FUSED EMBEDDING EDIT FOR ZERO-SHOT COMPOSED IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-Shot Composed Image Retrieval (ZS-CIR) retrieves target images using a reference image paired with a modification text, without training on costly real triplet annotations. Existing methods project images into text space or synthesize pseudo-visual features, ultimately performing cross-modal matching that suffers from an inherent *representation gap*, losing fine-grained visual details. We propose **FusionDiff**, a framework that addresses this representation gap through retrieval within a unified joint vision-language embedding space. Its conditional embedding diffusion model learns the distribution of target fusion embeddings conditioned on the composed query. For data efficiency, we design a lightweight Control-Adapter that adapts pre-trained diffusion backbones using only 200K synthetic triplets (*nearly two orders of magnitude less* than prior methods). Experiments on three benchmarks demonstrate state-of-the-art performance: +2.34 mAP@50 on CIRCO, +1.28 R@10 on CIRR, and +1.58 R@50 on FashionIQ over previous best methods.

1 INTRODUCTION

Composed Image Retrieval (CIR) retrieves target images using a query comprising a reference image and a natural-language modification, enabling precise, user-controllable search beyond single-modality retrieval (Baldrati et al., 2022; Lee et al., 2021; Liu et al., 2021). Consider a fashion designer seeking *"this dress but in emerald green with long sleeves"* or an interior architect searching for *"this living room but with modern furniture"*. Such compositional queries capture how humans conceptualize visual search: building upon existing examples while articulating desired modifications. Despite its promise for open-world search, fashion recommendation, and e-commerce, CIR faces two fundamental challenges: the *data gap* (supervised methods require costly real triplet annotations that generalize poorly (Delmas et al., 2022; Kim et al., 2021)) and the *representation gap* (aligning text-image compositions to visual targets remains difficult due to inherent modality differences (Wang et al., 2025; Li et al., 2025)).

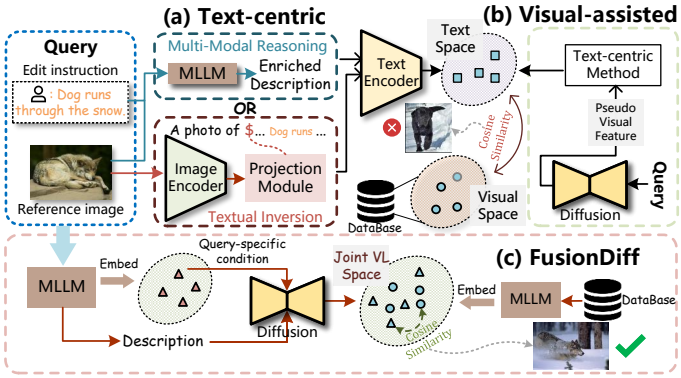


Figure 1: Comparison of Zero-Shot CIR paradigms. (a) *Text-centric* methods project images into text space, losing fine-grained visual details. (b) *Visual-assisted* methods synthesize pseudo-target features, but noise can misdirect retrieval. (c) Our *FusionDiff* uses a conditional embedding diffusion model for retrieval directly within the joint vision-language space.

Zero-Shot CIR (ZS-CIR) pursues several strategies to circumvent expensive annotations, as illustrated in Figure 1. *Text-centric* approaches include *inversion-based* methods (Saito et al., 2023; Baldrati et al., 2023; Tang et al., 2024) that map reference images into text space via textual inversion, simplifying the task to text-to-image retrieval but discarding fine-grained visual information; and *MLLM-based* methods (Karthik et al., 2024; Tang et al., 2025) that leverage multimodal large language models to generate enriched target descriptions, yet still perform cross-modal text-to-image matching. More recently, *visual-assisted* approaches (Wang et al., 2025; Li et al., 2025) synthesize

