Self-guided Semantic Inspection for Zero-Shot Composed Image Retrieval

Supplementary Material

A. Computational Complexity Analysis

Algorithm 1 Training Procedure of DiffComp

Require: Standard image-text pairs $\mathcal{D}_{\text{train}} = \{(I, T)\}$
Require: Pre-trained vision-language model
Initialize DiffComp parameters
while not converged **do**
  Sample a mini-batch from $\mathcal{D}_{\text{train}}$
  // Contextual Semantic Super-patch (CSS)
  Group patches into super-patches $\mathcal{SP}$ and extract features f_j^{SP} (1)-(2)
  // Phrase-guided Selective Masking (PSM)
  Compute phrase-superpatch similarity matrix and alignment scores z_j (4)-(5)
  Generate binary mask via Gumbel-Softmax (6)-(8)
  Apply mask to obtain differentiated feature f_{I_d} (11)
  // Difference-aware Composition (DAC)
  Calculate adaptive interpolation weights α_j based on semantic discrepancy (14)
  Perform hierarchical interpolation to form composed feature f_{comp} (15)-(16)
  Calculate total loss $\mathcal{L}$ and update parameters (17)
end while

The complete training process of DiffComp is summarized in Algorithm 1. We further analyze the computational complexity of its three core modules, including Contextual Semantic Super-patch (CSS), Phrase-guided Selective Masking (PSM), and Difference-aware Composition (DAC). **Computational Complexity.** We analyze the computational cost of each module in both training and inference stages:

- **CSS** divides an input image into N patches and reorganizes them into $N_s = N/S^2$ super-patches of size $S \times S$. Its complexity $\mathcal{O}(LNS^2d)$ is substantially lower than the standard global attention $\mathcal{O}(L(N+1)^2d)$.
- **PSM** computes phrase-superpatch similarities with complexity $\mathcal{O}(N_sPd)$, introducing negligible overhead.
- **DAC** fuses global and local features with complexity $\mathcal{O}(N_sd)$.

Among the three modules, CSS dominates the total cost but remains efficient in practice. For example, with ViT-L/14 and $S = 4$, it adds only about 7% overhead compared to vanilla global encoding. PSM is used only during training, while inference involves CSS and DAC, both of which scale linearly.

Table 1. Backbone details. Abbreviations: Img.=input image size; L=transformer layers; W=hidden width; FT-L=fine-tuned layers; Proj dim=shared embedding dimension.

Backbone	Visual encoder				Textual encoder			Proj
	Img.	*L*	*W*	*FT-L*	*L*	*W*	*FT-L*	*dim*
CLIP (ViT-L/14)	224	24	1024	12	12	768	12	768
CLIP (ViT-G/14)	224	48	1664	24	32	1280	24	1280
BLIP (ViT-L/16)	224	24	1024	12	24	768	12	256

B. Backbone Configurations

Table 1 summarizes the architectural settings of the backbones used in our experiments. For each backbone, we list the input image size, transformer depth (L), hidden width (W), and the number of fine-tuned layers (FT-L) for both visual and textual encoders. The last column (Proj dim) indicates the dimensionality of the shared embedding space after projection. We adopt partial fine-tuning to maintain computational efficiency while retaining sufficient adaptability for compositional retrieval tasks.

C. Dataset Details and Complete Results

C.1. Datasets and Evaluation Protocols

FashionIQ [23] contains three categories (Dress, Shirt, TopTee). Each query pairs a reference image with a modification text formed by two human-written relative captions. We follow the Original-Split protocol and report Recall@10,50 per category and their average.

CIRR [13] features natural objects and scene compositions, where each textual modification corresponds to a specific image pair. Besides overall retrieval, CIRR defines a fine-grained *subset* evaluation with visually similar hard negatives. We report overall Recall@1, 5, 10, 50 and subset $\text{Recall}_{\text{subset}}$@1, 2, 3 using the official online protocol.

CIRCO [1] targets large-scale, real-world composed retrieval and provides multiple ground truths per query (avg. ≈4.53). Its gallery comprises the full COCO [12] 120K images, making retrieval significantly more challenging than CIRR. We use the official validation (220 queries) and test (800 queries) splits and report mAP@5/10/25/50.

GeneCIS [21] evaluates conditional similarity under four task settings formed by crossing {Focus, Change}×{Attribute, Object}. Each task has about 2K queries; for each query, the model retrieves one ground-truth image from 10-15 visually similar candidates constructed from VAW/COCO-Panoptic. We report Recall@1, 2, 3 for each task and grouped averages

Table 2. Full comparison with state-of-the-art methods on the CIRR test set. Rows shaded in gray correspond to our method, and **bold** indicates the best scores among all methods.

