ReCALL: Recalibrating Capability Degradation for MLLM-based Composed Image Retrieval

Supplementary Material

6. Additional Experimental Results and Analysis

6.1. Data Scale Study on FashionIQ

To investigate the scalability of our Self-Guided Informative Instance Mining strategy, we conduct a quantitative analysis on the FashionIQ dataset by varying the mining hyperparameter K (denoted as top-K). This parameter determines the maximum number of informative instances mined for each failure query, directly controlling the volume of synthesized supervision.

Experimental Setup. To decouple data scaling from quality filtering, we conduct experiments *without* the VQA-Assisted Quality Control mechanism. The model is trained via the standard InfoNCE loss within our Grouped Contrastive Refinement framework.

Results. The quantitative results are visualized in Fig. 5. We employ a dual-axis plot to illustrate the relationship between the mining constraint K (bottom x-axis) and the resultant volume of synthesized training samples (top x-axis). As illustrated, increasing K from 1 to 5 significantly expands the training set from 13,351 to 57,125 samples. Crucially, this increase in data scale correlates with a consistent upward trend in retrieval performance. Specifically, Avg. R@10 (left axis) improves from 55.27% to 56.07%, and Avg. R@50 (right axis) rises from 75.70% to 76.29%. This positive scaling effect demonstrates that ReCALL can effectively leverage larger pools of informative instances to refine its discriminative boundaries, yielding continuous gains even in the absence of additional filtering.

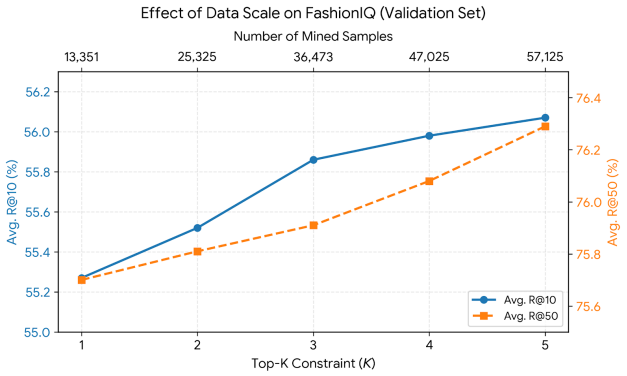

Figure 5. **Effect of data scale on the FashionIQ validation set.** The visualization employs dual x-axes to map the mining hyperparameter K (bottom) to the corresponding number of mined samples (top). The dual y-axes (left for R@10, right for R@50) with zoomed-in scales highlight the monotonic performance gains as the data scale increases.

6.2. Hyperparameter Analysis of Triplet Loss

To identify the optimal configuration for the targeted refinement stage, we conduct a grid search over two critical hyperparameters in the joint loss function: the triplet loss weight λ and the margin m. We evaluate the model on the FashionIQ validation set, identifying the optimal setting based on the Average R@10 metric. Specifically, the weight λ is varied within $\{0.1, 0.2, 0.3, 0.4, 0.5\}$, and the margin m within $\{0.05, 0.10, 0.20\}$.

Results. The sensitivity analysis is visualized in Fig. 6. First, concerning the loss weight λ, performance generally peaks at $\lambda = 0.3$. Lower weights (e.g., $\lambda = 0.1$) provide insufficient supervision for fine-grained discrimination, whereas excessive weights (e.g., $\lambda = 0.5$) tend to over-regularize the representation, potentially conflicting with the global alignment objective of the InfoNCE loss. Second, regarding the margin m, the model consistently favors a tighter constraint ($m = 0.05$). This preference suggests that the informative instances mined by our framework share high visual affinity with the ground truth targets. Consequently, a tighter margin compels the model to resolve these fine-grained ambiguities without disrupting the broader semantic structure of the embedding space. Based on these empirical findings, we adopt $\lambda = 0.3$ and $m = 0.05$ as the default configuration for FashionIQ, which yields the best Avg. R@10 of 57.04%.

6.3. Computational Cost and Efficiency Analysis

To ensure reproducibility and transparency regarding resource utilization, we detail the computational costs and data statistics **of** the ReCALL framework. All experiments were conducted on 8 NVIDIA H20 GPUs. Tab. 5 summarizes the training duration, generation latency, and filtering statistics for both the CIRR and FashionIQ datasets.

