversity. These properties make the resulting benchmark and finetuning dataset suitable for both diagnostic evaluation and robust model training.

E Focus visualizations and qualitative analyses of CIR models on the **FBCIR-Data** benchmark

In this section, we randomly select 50 samples from the **FBCIR-Data** benchmark, compute sample-wise focus balance ratios, and perform qualitative analyses of focus-related model behavior.

Figure 5 shows the behavior of the evaluated models on the samples. For better clarity in visualizations, image focus balance ratios r_I are sorted in a descending order.

Results show that focus imbalances are prevalent across all evaluated models, as reflected by the gaps between the image (blue) and text (orange) focus ratio curves. In addition, most models exhibit focus imbalances 1 on specific samples, as highlighted by red boxes, indicating that the models make retrieval decisions with one input modality completely ignored. Note that the BGE models Zhou et al. (2025) do not exhibit complete focus imbalances on the selected samples. Nevertheless, the gaps between their image and text focus ratios still suggest non-negligible imbalance.

For samples with focus imbalances 1, we present one representative visualization per model in Figure 5. For example, for the CLIP4CIR-RN50 Baldrati et al. (2023b) model, the image semantics of the samples highlighted in the red box are completely ignored. In such cases, model decisions are primarily based on textual semantics (e.g., “ignoring the image content, and searching for images

Table 9: Details of the trainable parameters of the tested CIR models.

Model	LoRA Ranks	Trainable Param Count (Proportion)	Trainable Param Range
CLIP-based Models			

CLIP4CIR-RN50	N/A	75.5M (29.8%)	`combiner.dynamic_scalar`
			`combiner.output_layer`
			`clip_model.logit_scale`

CLIP4CIR-RN50x4	N/A	29.5M (12.4%)	
SEARLE-base	N/A	1.0M (0.7%)	`searle.layers.6` `clip_model.logit_scale`
SEARLE-large	N/A	2.4M (0.5%)	

BGE-base	2	0.2M (0.2%)	`q_proj`
			`k_proj`
			`v_proj`
			`out_proj`

	4	0.5M (0.3%)	
	8	1.0M (0.7%)	
	16	2.0M (1.3%)	
BGE-large	2	0.5M (0.1%)	
	4	1.1M (0.3%)	
	8	2.2M (0.5%)	
	16	4.3M (1.0%)	
VLM-based Models			

GME-2B	2	2.3M (0.1%)	`q_proj`
			`k_proj`
			`v_proj`
			`o_proj`
			`gate_proj`
			`up_proj`
			`down_proj`

	4	4.6M (0.2%)	
	8	9.2M (0.4%)	
	16	18.5M (0.8%)	
GME-7B	2	5.0M (0.1%)	
	4	10.1M (0.2%)	
	8	20.2M (0.3%)	
	16	40.4M (0.5%)	
RzenEmbed-7B	2	5.0M (0.1%)	
	4	10.1M (0.2%)	
	8	20.2M (0.3%)	
	16	40.4M (0.5%)	
MM-Embed-7B	2	5.5M (0.1%)	
	4	11.1M (0.2%)	
	8	22.2M (0.3%)	
	16	44.3M (0.5%)	

with *a small rowboat floating near the island*"). While correct retrieval may still occur when the dataset distribution is favorable, such behavior indicates the presence of shortcut strategies learned during training.

Similar patterns can be observed across other models, suggesting that reliance on image-only or text-only shortcuts is a common outcome when training data and task settings are relatively simple. In contrast, the proposed **FBCIR-Data** workflow mitigates such shortcut behavior by hard negative construction. As illustrated in Figure 1 in the main paper, models trained with the augmented data are exposed to hard cases where neither modality alone is sufficient, thereby discouraging shortcut learning and encouraging more balanced cross-modal reasoning.

F Detailed configurations of model finetuning

As stated in Section 5.3 in the main paper, we apply an SFT-style finetuning procedure to all evaluated models. The lightweight finetuning strategy has the following properties:

- The finetuning dataset is relatively small in scale (5K-50K samples in this paper), and consists exclusively of the proposed augmented data.
- Most model parameters are frozen during finetuning, with only a small proportion ($<2\%$ in most cases in this paper) of modules set trainable.
- The models are trained for a small number of iterations (1 epoch in this paper).

For CLIP-based models, except for the BGE model family (which are implemented using the `AutoModel` wrapper from the `transformers` library), only the last few layers are set trainable. All VLM-based models are finetuned using LoRA modules with different ranks. Details of the trainable parameters for each model are listed in Table 9.

During finetuning, for each training sample, we randomly select three in-sample hard negatives from the corresponding local candidate pool. We adopt a standard contrastive loss with temperature set to 0.07, applied to both in-batch (common-case) negatives and in-sample (hard-case) negatives with different weights. The weight for in-batch negatives is fixed to 1, while the weight for in-sample negatives is gradually increased from 0.2 to 2 over the first 15% of finetuning steps. This schedule encourages the model to progressively focus on harder cases without destabilizing early training. In addition, we further incorporate a distillation loss to mitigate catastrophic forgetting. For each model, the pretrained checkpoint serves as the teacher, and the distillation loss regularizes the discrepancies between the logits produced by the teacher and the finetuned student:

$$\mathcal{L}_d = \tau^2 \cdot \mathrm{KL}(\mathrm{softmax}(\frac{p_{\text{student}}}{\tau}), \mathrm{softmax}(\frac{p_{\text{teacher}}}{\tau})) \tag{3}$$

where the temperature τ is set to 2. The weight of the distillation loss is set to 10^3 in all experiments.

All models are finetuned using 8x NVIDIA H20 (96GB) GPUs. We adopt a global batch size of 64 = per-GPU batch size (4) $\times$ number of GPUs (8) $\times$ number of gradient accumulation steps (2). A cosine learning rate scheduler is applied with an initial learning rate of 10^{-4}.

G Comprehensive finetuning performance analyses

G.1 Finetuned performance of the VLM-based models

The finetuned benchmark performance and focus imbalances of the evaluated VLM-based models under LoRA ranks 2, 4, 8, and 16 are respectively reported in Table 10 and 11, which are the extensions of Table 3 in the main paper.