Method	Venue	Backbone	Recall@K				$\text{Recall}_{\text{subset}}$@K		
			K=1	K=5	K=10	K=50	K=1	K=2	K=3
Pic2Word[17]	CVPR'23	CLIP ViT-L/14	23.90	51.70	65.30	87.80	–	–	–
SEARLE-XL[1]	ICCV'23		24.24	52.48	66.29	88.84	53.76	75.01	88.19
Context-I2W[19]	AAAI'24		25.60	55.10	68.50	89.80	–	–	–
LinCIR[7]	CVPR'24		25.04	53.25	66.68	–	57.10	77.40	88.90
KEDs[18]	CVPR'24		26.40	54.80	67.20	89.20	–	–	–
SlerpTAT[9]	ECCV'24		30.94	59.40	70.94	89.18	64.70	82.92	**92.31**
PLI[4]	ICME'25		26.15	56.82	69.30	89.76	56.22	77.52	89.74
PrediCIR[20]	CVPR'25		27.20	57.00	70.20	–	–	–	–
HIT[11]	ICCV'25		27.90	57.60	70.50	90.40	–	–	–
DiffComp	proposed		**32.36**	**62.90**	**74.88**	**93.23**	**64.92**	**83.20**	91.57
ISA[6]	ICLR'24	BLIP ViT-L/16	29.68	58.72	70.79	90.33	–	–	–
SlerpTAT[9]	ECCV'24		33.98	61.74	72.70	88.94	68.55	85.11	93.21
HIT[11]	ICCV'25		36.90	67.70	79.10	94.70	–	–	–
DiffComp	proposed		**39.72**	**68.64**	**79.42**	**95.43**	**69.62**	**85.51**	**93.88**

(Focus/Change and Attribute/Object).

C.2. Complete Comparisons.

Result on CIRR. We evaluate DiffComp on the CIRR benchmark to assess its ability to model fine-grained compositional relationships. As summarized in Table 1 of the main paper, it achieves state-of-the-art performance across all Recall@K metrics. Under the CLIP-L/14 backbone, DiffComp surpasses HIT by 4.46% in R@1 and 4.38% in R@10. Feature-composition approaches, including ours, consistently outperform pseudo-word methods, demonstrating stronger compositional generalization. This improvement stems from DiffComp's ability to explicitly model structural differences at the super-patch level rather than relying on synthetic token concatenation. Table 2 provides the full CIRR evaluation, including $\text{Recall}_{\text{subset}}$@K metrics. SlerpTAT [9] slightly outperforms HIT [11] under CLIP-L/14, benefiting from its multi-object pseudo-word strategy and fine-grained visual encoding. However, such pseudo-word methods rely on indirect token composition, which often distorts semantic alignment. **DiffComp**, in contrast, achieves the best overall and subset results by directly modeling super-patch–level relationships, yielding more faithful and transferable compositional representations.

Result on GeneCIS. We report averaged Recall@K results on GeneCIS in Table 3 of the main paper to provide a compact overview across semantic dimensions. DiffComp ranks first on *Focus* and *Object* averages at both R@1 and R@2, and achieves the second-best results on *Change* averages—showing strong robustness across different modification types. In attribute-centric tasks, DiffComp attains the best *Attribute* R@1 and competitive *Attribute* R@2 (slightly behind PrediCIR). The complete per-task results (Focus/Change × Attribute/Object) and grouped averages are listed in Tab. 3 and 4. These detailed results show that DiffComp is particularly strong on *Focus Attribute* and *Focus Object*, while maintaining balanced performance across *Change*-related subtasks.

D. Evaluation on Larger Backbone Models

To further assess the scalability of DiffComp, we conduct experiments using a larger backbone, CLIP-ViT-G/14 (denoted as CLIP-G). We follow the same training and evaluation protocols as described in the main paper, without introducing any additional supervision or architectural modification. To balance adaptability and computational efficiency, we fine-tune only the last 24 layers of the 48-layer visual encoder and the last 24 layers of the 32-layer text encoder. This design maintains efficiency while preserving sufficient capacity for compositional reasoning. For CLIP-G experiments, we fine-tune the models on two NVIDIA A800 GPUs for 10 epochs with a per-GPU batch size of 64, following the same optimizer and learning rate as in the base configuration.

As shown in Tab. 5, 6 and 7, DiffComp consistently improves performance across all benchmarks—FashionIQ, CIRR, and CIRCO—demonstrating strong scalability and compatibility with larger vision–language backbones. Even under this **partial fine-tuning** setup, DiffComp achieves substantial gains over recent state-of-the-art methods such as PrediCIR and LinCIR. These results confirm that our compositional modules effectively enhance representation alignment and reasoning ability without full model retrain-

Table 3. Quantitative full results on GeneCIS with CLIP ViT-L/14. † indicates the results are reproduced from [7]. Rows shaded in gray correspond to our method, and **bold** indicates the best scores among all methods.

Method	Venue	Focus Attribute			Change Attribute			Focus Object			Change Object		
		R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3
Pic2Word†	CVPR'23	15.65	28.16	38.65	13.87	24.67	33.05	8.42	18.01	25.77	6.68	15.05	24.03
SEARLE†	ICCV'23	17.00	29.65	40.70	16.38	25.28	34.14	7.76	16.68	25.31	7.91	16.84	25.05
Context-I2W	AAAI'24	17.20	30.50	41.70	16.40	28.30	**37.10**	8.70	17.90	26.90	7.70	16.00	25.40
LinCIR†	CVPR'24	16.90	29.95	41.45	16.19	27.98	36.84	8.27	17.40	26.22	7.40	15.71	25.00
PLI	ICME'25	20.85	33.40	43.15	14.63	26.14	35.46	12.55	21.07	30.77	11.48	21.68	32.50
PrediCIR	CVPR'25	18.20	31.90	42.60	**18.70**	**30.40**	35.40	12.70	19.00	31.20	**16.90**	25.50	34.10
DiffComp	proposed	**21.55**	**34.25**	**44.70**	16.44	26.52	36.93	**14.34**	**25.20**	**34.08**	15.45	**26.84**	**37.65**

Table 4. Quantitative average results on GeneCIS with CLIP ViT-L/14. † indicates the results are reproduced from [7]. Rows shaded in gray correspond to our method, and **bold** indicates the best scores among all methods.

