ConeSep: Cone-based Robust Noise-Unlearning Compositional Network for Composed Image Retrieval

Supplementary Material

This is the appendix of "ConeSep: Cone-based Robust Noise-Unlearning Compositional Network for Composed Image Retrieval".

- **Appendix A**: Sinkhorn-Knopp Algorithm for Targeted Unlearning
- **Appendix B**: Datasets
- **Appendix C**: Additional Quantitative Analyses
  - **Appendix C.1**: Efficiency Evaluation Analysis
  - **Appendix C.2**: Quantitative Analysis of Diagonal Negative Composition
  - **Appendix C.3**: Comparison of Sample Set Partition Purity
  - **Appendix C.4**: Additional Hyperparameter Analysis
- **Appendix D**: Additional Ablation Study
  - **Appendix D.1** GFQ Sampling Ablation Studies
- **Appendix E**: Algorithm of Training Procedure
- **Appendix F**: More Qualitative Results
  - **Appendix F.1**: Qualitative Analyses of Diagonal Negative Composition
  - **Appendix F.2**: NTC Identification Analysis
  - **Appendix F.3**: More Case Study
- **Appendix G**: More Related Work

A. Sinkhorn-Knopp Algorithm for Targeted Unlearning

In Section 3.4 of the main text, we formulate the boundary-based targeted unlearning as an entropy-regularized Optimal Transport (OT) problem (Eq. 10). Here, we provide the detailed derivation of the solution using the Sinkhorn-Knopp algorithm.

Problem Definition. We aim to find the optimal transport plan $\mathbf{P}^* \in \mathbb{R}^{B\times 2B}$ that minimizes the transport cost regarding $\mathbf{C}_{masked}$ while maximizing entropy. Following the notation in the main text, the objective is:

$$\mathbf{P}^* = \arg\min_{\mathbf{P}\in\Pi(\mathbf{u},\mathbf{v})} s(\mathbf{P}, \mathbf{C}_{masked}) - \epsilon H(\mathbf{P}), \tag{1}$$

where $s(\mathbf{P}, \mathbf{C}_{masked}) = \langle \mathbf{P}, \mathbf{C}_{masked}\rangle = \sum_{ij} \mathbf{P}_{ij}[\mathbf{C}_{masked}]_{ij}$ represents the total transport cost. The feasible set $\Pi(\mathbf{u}, \mathbf{v})$ is defined by the marginal constraints:

$$\Pi(\mathbf{u}, \mathbf{v}) = \{\mathbf{P} \in \mathbb{R}_+^{B\times 2B} \mid \mathbf{P}\mathbf{1}_{2B} = \mathbf{u}, \mathbf{P}^\top \mathbf{1}_B = \mathbf{v}\}, \tag{2}$$

where $\mathbf{u} \in \mathbb{R}^B$ and $\mathbf{v} \in \mathbb{R}^{2B}$ are the source and target marginal distributions (typically uniform, i.e., $\mathbf{u} = \frac{1}{B}\mathbf{1}_B, \mathbf{v} = \frac{1}{2B}\mathbf{1}_{2B}$).

Lagrangian and Gibbs Kernel. To solve this constrained optimization problem, we introduce Lagrange multipliers $\boldsymbol{\alpha} \in \mathbb{R}^B$ and $\boldsymbol{\beta} \in \mathbb{R}^{2B}$ for the marginal constraints. The Lagrangian is given by:

$$\begin{aligned}\mathcal{L}(\mathbf{P}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{ij} \mathbf{P}_{ij}\mathbf{C}_{ij} + \epsilon \sum_{ij} \mathbf{P}_{ij}(\log \mathbf{P}_{ij} - 1) \\ + \boldsymbol{\alpha}^\top(\mathbf{P}\mathbf{1} - \mathbf{u}) + \boldsymbol{\beta}^\top(\mathbf{P}^\top\mathbf{1} - \mathbf{v}).\end{aligned} \tag{3}$$

Taking the derivative with respect to $\mathbf{P}_{ij}$ and setting it to zero:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{P}_{ij}} = \mathbf{C}_{ij} + \epsilon \log \mathbf{P}_{ij} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j = 0. \tag{4}$$

Solving for $\mathbf{P}_{ij}$, we obtain the form of the optimal solution:

$$\mathbf{P}^*_{ij} = \exp\left(-\frac{\boldsymbol{\alpha}_i}{\epsilon}\right) \exp\left(-\frac{\mathbf{C}_{ij}}{\epsilon}\right) \exp\left(-\frac{\boldsymbol{\beta}_j}{\epsilon}\right). \tag{5}$$

We define the **Gibbs Kernel matrix** $\mathbf{K}$ as $\mathbf{K}_{ij} = \exp(-[\mathbf{C}_{masked}]_{ij}/\epsilon)$. By letting the scaling vectors be $\mathbf{a}_i = \exp(-\boldsymbol{\alpha}_i/\epsilon)$ and $\mathbf{b}_j = \exp(-\boldsymbol{\beta}_j/\epsilon)$, the optimal plan factorizes into:

$$\mathbf{P}^* = \mathrm{diag}(\mathbf{a})\mathbf{K}\mathrm{diag}(\mathbf{b}). \tag{6}$$

Iterative Scaling Updates. The unknown scaling vectors $\mathbf{a}$ and $\mathbf{b}$ must be determined such that $\mathbf{P}^*$ satisfies the marginal constraints. Substituting the factorization into the constraints yields:

$$\mathbf{a} \odot (\mathbf{K}\mathbf{b}) = \mathbf{u}, \quad \mathbf{b} \odot (\mathbf{K}^\top\mathbf{a}) = \mathbf{v}. \tag{7}$$

These equations are solved via the Sinkhorn-Knopp fixed-point iteration. In each step t, we update:

$$\mathbf{a}^{(t+1)} \leftarrow \frac{\mathbf{u}}{\mathbf{K}\mathbf{b}^{(t)}}, \tag{8}$$

$$\mathbf{b}^{(t+1)} \leftarrow \frac{\mathbf{v}}{\mathbf{K}^\top\mathbf{a}^{(t+1)}}, \tag{9}$$

where division is element-wise. The algorithm typically converges within 10-20 iterations. The final smooth transport plan $\mathbf{P}^*$ is then used as the soft target Y (Eq. 11) to guide the targeted unlearning.

B. Datasets

We evaluate the performance of ConeSep on two widely recognized benchmarks. The detailed descriptions are as follows:

- **FashionIQ** [36] is a premier CIR benchmark dedicated to the fashion domain. It comprises 77,684 high-resolution fashion images and 30,134 annotated triplets. These triplets are explicitly categorized into three distinct subsets: *Dresses*, *Shirts*, and *Tops&Tees*. The primary challenge of this dataset lies in its modification texts, which focus on describing fine-grained visual attribute manipulations (e.g., "change to V-neck" or "denser stripes"), requiring the model to capture subtle semantic differences.
- **CIRR** [37] is a large-scale, open-domain benchmark with images derived from the classic NLVR2 [102] dataset. It contains a total of 21,552 unique images and 36,554 annotated triplets. In contrast to FashionIQ, CIRR presents distinct challenges: (1) its images depict complex real-world scenes containing multiple objects; (2) the modification texts involve not only attribute changes but also diverse compositional logic such as spatial relationships and object addition/deletion. Furthermore, to mitigate the issue of false negatives, CIRR provides a specific candidate subset for each query during evaluation.