Across all LoRA configurations, the observed trends are consistent with those reported in the main experiments. Finetuned models exhibit improved hard-case performance and reduced focus imbalances, while maintaining competitive performance on standard benchmarks. These results further validate the effectiveness of the proposed **FBCIR-Data** workflow in encouraging balanced cross-modal reasoning and improving robustness under challenging retrieval scenarios. Moreover, results indicate that the consistent model performance improvements across different LoRA ranks are not sensitive to specific parameterization choices.

To further validate its generalization capabilities, we apply the identical augmentation strategy to CIRR to construct a distribution-shifted zero-shot hard-case benchmark (**FBCIR**-CIRR). As shown in Table 12, the models finetuned with the proposed dataset consistently outperform their pretrained counterparts, confirming that the observed improvements are stable and transferable.

G.2 Finetuned performance of the CLIP-based models

We apply the identical finetuning procedure to CLIP-based CIR models, and report the results in Table 13. Compared to VLM-based models, most CLIP-based models show less pronounced performance gains, and in some cases exhibit performance drops on standard benchmarks. We discuss several possible contributing factors below:

- **Differences in benchmark objectives.** As stated in Section 4.1 in the main paper, the **FBCIR-Data** benchmark is designed to remove potential "shortcuts" that existing CIR models may exploit during training. In contrast, many standard benchmarks implicitly reward such shortcuts (e.g., background or low-level correlation cues). When finetuned to abandon shortcut strategies, models may experience apparent performance drops on easier benchmarks, even though their robustness under challenging cases improves. Notably, as shown in the main experiments, this trade-off disappears for higher-capacity VLM-based models, demonstrating the effectiveness of the **FBCIR-Data** workflow.

Table 10: Finetuned performance of the VLM-based CIR models trained with different data scales and LoRA ranks on the standard and the **FBCIR-Data** benchmarks (extension of Table 3 in the main paper). Higher is better.

Metrics: CIRR R@1 / CIRR Rs@1 / FashionIQ R@10 / GeneCIS Rs@1 || Average (Gain) || **FBCIR-Data** Rs@1 (Gain)						
Data / Rank	GME-2B			GME-7B		
Pretrained	32.8 / 41.5 / 39.8 / 21.1	33.8	29.7	35.2 / 44.3 / 43.3 / 21.7	36.1	32.9
5K / 2	36.8 / 48.3 / 40.0 / 20.9	36.5 (+2.7)	37.6 (+7.9)	41.1 / 51.1 / 43.2 / 21.7	39.3 (+3.1)	42.2 (+9.3)
10K / 2	37.1 / 48.4 / 39.7 / 21.3	36.6 (+2.8)	37.7 (+8.0)	44.1 / 55.3 / 43.2 / 21.8	41.1 (+5.0)	42.7 (+9.8)
20K / 2	35.3 / 45.6 / 39.9 / 20.5	35.3 (+1.5)	36.6 (+6.9)	43.5 / 54.6 / 43.2 / 21.7	40.7 (+4.6)	44.2 (+11.3)
50K / 2	36.8 / 47.7 / 40.1 / 20.8	36.4 (+2.6)	37.9 (+8.2)	44.5 / 56.2 / 43.0 / 21.4	41.3 (+5.1)	46.6 (+13.7)
5K / 4	35.5 / 45.7 / 40.1 / 21.3	35.6 (+1.9)	35.8 (+6.1)	43.6 / 54.8 / 43.4 / 22.0	40.9 (+4.8)	43.1 (+10.2)
10K / 4	35.7 / 45.7 / 39.8 / 20.4	35.4 (+1.6)	36.1 (+6.4)	42.9 / 53.8 / 43.2 / 21.5	40.3 (+4.2)	43.4 (+10.5)
20K / 4	37.1 / 48.2 / 40.1 / 21.1	36.6 (+2.8)	40.7 (+11.0)	42.5 / 52.9 / 43.2 / 21.9	40.1 (+4.0)	45.0 (+12.1)
50K / 4	36.5 / 47.8 / 39.6 / 20.5	36.1 (+2.3)	38.8 (+9.1)	46.4 / 59.2 / 43.3 / 21.9	42.7 (+6.6)	47.3 (+14.4)
5K / 8	35.8 / 46.8 / 40.0 / 21.0	35.9 (+2.1)	37.5 (+7.8)	43.1 / 54.1 / 43.1 / 22.1	40.6 (+4.5)	43.8 (+10.9)
10K / 8	37.1 / 48.5 / 39.9 / 20.8	36.6 (+2.8)	38.4 (+8.7)	44.0 / 55.6 / 43.1 / 21.7	41.1 (+5.0)	45.6 (+12.7)
20K / 8	37.1 / 48.7 / 40.0 / 20.7	36.6 (+2.8)	41.1 (+11.4)	44.1 / 55.6 / 43.0 / 21.4	41.0 (+4.9)	44.9 (+12.0)
50K / 8	36.8 / 47.8 / 39.8 / 21.3	36.4 (+2.7)	41.5 (+11.8)	42.8 / 54.1 / 42.5 / 21.4	40.2 (+4.1)	45.9 (+13.0)
5K / 16	35.8 / 46.5 / 39.7 / 20.9	35.7 (+1.9)	36.9 (+7.2)	40.7 / 50.3 / 42.9 / 21.3	38.8 (+2.7)	43.4 (+10.5)
10K / 16	35.1 / 45.8 / 39.5 / 21.0	35.3 (+1.6)	36.3 (+6.6)	40.2 / 48.7 / 42.9 / 21.4	38.3 (+2.2)	42.3 (+9.4)
20K / 16	36.2 / 46.5 / 39.6 / 20.3	35.7 (+1.9)	39.7 (+10.0)	42.3 / 52.9 / 43.4 / 21.6	40.0 (+3.9)	45.4 (+12.5)
50K / 16	37.5 / 48.6 / 39.8 / 20.9	36.7 (+2.9)	43.5 (+13.8)	42.0 / 51.9 / 42.8 / 21.3	39.5 (+3.4)	46.1 (+13.2)
Data / Rank	RzenEmbed-7B			MM-Embed-7B		
Pretrained	40.1 / 47.5 / 41.2 / 21.4	37.6	39.3	24.5 / 27.2 / 38.2 / 18.9	27.2	26.7
5K / 2	41.5 / 49.8 / 41.3 / 22.1	38.6 (+1.1)	42.2 (+2.9)	26.6 / 30.0 / 38.3 / 19.1	28.5 (+1.3)	30.9 (+4.2)
10K / 2	43.7 / 53.5 / 40.7 / 23.6	40.4 (+2.8)	44.7 (+5.4)	28.3 / 32.2 / 38.6 / 19.1	29.6 (+2.4)	34.2 (+7.5)
20K / 2	45.5 / 55.9 / 40.8 / 23.0	41.3 (+3.7)	44.1 (+4.8)	30.2 / 35.0 / 39.0 / 20.7	31.2 (+4.0)	35.0 (+8.3)
50K / 2	44.6 / 54.8 / 40.6 / 22.9	40.7 (+3.2)	45.4 (+6.1)	31.3 / 36.8 / 38.9 / 20.6	31.9 (+4.7)	38.5 (+11.8)
5K / 4	42.2 / 51.1 / 41.5 / 22.1	39.2 (+1.7)	42.4 (+3.1)	24.5 / 27.4 / 38.8 / 18.2	27.2 (+0.0)	30.2 (+3.5)
10K / 4	41.6 / 49.8 / 41.6 / 22.3	38.9 (+1.3)	42.5 (+3.2)	27.4 / 30.8 / 38.4 / 20.6	29.3 (+2.1)	32.8 (+6.1)
20K / 4	40.5 / 47.4 / 41.4 / 21.5	37.7 (+0.1)	42.8 (+3.5)	31.3 / 36.6 / 38.4 / 21.4	31.9 (+4.7)	35.7 (+9.0)
50K / 4	43.0 / 50.8 / 40.9 / 22.4	39.3 (+1.7)	45.5 (+6.2)	26.3 / 29.2 / 37.8 / 20.1	28.4 (+1.2)	35.7 (+9.0)
5K / 8	42.2 / 51.0 / 41.7 / 22.6	39.4 (+1.8)	42.5 (+3.2)	30.3 / 34.7 / 38.8 / 20.9	31.2 (+4.0)	33.0 (+6.3)
10K / 8	42.4 / 51.3 / 41.5 / 23.0	39.5 (+1.9)	42.5 (+3.2)	27.9 / 31.9 / 38.6 / 20.5	29.7 (+2.5)	34.0 (+7.3)
20K / 8	39.5 / 46.6 / 41.5 / 22.3	37.5 (-0.1)	42.5 (+3.2)	26.8 / 30.2 / 38.2 / 20.6	28.9 (+1.7)	36.9 (+10.2)
50K / 8	41.0 / 48.5 / 41.3 / 21.4	38.0 (+0.5)	44.6 (+5.3)	29.6 / 34.5 / 38.2 / 20.6	30.7 (+3.5)	38.4 (+11.7)
5K / 16	40.4 / 48.4 / 41.1 / 22.0	38.0 (+0.4)	42.9 (+3.6)	29.5 / 33.9 / 38.7 / 19.9	30.5 (+3.3)	34.1 (+7.4)
10K / 16	40.2 / 47.8 / 42.0 / 22.9	38.2 (+0.6)	44.0 (+4.7)	33.9 / 40.0 / 39.5 / 21.8	33.8 (+6.6)	38.0 (+11.3)
20K / 16	42.4 / 51.0 / 41.3 / 22.4	39.3 (+1.7)	44.5 (+5.2)	27.0 / 30.3 / 37.0 / 19.8	28.5 (+1.3)	34.6 (+7.9)
50K / 16	43.1 / 51.8 / 41.6 / 23.2	39.9 (+2.4)	46.5 (+7.2)	24.6 / 27.1 / 38.2 / 18.4	27.1 (-0.1)	35.7 (+9.0)