Method	Venue	Focus Avg			Change Avg			Attribute Avg			Object Avg		
		R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3
Pic2Word†	CVPR'23	12.04	23.09	32.21	10.28	19.86	28.54	14.76	26.42	35.85	7.55	16.53	24.90
SEARLE†	ICCV'23	12.38	23.17	33.01	12.15	21.06	29.60	16.69	27.47	37.42	7.84	16.76	25.18
Context-I2W	AAAI'24	12.95	24.20	34.30	12.05	22.15	31.25	16.80	29.40	39.40	8.20	16.95	26.15
LinCIR†	CVPR'24	12.59	23.68	33.84	11.80	21.85	30.92	16.55	28.97	39.15	7.84	16.56	25.61
PLI	ICME'25	16.70	27.24	36.96	13.06	23.91	33.98	17.74	29.77	39.31	12.02	21.38	31.64
PrediCIR	CVPR'25	15.45	25.45	36.90	**17.80**	**27.95**	34.75	18.45	**31.15**	39.00	14.80	22.25	32.65
DiffComp	proposed	**17.95**	**29.73**	**39.39**	15.95	26.68	**37.29**	**19.00**	30.39	**40.82**	**14.90**	**26.02**	**35.87**

ing, achieving both **efficiency and generalization** when scaled to larger models.

E. Additional Ablation Studies

This section provides the complete quantitative results corresponding to the compact summary in Table 4 of the main paper. We include the full tables for module combinations, super-patch size, masking strategy and modeling variants, as well as the detailed trend of masking ratio.

Effectiveness of the proposed modules We conduct ablation studies on FashionIQ and CIRR to evaluate the contribution of each core component: Contextual Semantic Super-patch **(CSS)**, Phrase-guided Selective Masking **(PSM)**, and Difference-aware Composition **(DAC)**. Tab. 8 reports the results across different module configurations. **Row 1** serves as the baseline, using standard ViT patches with random masking and global linear interpolation, without incorporating any proposed modules. **Row 2**, which adds CSS alone, provides strong and consistent improvements across datasets: +0.6% on FashionIQ R@10 and +1.1% on CIRR R@1. This highlights the benefit of aggregating local patches into semantically coherent units. In contrast, **Row 3** (PSM) yields modest improvements, while **Row 4** (DAC) shows unstable performance, likely due to the limited semantic granularity of vanilla ViT patches, which undermines both masking and interpolation effectiveness. **Row 5** (CSS+PSM) and **Row 6** (CSS+DAC) achieve further substantial gains over Row 2, demonstrating the complementary benefits of PSM and DAC when supported by a structured semantic representation, where PSM improves phrase alignment and DAC facilitates more precise semantic integration. **Row 7** (PSM+DAC), without CSS, shows moderate improvements but still underperforms combinations involving CSS, reaffirming the necessity of semantically coherent units. **Row 8**, our full model that integrates CSS, PSM, and DAC, delivers the best performance across all metrics. Compared to using the foundational CSS module alone (Row 2), our full model achieves notable improvements of +2.5% on FashionIQ R@10 and +2.9% on CIRR R@1. These results validate our *Differentiate-then-Compose* framework, in which CSS provides a structured semantic basis, PSM introduces targeted differences, and DAC reconciles them through adaptive cross-modal composition.

Effect of super-patch aggregation variants. Tab. 9 reports the results under different aggregation configurations in the CSS module. The 1×1 setting (i.e., original CLIP patches) yields the lowest performance, suggesting that although CLIP patches encode strong global semantics, their lack of local spatial coherence hinders precise correspondence and difference modeling. As the aggregation scale increases, performance improves steadily up to the 4×4

Table 5. Performance comparison using CLIP-ViT-G backbone on the FashionIQ validation set. **Bold** indicates the best results.

Method	Venue	Dress		TopTee		Shirt		Average	
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Pic2Word	CVPR'23	25.43	47.65	35.24	57.62	33.17	50.39	31.28	51.89
SEARLE	ICCV'23	28.16	50.32	39.83	61.45	36.46	55.35	34.81	55.71
LinCIR	CVPR'24	38.08	60.88	50.48	71.09	46.76	65.11	45.11	65.70
PrediCIR	CVPR'25	39.70	62.40	53.70	73.60	48.20	67.40	47.20	67.80
DiffComp	proposed	**41.30**	**63.95**	**54.50**	**74.70**	**49.25**	**67.50**	**48.35**	**68.72**

Table 6. Performance comparison using CLIP-ViT-G backbone on the CIRR validation set. **Bold** indicates the best results.