C. Additional Quantitative Analysis

C.1. Efficiency Evaluation Analysis

To comprehensively evaluate the feasibility and cost-effectiveness of the model in practical scenarios, we employed multi-dimensional evaluation metrics in Table 4 to conduct a detailed comparison of various methods. Specifically, **FLOPs(G)** quantifies the theoretical computational complexity during a single forward pass, while **Parameters(M)** reflects the model's parameter size and storage footprint. Regarding time and resource dimensions, **Test time(s/sample)** records the average latency for processing a single query sample during inference, directly correlating with the real-world deployment experience; **Train Time(s/iteration)** measures the time cost per iteration during training; and **GPU Memory(MiB)** indicates the peak GPU memory usage under a fixed batch size, reflecting hardware entry barriers. Furthermore, we incorporate **FashionIQ-Avg** and **CIRR-Avg** as core performance indicators to comprehensively assess the Efficiency-Performance Trade-off.

Comparative analysis based on the aforementioned metrics demonstrates that ConeSep achieves improvements in both inference efficiency and retrieval performance while maintaining a comparable parameter scale and computational load. Regarding model complexity, ConeSep 's parameter count (915.69M) and FLOPs (411.51G) are comparable to the ordinary baseline SPRC and the robust baseline TME, indicating that the performance gains do not stem from a blind expansion of model scale. Notably, ConeSep exhibits superior efficiency during the inference phase, with a single-sample test time of only 0.0091 seconds. Benefiting from our core modules (GFQ, NBL, BTU) applying constraints solely during training without participating in inference calculations, ConeSep 's inference speed is slightly faster than SPRC (0.011s) and approximately $13.6\times$ faster than TME (0.124s). Regarding training overhead, although introducing robustness constraints results in higher GPU memory usage (25807 MiB) and iteration time (5.15s) for ConeSep compared to the simple SPRC, our training efficiency is improved by approximately 34.5% compared to the robust counterpart TME. Crucially, ConeSep exchanges reasonable training resources for significant performance returns, achieving the best retrieval accuracy on both benchmark datasets (FashionIQ-Avg=64.97%, CIRR-Avg=80.43%). In summary, ConeSep successfully unifies SOTA-level robustness with inference efficiency, demonstrating substantial value for real-world deployment.

C.2. Quantitative Analysis of Diagonal Negative Composition

In order to further verify whether the "Diagonal Negative Composition" ($\mathbf{F}_{neg}$) constructed by the Negative Boundary Learning (NBL) module truly serves as an effective negative anchor, we perform quantitative analyses on its geometric separability from the original composed feature ($\mathbf{F}_c$). Specifically, we calculate the cosine similarity distribution for all sample pairs $(\mathbf{F}_c, \mathbf{F}_{neg})$ on the FashionIQ validation set and visualize it as a histogram and a Kernel Density Estimation (KDE) curve, as shown in Figure 5. Observing this figure, we find that the similarity values exhibit extremely high concentration. The similarities for the vast majority of samples are tightly distributed within the narrow range of $[0, 0.1]$, and the distribution peak significantly approaches 0. This statistical result strongly demonstrates that $\mathbf{F}_{neg}$ maintains a strictly Approximate Orthogonality relationship with $\mathbf{F}_c$ in the high dimensional feature space. This phenomenon is consistent with the constraint optimization objective that we design through $\mathcal{L}_{intra}$ in the methodology section (Section 3.3), which is to force the learned negative anchor to be orthogonal to the query in terms of semantic direction. This orthogonality ensures that $\mathbf{F}_{neg}$ neither contains the query's effective semantics (thus avoiding the loss of positive knowledge) nor overlaps with the target image features, thereby successfully constructing a clear, independent structured boundary that provides precise directional guidance for the Targeted Unlearning (BTU) module.

C.3. Comparison of Sample Set Partition Purity

In Section 3.2, we propose the Geometric Fidelity Quantization (GFQ) module, which aims to distinguish the high-fidelity clean set ($\mathcal{T}_{clean}$) from the low-fidelity noisy set ($\mathcal{T}_{noisy}$) by estimating geometric noise boundaries. However, our subsequent Boundary-based Targeted Unlearning

Table 4. Comparison of computational complexity and efficiency among SPRC, TME, and ConeSep with its ablation variants.

Type	Method	FLOPs(G)	Parameters(M)	GPU Memory(MiB)	Test time(s/sample)	Train Time(s/iteration)	FashionIQ-Avg	CIRR-Avg
Ordinary	SPRC	413.38	915.69	24478(bs=128)	0.011	**2.624(bs=128)**	56.33	76.98
Robust	TME	**405.2**	**915.68**	**12405(bs=128)**	0.124	7.858(bs=128)	63.97	79.74
	ConeSep (Ours)	411.51	915.69	25807(bs=128)	**0.0091**	5.15(bs=128)	**64.97**	**80.43**

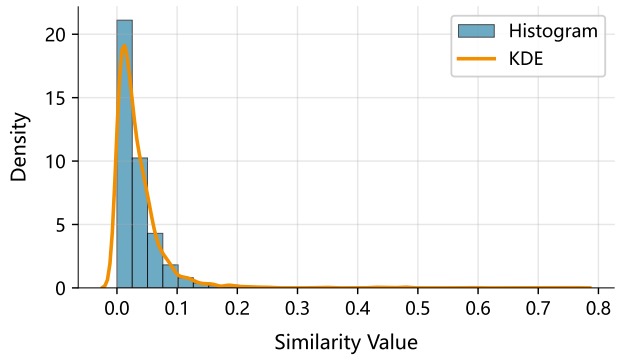

Figure 5. Visualization of the cosine similarity distribution between the composed feature $\mathbf{F}_c$ and the learned diagonal negative composition $\mathbf{F}_{neg}$. The distribution is highly concentrated around 0, indicating an orthogonal relationship.

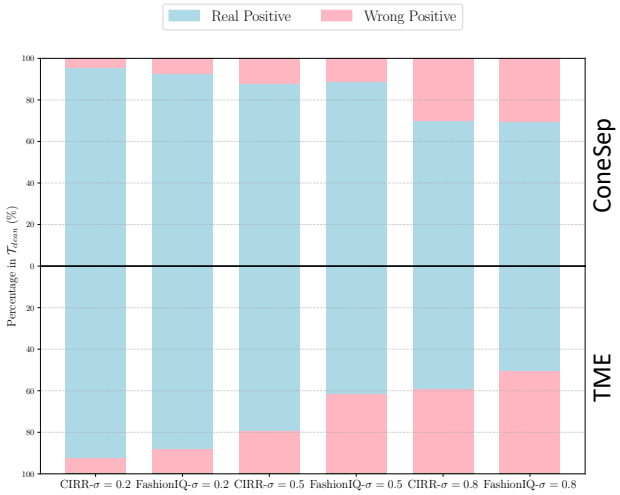

Figure 6. Quantitative analyses and comparison of the internal purity of the $\mathcal{T}_{clean}$ set. This figure demonstrates the composition of the "high-fidelity" set selected by the model under different initial noise rates σ. The blue bars (Real Positive) represent the genuine clean samples correctly identified, while the pink bars (Wrong Positive) represent the hard noise mistakenly classified as clean samples. The upper part shows our **ConeSep** model, and the lower part shows the **TME** baseline model. A smaller proportion of the pink area indicates a higher purity of the selected set. The results show that ConeSep maintains a high purity even under high noise and significantly outperforms TME.