Table 11: Finetuned focus imbalances of the VLM-based CIR models trained with different data scales and LoRA ranks (extension of Table 3 in the main paper). Lower is better.

Metrics: Focus Imbalance (Reduction)								
Data	5K	10K	20K	50K	5K	10K	20K	50K
Rank	GME-2B (Pretrained: 0.54)				GME-7B (Pretrained: 0.55)			
2	0.51 (-0.03)	0.58 (+0.04)	0.21 (-0.33)	0.28 (-0.26)	0.54 (-0.01)	0.54 (-0.02)	0.23 (-0.33)	0.13 (-0.42)
4	0.56 (+0.02)	0.40 (-0.14)	0.17 (-0.38)	0.32 (-0.22)	0.54 (-0.01)	0.54 (-0.02)	0.23 (-0.33)	0.13 (-0.42)
8	0.58 (+0.04)	0.16 (-0.38)	0.20 (-0.34)	0.17 (-0.37)	0.53 (-0.03)	0.26 (-0.29)	0.27 (-0.29)	0.31 (-0.24)
16	0.56 (+0.02)	0.34 (-0.20)	0.32 (-0.22)	0.27 (-0.27)	0.56 (-0.00)	0.22 (-0.33)	0.35 (-0.21)	0.28 (-0.28)
Rank	RzenEmbed-7B (Pretrained: 0.43)				MM-Embed-7B (Pretrained: 0.60)			
2	0.44 (+0.01)	0.23 (-0.21)	0.10 (-0.33)	0.11 (-0.32)	0.56 (-0.04)	0.05 (-0.55)	0.06 (-0.54)	0.16 (-0.44)
4	0.44 (+0.01)	0.08 (-0.35)	0.21 (-0.22)	0.04 (-0.39)	0.67 (+0.07)	0.17 (-0.43)	0.05 (-0.55)	0.14 (-0.45)
8	0.46 (+0.03)	0.19 (-0.24)	0.12 (-0.31)	0.04 (-0.39)	0.62 (+0.02)	0.01 (-0.59)	0.10 (-0.50)	0.06 (-0.54)
16	0.42 (-0.01)	0.18 (-0.25)	0.10 (-0.33)	0.06 (-0.37)	0.07 (-0.53)	0.02 (-0.57)	0.17 (-0.43)	0.19 (-0.41)

- **Relatively limited model capacity.** Due to the highly-entangled multi-modal feature spaces of CLIP Radford et al. (2021) models Chen et al. (2024), CLIP-based CIR models may struggle to simultaneously capture fine-grained semantic distinctions required by the hard negatives in **FBCIR-Data** and preserve the coarse-grained features effective for

Table 12: Finetuned zero-shot performance of the VLM-based CIR models trained with different data scales and LoRA ranks on the **FBCIR**-augmented CIRR benchmark. Higher is better.