Analysis. We analyze the computational overhead across the three primary phases of our pipeline:

Comparable Training Latency (Stage 1 vs. Stage 4). The training duration for Targeted Refinement (Stage 4) is virtually identical to that of the Baseline Adaptation (Stage 1). For instance, on CIRR, Stage 4 requires approximately 3.6 hours, matching the 3.6 hours of Stage 1. This equivalence demonstrates that our Grouped Contrastive Refinement strategy (Sec. 3.5) effectively recalibrates the model

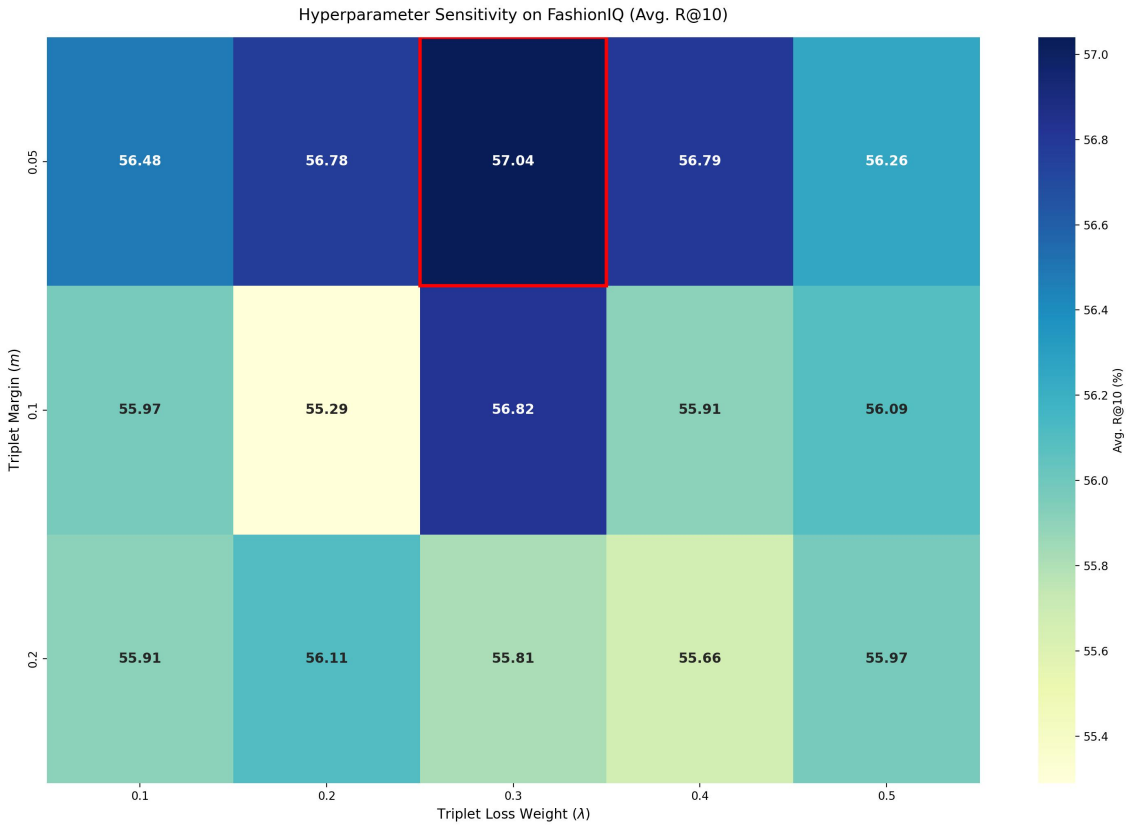

Figure 6. **Hyperparameter sensitivity analysis on the FashionIQ validation set.** We report Avg. R@10 (%) under varying triplet loss weights (λ) and margins (m). The red box highlights the optimal configuration adopted in our final model.

without introducing significant computational overhead to the online training loop.

One-off Offline Synthesis (Stages 2 & 3). The combined process of mining informative instances and synthesizing corrective instructions constitutes the primary computational cost. Specifically, CoT-assisted generation accounts for approximately 14.2 hours on CIRR and 10.9 hours on FashionIQ. Crucially, however, this represents a *one-time, offline investment*. Once synthesized, these high-quality triplets serve as a permanent asset that can be reused indefinitely for subsequent training runs or hyperparameter tuning, rendering the amortized cost negligible.

Efficient Quality Assurance. The VQA-Assisted Quality Control mechanism demonstrates high efficiency. It effectively purges noisy data—removing 5,455 instances (8.5%) for CIRR and 5,947 instances (10.4%) for FashionIQ—while consuming only $\sim$1 hour of processing time. This ensures that the final refinement is driven by high-fidelity supervision with minimal time penalty.

6.4. Generalization and Transferability Across Backbones

To evaluate the generalizability and transferability of the ReCALL framework, we extend our experiments to a different model family, LLaVA-NeXT. Furthermore, we investigate whether the informative instances and corrective instructions synthesized by one model can benefit another.

As shown in Tab. 6, cross-model transfer (training LLaVA-NeXT using triplets synthesized by Qwen2.5-VL) yields a +1.15% gain (51.93% $\rightarrow$ 53.08%) on CIRR. This confirms that different MLLMs share certain common cognitive blind spots, allowing synthesized corrective data to be highly transferable. However, the full ReCALL pipeline—where LLaVA-NeXT diagnoses and refines its own specific failure cases—achieves a more significant gain of +2.72% (54.65% on R@1). This indicates that while cross-model transfer is effective, the self-diagnosis phase remains essential to optimally address model-specific cognitive gaps.