(BTU) module does not directly treat the filtered $\mathcal{T}_{clean}$ as absolute Ground Truth. Instead, it employs an optimal transport (OT) based soft correction mechanism. The core premise of this design is that, constrained by the "Modality Suppression" challenge, a portion of "hard noise" samples are theoretically inevitable within $\mathcal{T}_{clean}$. These hard noise samples are instances where the reference image and the target image exhibit high visual similarity, but the modification text is incorrect. To validate this premise and simultaneously evaluate the efficacy of our filtering strategy compared to existing methods, we conduct an in-depth quantitative analyses of the purity of $\mathcal{T}_{clean}$ in this section.

Since real-world datasets lack Ground Truth annotations for Noisy Triplet Correspondence (NTC), we conduct experiments on the FashionIQ and CIRR datasets using synthetic noise settings based on $\sigma \in \{0.2, 0.5, 0.8\}$. This setup allows us to precisely track which samples are artificially injected noise. Figure 6 intuitively illustrates the composition of the sample set identified by the model as high fidelity. The blue region represents the genuinely clean samples that are correctly identified (Real Positive), while the pink region represents the noise triplets that are incorrectly classified as clean samples (Wrong Positive, i.e., hard noise). The upper part of the figure shows the performance of our ConeSep at different noise rates, and the lower part shows the performance of TME [1], the current state-of-the-art robust baseline model.

By observing Figure 6, we clearly find a significant difference in noise resistance between ConeSep (upper part) and the baseline method TME (lower part). TME exhibits noticeable performance degradation in high-noise environments. In the $\sigma = 0.8$ setting, the pink region in its filtered set expands, even approaching the proportion of the blue region on FashionIQ. This indicates that the Gaussian Mixture Model (GMM) upon which TME relies cannot effectively distinguish samples that are visually similar but semantically inconsistent when facing substantial hard noise. In contrast, our ConeSep demonstrates superior robustness. Benefiting from the effective penetration of geometric noise boundary against Modality Suppression, the blue region remains dominant in ConeSep's filtering results, even under extremely high noise interference. This strongly proves the effectiveness of the GFQ module in decouple features and noise identification, and the purity of its filtered samples is significantly better than existing SOTA methods.

Although ConeSep's filtering purity is clearly superior to others, a small amount of pink residue is still observable in the figure. This objectively reflects the inherent challenge of "hard noise" in the NTC task: some samples are dif-

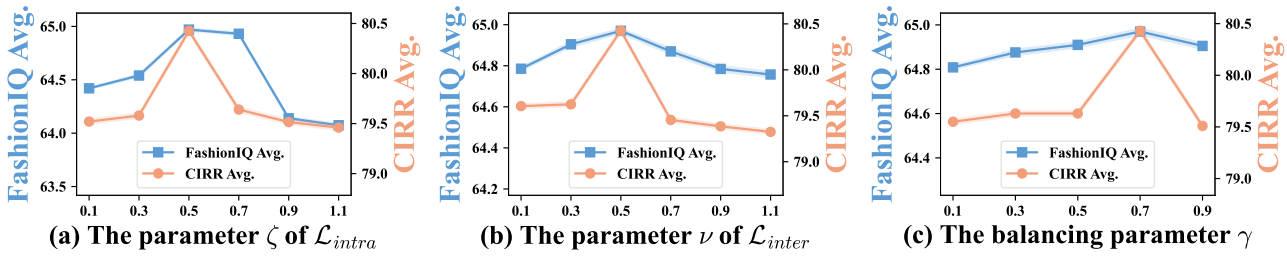

Figure 7. Sensitivity analysis of hyperparameters on FashionIQ and CIRR datasets: (a) the intra-modal loss weight ζ, (b) the inter-modal loss weight ν, and (c) the balancing parameter γ.

ficult to be completely "hard" cut off by any filter due to their extremely high visual similarity. This observation provides solid empirical evidence for the Boundary-based Targeted Unlearning (BTU) module designed in Section 3.4. Precisely because we recognize that Hard Filtering cannot theoretically achieve 100% perfection, we do not directly treat $\mathcal{T}_{clean}$ as absolute Ground Truth. Instead, we introduce an optimal transport-based soft correction mechanism. This mechanism robustly tolerates and handles the residual impurities that are unavoidable during the filtering stage, thereby maximizing the reduction of the risk of overfitting the residual noise while utilizing the clean sample information, achieving training robustness in a noisy environment.

C.4. Additional Hyperparameter Analysis

In this section, we further discuss the parameter sensitivity of three additional crucial hyperparameters in ConeSep: the intra-modal loss weight ζ and inter-modal loss weight ν within the NBL module, and the balancing parameter γ within the BTU module. We evaluate the impact of these parameters on the FashionIQ and CIRR datasets, and the results are illustrated in Figure 7.

a) Intra-modal loss weight ζ**.** As shown in subplot (a), the model performance exhibits a trend of increasing initially and then decreasing as the value of ζ increases, peaking at 0.5. ζ controls the weight of $\mathcal{L}_{intra}$, a loss designed to constrain the similarity between the composed feature $\mathbf{F}_c$ and its diagonal negative composition $\mathbf{F}_{neg}$, ensuring they remain orthogonal in the metric space. When ζ is too low, the model fails to effectively shape $\mathbf{F}_{neg}$ as the opposite of the query semantics, resulting in a blurred negative boundary. Conversely, when ζ is too high, the enforced orthogonal constraint may distort the geometric structure of the feature space, thereby interfering with the core retrieval paradigm learning in the Positive Alignment Path.

b) Inter-modal loss weight ν**.** As shown in subplot (b), the model achieves optimal performance at $\nu = 0.5$. ν regulates the weight of $\mathcal{L}_{inter}$, which is responsible for pushing $\mathbf{F}_{neg}$ away from its corresponding target image $\mathbf{F}_t$ through *Target-oriented Learning*. This trend indicates that a moderate ν value facilitates the construction of a high-quality Diagonal Negative Composition, serving as an effective directional guide. However, an excessively high ν causes the optimization process to over-focus on constructing the negative boundary, distracting the model from pulling positive pairs closer via $\mathcal{L}_{rank}$, thus compromising the final retrieval accuracy.

c) The balancing parameter γ**.** As shown in subplot (c), the model performance improves steadily as γ increases from 0.1 to 0.7, followed by a decline. γ balances the global optimal plan $\mathbf{P}$ derived from Optimal Transport (OT) and the original hard label $\mathbf{L}$ when constructing the smooth soft label $\mathbf{Y}$. A lower γ implies that the model over-relies on the hard label $\mathbf{L}$, ignoring the geometric structural information provided by OT, which leads to a lack of smoothness in the unlearning process. On the other hand, an overly high γ may weaken the explicit "targeted unlearning" instruction for noisy samples inherent in the hard labels (i.e., the strong supervisory signal in $\mathbf{L}$ where the diagonal for noisy samples is set to 0), resulting in insufficient correction strength for the noise.