Metrics: **FBCIR**-CIRR Rs@1 (Gain)								
Data	5K	10K	20K	50K	5K	10K	20K	50K
Rank	GME-2B (Pretrained: 28.5)				GME-7B (Pretrained: 25.0)			
2	29.8 (+1.3)	29.6 (+1.1)	28.0 (-0.5)	30.2 (+1.7)	27.4 (+2.4)	27.7 (+2.7)	27.3 (+2.3)	29.0 (+4.0)
4	29.2 (+0.7)	28.7 (+0.2)	29.7 (+1.2)	30.0 (+1.5)	27.3 (+2.3)	28.7 (+3.7)	29.4 (+4.4)	30.1 (+5.1)
8	28.8 (+0.3)	29.4 (+0.9)	30.3 (+1.8)	30.5 (+2.0)	29.0 (+4.0)	27.7 (+2.7)	28.7 (+3.7)	28.5 (+3.5)
16	29.3 (+0.8)	28.0 (-0.5)	29.3 (+0.8)	31.5 (+3.0)	27.7 (+2.7)	28.5 (+3.5)	28.9 (+3.9)	30.4 (+5.4)
Rank	RzenEmbed-7B (Pretrained: 30.1)				MM-Embed-7B (Pretrained: 27.5)			
2	30.5 (+0.4)	31.5 (+1.4)	30.4 (+0.3)	33.2 (+3.1)	29.4 (+1.9)	31.5 (+4.0)	29.5 (+2.0)	33.3 (+5.8)
4	30.8 (+0.7)	32.0 (+1.9)	32.2 (+2.1)	34.0 (+3.9)	30.7 (+3.2)	30.8 (+3.3)	31.9 (+4.4)	30.9 (+3.4)
8	30.5 (+0.4)	30.8 (+0.7)	33.0 (+2.9)	33.8 (+3.7)	31.4 (+3.9)	30.9 (+3.4)	32.4 (+4.9)	31.8 (+4.3)
16	32.3 (+2.2)	32.5 (+2.4)	32.4 (+2.3)	32.0 (+1.9)	30.3 (+2.8)	32.6 (+5.1)	33.0 (+5.5)	33.3 (+5.8)

Table 13: Finetuned performance of the CLIP-based CIR models trained with different data scales and LoRA ranks on the standard and the **FBCIR-Data** benchmarks. Higher is better.

Metrics: Average (Gain) / **FBCIR-Data** Rs@1 (Gain)								
Data	5K	10K	20K	50K	5K	10K	20K	50K
Rank	CLIP4CIR-RN50 (Pretrained: 24.9 / 33.4)				CLIP4CIR-RN50x4 (Pretrained: 26.0 / 31.2)			
N/A	20.3 (-4.6) 42.2 (+8.8)	20.2 (-4.6) 47.0 (+13.6)	17.6 (-7.3) 41.7 (+8.3)	19.1 (-5.7) 48.4 (+15.0)	27.8 (+1.8) 50.7 (+19.5)	26.5 (+0.5) 51.0 (+19.8)	25.6 (-0.4) 53.8 (+22.6)	25.4 (-0.6) 51.4 (+20.2)
Rank	SEARLE-base (Pretrained: 23.6 / 29.3)				SEARLE-large (Pretrained: 24.7 / 30.2)			
N/A	23.0 (-0.6) 31.5 (+2.2)	22.7 (-0.9) 32.2 (+2.9)	22.9 (-0.7) 35.4 (+6.1)	22.5 (-1.0) 40.0 (+10.7)	24.7 (+0.0) 37.9 (+7.7)	24.4 (-0.3) 39.3 (+9.1)	23.9 (-0.7) 44.9 (+14.7)	23.3 (-1.4) 51.3 (+21.1)
Rank	BGE-base (Pretrained: 36.5 / 41.6)				BGE-large (Pretrained: 38.3 / 38.8)			
2	36.1 (-0.3) 42.4 (+0.8)	35.9 (-0.5) 41.8 (+0.2)	35.8 (-0.7) 43.0 (+1.4)	35.0 (-1.4) 44.6 (+3.0)	37.5 (-0.8) 40.7 (+1.9)	37.4 (-0.8) 40.7 (+1.9)	37.6 (-0.7) 40.8 (+2.0)	37.4 (-0.9) 42.3 (+3.5)
4	36.1 (-0.4) 42.4 (+0.8)	35.6 (-0.9) 42.9 (+1.3)	35.9 (-0.5) 42.3 (+0.7)	35.8 (-0.7) 44.2 (+2.6)	37.6 (-0.6) 41.0 (+2.2)	37.6 (-0.7) 40.8 (+2.0)	37.6 (-0.6) 40.9 (+2.1)	37.3 (-1.0) 42.1 (+3.3)
8	35.9 (-0.6) 41.9 (+0.3)	35.9 (-0.6) 43.6 (+2.0)	35.9 (-0.5) 43.6 (+2.0)	36.0 (-0.5) 44.4 (+2.8)	37.8 (-0.4) 40.2 (+1.4)	38.0 (-0.3) 41.0 (+2.2)	37.6 (-0.7) 41.9 (+3.1)	37.6 (-0.7) 42.2 (+3.4)
16	36.0 (-0.4) 42.4 (+0.8)	35.9 (-0.6) 43.5 (+1.9)	35.7 (-0.7) 42.8 (+1.2)	35.8 (-0.6) 45.3 (+3.7)	37.9 (-0.3) 40.6 (+1.8)	37.6 (-0.7) 40.9 (+2.1)	37.6 (-0.7) 41.5 (+2.7)	37.3 (-1.0) 41.4 (+2.6)

standard benchmarks. In contrast, higher-capacity VLM-based models are better equipped to handle this trade-off.

- **Constraints of the finetuning strategy.** Unlike VLM-based models, CLIP-based models are generally not finetuned with LoRA. In addition, for CLIP4CIR Baldrati et al. (2023b) and SEARLE Baldrati et al. (2023a), the selection of trainable layers is heuristic rather than optimized for maximal performance gains. In this work, we focus on evaluating the quality and effectiveness of the proposed data augmentation rather than exploring model-specific finetuning strategies. The consistent improvements of VLM-based models demonstrate the effectiveness of the data itself. For CLIP-based models, the proposed data can be combined with more advanced or tailored finetuning techniques to improve their specific capabilities according to task requirements.