6.5. Comparison with Alternative Mining Strategies

We further compare our Self-Guided Informative Instance Mining against a standard Hard Negative Mining baseline.

Table 5. **Detailed statistics of computational cost and data generation across the ReCALL pipeline.** Generation denotes the CoT-assisted synthesis in Stage 3, while Filtering refers to the VQA-based quality control. Note that the costs associated with Stages 2 and 3 are **one-time and offline**.

Dataset	Stage 1	Stages 2 & 3: Diagnose & Generate (Offline)			Stage 4
	Adaptation Time	*Generation Time*	*Samples (Generated → **Kept**)*	*Filtering Time*	*Refinement Time*
CIRR	3h 34m	∼14.2h	64,105 → **58,650**	1h 13m	3h 35m
FashionIQ	2h 43m	∼10.9h	57,125 → **51,155**	1h 04m	2h 45m

Table 6. **Generalization & Transferability Analysis on the CIRR test set.**

Method Setting	R@1	R@5	R@10	R@50
LLaVA Baseline ($\mathcal{R}_{base}$)	51.93	81.87	88.95	97.58
+ ReCALL (Transfer: Qwen Data)	53.08	83.40	91.21	98.46
+ ReCALL (Full Pipeline)	**54.65**	**84.02**	**91.33**	**98.41**

For the Hard Negative baseline, we re-finetune $\mathcal{R}_{base}$ directly using the mined informative instances as hard negatives without any textual refinement.

As reported in Tab. 7, the Hard Negative strategy achieves an R@1 of 52.57%, comparable to a Random Mining strategy (52.07%), and shows a notable decline in broader metrics (R@5/10/50) compared to the baseline. This implies that blindly enforcing repulsion on visually ambiguous negatives—without explicitly defining *why* they differ—introduces contradictory gradients that distort the learned manifold. In contrast, ReCALL resolves this via *semantic correction*: we generate $\tilde{T}_m$ to explicitly describe the hard negative, converting it into a constructive positive pair $(I_r, \tilde{T}_m, I_h)$. This precise semantic direction explains the superior capability calibration (+4.29% on R@1) achieved by the full ReCALL pipeline.

Table 7. **Comparison of Mining Strategies on CIRR.**

Method Setting	R@1	R@5	R@10	R@50
Baseline ($\mathcal{R}_{base}$)	51.23	82.15	90.20	98.20
+ Random Mining	52.07	81.64	90.02	97.84
+ Hard Neg. Mining (No Edit)	52.57	81.56	89.33	97.70
+ ReCALL (Full Pipeline)	**55.52**	**84.07**	**91.83**	**98.55**

6.6. Ablation on Model Capacity and Adaptation

To investigate whether capability degradation stems from limited parameter capacity during adaptation, we conducted ablations by scaling the LoRA rank ($r = 32, 64$) and performing Full Fine-tuning.

As shown in Tab. 8, increasing the number of trainable parameters paradoxically worsens retrieval performance. This confirms that capability degradation is not a consequence of limited parameter capacity. Instead, it originates from the intrinsic paradigm conflict between the MLLM's generative pre-training and the discriminative retrieval adaptation. Under a fixed training dataset, expanding trainable parameters accelerates overfitting to the coarse-grained retrieval task, thereby exacerbating the suppression of native fine-grained reasoning priors.

Table 8. **Ablation on LoRA Rank & Full Fine-tuning on CIRR.**

Setting	R@1	R@5	R@10	R@50
LoRA $r = 16$ (Ours Baseline)	**51.23**	**82.15**	**90.20**	**98.20**
LoRA $r = 32$	51.04	81.69	89.54	98.22
LoRA $r = 64$	49.74	80.58	89.35	98.05
Full Fine-tuning	48.70	80.55	89.64	97.98

6.7. Further Methodological Discussions

Distribution Integrity and Label Space. It is crucial to emphasize that ReCALL operates strictly as an informative instance augmentation strategy rather than altering or re-labeling the original ground-truth targets. The original triplets (I_r, T_m, I_t) are strictly retained to anchor the model to the source distribution. Furthermore, our Minimal Edit Principle (Sec. 3.4) guarantees that the synthesized text $\tilde{T}_m$ matches the original style and length. This design ensures that ReCALL provides additive regularization to sharpen decision boundaries without shifting the training label space.

Reliability of VQA-Assisted Filtering. To quantitatively ensure the reliability of the generated corrective supervision, we conducted a rigorous human evaluation of the VQA-Assisted Quality Control mechanism (Sec. 3.4). We employed three human evaluators to verify 300 randomly sampled triplets that passed the VQA filter (confidence threshold ≥ 0.95). The evaluation yielded a high average accuracy of 92%, confirming that the VQA-based check serves as a highly reliable proxy for filtering valid textual modifications.