D. Additional Ablation Study

D.1. GFQ Sampling Ablation Studies

Effectiveness of Gaussian Sampling Strategy. Impact of Boundary Sampling Strategies. To strictly validate the theoretical premise of our *Geometric Fidelity Quantization*, we conduct a comprehensive comparison of boundary estimation strategies on both FashionIQ and CIRR datasets. As shown in Table 5 and Table 6, we compare our Gaussian strategy with three alternatives: (1) **w/ Empirical**, which derives the boundary from the shuffled training batch; (2) **w/ Uniform Dist**, which samples vectors from a bounded Uniform distribution; and (3) **w/ Laplace Dist**, which samples from a Laplace distribution characterized by sharper peaks and heavier tails.

The results across both benchmarks consistently demonstrate that ConeSep (Gaussian) yields the best performance (e.g., 65.31% Avg on FashionIQ and 80.43% Avg on CIRR). Specifically, w/ Empirical exhibits performance degradation due to the high variance of mini-batch statistics, which leads to unstable boundary estimation. Furthermore, while w/ Laplace Dist. generally outperforms w/ Uniform

Table 5. Ablation study on different sampling strategies for boundary estimation on FashionIQ (σ=0.2).

Strategy	Dress		Shirt		Toptee		Average		
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	AVG.
w/ Empirical	49.55	72.10	55.02	74.35	57.50	77.05	54.02	74.50	64.26
w/ Uniform Dist	49.21	71.85	54.86	74.01	57.33	76.82	53.80	74.22	64.01
w/ Laplace Dist	49.80	72.45	55.15	74.20	57.65	77.20	54.20	74.62	64.41
ConeSep (Gaussian)	**50.23**	**72.19**	**56.28**	**74.45**	**58.29**	**78.38**	**54.93**	**75.01**	**64.97**

Table 6. Ablation study on different sampling strategies for boundary estimation on CIRR (σ=0.2).

Strategy	R@K				R_{sub}@K			Avg
	K=1	K=5	K=10	K=50	K=1	K=2	K=3	
w/ Empirical	51.65	81.50	89.45	97.80	77.55	91.20	96.45	79.85
w/ Uniform Dist	51.30	81.25	89.10	97.65	77.20	90.95	96.30	79.65
w/ Laplace Dist	51.95	81.85	89.70	98.05	77.90	91.45	96.60	80.15
ConeSep (Gaussian)	**52.29**	**82.19**	**89.98**	**98.19**	**78.66**	**91.76**	**96.75**	**80.43**

Dist. by better approximating the central tendency of high-dimensional features, it still falls short of the Gaussian strategy. This empirical evidence confirms that the aggregated deep embeddings asymptotically converge to a Gaussian distribution. Consequently, our Gaussian sampling strategy provides the most accurate and stable geometric estimation for the noise boundary in the metric space.

Number of Random Samples K of GFQ. Figure 8 illustrates the impact of the number of random samples K on the model's performance. As K increases from 1 to 4, we observe a consistent improvement in retrieval accuracy on both FashionIQ and CIRR datasets. This trend validates that a single sample ($K = 1$) is insufficient to robustly estimate the geometric noise boundary $\mathbb{B}$, as it is susceptible to random fluctuations in the high-dimensional space. Increasing K allows for a more accurate approximation of the boundary's expectation, thereby stabilizing the Fidelity Quantization process. However, the performance saturates and exhibits a slight decline when K exceeds 4. This suggests that $K = 4$ provides a sufficient statistical sample to capture the distribution characteristics of the noise boundary. Further increasing K yields diminishing returns and may lead to an overly smoothed boundary estimation that lacks the flexibility to handle specific hard samples, while also incurring unnecessary computational overhead. Therefore, we adopt $K = 4$ as the optimal setting to balance estimation stability and computational efficiency.

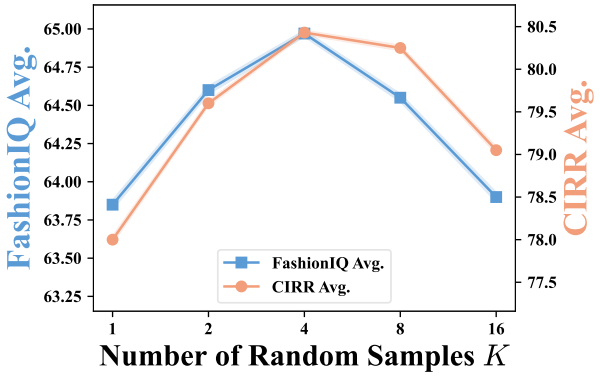

Figure 8. Sensitivity analysis of the number of random samples on FashionIQ and CIRR datasets.

E. Algorithm of Training Procedure

To support the methodology discussion in the main text and comprehensively display the implementation logic of ConeSep , we provide detailed training algorithm pseudocode in **Algorithm 1**. This training process constructs a closed-loop optimization system that organically integrates Geometric Fidelity Quantization (GFQ), Negative Boundary Learning (NBL), and Boundary-based Targeted Unlearning (BTU) into a unified training pipeline.

The execution of the algorithm initiates with a feature extraction and perception phase, where the model utilizes Gaussian random sampling to dynamically estimate the geometric noise boundary $\mathbb{B}$. This process quantifies the matching fidelity of samples and explicitly partitions the current batch into a high-fidelity clean set $\mathcal{T}_{\text{clean}}$ and a low-fidelity noisy set $\mathcal{T}_{\text{noisy}}$. Subsequently, the training adheres to a progressive dual-stage strategy. During the initial warm-up phase (the first N Epochs), the algorithm focuses on the construction of the NBL path. By optimizing the Target-oriented and Query-oriented loss functions, the model constructs a robust "Diagonal Negative Composition" $\mathbf{F}_{\text{neg}}$ for each query to serve as a semantic anchor. Upon the establishment of the boundaries, the process transitions smoothly to the joint optimization phase based on BTU. In this phase, the algorithm employs a specifically designed mask matrix $\mathbf{M}$, which blocks the positive path of noisy samples while preserving the negative path of clean

samples, to solve the entropy-regularized optimal transport problem. This operation generates smooth soft labels $\mathbf{Y}$ to execute precise Targeted Unlearning. Ultimately, the entire framework undergoes end-to-end updates by jointly minimizing the Robust Contrastive Loss, Query-oriented Learning loss, and targeted forgetting loss. Consequently, the model achieves robust and accurate Composed Image Retrieval (CIR) within complex Noisy Triplet Correspondence (NTC) environments.

F. More Qualitative Results

F.1. Qualitative Analyses of Diagonal Negative Composition

To intuitively verify whether our proposed "Diagonal Negative Composition" ($\mathbf{F}_{neg}$) truly learns the semantic opposition required as a "negative anchor", we conduct visualized comparative retrieval experiments. Specifically, for a given query $\langle x_r, x_m \rangle$, we use the positive composed feature $\mathbf{F}_c$ generated by the model and the Diagonal Negative Composition $\mathbf{F}_{neg}$ to perform nearest neighbor searches in the retrieval gallery and display the Top-5 retrieval results. In this context, the "semantic opposition" we define does not refer to completely unrelated random images but specifically refers to a semantic state that violates or ignores the instructions of the modification text x_m. For example, if the x_m instruction is "change to red", then "keeping the original blue" or "wrong green" constitutes "hard negative" semantics that are not only visually confusing but also logically contrary to the instruction.