G.3 Specific types of model abilities improved through finetuning

In this section, we provide quantitative evidence on the retrieval quality of the finetuned CIR models, and offer in-depth analyses of the specific types of abilities that are improved through finetuning. Specifically, we analyze global R@K improvements on **FBCIR-Data** after finetuning (with LoRA rank 16) with 50K samples with 100% augmentation. Quantitative results are presented in Table 14. Note that R@1 directly reflects the ability to distinguish positives from in-sample hard negatives; recalls with larger Ks have weaker reflections of such abilities, and recalls with K$\geq$5 reflect models' general retrieval capabilities. From the results, gains on R@1 are consistently much larger than those

Table 14: R@K gains of the tested CIR models finetuned (with LoRA rank 16) on **FBCIR-Data** with 50K samples with 100% augmentation. Higher is better.

Model	R@1 Gain	R@3 Gain	R@5 Gain
CLIP-based Models			
CLIP4CIR-RN50	+16.7	+8.1	+0.3
CLIP4CIR-RN50x4	+16.0	+5.3	+0.2
SEARLE-base	+20.4	+10.7	+0.4
SEARLE-large	+23.3	+13.6	+0.5
BGE-base	+5.0	+0.9	+0.0
BGE-large	+5.0	+1.4	+0.0
VLM-based Models			
GME-2B	+15.8	+2.1	+0.0
GME-7B	+17.5	+5.1	+0.1
RzenEmbed-7B	+10.9	+3.3	+0.0
MM-Embed-7B	+8.3	+0.3	+0.0

on R@3 and R@5 across all tested models, demonstrating that **FBCIR-Data** primarily improves fine-grained positive-negative differentiation, which is our target objective. On the other hand, non-negative gains on R@5 indicate that the models' general retrieval abilities are preserved.

G.4 Additional ablation studies

Table 15: Finetuned (with LoRA rank 2) performance of the VLM-based CIR models trained with different data scales and negative ratios on the **FBCIR-Data** benchmark (extension of Table 4 in the main paper). Higher is better.

Metrics: Average (Gain) / **FBCIR-Data** Rs@1 (Gain)								
Neg	5K	10K	20K	50K	5K	10K	20K	50K
Ratio	GME-2B (Pretrained: 33.8 / 29.7)				GME-7B (Pretrained: 36.1 / 32.9)			
0%	35.7 / 33.6	36.0 / 33.1	35.3 / 34.5	35.6 / 33.3	38.8 / 38.6	39.6 / 38.6	39.2 / 39.3	39.4 / 39.1
25%	35.9 (+0.2) 34.6 (+1.0)	36.2 (+0.2) 36.2 (+3.1)	37.2 (+1.8) 37.2 (+2.7)	36.2 (+0.6) 37.8 (+4.5)	40.4 (+1.6) 40.5 (+1.9)	39.3 (-0.2) 41.5 (+2.9)	38.5 (-0.7) 40.3 (+1.0)	39.6 (+0.1) 41.1 (+2.0)
50%	36.5 (+0.8) 37.6 (+4.0)	36.6 (+0.6) 37.7 (+4.6)	35.3 (+0.0) 36.6 (+2.1)	36.4 (+0.7) 37.9 (+4.6)	39.3 (+0.5) 42.2 (+3.6)	41.1 (+1.6) 42.7 (+4.1)	40.7 (+1.6) 44.2 (+4.9)	41.3 (+1.8) 46.6 (+7.5)
75%	36.7 (+1.0) 39.4 (+5.8)	37.0 (+1.0) 39.5 (+6.4)	35.4 (+0.0) 40.7 (+6.2)	35.0 (-0.6) 42.0 (+8.7)	41.1 (+2.3) 44.9 (+6.3)	41.3 (+1.7) 45.7 (+7.1)	41.2 (+2.1) 47.3 (+8.0)	39.9 (+0.5) 47.7 (+8.6)
100%	36.6 (+0.9) 40.7 (+7.1)	36.4 (+0.4) 42.6 (+9.5)	34.4 (-0.9) 41.1 (+6.6)	36.5 (+0.9) 45.5 (+12.2)	40.2 (+1.4) 45.0 (+6.4)	41.7 (+2.2) 49.3 (+10.7)	39.1 (+0.0) 48.7 (+9.4)	39.8 (+0.4) 50.4 (+11.3)
	RzenEmbed-7B (Pretrained: 37.6 / 39.3)				MM-Embed-7B (Pretrained: 27.2 / 26.7)			
0%	37.3 / 38.9	36.5 / 37.7	36.8 / 37.8	36.7 / 38.3	24.5 / 25.4	24.6 / 25.7	26.1 / 26.5	21.9 / 24.0
25%	38.4 (+1.0) 41.8 (+2.9)	37.1 (+0.6) 41.2 (+3.5)	38.5 (+1.6) 41.3 (+3.5)	38.8 (+2.2) 41.3 (+3.0)	26.7 (+2.2) 29.1 (+3.7)	27.3 (+2.7) 29.3 (+3.6)	28.8 (+2.7) 31.5 (+5.0)	27.6 (+5.7) 32.8 (+8.8)
50%	38.6 (+1.3) 42.2 (+3.3)	40.4 (+3.9) 44.7 (+7.0)	41.3 (+4.4) 44.1 (+6.3)	40.7 (+4.1) 45.4 (+7.1)	28.5 (+4.0) 30.9 (+5.5)	29.6 (+5.0) 34.2 (+8.5)	31.2 (+5.1) 35.0 (+8.5)	31.9 (+10.0) 38.5 (+14.5)
75%	40.0 (+2.7) 42.6 (+3.7)	39.4 (+2.9) 45.9 (+8.2)	38.2 (+1.4) 47.5 (+9.7)	39.7 (+3.0) 48.5 (+10.2)	34.1 (+9.5) 41.0 (+15.6)	32.4 (+7.8) 39.4 (+13.7)	32.0 (+5.8) 41.2 (+14.7)	34.8 (+12.9) 41.1 (+17.1)
100%	38.9 (+1.6) 45.7 (+6.8)	40.1 (+3.6) 47.8 (+10.1)	39.2 (+2.3) 48.2 (+10.4)	41.4 (+4.7) 50.1 (+11.8)	29.4 (+4.8) 34.7 (+9.3)	33.2 (+8.6) 40.3 (+14.6)	33.0 (+6.9) 43.0 (+16.5)	32.8 (+10.9) 45.4 (+21.4)

In this subsection, we conduct additional ablation studies similar to those in Section 5.4 in the main paper. Apart from more quantitative results, we additionally focus on the changes in focus imbalances. Table 15 presents additional results w.r.t. different data scales and negative ratios, while Table 16 reports the focus imbalance values under varying data scales and negative ratios.