Prompt for Composed Query on CIRR

[System Instruction] You are a multimodal retrieval encoder. Your sole function is to map any input into a compact embedding for nearest-neighbor retrieval.

[User Input] <Reference Image> Modify this image with {Modification Text} Represent the modified image in one word:

(a) Prompt for Composed Query on CIRR

Prompt for Candidate Image on CIRR

[System Instruction] You are a multimodal retrieval encoder. Your sole function is to map any input into a compact embedding for nearest-neighbor retrieval.

[User Input] <Target Image> Represent the given image in one word:

(b) Prompt for Candidate Image on CIRR

Prompt for Composed Query on Fashion-IQ

[System Instruction] You are a multimodal retrieval encoder. Your sole function is to map any input into a compact embedding for nearest-neighbor retrieval.

[User Input] <Reference Image> Change the style of this {Category} to {Modification Text}\n Represent this modified {Category} in one word:

(c) Prompt for Composed Query on FashionIQ

Prompt for Candidate Image on FashionIQ

[System Instruction] You are a multimodal retrieval encoder. Your sole function is to map any input into a compact embedding for nearest-neighbor retrieval.

[User Input] <Target Image> Represent the given image in one word:

(d) Prompt for Candidate Image on FashionIQ

Figure 7. Full prompt templates for retrieval encoding on CIRR and FashionIQ. The structure utilizes an integrated System Instruction to enforce the role of a discriminative encoder. The query prompts are specialized: CIRR uses a general modification instruction, while FashionIQ incorporates category information for fine-grained attribute manipulation.

7. Prompt Details

7.1. Retrieval Prompts for Query and Candidate Encoding

To counteract the Capability Degradation identified in Sec. 1, we engineer specialized prompt templates that explicitly condition the MLLM to operate as a discriminative retrieval encoder ($\mathcal{R}_{base}$ and $\mathcal{R}_{refine}$), effectively suppressing its default conversational tendencies.

Fig. 7 illustrates the prompt architectures employed for encoding inputs on both the CIRR and FashionIQ datasets. Our design adheres to two governing principles:

Role Enforcement via System Instruction. A mandatory system instruction is embedded in every prompt instance. This directive explicitly constrains the model's output space, enforcing a retrieval-oriented role and inhibiting open-ended generative behaviors.

Dataset-Specific Attention Guidance. The user input instruction is tailored to steer the model's attention mechanism towards feature fusion strategies appropriate for each dataset. We highlight a critical distinction in the Composed Query prompt: whereas the CIRR template employs a generalized modification instruction suitable for open-domain objects, the FashionIQ template integrates category-aware phrasing (e.g., *"Change the style of this {Category}..."*) to enhance domain specificity and attribute sensitivity.

7.2. Prompts for VQA-Assisted Quality Control

The generative calibration process described in Sec. 3.4 entails an inherent risk of synthesizing hallucinated or visually ungrounded corrective triplets. To attenuate this noise, we implement a VQA-Assisted Quality Control mechanism, repurposing the Foundation Model ($\mathcal{F}$) to function as a rigorous visual verifier. This step necessitates a specialized VQA prompt designed to validate the semantic alignment between the synthesized modified instruction ($\tilde{T}_m$) and the actual informative instance (I_h).

Fig. 8 illustrates the prompt structure engineered for this verification task. Our design relies on two key mechanisms:

Strict Binary Constraint. The prompt explicitly constrains the model's output space, mandating a single, lowercase token response (`yes` or `no`). This binary restriction inhibits the model's open-ended generative tendencies.

Discriminative Reasoning Activation. By disabling the generative mode, the constraint compels the model to perform critical discriminative reasoning to verify semantic

Prompt for VQA-Assisted Quality Control

[User Input] You are a strict visual verifier. Output exactly one token: yes or no (lowercase).Do not add punctuation or explanations.
Reference image: <Reference Image> Candidate image: <Candidate Image>
Instruction:{Modification Text}
Decide if the candidate image matches the result of applying the instruction to the reference image.
Return yes if all required elements implied by the instruction are satisfied (like counts, categories, attributes, spatial relations). If any required element is missing or contradicted, answer no.
Answer:

Figure 8. Prompt template for the VQA-Assisted Quality Control mechanism. This zero-shot prompt conditions the Foundation Model to act as a strict binary verifier, ensuring the quality of the synthesized informative triplets.

consistency. This serves as a robust filter, ensuring that only high-fidelity informative instances are admitted into the final refinement stage.

7.3. Prompts for CoT-Assisted Instruction Synthesis

We provide the complete Chain-of-Thought (CoT) prompts utilized in Stage 3: Generative Calibration (see Sec. 3.4) to synthesize high-fidelity corrective supervision. Fig. 15 visualizes the prompt architectures for both datasets.