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

pseudo-target visual features via diffusion, but generative noise can misdirect retrieval. Despite their methodological differences, all paradigms share a fundamental limitation: *they rely on text or text-enhanced representations to match against visual features*, inevitably discarding fine-grained visual attributes that language cannot fully express. This representation gap becomes particularly detrimental for queries requiring fine-grained visual understanding, and motivates our central question: *Can we perform ZS-CIR directly within a unified vision-language space, addressing the cross-modal matching bottleneck?*

We present **FusionDiff**, a framework that reconceptualizes ZS-CIR as *conditional distribution modeling in a joint vision-language embedding space* (Figure 1(c)). Rather than projecting into text space or synthesizing pixel-level pseudo-targets, FusionDiff models the conditional distribution of target *fusion embeddings*—unified multimodal representations integrating both visual and linguistic semantics from a pre-trained MLLM within a deeply fused space, distinct from traditionally alignment. At inference, we sample multiple hypotheses from this distribution for robust ensemble retrieval directly within the joint VL space, preserving fine-grained visual information while respecting textual modifications. Realizing this vision presents two technical challenges:

- **Distribution Modeling.** How can we model the conditional distribution of target fusion embeddings during training? We address this with a *Conditional Embedding Diffusion Model* operating within the MLLM joint space, learning to transform query fusion embeddings toward the target distribution.
- **Data Efficiency.** How can we train a conditional diffusion model without large-scale annotated triplets? We propose a *Control-Adapter*—a lightweight, parameter-efficient module that freezes the pre-trained backbone and trains only adapter layers, achieving strong generalization with only 200K synthetic triplets, *nearly two orders of magnitude* less than prior diffusion-based methods.

Contribution. (1) We propose **FusionDiff**, the first ZS-CIR framework that performs retrieval within a unified vision-language space via conditional embedding diffusion, addressing the representation gap; (2) We design a parameter-efficient **Control-Adapter** that preserves the pre-trained generative prior while specializing for compositional editing with only 200K synthetic triplets. (3) We achieve **state-of-the-art** performance on three ZS-CIR benchmarks (CIRCO, CIRR, FashionIQ) with comprehensive validation of our design choices.

2 Related Work

Zero-Shot Composed Image Retrieval. Early CIR relied on full triplet supervision, but the prohibitive annotation cost has driven the community toward zero-shot variants. *Inversion-based* methods (Saito et al., 2023; Baldrati et al., 2023) compress the reference image into pseudo-tokens that are concatenated with the edit text for text-to-image retrieval; this textual compression discards perceptual details. *MLLM-based* methods (Karthik et al., 2024; Tang et al., 2025; Bao et al., 2025) prompt MLLMs to generate target descriptions that drive the search. Despite distinct mechanisms, both paradigms enrich the query primarily on the textual side and never incorporate explicit visual evidence of the target, leaving the representation gap unaddressed.

Multimodal Fusion. Dual-encoder models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) cannot represent compositional concepts in a single vector. MLLMs—e.g., BLIP-2 (Li et al., 2023) and PaLI-3 (Chen et al., 2023)—achieve deep cross-modal fusion but optimize for text generation rather than retrieval. Recent unified-embedding models (Girdhar et al., 2023; Liu et al., 2023; Lin et al., 2025; Zhang et al., 2025) model the static joint distribution of vision and language but do not address the CIR-specific challenge of capturing the *semantic transformation* from a composed query to its target. FusionDiff introduces a conditional embedding diffusion model that explicitly learns this transformation within the joint space.

Diffusion Models and CIR. Diffusion models (Ho et al., 2020; Song et al., 2021; Chen et al., 2024; Xie et al., 2025) have become state-of-the-art for generative modeling, with latent variants (Rombach et al., 2022) powering text-to-image synthesis (Saharia et al., 2022; Zhang et al., 2023) and editing (Brooks et al., 2023; Meng et al., 2022). Prior CIR adaptations (Wang et al., 2025; Li et al., 2025) synthesize pseudo-target images via pixel-space diffusion—an expensive two-stage pipeline bottlenecked by visual artifacts. While CompoDiff (Gu et al., 2024a) performs feature-space diffusion, it isolates captions as mere textual conditions, precluding deep joint understanding. FusionDiff