Table 16: Finetuned (with LoRA rank 2) focus imbalances of the VLM-based CIR models trained with different data scales and negative ratios. Lower is better.

Metrics: Focus Imbalance (Reduction)								
Neg	5K	10K	20K	50K	5K	10K	20K	50K
Ratio	GME-2B (Pretrained: 0.54)				GME-7B (Pretrained: 0.55)			
0%	0.58	0.53	0.15	0.28	0.49	0.53	0.32	0.28
25%	0.52 (-0.05)	0.37 (-0.16)	0.28 (+0.13)	0.20 (-0.08)	0.52 (+0.03)	0.55 (+0.02)	0.28 (-0.04)	0.12 (-0.17)
50%	0.51 (-0.07)	0.58 (+0.05)	0.21 (+0.06)	0.28 (+0.01)	0.54 (+0.05)	0.54 (+0.01)	0.23 (-0.09)	0.13 (-0.15)
75%	0.48 (-0.10)	0.30 (-0.22)	0.23 (+0.08)	0.25 (-0.03)	0.46 (-0.03)	0.36 (-0.17)	0.29 (-0.02)	0.26 (-0.03)
100%	0.45 (-0.13)	0.53 (-0.00)	0.17 (+0.02)	0.21 (-0.07)	0.53 (+0.04)	0.19 (-0.34)	0.12 (-0.20)	0.32 (+0.03)
	RzenEmbed-7B (Pretrained: 0.43)				MM-Embed-7B (Pretrained: 0.60)			
0%	0.45	0.46	0.34	0.22	0.68	0.18	0.17	0.24
25%	0.45 (-0.00)	0.54 (+0.08)	0.05 (-0.29)	0.22 (-0.00)	0.02 (-0.66)	0.12 (-0.07)	0.12 (-0.05)	0.07 (-0.16)
50%	0.44 (-0.01)	0.23 (-0.24)	0.10 (-0.24)	0.11 (-0.10)	0.56 (-0.12)	0.05 (-0.14)	0.06 (-0.11)	0.16 (-0.08)
75%	0.44 (-0.01)	0.04 (-0.42)	0.05 (-0.29)	0.13 (-0.09)	0.57 (-0.11)	0.01 (-0.18)	0.04 (-0.13)	0.04 (-0.19)
100%	0.44 (-0.01)	0.07 (-0.40)	0.24 (-0.11)	0.01 (-0.20)	0.63 (-0.05)	0.08 (-0.11)	0.04 (-0.14)	0.11 (-0.12)

Across most configurations, finetuning with the proposed hard negatives consistently improves model performance and reduces focus imbalances compared to the baseline data without augmentation. These results provide further evidence that the crafted hard negatives in **FBCIR-Data** are effective in encouraging more balanced cross-modal focuses.

H BENCHMARK-WISE ANALYSES OF MODELS' ZERO-SHOT PERFORMANCE GAINS

Table 3 in the main paper and Table 10 report the performance gains of the models finetuned with the proposed data on standard CIR benchmarks. In this section, we provide more detailed benchmark-wise analyses of their zero-shot hard-case performance.

The standard benchmarks adopted in this work exhibit distinct characteristics:

- CIRR Liu et al. (2021): A real-world dataset that additionally defines a candidate subset for each query, where the candidates are semantically closer to one another than those in the global pool. Therefore, subset retrieval on CIRR (measured by Rs@1) poses a more challenging setting than standard global retrieval tasks, making it particularly suitable for evaluating hard-case robustness.
- FashionIQ Wu et al. (2021): A fashion clothes dataset with relatively narrow visual categories and limited diversity. Consequently, it exhibits noticeable distribution differences from other benchmarks and from the datasets used to construct **FBCIR-Data**.
- GeneCIS Vaze et al. (2023): A real-world dataset that separates queries into object-related and attribute-related categories.

Note that CIRCO Baldrati et al. (2023a) is another widely-adopted CIR benchmark. However, CIRCO follows a multi-positive paradigm, where positives are relatively more "softly" aligned with queries. Under our setting, such positives behave similarly to hard negatives, making the evaluation protocol not directly compatible. The user studies presented in Appendix D validate that the quality of CIRCO is lower than **FBCIR-Data**, supporting our claim to some extent. For this reason, we do not include CIRCO in our evaluation. Possible adaptations of **FBCIR** for fair evaluation on CIRCO remain to be explored.

The respective zero-shot performance gains of the models on the standard benchmarks are reported in Table 17 and 18. Results are highly consistent with the benchmark characteristics outlined above:

- **Stronger gains on CIRR Rs@1.** Performance improvements on CIRR Rs@1 are generally larger than those observed on other benchmark metrics. Since **FBCIR-Data** is explicitly designed to target hard cases, models show the most pronounced gains on the most challenging evaluation settings. This provides additional evidence for the effectiveness of the proposed workflow.

Table 17: Finetuned performance of the VLM-based GME models trained with different data scales and LoRA ranks on the standard benchmarks (extension of Table 3 in the main paper). Higher is better.