As shown in Figure 9, the visualization results clearly reveal the significant differences in semantic direction between $\mathbf{F}_{neg}$ and $\mathbf{F}_c$. In the case of FashionIQ (top row), the reference image is a black tight dress, and the modification text requires changing it to "red, looser, and with shorter sleeves". Observing the retrieval results reveals that the results of $\mathbf{F}_c$ accurately align with the target semantics, and the Top-5 images are all red dresses conforming to the description. In sharp contrast, the results retrieved using $\mathbf{F}_{neg}$ mainly consist of black dresses with similar styles. This indicates that $\mathbf{F}_{neg}$ successfully captures the semantic state of "ignoring text instructions", which implies retaining the visual features of the reference image such as the black color while failing to execute the color conversion. This state represents the most typical "hard noise" in the Noisy Triplet Correspondence (NTC) problem, characterized by visual similarity but semantic mismatch, and serves as the direction that the model aims to "push away" in the Boundary-based Targeted Unlearning (BTU) module.

Algorithm 1 Training Procedure of ConeSep

Input: Reference image x_r, target image t, modification text y_m.
Parameters: Total epochs N_{total}, NBL epochs N, Learning Rate η, Fidelity threshold ω, Loss weights $(\zeta, \kappa, \nu, \gamma)$.
Output: Fine-tuned model parameters Ψ^*.

Require: Training Dataset $\mathcal{D}$, Batch size B.
Initialize all model components Ψ.
for $epoch = 1$ to N_{total} **do**
  for batch $\mathcal{B}$ in $\mathcal{D}$ **do**
    // 1. Feature Extraction and Fidelity Quantization
    Compute Composed Query Feature $\mathbf{F}_c$ and Target Feature $\mathbf{F}_t$.
    Compute Similarity $\mathbf{S}$ and Quantify Fidelity $\mathcal{F}(\mathbf{F}_c, \mathbf{F}_t)$.
    Separate batch into $\mathcal{T}_{clean}$ ($\mathcal{F} \geq \omega$) and $\mathcal{T}_{noisy}$ ($\mathcal{F} < \omega$).
    if $epoch \leq N$ **then**
      // 2. Phase I: Negative Boundary Learning (NBL)
      Compute Diagonal Negative Composition feature $\mathbf{F}_{neg}$.
      Compute Robust Contrastive Loss $\mathcal{L}_{robust}$.
      Compute Query-oriented loss($\mathcal{L}_{intra}$) and Target-oriented ($\mathcal{L}_{inter}$).
      Compute NBL Joint Objective: $\mathcal{L}_{NBL} = \mathcal{L}_{robust} + \zeta\mathcal{L}_{intra} + \nu\mathcal{L}_{inter}$.
      Update model parameters Θ using $\nabla_\Theta \mathcal{L}_{NBL}$.
    else
      // 3. Phase II: Boundary-based Targeted Unlearning (BTU)
      // Optimal Transport for Unlearning
      Compute Cost Matrix $\mathbf{C}$ and Mask Matrix $\mathbf{M}$ using features $\mathbf{F}_c, \mathbf{F}_t, \mathbf{F}_{neg}$.
      Solve Masked Entropy-Regularized OT problem for optimal transport plan $\mathbf{P}^*$.
      Construct Smooth Soft Label $\mathbf{Y}$ (using $\mathbf{P}^*$ and Hard Label $\mathbf{L}$).
      Compute Targeted Unlearning Loss $\mathcal{L}_{ul}$ (using $KL(LogitMatrix||\mathbf{Y})$).
      // Joint Optimization
      Compute Robust Contrastive Loss $\mathcal{L}_{robust}$ and $\mathcal{L}_{intra}$.
      Compute BTU Joint Objective: $\Psi^* = \mathcal{L}_{robust} + \kappa\mathcal{L}_{ul} + \zeta\mathcal{L}_{intra}$.
      Update model parameters Ψ using $\nabla_\Theta \mathcal{L}_{BTU}$.
    end if
  end for
end for
Return Trained model parameters Ψ^*

In the case of CIRR (bottom row), the reference image shows two monkeys, and the text instruction is "more monkeys". The retrieval results of $\mathbf{F}_c$ correctly point to images of monkey groups containing multiple monkeys, which reflects an understanding of the semantics regarding quantity increase. However, the retrieval results of $\mathbf{F}_{neg}$ show single monkeys or very few monkeys, which directly constitutes a semantic reversal of the instruction "more" in terms of quantity relationships. This phenomenon further confirms that $\mathbf{F}_{neg}$ is not a random vector in the feature space but a structured "reverse signpost" with explicit semantic direction. By explicitly constructing and utilizing it as a negative boundary, ConeSep employs the Boundary-based Targeted Unlearning (BTU) module to define a clear boundary within the continuous metric space. Consequently, this achieves precise repulsion and unlearning of noise patterns without damaging neighboring clean samples.

F.2. NTC Identification Analysis

As shown in Figure 10, we present the discrimination of ConeSep on Clean and Noisy triplets under NTC scenarios, along with the Fidelity scores calculated by the Geometric Fidelity Quantization (GFQ) module. The results demonstrate that ConeSep is able to accurately distinguish between clean and noisy matches, outputting reasonable fidelity estimates, thereby exhibiting strong discriminative capability in overcoming Modality Suppression.

Specifically, the green area on the left highlights triplets identified as Clean by ConeSep , which are generally assigned high fidelity scores. For example, in the top-left case, both the reference and target images depict fluffy dogs, with the modification text "Pomeranian is sitting on a white surface instead of a gray one". ConeSep assigns a high fidelity estimation of 0.249, accurately reflecting that it captures the fine-grained semantic change of the background surface without being confused by the visual similarity of the dogs. Similarly, in the third row on the left, for the query "has both shoulders covered and is black and is v necked", ConeSep correctly associates the reference red dress with the target black dress, assigning a high score of 0.247, indicating precise alignment with complex attribute modifications.

In contrast, the red areas in the middle and right columns display Noisy Correspondence, which are given noticeably lower fidelity scores. For instance, the example at the top of the middle column shows a reference image of a dog and a target image of a smoothie, with the text "Blend the fruits into a drink". This represents a clear cross-category semantic mismatch, and ConeSep correctly classifies it as noise, with a fidelity of only 0.063. More importantly, ConeSep effectively identifies Hard Noise where visual similarity might mask semantic inconsistency. In the top-right example, both the reference and target images depict indoor room scenes (high visual similarity), but the text describes "Two puppy dogs playing together". Despite the strong visual correlation between the images, ConeSep successfully overcomes the Modality Suppression caused by visual dominance, assigning an extremely low fidelity score of 0.011 due to the complete textual misalignment.

Overall, this visualization strongly supports ConeSep 's capability for accurate Geometric Fidelity Quantization of triplets in NTC settings. ConeSep not only identifies semantically consistent samples but also reliably assigns low scores to various types of noise (including subtle Hard Noise) by explicitly locating the geometric noise boundary. Such fine-grained geometric discrimination effectively provides high-quality signals for the subsequent Boundary-based Targeted Unlearning (BTU), thereby significantly enhancing the model's robustness.