Structured Reasoning Constraints. Unlike standard open-ended captioning, our templates impose rigorous constraints through explicit *Key Principles* and a mandatory *JSON Output Schema*. This structured design compels the Foundation Model to engage in a sequential reasoning process: it must first perform Intent Decomposition & Verification before executing Minimal Edit Synthesis. This mechanism ensures that the generated instruction is not merely a hallucinated caption, but a precise modification strictly grounded in the observed visual discrepancies.

Domain-Specific Adaptation. To accommodate the distinct characteristics of the benchmarks, the prompts are domain-adapted. The CIRR prompt is engineered to reason about complex object relations, cardinalities, and spatial states, whereas the FashionIQ prompt is optimized for fine-grained attribute manipulation, focusing on nuanced details such as texture, silhouette, and pattern.

8. Additional Qualitative Analysis and Visualization

8.1. Additional Baseline Comparisons

In this section, we present an expanded qualitative comparison between the baseline retriever ($\mathcal{R}_{base}$) and our refined model ($\mathcal{R}_{refine}$) to further illustrate the impact of capability recalibration. Figs. 9 and 10 showcase top-ranked retrieval results on the CIRR and FashionIQ datasets, respectively. In each panel, the left column displays the multimodal query, highlighting the critical modification instructions, while the right columns compare the top retrieved candidates from both models. The ground-truth targets are highlighted with green bounding boxes.

The results on CIRR (Fig. 9) clearly expose the coarse-grained tendency of the baseline model. While $\mathcal{R}_{base}$ correctly identifies the main object category (e.g., food, llamas, or safety pins), it frequently collapses on fine-grained spatial or state-based constraints. A striking example is Case 3, where the instruction demands a specific arrangement of safety pins ("opened and closed... side by side"). The baseline merely retrieves isolated pins or incorrect states, whereas ReCALL accurately reasons about the requested object configuration. Similarly, in Case 2, ReCALL respects the contextual constraint ("mountainous area"), whereas the baseline retrieves semantically relevant but visually inconsistent backgrounds. This validates that our framework effectively internalizes the complex logic required for open-domain compositional reasoning.

Parallel observations on FashionIQ (Fig. 10) demonstrate ReCALL's superiority in fine-grained attribute manipulation. The baseline often succumbs to visual biases, retrieving images that match the reference image's dominant features (such as color or shape) but ignoring the text modifier. For instance, in Case 1, although the instruction explicitly specifies "striped", the baseline is dominated by the solid green color of the reference. ReCALL, having been trained on generated hard negatives, successfully suppresses this bias to retrieve the correct textured garment. Furthermore, Case 3 highlights the model's ability to handle rigorous category shifts ("is a scarf and not a long dress"), where the baseline fails to disengage from the visual semantics of the reference dress. These comparisons confirm that ReCALL successfully recalibrates capability degradation, restoring the model's native ability to adhere to precise textual instructions.

8.2. Visualization of Informative Instance Mining and Triplet Synthesis

In this section, we provide additional qualitative visualizations to further substantiate the efficacy of the ReCALL