Data / Rank	CIRR R@1	CIRR Rs@1	FashionIQ R@10	GeneCIS Rs@1
		GME-2B		
Pretrained	32.8	41.5	39.8	21.1
5K / 2	36.8 (+4.0)	48.3 (+6.8)	40.0 (+0.2)	20.9 (-0.2)
10K / 2	37.1 (+4.4)	48.4 (+6.9)	39.7 (-0.1)	21.3 (+0.2)
20K / 2	35.3 (+2.5)	45.6 (+4.1)	39.9 (+0.0)	20.5 (-0.6)
50K / 2	36.8 (+4.1)	47.7 (+6.3)	40.1 (+0.3)	20.8 (-0.3)
5K / 4	35.5 (+2.8)	45.7 (+4.2)	40.1 (+0.3)	21.3 (+0.2)
10K / 4	35.7 (+2.9)	45.7 (+4.3)	39.8 (+0.0)	20.4 (-0.8)
20K / 4	37.1 (+4.4)	48.2 (+6.7)	40.1 (+0.3)	21.1 (+0.0)
50K / 4	36.5 (+3.8)	47.8 (+6.3)	39.6 (-0.2)	20.5 (-0.6)
5K / 8	35.8 (+3.0)	46.8 (+5.4)	40.0 (+0.2)	21.0 (-0.1)
10K / 8	37.1 (+4.3)	48.5 (+7.1)	39.9 (+0.1)	20.8 (-0.3)
20K / 8	37.1 (+4.3)	48.7 (+7.2)	40.0 (+0.2)	20.7 (-0.4)
50K / 8	36.8 (+4.1)	47.8 (+6.4)	39.8 (+0.0)	21.3 (+0.2)
5K / 16	35.8 (+3.0)	46.5 (+5.0)	39.7 (-0.1)	20.9 (-0.2)
10K / 16	35.1 (+2.4)	45.8 (+4.3)	39.5 (-0.3)	21.0 (-0.2)
20K / 16	36.2 (+3.5)	46.5 (+5.0)	39.6 (-0.2)	20.3 (-0.8)
50K / 16	37.5 (+4.7)	48.6 (+7.2)	39.8 (+0.0)	20.9 (-0.3)
Average Gain	+3.6	+5.8	+0.0	-0.3
		GME-7B		
Pretrained	35.2	44.3	43.3	21.7
5K / 2	41.1 (+5.9)	51.1 (+6.7)	43.2 (-0.1)	21.7 (+0.0)
10K / 2	44.1 (+8.9)	55.3 (+11.0)	43.2 (-0.1)	21.8 (+0.2)
20K / 2	43.5 (+8.3)	54.6 (+10.2)	43.2 (-0.1)	21.7 (+0.0)
50K / 2	44.5 (+9.3)	56.2 (+11.8)	43.0 (-0.3)	21.4 (-0.3)
5K / 4	43.6 (+8.4)	54.8 (+10.5)	43.4 (+0.0)	22.0 (+0.3)
10K / 4	42.9 (+7.7)	53.8 (+9.5)	43.2 (-0.2)	21.5 (-0.2)
20K / 4	42.5 (+7.3)	52.9 (+8.6)	43.2 (-0.1)	21.9 (+0.3)
50K / 4	46.4 (+11.2)	59.2 (+14.9)	43.3 (+0.0)	21.9 (+0.2)
5K / 8	43.1 (+7.9)	54.1 (+9.7)	43.1 (-0.2)	22.1 (+0.4)
10K / 8	44.0 (+8.9)	55.6 (+11.2)	43.1 (-0.2)	21.7 (+0.0)
20K / 8	44.1 (+8.9)	55.6 (+11.2)	43.0 (-0.3)	21.4 (-0.3)
50K / 8	42.8 (+7.7)	54.1 (+9.8)	42.5 (-0.8)	21.4 (-0.3)
5K / 16	40.7 (+5.6)	50.3 (+5.9)	42.9 (-0.4)	21.3 (-0.4)
10K / 16	40.2 (+5.0)	48.7 (+4.4)	42.9 (-0.5)	21.4 (-0.3)
20K / 16	42.3 (+7.2)	52.9 (+8.6)	43.4 (+0.0)	21.6 (-0.1)
50K / 16	42.0 (+6.8)	51.9 (+7.5)	42.8 (-0.5)	21.3 (-0.4)
Average Gain	+7.8	+9.5	-0.2	-0.1

- **Higher gains on CIRR than on other benchmarks.** Performance improvements on CIRR are generally larger than those on other benchmarks. CIRR focuses on general image retrieval and shares greater similarity with the base datasets used in constructing **FBCIR-Data** (MegaPairs Zhou et al. (2025), SynthTriplets18M Gu et al. (2023), and GPT-Image-Edit-1.5M Wang et al. (2025)). In contrast, models' zero-shot performance on FashionIQ and GeneCIS shows more variability, particularly for GME models Zhang et al. (2024). This behavior likely reflects the increased specialization of these benchmarks and their distribution differences from **FBCIR-Data**, rather than a limitation of the proposed augmentation strategy.

Table 18: Finetuned performance of the VLM-based RzenEmbed-7B and MM-Embed-7B models trained with different data scales and LoRA ranks on the standard benchmarks (extension of Table 3 in the main paper). Higher is better.