F.3. More Case Study

To further comprehensively evaluate the effectiveness of ConeSep in handling Noisy Triplet Correspondence (NTC) and complex semantic modification tasks, this section provides extended qualitative case studies on the FashionIQ and CIRR benchmark datasets. We present a comparison of the Top-5 retrieval results between ConeSep and the current SOTA robust baseline, TME, in Figure 11 and Figure 12. The images outlined in blue represent the ground-truth (GT) targets, whose recall rankings intuitively reflect the retrieval accuracy and robustness of the models.

Analysis of Success Cases. Figure 11 illustrates the retrieval results on the FashionIQ dataset, which emphasizes fine-grained attribute modifications. The comparative results clearly reveal the significant advantage of ConeSep in overcoming "Modality Suppression". In cases (a) and (b), the user instructions require changing the color (to *dark blue*) or the pattern (to *striped*) of the reference image. The TME model exhibits strong "visual inertia", with its retrieval results often retaining the textual logo or the original style of the reference image, leading to semantic mismatch. In contrast, ConeSep precisely decouples visual features from textual instructions, retrieving targets that match descriptions like "*solid dark blue*" or "*striped scoop neck*". Furthermore, when facing the complex multi-attribute instruction in case (c) involving "*lighter color*", "*long sleeves*", and "*floral patterns*", ConeSep demonstrates exceptional compositional reasoning capability. It accurately hits the target that satisfies all constraints simultaneously, whereas TME remains dominated by the dark visual features of the reference image, failing to effectively execute the color transformation instruction.

Similarly, Figure 12 presents results on the CIRR dataset, which involves more drastic semantic spans and spatial relationship reasoning. In case (a), the query requests changing "*muffins*" to a "*vegetable platter*". Due to the immense difference in visual features, traditional methods are prone to

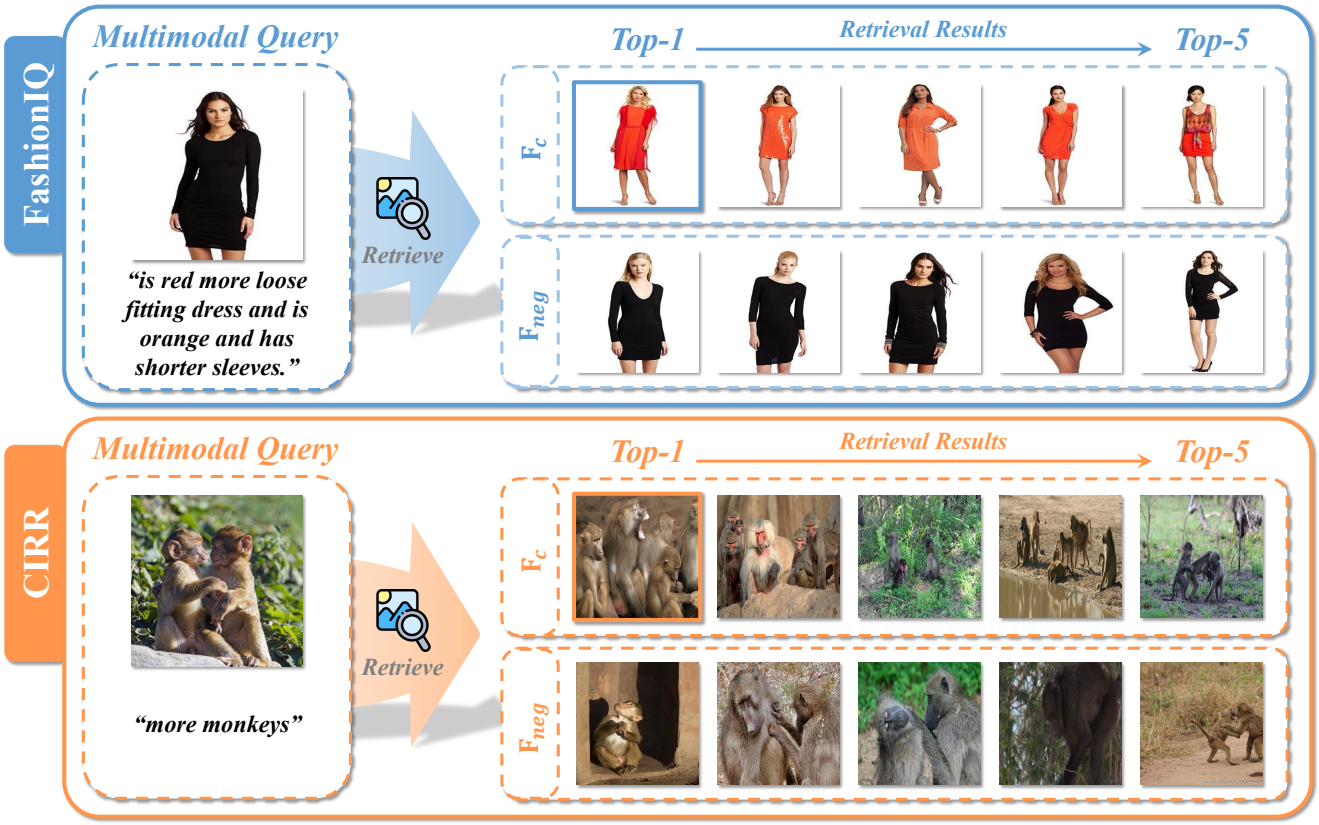

Figure 9. Visualization of retrieval results using the composed feature $\mathbf{F}_c$ and the diagonal negative composition $\mathbf{F}_{neg}$. **Top (FashionIQ):** While $\mathbf{F}_c$ correctly retrieves red dresses following the text instruction, $\mathbf{F}_{neg}$ retrieves black dresses that retain the reference visual semantics but violate the modification logic. **Bottom (CIRR):** $\mathbf{F}_c$ retrieves images with "more monkeys", whereas $\mathbf{F}_{neg}$ retrieves images with fewer or single monkeys, effectively capturing the semantic opposite of the instruction.

failure. TME retrieves mashed potatoes visually similar to muffins at the Top-1 rank, remaining trapped by the visual appearance of the reference image; conversely, ConeSep successfully bridges the semantic gap, retrieving the correct vegetable category. In cases (b) and (c), whether it is the object category transition from "*sofa*" to "*bed*", or the dynamic action modification from "*holding fish*" to "*walking out of water*", ConeSep accurately captures the core semantic changes. This further confirms that our proposed *Geometric Fidelity Quantization (GFQ)* strategy can effectively identify and suppress hard noise interference that is visually similar but semantically mismatched, thereby guiding the model to focus on critical semantic information within the modification text.

In-depth Analysis of Failure Cases. Although ConeSep exhibits strong robustness, retrieval misses can still occur in certain extreme scenarios. We conducted an in-depth analysis of typical "failure cases" in Figure 11(d) and Figure 12(d) and found that these cases actually reflect the "False Negative" challenge present in the current CIR evaluation system, which conversely corroborates the strong semantic understanding capability of ConeSep .

Specifically, in case (d) of FashionIQ, the instruction requires changing a "*floral top*" to one that is "*longer, grey, flowing, and plain*". Although ConeSep failed to recall the GT at Top-1 (the GT is at Top-3), observing its Top-1 and Top-2 results reveals that they are perfect "*grey plain long tops*", which fully comply with the text description visually and semantically. In contrast, TME's results are mixed with garments containing pink patterns, clearly suffering from residual interference from the reference image. Similarly, in case (d) of CIRR, the core instruction is "*add a loving friend*" (i.e., adding a dog). The Top-5 results of ConeSep all display scenes of "*two dogs*" interacting, demonstrating extremely high semantic consistency. However, TME's results waver between "*single dog*" and "*two dogs*", indicating its failure to stably capture the instruction regarding quantity change.