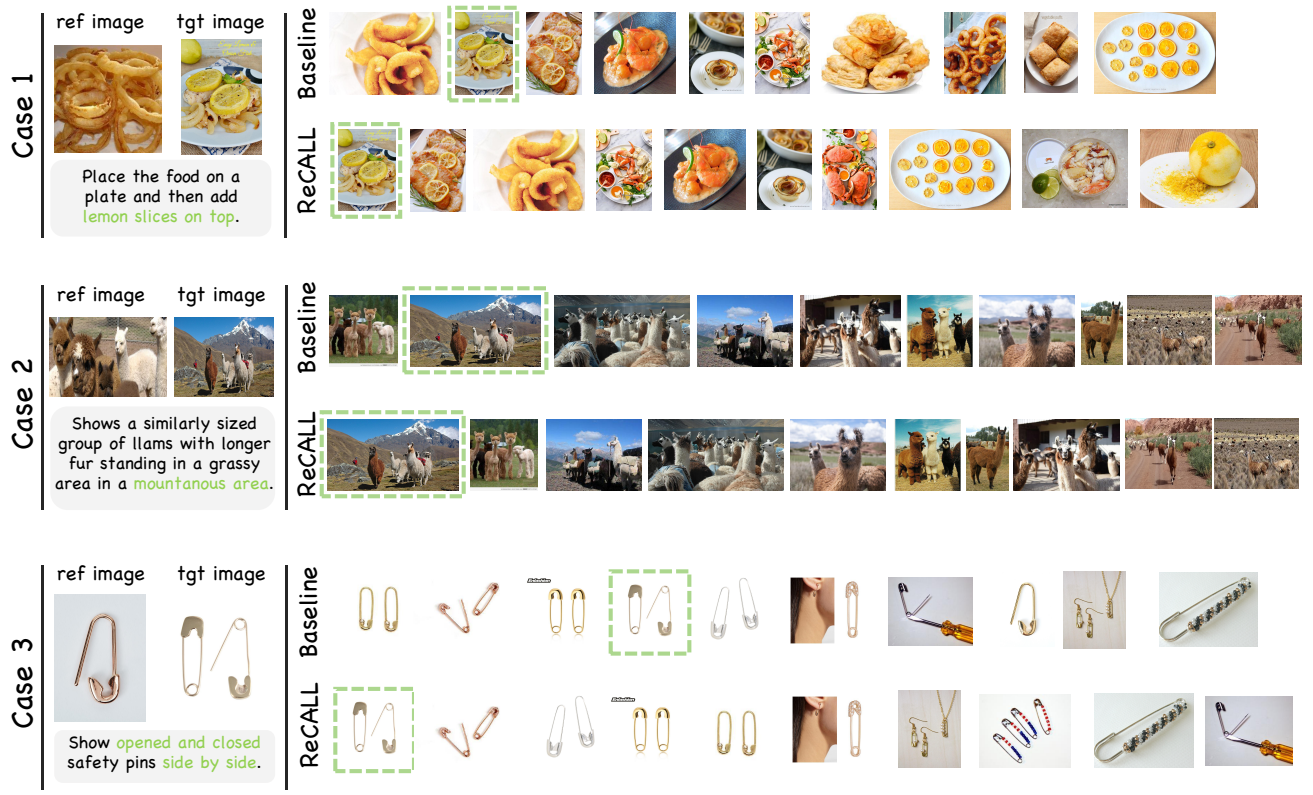

Figure 9. **Qualitative comparison on the CIRR dataset.** We compare the top retrieved images from the baseline ($\mathcal{R}_{base}$) and ReCALL ($\mathcal{R}_{refine}$). The green dashed boxes indicate the ground-truth targets. The baseline model tends to focus on the primary object but misses specific constraints (e.g., the "mountainous area" in Case 2 or the specific "opened and closed" state in Case 3). In contrast, ReCALL successfully retrieves the correct targets by reasoning about the fine-grained details in the modification text.

framework. Figs. 13 and 14 present a detailed breakdown of the data construction pipeline on the CIRR and FashionIQ datasets, respectively. Unlike the schematic overview in the main paper, these figures showcase specific real-world examples where the baseline model ($\mathcal{R}_{base}$) initially fails, tracing the complete trajectory from failure diagnosis to the synthesis of corrective training signals.

The visualizations are organized to reflect the *Diagnose-Generate-Refine* workflow. As shown in the *Original Triplet* panel, the highlighted text (marked in red) indicates specific fine-grained constraints that the baseline retriever ignored, leading to the retrieval of the false positives shown in the *Informative Instances* panel. Crucially, these mined instances reveal distinct failure modes: while some queries are confused by a single distinct distractor (e.g., Case 1 in Fig. 13), others suffer from multiple high-confidence hard negatives (e.g., Case 2 in Fig. 13), necessitating the generation of multiple targeted corrective triplets.

By employing CoT-assisted generation, ReCALL explicitly verbalizes these visual discrepancies. The *Synthesized Corrective Triplet* panel demonstrates the precision of this process, where the generated instructions (with modifications highlighted in green) strictly adhere to the visual evidence of the mined instances. For example, in the CIRR dataset (Fig. 13), the model successfully disambiguates complex spatial relations ("stands" vs. "sits" vs. "lies") and fine-grained object categories ("ball" vs. "stuffed toy"). Similarly, in the FashionIQ dataset (Fig. 14), the synthesized triplets capture subtle attribute nuances, such as distinguishing "white polka dots" from a "white floral print" despite similar dress silhouettes. These qualitative results confirm that the synthesized supervision is both semantically dense and visually grounded, effectively guiding the model to recalibrate its decision boundaries.

8.3. Failure Case Analysis

To provide a comprehensive understanding of limitations, we visualize representative failure cases of ReCALL on FashionIQ and CIRR in Figs. 11 and 12. An analysis of these instances reveals that the "failures" often stem from the inherent ambiguity of natural language instructions and the incompleteness of ground-truth annotations, rather than a fundamental breakdown of the model's reasoning.

False Negatives and Annotation Issues. A signifi-

Figure 10. **Qualitative comparison on the FashionIQ dataset.** Comparison of top retrieval results between the baseline and ReCALL. Ground-truth targets are highlighted in green. These examples illustrate how ReCALL overcomes the baseline's tendency to ignore textual modifiers. For instance, in Case 1, ReCALL correctly attends to the "striped" pattern attribute, and in Case 3, it successfully executes a category shift from a dress to a scarf, whereas the baseline remains fixated on the reference image's category.

cant portion of retrieval errors, particularly on FashionIQ (Fig. 11), can be attributed to the False Negative problem. In CIR tasks, datasets typically annotate a single ground-truth target per query. However, in large-scale galleries, multiple images may validly satisfy the modification instruction. For instance, in Case 2 of Fig. 11, the instruction requests a dress with "no sleeves" that is "white and short". ReCALL retrieves several valid candidates (Rank 1-4) that perfectly match this description. Yet, because they differ from the specific ground-truth instance (which is not in the top-10), they are penalized as errors. Similarly, in Case 3, the model retrieves multiple "red shirts with printed words", all semantically correct despite not being the annotated target. This suggests that the reported performance metrics may underestimate the model's actual retrieval utility.