Method	Venue	Recall@K			Recall$_{subset}$@K		
		K=1	K=5	K=10	K=1	K=2	K=3
Pic2Word [17]	CVPR'23	30.41	58.12	69.23	68.92	85.45	93.04
SEARLE [1]	ICCV'23	34.80	64.07	75.11	68.72	84.70	93.23
LinCIR [7]	CVPR'24	35.25	64.72	76.05	63.35	82.22	91.98
PrediCIR [20]	CVPR'25	37.00	66.10	77.90	–	–	–
DiffComp	proposed	**38.56**	**68.25**	**79.62**	**69.89**	**89.05**	**93.44**

Table 7. Performance comparison using CLIP-ViT-G backbone on the CIRCO validation set. **Bold** indicates the best results.

Method	Venue	mAP@5	mAP@10	mAP@25	mAP@50
Pic2Word [17]	CVPR'23	5.54	5.59	6.60	7.29
SEARLE [1]	ICCV'23	13.20	13.85	15.32	16.04
LinCIR [7]	CVPR'24	19.71	21.01	23.13	24.18
PrediCIR [20]	CVPR'25	23.70	24.60	25.40	26.00
DiffComp	proposed	**24.43**	**25.52**	**26.66**	**27.31**

grid, where moderate spatial grouping encourages the formation of semantically coherent regions while still retaining sufficient fine-grained detail for reasoning. However, when the grid expands to 8×8, the performance begins to degrade, likely because excessively coarse aggregation blurs object boundaries and reduces the model's sensitivity to subtle local differences. Beyond uniform grid-based aggregation, we also examine a non-uniform variant based on K-Means clustering, which adaptively groups spatially and semantically similar tokens into clusters. This adaptive variant achieves slightly higher results on CIRR but lower on FashionIQ, indicating that while clustering can capture flexible semantic structures, it may also merge fine-grained part-level variations into broader clusters, an effect that is particularly detrimental for fashion images with subtle attribute-level cues. In addition, clustering incurs higher computational cost and produces clusters with variable sizes across images. The differing numbers of patches per cluster make per-cluster local feature extraction difficult to batch efficiently, reducing training throughput. Overall, these results suggest that our grid-based super-patch aggregation provides a favorable balance between semantic abstraction and spatial consistency, while adaptive clustering, although conceptually appealing, remains computationally demanding and less convenient for large-scale training with fixed-length inputs.

Analysis of masking strategies and modeling approaches. We analyze three key design factors in the PSM module: **masking strategy**, **mask probability modeling**, and **textual guidance granularity**. Tab. 10 presents the ablation results on the FashionIQ validation set. **First**, among masking strategies, random masking and cropping underperform, whereas visually guided masking achieves stronger results. Random masking, as the baseline, achieves 30.8% R@10. Random cropping performs slightly worse (29.8%) due to its lack of semantic selectivity, often discarding regions relevant to phrase modifications. CAM-based masking shows more reasonable effectiveness by emphasizing visually salient, class-discriminative areas, improving R@10 to 31.2%. However, since it is guided purely by visual saliency, CAM-based regions may not always align with phrase-level textual cues, which limits its precision in fine-grained correspondence. Overall, structured and visually guided masking performs better than heuristic random strategies, but its lack of explicit text grounding constrains performance in cross-modal alignment. **Second**, we compare different masking modeling approaches. Hard Binary modeling makes discrete selections, simpli-

Table 8. Ablation study on FashionIQ and CIRR of three key modules: CSS, PSM, and DAC. Note that '✓' in tables means retaining it otherwise removing.

	Components			FashionIQ		CIRR		
	CSS	PSM	DAC	R@10	R@50	R@1	R@5	R@10
1.				29.5	49.7	28.4	58.6	70.4
2.	✓			30.1	50.9	29.5	61.2	72.7
3.		✓		29.8	50.2	28.9	60.5	71.9
4.			✓	29.3	50.4	29.6	59.8	72.3
5.	✓	✓		31.2	52.0	30.4	61.3	72.6
6.	✓		✓	30.8	52.3	31.0	61.8	73.2
7.		✓	✓	30.7	51.5	29.9	60.7	72.8
8.	✓	✓	✓	**32.6**	**53.9**	**32.4**	**62.9**	**74.9**

Table 9. Impact of different super-patch aggregation variants in CSS on performance.

Variant	FashionIQ		CIRR			CIRCO
	R@10	R@50	R@1	R@5	R@10	mAP@5
grid 1×1	30.7	51.5	29.9	60.3	72.8	11.8
grid 2×2	31.9	52.2	31.5	61.4	72.3	14.7
grid 4×4	32.6	53.9	32.4	62.9	74.9	16.2
grid 8×8	29.6	49.4	29.7	58.4	69.2	13.8
clustering	31.6	52.7	32.8	61.8	73.5	15.5

fying decision-making but preventing gradient flow, which hinders optimization and leads to weaker results. Soft Sigmoid enables continuous gradients but retains partial information from most regions, reducing the model's ability to focus on truly distinctive differences. Gumbel-Softmax provides a differentiable approximation to discrete sampling but may introduce stochastic noise. In our implementation, we integrate Gumbel-Softmax with a straight-through estimator, allowing discrete decisions during the forward pass and stable gradients during backpropagation, which effectively supports region suppression and yields the best performance. **Finally**, we examine the impact of textual granularity on mask generation. Noun-level guidance focuses primarily on object entities and often overlooks contextual cues such as attributes or relationships, resulting in limited semantic coverage (30.9% R@10). Sentence-level descriptions offer broader contextual information but can introduce unnecessary noise and mismatch the localized nature of visual phrases (31.6% R@10). Phrase-level guidance provides a balanced middle ground, being sufficiently rich to capture modification intent while maintaining spatial precision, and achieves the highest performance of 32.6%. These results confirm the effectiveness of phrase-level supervision as an intermediate semantic unit for fine-grained vision–language alignment.