In summary, these so-called "failure cases" reveal that ConeSep actually generates retrieval results that are more aligned with human intent than the Baseline. The semantic consistency demonstrated by the model in the Top-k candidate list is attributed to the negative boundary constructed by the *Negative Boundary Learning (NBL)* module and the targeted unlearning mechanism of the *Boundary-based Targeted Unlearning (BTU)* module. This enables the model to

Figure 10. **Visualization of NTC Identification Analysis.** We display the discrimination between **Clean** (Left) and **Noisy** (Middle/Right) triplets by ConeSep in the NTC scenario, along with the **Fidelity** scores computed by the **Geometric Fidelity Quantization (GFQ)** module. ConeSep successfully distinguishes clean samples (assigned high fidelity scores, e.g., 0.249) from noisy ones. Notably, it effectively overcomes **Modality Suppression** in "Hard Noise" cases (e.g., Top-Right: visually similar rooms but mismatched text "Two puppy dogs"), assigning them extremely low fidelity scores (e.g., 0.011). This precise geometric differentiation provides high-quality signals for the subsequent **Boundary-based Targeted Unlearning (BTU)**.

firmly "push away" erroneous visual features from the reference image (e.g., floral patterns, single dog) and precisely align with the user's true modification intent. This not only proves the robustness of ConeSep in dealing with semantic uncertainty but also suggests directions for improvement in CIR dataset annotation and evaluation metrics for future work.

G. More Related Work

To position this work more comprehensively within the field of Composed Image Retrieval (CIR), particularly regarding the core issue of Noisy Triplet Correspondence (NTC), this section provides a supplementary review and discussion of several recent studies with significant reference value. These works encompass the latest surveys in the CIR domain, robustness benchmarks, consensus learning targeting noisy annotations, and novel retrieval paradigms based on instruction tuning and re-ranking.

CIR Survey and Robustness Benchmarking With the rapid development of the Composed Image Retrieval (CIR) task, Wan et al. [4] provide a comprehensive review of the latest advancements in this field, systematically classifying and summarizing the pipeline from feature extraction, alignment, and fusion to the retrieval process. This survey not only delineates the technological trajectory but also highlights key challenges currently facing the field, offer-

ing a macroscopic background for understanding the Noisy Triplet Correspondence (NTC) issue discussed in this paper. Regarding robustness evaluation, Sun et al. [1] contribute pioneering work by establishing the first robustness benchmark specifically for the CIR task. This work systematically analyzes the performance of CIR models when confronting visual natural corruptions, such as noise and blur, and variations in text understanding. Unlike Sun et al., who focus on robustness against external perturbations during the inference stage, ConeSep in this paper primarily addresses the internal robustness issues arising from data annotation errors, specifically the NTC setting, during the training stage. Consequently, these two approaches complement each other across different dimensions of robustness research.

Handling Noisy Annotations and Correspondence Addressing the issues of noise and ambiguity in data, several recent studies with motivations similar to ours emerge. Zhang et al. [5] propose the Collaborative Group framework, which aims to resolve the problem of triplet ambiguity. This method leverages the psychological concept that the "group outperforms the individual" to mitigate the impact of noisy annotations through consensus learning among multiple compositors. Unlike the Collaborative Group approach, which seeks consensus through model ensembling, ConeSep explicitly models and forgets noise within the feature space via Geometric Fidelity Quantization (GFQ) and Negative Boundary Learning (NBL), thereby providing a solution based on a geometric perspective. Furthermore, Lyu et al. [6] propose TSVC for the image-text retrieval task, utilizing semantic variation consistency to handle noisy correspondence. However, TSVC targets cross-modal retrieval rather than the CIR task, and its methodology for handling noisy labels is difficult to apply directly to NTC samples.

Instruction Tuning and Re-ranking Strategies Beyond noise processing, the field of Composed Image Retrieval (CIR) also sees new explorations in retrieval paradigms. Zhong et al. [2] introduce Instruction Contrastive Tuning, which utilizes Multimodal Large Language Models (MLLM) tuned with instruction prompts to generate unified embeddings. This technique enhances the model's ability to follow modification instructions. This work represents a frontier direction that leverages generative models to improve CIR feature representation. On the other hand, re-ranking is also widely studied as an effective training-free method for boosting retrieval accuracy. Wu et al. [3] propose a Chain-of-Thought Re-ranking method that directly involves the reasoning capability of MLLMs in the re-ranking process of candidate images. Sun et al. [7] propose a training-free method based on Local Concept Reranking that optimizes the initial retrieval results by focusing on the local discriminative information within the modification text. While re-ranking strategies typically act as a postprocessing stage and are orthogonal to the feature learning framework proposed in this work (ConeSep), they have complementary potential for mitigating biases caused by noise in the early retrieval stage, serving as an effective supplementary strategy for ConeSep during the inference phase. We explore this in future work.

References

[1] Sun S, Gu J, Gong S. Benchmarking robustness of text-image composed retrieval[J]. arXiv preprint arXiv:2311.14837, 2023, 6. 10

[2] Zhong W, An W, Jiang F, et al. Instruction Contrastive Tuning for Zero-shot Composed Image Retrieval[J]. 10

[3] Wu S, Zhou Y, Chen Y, et al. Chain-of-Thought Reranking for Image Retrieval Tasks[J]. arXiv preprint arXiv:2509.14746, 2025. 10

[4] Wan Y, Zou G, Zhang B. Composed image retrieval: a survey on recent research and development[J]. Applied Intelligence, 2025, 55(6): 482. 9

[5] Zhang X, Zheng Z, Zhu L, et al. Collaborative group: Composed image retrieval via consensus learning from noisy annotations[J]. Knowledge-Based Systems, 2024, 300: 112135. 10

[6] Lyu S, Tian Z, Ou Z, et al. TSVC: Tripartite Learning with Semantic Variation Consistency for Robust Image-Text Retrieval[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(18): 19269-19277. 10

[7] Sun S, Ye F, Gong S. Training-free zero-shot composed image retrieval with local concept reranking[J]. arXiv preprint arXiv:2312.08924, 2023. 10