Data / Rank	CIRR R@1	CIRR Rs@1	FashionIQ R@10	GeneCIS Rs@1
RzenEmbed-7B				
Pretrained	40.1	47.5	41.2	21.4
5K / 2	41.5 (+1.3)	49.8 (+2.3)	41.3 (+0.1)	22.1 (+0.7)
10K / 2	43.7 (+3.5)	53.5 (+6.0)	40.7 (-0.5)	23.6 (+2.2)
20K / 2	45.5 (+5.3)	55.9 (+8.4)	40.8 (-0.4)	23.0 (+1.5)
50K / 2	44.6 (+4.5)	54.8 (+7.3)	40.6 (-0.6)	22.9 (+1.4)
5K / 4	42.2 (+2.1)	51.1 (+3.6)	41.5 (+0.2)	22.1 (+0.7)
10K / 4	41.6 (+1.5)	49.8 (+2.3)	41.6 (+0.4)	22.3 (+0.9)
20K / 4	40.5 (+0.4)	47.4 (-0.1)	41.4 (+0.2)	21.5 (+0.1)
50K / 4	43.0 (+2.8)	50.8 (+3.3)	40.9 (-0.3)	22.4 (+1.0)
5K / 8	42.2 (+2.1)	51.0 (+3.5)	41.7 (+0.4)	22.6 (+1.2)
10K / 8	42.4 (+2.2)	51.3 (+3.8)	41.5 (+0.2)	23.0 (+1.5)
20K / 8	39.5 (-0.6)	46.6 (-0.9)	41.5 (+0.3)	22.3 (+0.9)
50K / 8	41.0 (+0.8)	48.5 (+1.1)	41.3 (+0.1)	21.4 (+0.0)
5K / 16	40.4 (+0.3)	48.4 (+0.9)	41.1 (-0.1)	22.0 (+0.6)
10K / 16	40.2 (+0.1)	47.8 (+0.3)	42.0 (+0.7)	22.9 (+1.4)
20K / 16	42.4 (+2.2)	51.0 (+3.5)	41.3 (+0.1)	22.4 (+1.0)
50K / 16	43.1 (+2.9)	51.8 (+4.3)	41.6 (+0.4)	23.2 (+1.7)
Average Gain	+2.0	+3.1	+0.1	+1.1
MM-Embed-7B				
Pretrained	24.5	27.2	38.2	18.9
5K / 2	26.6 (+2.0)	30.0 (+2.8)	38.3 (+0.1)	19.1 (+0.3)
10K / 2	28.3 (+3.7)	32.2 (+5.0)	38.6 (+0.4)	19.1 (+0.3)
20K / 2	30.2 (+5.6)	35.0 (+7.8)	39.0 (+0.8)	20.7 (+1.8)
50K / 2	31.3 (+6.7)	36.8 (+9.5)	38.9 (+0.7)	20.6 (+1.8)
5K / 4	24.5 (+0.0)	27.4 (+0.2)	38.8 (+0.6)	18.2 (-0.7)
10K / 4	27.4 (+2.8)	30.8 (+3.6)	38.4 (+0.2)	20.6 (+1.7)
20K / 4	31.3 (+6.8)	36.6 (+9.4)	38.4 (+0.2)	21.4 (+2.6)
50K / 4	26.3 (+1.8)	29.2 (+2.0)	37.8 (-0.4)	20.1 (+1.2)
5K / 8	30.3 (+5.8)	34.7 (+7.5)	38.8 (+0.7)	20.9 (+2.0)
10K / 8	27.9 (+3.3)	31.9 (+4.7)	38.6 (+0.5)	20.5 (+1.7)
20K / 8	26.8 (+2.2)	30.2 (+3.0)	38.2 (+0.0)	20.6 (+1.7)
50K / 8	29.6 (+5.1)	34.5 (+7.3)	38.2 (+0.1)	20.6 (+1.7)
5K / 16	29.5 (+5.0)	33.9 (+6.7)	38.7 (+0.5)	19.9 (+1.0)
10K / 16	33.9 (+9.4)	40.0 (+12.8)	39.5 (+1.3)	21.8 (+3.0)
20K / 16	27.0 (+2.5)	30.3 (+3.1)	37.0 (-1.2)	19.8 (+1.0)
50K / 16	24.6 (+0.1)	27.1 (-0.1)	38.2 (+0.0)	18.4 (-0.5)
Average Gain	+3.9	+5.3	+0.3	+1.3

Table 19: Inference time of the tested CIR models on a batch of size 12 on one single NVIDIA H20 GPU with `bfloat16` precision.

Model	CLIP4CIR-RN50	SEARLE-base	BGE-base	GME-2B	RzenEmbed-7B
Time / s	0.24	0.23	0.05	0.53	0.90
Model	CLIP4CIR-RN50x4	SEARLE-large	BGE-large	GME-7B	MM-Embed-7B
Time / s	0.25	0.29	0.06	0.89	1.29

Table 20: Focus imbalance statistics conditioned on different beam sizes. The results are averaged on 100 held-out samples.

Beam Size	1	5	10	Exhaustive
CLIP-based Models				
CLIP4CIR-RN50	0.78	0.53	0.56	0.55
CLIP4CIR-RN50x4	0.75	0.65	0.64	0.62
SEARLE-base	0.20	0.27	0.26	0.25
SEARLE-large	0.84	0.70	0.72	0.76
BGE-base	0.03	0.10	0.08	0.10
BGE-large	0.02	0.02	0.02	0.02
VLM-based Models				
GME-2B	0.61	0.54	0.57	0.56
GME-7B	0.70	0.55	0.51	0.50
RzenEmbed-7B	0.35	0.43	0.39	0.41
MM-Embed-7B	0.66	0.60	0.61	0.60

I COMPUTATION ANALYSIS

The **FBCIR** focus interpretation method stated in Section 3 involves iterative pruning over multi-modal tokens, resulting in a combinatorial growth in time complexity. In this section, we provide a quantitative analysis of the computational costs of the process and discuss its practical implications.

As shown in Figure 3, the majority of computations occur in the model inferences during the state validation stage. In the kth iteration of the refinement process, each input state has $k-1$ tokens pruned. Following the beam-search strategy stated in Section 3.1, we set the maximum number of valid states in each iteration to a constant w. Consequently, at most $w \times (n_I + n_T - k + 1)$ new raw states can be derived in the kth iteration.

Therefore, the maximum number of model inferences required by the entire refinement process can therefore be upper-bounded as:

$$\mathcal{N} = \sum_{k=1}^{n_I+n_T} w \times (n_I + n_T - k + 1) = \frac{w}{2} \times (n_I + n_T + 1) \times (n_I + n_T) \tag{4}$$

which grows quadratically with respect to the total number of image and text tokens. This bound corresponds to a worst-case scenario in which all generated states remain valid and the refinement proceeds until all tokens are pruned. In practice, the refinement process typically terminates earlier, resulting in fewer iterations and inferences.

In our implementation, we adopt the SAM-ViT-Huge Kirillov et al. (2023) image segmentation model with moderate granularity (`points_per_side=15`) to determine image tokens. On average, the constructed dataset contains 15.7 image tokens and 10.3 text tokens per sample. With $w = 5$, the theoretical worst case corresponds to approximately 1690 model inferences per sample.

Table 19 reports the average inference time measured on a batch size of 12 using 1x NVIDIA H20 GPU with `bfloat16` precision. Based on the above analyses, the estimated worst-case runtime ranges from 7.0 to 181.7 seconds per sample, depending on the parameter scales of the models. While such computational costs may be non-negligible in extreme cases, **FBCIR** is explicitly designed for offline model diagnosis and guidance for the data augmentation workflow instead of real-time analyses over large-scale datasets. As a result, the computational overhead of **FBCIR** does not affect the efficiency of model training, evaluation, or deployment, and remains acceptable within its intended diagnostic scope.

To further validate that the proposed beam-search strategy does not sacrifice enumeration precision, we conduct extensive experiments with different beam sizes. Results in Table 20 show that beam size 5 reaches relatively the best balance between computational efficiency and diagnostic fidelity. The focus-oriented analyses can therefore be guaranteed.

J IMPACT STATEMENT

The primary goal of this work is to improve robustness and interpretability of machine learning models in the field of composed image retrieval. The techniques introduced in this work have positive impacts on practical applications such as recommendation systems and human-machine interaction. This work introduces no new ethical concerns beyond those commonly raised by existing vision-language models and AI-generated content, and does not need specific highlights or additional discussions in this paper.