Effect of masking ratio. To investigate the influence of masking strength, we vary the target masking ratio ρ in Eq. (20) from 0.1 to 0.9 during training. As shown in Fig. 1, model performance on the CIRCO test set exhibits a rise–then–decline pattern, reaching the best results around $\rho = 0.7$ (e.g., 31% R@1 and 67.5% R@50). At low masking ratios (0.1–0.3), most visual content remains intact, yielding limited visual–textual discrepancy and leading to trivial contrastive learning signals. As the ratio increases (0.6–0.7), more text-relevant regions are masked out, introducing stronger cross-modal contrast and encouraging the model to rely on complementary textual cues, thereby improving alignment and compositional reasoning. However, when the masking ratio exceeds 0.8, excessive suppression of visual content disrupts structural cues and weakens discriminative capacity. The resulting representation becomes overly text-biased, impairing its ability to match the target image and causing a noticeable performance drop.

Table 10. Ablation study of different mask generation variants and modeling approaches on FashionIQ validation set.

Category	Method	R@10			
		Dress	TopTee	Shirt	Avg
Masking Strategy	Random Mask	25.4	34.0	32.6	30.8
	Random Cropping	24.1	33.2	32.1	29.8
	CAM-based	25.8	34.8	33.0	31.2
Masking Modeling	Hard Binary	24.8	33.5	32.2	30.2
	Soft Sigmoid	26.0	34.8	33.5	31.4
	Gumbel-Softmax	26.8	35.4	34.2	32.1
Text Granularity	Noun-level	25.4	34.0	33.9	30.9
	Sentence-level	25.7	34.9	34.1	31.6
Ours	PSM	**27.0**	**36.1**	**34.8**	**32.6**

Effect of composition control parameters. Fig. 2 illustrates the effects of two control parameters in DAC, α_{base} and λ, on FashionIQ and CIRCO benchmarks. As shown in the left panel, performance improves consistently as α_{base} increases, peaking around 0.8. This suggests that while textual guidance is crucial for effective composition, retaining a moderate proportion of visual features helps preserve spatial grounding and complementary semantics. When α_{base} approaches 1.0 (i.e., relying purely on text), performance declines, confirming that overemphasizing textual cues weakens multimodal balance. In the right panel, varying λ adjusts the weighting between global and local visual representations. Using only global features ($\lambda = 0$) already provides a strong baseline, whereas relying exclusively on local super-patch features ($\lambda = 1$) leads to a noticeable drop due to the loss of holistic structure. Empirically, setting λ around 0.4 yields stable and competitive performance, suggesting that the global context plays a dominant role in composition, while incorporating local cues offers complementary refinements without disrupting global coherence.

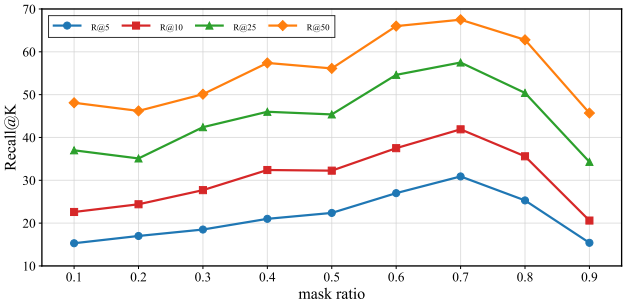

Figure 1. Effect of mask ratio on Recall@k using CLIP-ViT-L/14 on CIRCO test set. Evaluation based on first annotated ground truth per query.

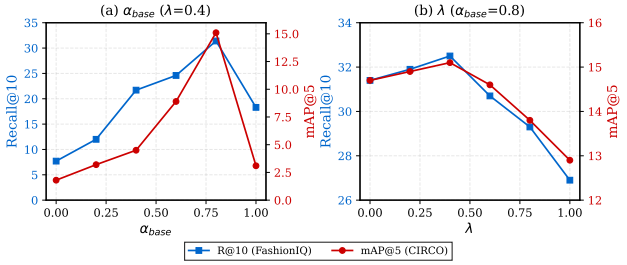

Figure 2. Impact of DAC hyperparameters α_{base} and λ on retrieval performance. Optimal values balance visual grounding and global–local feature fusion.

F. Additional Visualization and Analysis

In this section, we provide additional qualitative and distributional analysis to better illustrate how our Phrase-guided Selective Masking (PSM) and Discrepancy-Aware Composition (DAC) jointly shape the learning dynamics of DiffComp. We begin by examining how PSM constructs discrepancy at the super-patch level, followed by an analysis of how such perturbations reshape the global image–text consistency distribution. We then present extended visualizations of inference-time behavior across CIRR, CIRCO, and FashionIQ benchmarks.