Ambiguity in Instructions. Certain directives such as "different pattern" (Case 1 in Fig. 11) or "fewer animals" (Case 1 in Fig. 12) are inherently subjective. In the latter case, ReCALL retrieves images with small groups of birds, which is a valid interpretation of "fewer" compared to a large flock, even if it doesn't match the exact count of the ground truth. The model struggles to align its threshold for these relative terms with the annotator's intent.

Fine-grained Spatial Reasoning. While ReCALL significantly improves spatial understanding, it still faces challenges with complex geometric transformations. As shown in Case 3 of Fig. 12, the instruction requires rotating a stingray so its head faces "upward". While the model retrieves stingrays with varying orientations, it fails to consistently isolate the specific "upward" pose. This limitation likely stems from the Foundation Model, which, despite its strength, may still have residual weaknesses in zero-shot spatial rotation reasoning that are inherited by the retriever.

In summary, while ReCALL effectively recalibrates compositional reasoning, future work could focus on mitigating label noise through one-to-many evaluation pro-

Figure 11. **Failure cases on the FashionIQ dataset.** We display the top retrieved candidates by ReCALL for queries where the ground-truth (GT) target was not found in the top-10. In many instances (e.g., Case 2 and Case 3), the retrieved images are actually valid matches that satisfy the text modification (False Negatives), highlighting the issue of sparse ground-truth annotations in the dataset. Text in red indicates the key modification constraints.

Figure 12. **Failure cases on the CIRR dataset.** Representative errors showing challenges with ambiguous instructions (e.g., "fewer" in Case 1) and complex spatial rotations (e.g., "facing upward" in Case 3). The green dashed boxes indicate the ground-truth target if it appears in the top candidates; otherwise, the text "GT not in Top-10" is displayed.

tocols and further enhancing the spatial geometric understanding of the backbone itself.

Figure 13. **Visualization of the informative instance mining and triplet synthesis process on the CIRR dataset.** This figure illustrates representative failure cases of the baseline retriever. From left to right, the panels display: (1) the Original Triplet, where red text highlights constraints violated by hard negatives; (2) the mined Informative Instances (I_h), representing the model's cognitive blind spots; (3) the Corrective Instruction generated via CoT; and (4) the final Synthesized Corrective Triplet, where green text denotes the minimal semantic edits required to align the instruction with the mined instance.

Figure 14. **Visualization of the informative instance mining and triplet synthesis process on the FashionIQ dataset.** This figure illustrates representative failure cases of the baseline retriever. From left to right, the panels display: (1) the Original Triplet, where red text highlights constraints violated by hard negatives; (2) the mined Informative Instances (I_h), representing the model's cognitive blind spots; (3) the Corrective Instruction generated via CoT; and (4) the final Synthesized Corrective Triplet, where green text denotes the minimal semantic edits required to align the instruction with the mined instance.

Prompt for CoT-Assisted Generation on CIRR

[System Instruction] You are a multimodal edit refiner. You receive TWO images (Picture 1 = REFERENCE, Picture 2 = TARGET) and ONE original edit text.
Goal: produce a minimally edited rewrite that stays stylistically close to the original text while reflecting the TARGET image facts.
Key principles:
- Look directly at Picture 1 and Picture 2 to gather evidence.
- Preserve all parts of the original text that remain correct; only replace spans that contradict the TARGET image.
- Rewritten spans must remain short and natural, matching the tone and grammar of the original sentence.
- When replacing an object (animal, item, etc.), include a visible distinguishing detail such as color, pattern, or breed if it is clear in the TARGET image.
- When uncertainty exists, explicitly state "uncertain" in the relevant rewritten span.
- Output strict JSON exactly matching the required schema; do not emit explanations outside the JSON.
[User Input]

```
{
  "input": {
    "reference_image": <Reference Image>,
    "target_image": <Target Image>,
    "original_edit_text": <Modification Text>  },
  "tasks": [
    "Describe concise visual facts for each picture",
    "List spans from the original edit text that must change",
    "Produce minimal replacements based on TARGET evidence",
    "Return the final rewritten text by replacing those spans only"
  ],
  "output_schema": {
    "visual_summary": {
      "reference": "short description focusing on key objects/attributes in Picture 1",
      "target": "short description focusing on key objects/attributes in Picture 2",
      "differences": ["bullet-like strings highlighting major contrasts"]
    },
    "rewrite_segments": [
      {
        "original_span": "exact substring copied from original_edit_text",
        "new_span": "replacement phrase grounded in TARGET",
        "reason": "one-line justification referencing visible evidence"
      }
    ],
    "final_text": "string"
  },
  "constraints": [
    "Every original_span must appear verbatim in original_edit_text",
    "Do not invent spans that are not in the original text",
    "If no change needed, return an empty list for rewrite_segments and set final_text equal to original_edit_text",
    "When changes are required, final_text must equal original_edit_text with each original_span replaced by new_span in order",
    "If the text contains 'instead of', 'replace ... with ...', or 'swap ... for ...', treat the whole contrast clause as a single span and rewrite it so the comparison is correct for the TARGET image (or remove it if no contrast remains)",
    "Do not leave contradictory contrast phrases (e.g., 'instead of ...') when the rewritten subject and contrast refer to the same object",
    "If the contrast is no longer needed after rewriting, remove the entire 'instead of/replace/swap' clause",
    "When replacing an animal or object noun, include a specific visible attribute (e.g., color, pattern, breed) if the TARGET image makes it clear",
    "Keep wording style consistent with the original text",
    "Return valid JSON only"
  ]
}
```