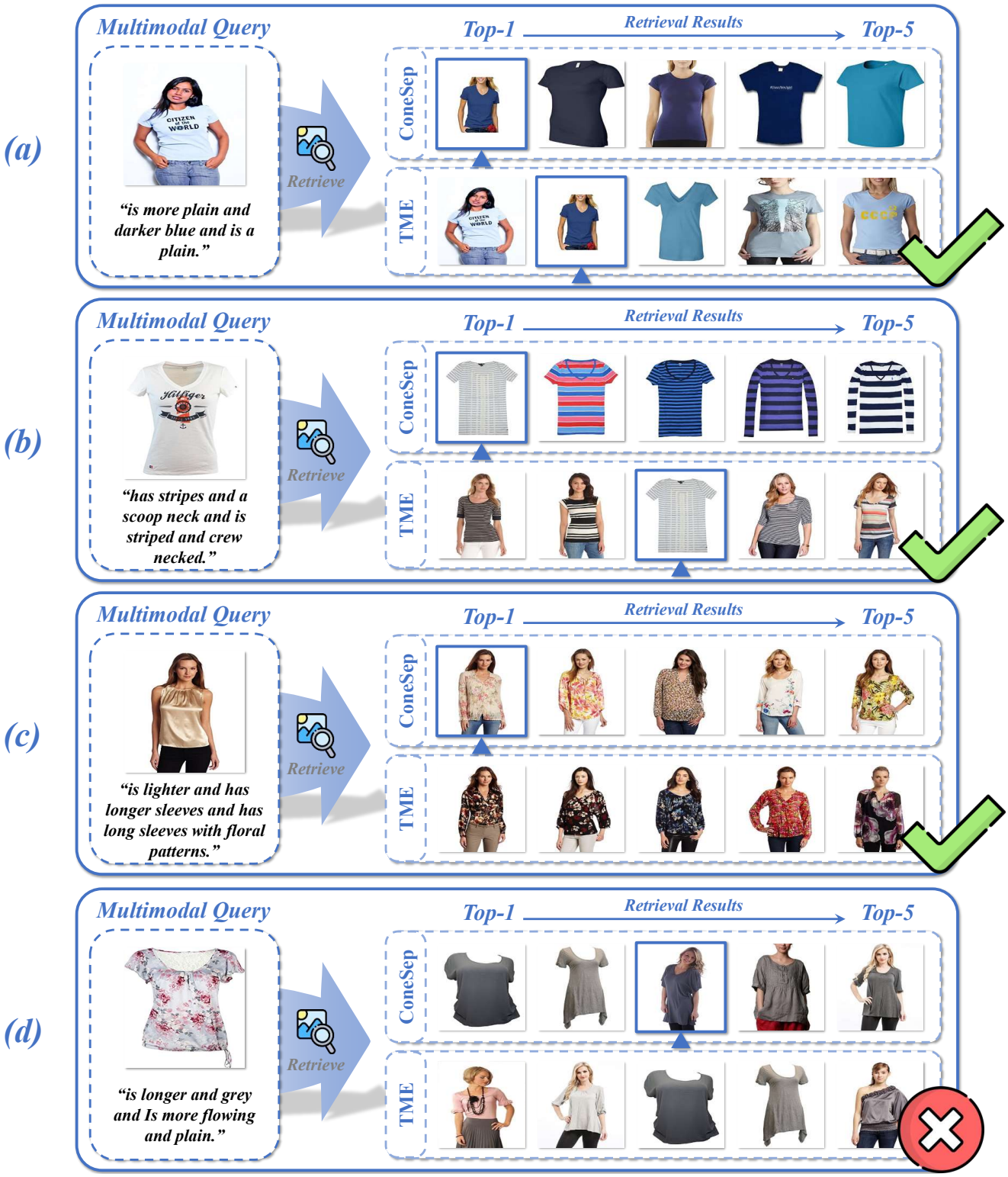

Figure 11. Additional retrieval comparisons on **FashionIQ**. ConeSep accurately follows fine-grained attribute changes (e.g., patterns, sleeve lengths), while TME often suffers from visual inertia from the reference image.

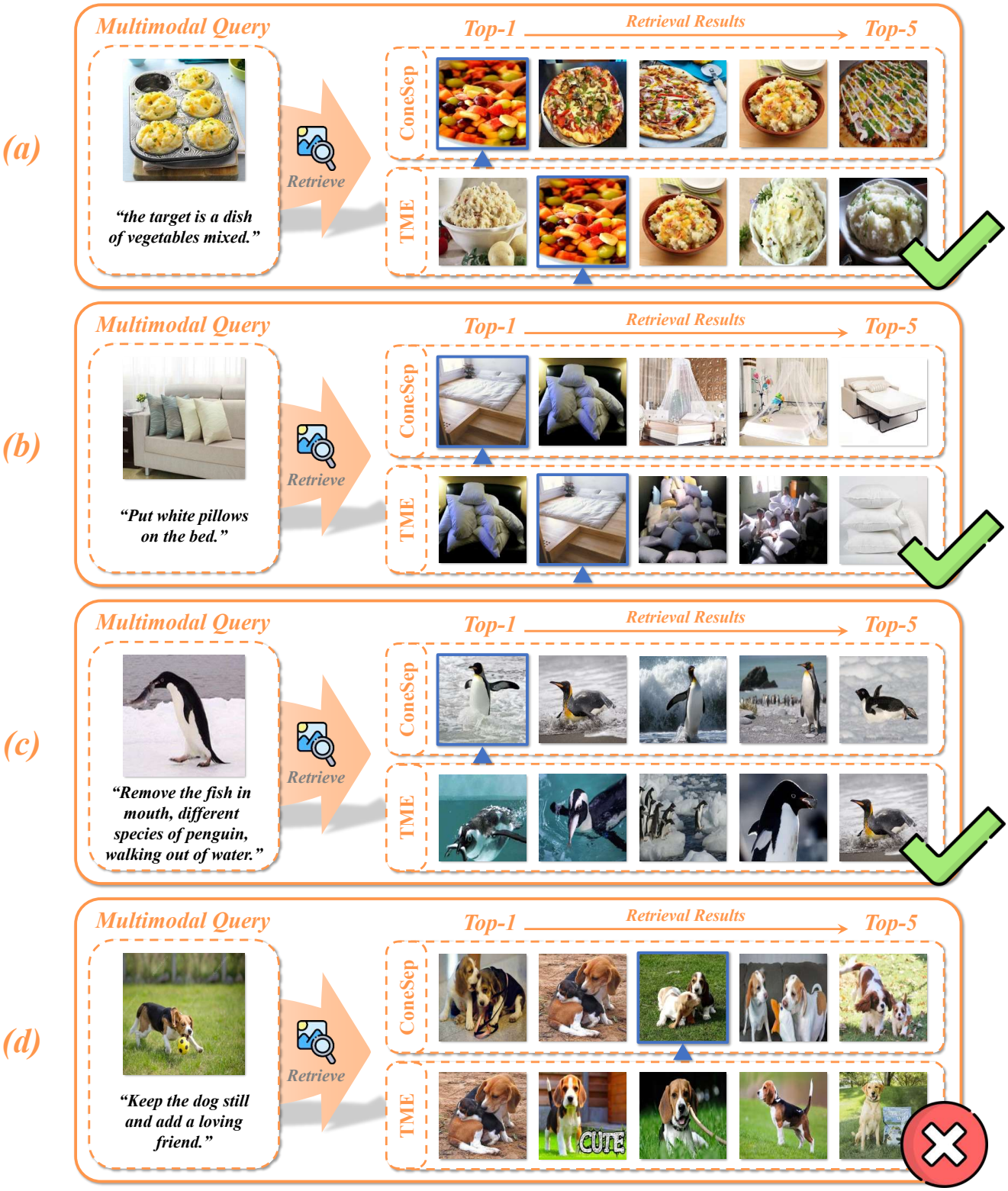

Figure 12. Additional retrieval comparisons on **CIRR**. ConeSep demonstrates superior capability in handling large semantic shifts (e.g., Muffin → Vegetable) and complex spatial/action modifications, whereas TME struggles to break away from the reference image's visual dominance.