F.1. Visualization of Phrase-guided Masking

PSM evaluates the relevance of each super-patch with respect to the modification phrase and assigns a masking probability accordingly. As shown in Figure 3, the highest probabilities (dark regions) consistently align with regions semantically tied to the modifying phrase—such as the ball and tackling player in action-oriented examples, or fine-grained garment regions for attribute-level descriptions. Masking these phrase-aligned regions deliberately injects semantic discrepancy, forcing DAC to reason over the remaining evidence and encouraging robust difference-aware representation learning. This mechanism is central to reducing shortcut behavior and improving generalization to ZS-CIR scenarios.

F.2. Image-text consistency distributions.

To better understand how PSM reshapes the training signal, we analyze the cosine similarity between image and text embeddings under three types of pairs, as shown in Figure 4. First, the original CC3M training pairs (*Train (Aligned)*) exhibit relatively high similarity, indicating that most captions closely match their images. Second, zero-shot composed retrieval benchmarks (*Test (ZS-CIR)*, aggregating CIRR, CIRCO, and FashionIQ) naturally lie in a lower-similarity regime because the composed query explicitly describes a *different* target image. This mismatch introduces a structural distribution gap between training and inference. Third, when the CC3M images are perturbed by PSM and paired with the same texts (*Train (PSM-masked)*), the resulting similarity distribution shifts leftwards and nearly matches the inference regime. This demonstrates that PSM effectively constructs discrepancy-aware training pairs whose consistency resembles that of real composed queries, thereby narrowing the train-test gap and stabilizing the optimization of DAC.

F.3. Inference-time visual modulation.

To further illustrate how DiffComp adjusts visual retention under various compositional scenarios, Figure 5 presents additional examples from both object/scene-level (CIRR) and attribute-level (FashionIQ) benchmarks. For CIRR, generic prompts such as "human and wild animal interaction" yield balanced emphasis across subjects, whereas contrastive or role-switching modifiers (*e.g.* "instead of man" or "posing for camera") produce strong, localized modulation. For FashionIQ, concrete attributes (*e.g.* "gray designs" or "collar") lead to sharply localized retention, while abstract descriptions (*e.g.* "more casual") induce broader structural adjustments. These examples confirm that DiffComp dynamically adapts its interpolation strength according to semantic intent, supporting generalizable, fine-grained compositional reasoning across visual and textual modalities.

Discussion. PSM is used during training to intentionally construct image-text discrepancy in a phrase-guided manner. For concrete semantics with clear visual correspondences, it produces focused masking patterns that perturb aligned training pairs and highlight phrase-relevant differences. For more abstract or higher-level semantics, the behavior depends on how the meaning manifests visually. Some abstract semantics can still be grounded in identifiable cues (e.g., "a warmer atmosphere" conveyed through lighting or contextual objects), allowing PSM to generate meaningful relevance signals. For abstract semantics that cannot be associated with any specific spatial region, the masking probabilities naturally become more diffuse be-

Figure 3. Visualization of phrase-guided masking probabilities produced by PSM on CC3M training images. Darker super-patches correspond to *higher masking probability*, indicating stronger alignment between the super-patch and the modification phrase. By intentionally masking these phrase-relevant regions, PSM breaks the strong alignment in original CC3M pairs and synthesizes controlled image–text discrepancy for training. DAC subsequently reconciles this discrepancy, guiding the model toward modification-centric visual reasoning rather than trivially copying aligned content.

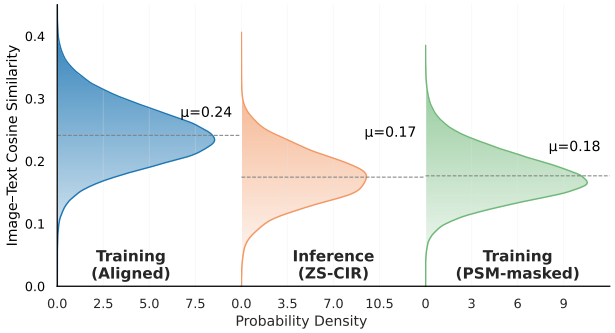

Figure 4. Image-text cosine similarity distributions under different pair constructions. **Train (Aligned)** denotes the original CC3M training pairs. **Test (ZS-CIR)** aggregates composed queries from the CIRR, CIRCO, and FashionIQ benchmarks at inference time. **Train (PSM-masked)** measures the similarity between PSM-masked images and their corresponding texts right before the DAC module. PSM shifts the training distribution from the highly aligned regime (*e.g.* mean $\mu \approx 0.24$) towards the harder ZS-CIR regime (*e.g.* $\mu \approx 0.17$), producing perturbed pairs (*e.g.* $\mu \approx 0.18$) that closely match the test distribution. This reduces the train-test consistency gap and provides more informative supervision for learning discrepancy-aware composition.

cause the semantics are expressed at a global rather than localized level. This reduced spatial selectivity does not hinder the method, because PSM serves primarily as a mechanism to induce discrepancy during training, while the core compositional reasoning is handled by DAC. Abstract semantics in ZS-CIR also rely on global, rather than localized, differences. In such scenarios, DAC integrates these global shifts by comparing the full visual feature field with the target representation implied by the text. Consequently, DiffComp can accommodate abstract semantics through global feature adjustments while still benefiting from localized perturbations when the semantics are concrete.