(a) Prompt for CoT-Assisted Generation on CIRR

Prompt for CoT-Assisted Generation on Fashion-IQ

[System Instruction] You are a multimodal fashion edit refiner. You receive TWO clothing images (Picture 1 = REFERENCE garment, Picture 2 = TARGET garment) and ONE original edit text.
Goal: produce a minimally edited rewrite that stays stylistically close to the original text while reflecting the TARGET garment facts.
Key principles:
- Look directly at Picture 1 and Picture 2 to gather evidence.
- Focus strictly on visible garment attributes: category (dress, shirt, toptee, outerwear, etc.), color/shade, pattern/print (solid, striped, plaid, floral, polka dots, graphic), silhouette/fit (slim, regular, oversized, bodycon, A-line), length (sleeve length and hemline), neckline/collar (crew, v-neck, turtleneck, polo, collared, collarless), materials/fabric (denim, knit, cotton, silk, chiffon, leather), construction/details (buttons, zipper, pockets, ruffles, pleats, lace, bow, trim), layering if clearly visible.
- Ignore background, human identity, face, pose, lighting, and camera angle unless they directly change how the garment appears (e.g., sleeve visibility) or fit.
- Do not mention brand names or logos unless they are clearly visible and legible; otherwise mark as "uncertain".
- Preserve all parts of the original text that remain correct; only replace spans that contradict the TARGET garment.
- Rewritten spans must remain short and natural, matching the tone and grammar of the original sentence.
- When replacing a garment/object noun, include a visible distinguishing detail such as color, pattern,or material if it is clear in the TARGET image.
- When uncertainty exists, explicitly state "uncertain" in the relevant rewritten span.
- Output strict JSON exactly matching the required schema; do not emit explanations outside the JSON.
[User Input]

```
{
  "input": {
    "reference_image": <Reference Image>,
    "target_image": <Target Image>,
    "original_edit_text": <Modification Text>
  },
  "tasks": [
    "Describe concise visual facts for each garment (category, color, pattern, silhouette/fit, length, neckline/collar, materials, and key details)",
    "List spans from the original edit text that must change to match the TARGET garment",
    "Produce minimal replacements grounded in TARGET garment evidence using concise fashion terminology",
    "Return the final rewritten text by replacing only those spans"
  ],
  "output_schema": {
    "visual_summary": {
      "reference": "short garment description for Picture 1 (attributes listed above as applicable)",
      "target": "short garment description for Picture 2 (attributes listed above as applicable)",
      "differences": ["bullet-like strings highlighting major garment contrasts"]
    },
    "rewrite_segments": [
      {
        "original_span": "exact substring copied from original_edit_text",
        "new_span": "replacement phrase grounded in TARGET garment attributes",
        "reason": "one-line justification referencing visible garment evidence"
      }
    ],
    "final_text": "string"
  },
  "constraints": [
    "Every original_span must appear verbatim in original_edit_text",
    "Do not invent spans that are not in the original text",
    "If no change needed, return an empty list for rewrite_segments and set final_text equal to original_edit_text",
    "When changes are required, final_text must equal original_edit_text with each original_span replaced by new_span in order",
    "Keep wording style consistent with the original text",
    "Return valid JSON only"
  ]
}
```

(b) Prompt for CoT-Assisted Generation on FashionIQ

Figure 15. CoT prompts for Generative Calibration. To implement the diagnose-generate-refine pipeline, we design structured prompts that guide the Foundation Model to explicitly reason about visual discrepancies between the target and the informative instance. The enforced JSON output format ensures that the generated corrective instructions ($\tilde{T}_m$) are both stylistically natural and semantically precise.