G. Discussion of DiffComp's Innovations

A central challenge in zero-shot composed image retrieval (ZS-CIR) is the training-inference gap: during training, only aligned image-caption pairs are available, whereas inference requires composing a reference image with a modification text to retrieve a semantically different target. Several recent methods attempt to bridge this gap by introducing visual perturbations, but they differ substantially in how discrepancies are constructed and resolved. We provide an extended analysis of DiffComp's design philosophy and its distinctions from prior work.

The Differentiate-then-Compose Paradigm. PLI [4] applies random binary masks to image patches, and PrediCIR [20] performs random cropping with a reconstruction-

Figure 5. Visualization of $(1-\alpha_j)$ under different modification texts and corresponding retrieval results. Darker super-patches indicate stronger visual retention. Text segments in blue denote the semantic components that most strongly influence visual preservation. For instance, in (b), the phrase "instead of man" suppresses the human region while emphasizing the animal, illustrating text-driven spatial modulation.

based predictor. Both produce perturbations agnostic to the text, so models are trained to recover missing visual information generally rather than to reason about specific textual modifications. Moreover, PrediCIR's predictor is computationally heavy and lacks mechanisms to explicitly resolve conflicts between text and visual regions. DiffComp shifts this formulation. Phrase-guided Selective Masking (PSM) leverages phrase-level textual guidance to mask the most text-aligned image regions, creating controlled, text-grounded discrepancies that simulate compositional modifications encountered at inference. Difference-Aware Composition (DAC) resolves these discrepancies via hierarchical feature modulation: it anchors valid visual content, suppresses conflicting cues, and injects textual semantics precisely where visual information has been removed. This tightly coupled loop distinguishes DiffComp from prior perturbation-recovery pipelines that lack both text grounding and spatial adaptivity. Trained on only 10% of CC3M for 4 hours, DiffComp outperforms PrediCIR (28h full dataset) on FashionIQ R@10 (32.6 vs. 30.1) and CIRR R@1 (32.4 vs. 27.2).

CSS: Semantic Coherence as a Foundation. Contextual Semantic Super-patch (CSS) provides semantically coherent visual units that serve as the foundation for difference induction in PSM and localized perception in DAC. Patch-level approaches such as LAPs [28] operate on fragmented ViT tokens, leading to quadratic complexity and lacking spatial coherence for phrase-level grounding. CSS groups adjacent patches into super-patches and encodes them efficiently in parallel, incurring only 7% training overhead with ViT-L/14 (Sec. A). Non-grid alternatives including K-Means clustering and CAM-based grouping increase computational cost without improving performance (Tab. 9), as clustering produces variable-sized groups that are difficult to batch and CAM produces image- or caption-level activation maps rather than phrase-level correspondences, limiting its ability to guide fine-grained masking.

PSM: Text-Guided Discrepancy Construction. PSM deliberately masks the most text-aligned visual regions to induce semantic-grounded discrepancies that mirror inference conditions. As visualized in Fig. 4, PSM shifts the training similarity distribution from the highly aligned regime of CC3M pairs ($\mu \approx 0.24$) toward the lower-similarity regime characteristic of ZS-CIR benchmarks ($\mu \approx 0.17$), producing perturbed pairs ($\mu \approx 0.18$) that effectively narrow the train-test gap. For concrete phrases with clear visual correspondences, masking is focused. For abstract or global semantics, masking becomes diffuse, and DAC (as detailed below) models holistic discrepancies, ensuring DiffComp accommodates both localized attribute changes and broader stylistic modifications.

DAC: Spatially Adaptive Composition. Existing composition strategies vary in fusion granularity and difference awareness. SlerpTAT [9] applies a single global interpolation weight, ignoring spatial heterogeneity. HIT [11] fuses tokens in a discrepancy-agnostic manner, and PLI relies on global feature arithmetic with uniform weighting. DAC addresses these limitations by weighting fusion according to measured cross-modal differences. At the super-patch level, it assigns stronger textual weights to regions where visual content has been suppressed and weaker weights to preserved regions. At the global level, DAC captures holistic distributional shifts to complement localized adjustments, enabling both fine-grained attribute modifications and broader structural changes within a unified framework.

Module Synergy and Generalization. DiffComp's gains arise from the tight integration of CSS, PSM, and DAC. PSM constructs controlled semantic discrepancies, DAC re-

solves them adaptively, and CSS provides a structured semantic basis. This synergy allows robust compositional reasoning with minimal overhead: CSS adds only 7% training cost with ViT-L/14, and DAC scales linearly. These innovations generalize across backbones and datasets. DiffComp consistently outperforms strong baselines on CLIP-L/14, BLIP-L/16, and CLIP-G/14, and across FashionIQ, CIRR, CIRCO, and GeneCIS, showing that the observed gains stem from difference-aware composition rather than backbone scaling. This combination of paradigm-level novelty, module-level synergy, and empirical validation establishes DiffComp as a robust and efficient framework for zero-shot composed image retrieval.

H. Extended Discussion of Related Work

H.1. Zero-shot Composed Image Retrieval

Recently, CIR models have been trained solely on large-scale image–caption pairs or unlabeled images, enabling the zero-shot setting. A common strategy reformulates the retrieval task as text-to-image retrieval through visual-to-linguistic transformation. Methods such as [6, 17, 19] implicitly transform visual inputs into pseudo-word tokens via a learned mapping network, which are then concatenated with the modification text to form the composed query. To improve pseudo-word quality, some methods incorporate pre-mapping modules, such as extracting context-aware tokens [19] or filtering redundant regions. Others [6, 11] generate multiple pseudo-words to capture fine-grained details. Alternatively, LLM-based approaches [26, 27] reformulate CIR as a natural language inference task, where the reference image is described in natural language and combined with the modification text to form prompts for LLMs, which then generate the target image descriptions for retrieval. Despite promising zero-shot generalization, such language-driven paradigms depend heavily on external models and text generation quality. Our DiffComp avoids these limitations by operating directly in the feature space, modeling visual and textual discrepancies without auxiliary supervision or language generation.

H.2. Vision-Language Models for Fine-Grained Understanding

Vision-language models (VLMs) such as CLIP [16] have been widely adopted in zero-shot composed image retrieval (ZS-CIR) [11, 17, 19], owing to their strong cross-modal alignment and generalization capability. However, CLIP-style representations are inherently global, prioritizing overall scene semantics while underrepresenting localized structures [3, 5, 14, 24]. This weakness hinders the ability to reason about region-specific modifications, which is crucial in CIR tasks. Several efforts [14, 28] attempt to strengthen local alignment by associating vision patches with text tokens, but these methods typically require retraining large backbones and suffer from high computational cost. Our CSS module addresses this by grouping adjacent patches into semantic super-patches, enabling fine-grained composition while maintaining efficiency and compatibility with existing VLMs.

H.3. Masking Modeling in Vision-Language Pretraining

Masked Image Modeling (MIM) has emerged as a fundamental paradigm for visual pretraining, as demonstrated by MAE [8] and BEiT [2]. These approaches improve data efficiency and representation quality by reconstructing masked image regions. Inspired by this, masking strategies have also been extended to vision-language pretraining (VLP). For instance, FLIP [10] adopts random patch masking to accelerate training; A-CLIP [25] uses attention-based masking to retain patches most relevant to the text; E-CLIP [22] applies clustering-based masking to preserve structural visual information; and CLIP-PGS [15] integrates edge detection to protect object contours. While these approaches improve image–text alignment and efficiency, most of them emphasize preserving alignment rather than exploring misalignment cues that are crucial for compositional reasoning. Our Phrase-guided Selective Masking (PSM) differs by explicitly modeling semantic discrepancy: it leverages phrase-level textual guidance to mask image regions most relevant to the modification, thereby inducing cross-modal contrast and enabling the model to learn discrepancy-aware compositional features.

References

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

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

[3] Ioana Bica, Anastasija Ilic, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A. Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, and Jovana Mitrovic. Improving fine-grained understanding in image-text pre-training. In *Forty-first International Conference on Machine Learning*, 2024.

[4] Junyang Chen and Hanjiang Lai. Pretrain like your inference: Masked tuning improves zero-shot composed image retrieval. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2025.

[5] Gefen Dawidowicz, Elad Hirsch, and Ayellet Tal. Limitr: Leveraging local information for medical image-text representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21165–21173, 2023.

[6] Yongchao Du, Min Wang, Wengang Zhou, Shuping Hui, and Houqiang Li. Image2sentence based asymmetrical zero-shot composed image retrieval. In *The Twelfth International Conference on Learning Representations*, 2024.

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

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

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

[10] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23390–23400, 2023.

[11] Zhe Li, Lei Zhang, Zheren Fu, Kun Zhang, and Zhendong Mao. Hierarchy-aware pseudo word learning with text adaptation for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24319–24329, 2025.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[13] 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*, pages 2125–2134, 2021.

[14] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip HS Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19413–19423, 2023.

[15] Gensheng Pei, Tao Chen, Yujia Wang, Xinhao Cai, Xiangbo Shu, Tianfei Zhou, and Yazhou Yao. Seeing what matters: Empowering clip with patch generation-to-selection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24862–24872, 2025.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021.

[17] 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*, pages 19305–19314, 2023.

[18] Yucheng Suo, Fan Ma, Linchao Zhu, and Yi Yang. Knowledge-enhanced dual-stream zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26951–26962, 2024.

[19] Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5180–5188, 2024.

[20] Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Gaopeng Gou, and Qi Wu. Missing target-relevant information prediction with world model for accurate zero-shot composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24785–24795, 2025.

[21] 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*, pages 6862–6872, 2023.

[22] Zihao Wei, Zixuan Pan, and Andrew Owens. Efficient vision-language pre-training by cluster masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26815–26825, 2024.

[23] 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*, pages 11307–11317, 2021.

[24] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. In *International Conference on Learning Representations*, 2024.

[25] Yifan Yang, Weiquan Huang, Yixuan Wei, Houwen Peng, Xinyang Jiang, Huiqiang Jiang, Fangyun Wei, Yin Wang, Han Hu, Lili Qiu, et al. Attentive mask clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2771–2781, 2023.

[26] Zhenyu Yang, Shengsheng Qian, Dizhan Xue, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. Semantic editing increment benefits zero-shot composed image retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1245–1254, 2024.

[27] Zhenyu Yang, Dizhan Xue, Shengsheng Qian, Weiming Dong, and Changsheng Xu. Ldre: Llm-based divergent reasoning and ensemble for zero-shot composed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 80–90, 2024.

[28] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022.