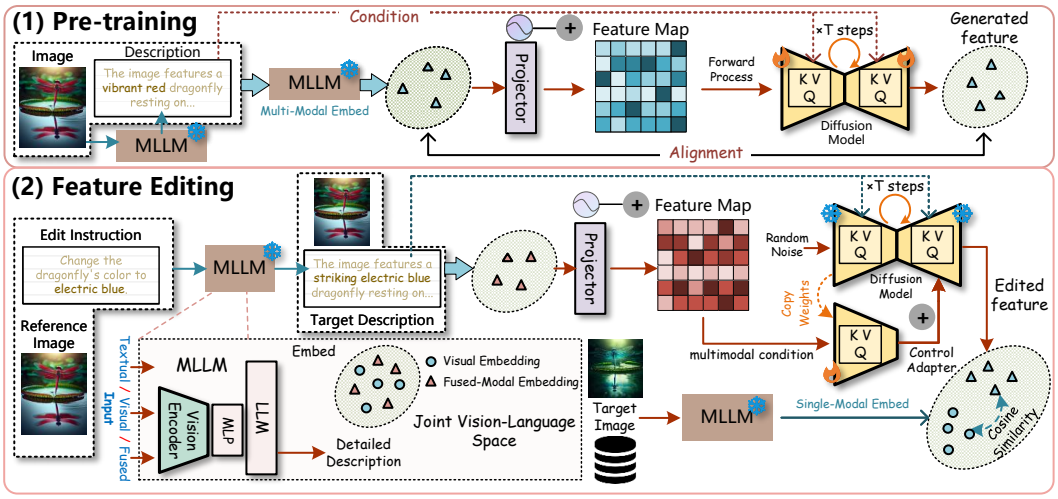


Figure 2: Overview of FusionDiff. **(1) Pre-training:** The MLLM produces fused embeddings in joint space $\mathcal{Z}$; a projector reshapes them into pseudo-spatial maps that are corrupted via diffusion, with the model learning to denoise under text conditioning. **(2) Compositional Editing:** For a composed query, the MLLM yields the query embedding and a target-oriented description; the frozen backbone with Control-Adapter denoises a target-embedding hypothesis aligned with gallery embeddings.

instead diffuses *within* the joint embedding space, sampling feature hypotheses rather than pixels, bypassing pixel artifacts and enabling retrieval directly within the joint space.

3 Preliminaries

Multimodal Large Language Models An MLLM $f_{\mathcal{M}}$ comprises a visual encoder $\mathcal{V}$, a modality alignment module $\mathcal{A}$, and a language model backbone $\mathcal{L}$. Given image I, visual patch embeddings $\mathbf{V} = \mathcal{V}(I) \in \mathbb{R}^{P \times d_v}$ are projected into the language model's token space via $\tilde{\mathbf{V}} = \mathcal{A}(\mathbf{V}) \in \mathbb{R}^{P \times d_l}$ and concatenated with text tokens to form a unified input sequence, which $\mathcal{L}$ processes to produce contextualized hidden states (full formulation in Appendix B).

Joint Vision-Language Embeddings. We extract a fixed-dimensional representation by ℓ_2-normalizing the final hidden state:

$$\frac{\mathbf{h}_{P+K}}{\|\mathbf{h}_{P+K}\|_2} \in \mathcal{Z} \subset \mathbb{R}^d, \tag{1}$$

where P is the number of visual patch tokens output by the visual encoder, K is the number of text tokens, and $\mathbf{h}_{P+K}$ is the hidden state at the final position after the language-model backbone jointly processes the P visual tokens and K text tokens. $\mathcal{Z}$ denotes the *joint vision-language embedding space* which maps images, text, and multimodal compositions to a shared manifold with semantically meaningful geometry, enabling direct similarity computation across heterogeneous inputs.

Diffusion Probabilistic Models Diffusion models define a generative process through paired forward corruption and learned reverse denoising chains. Given a clean sample $\mathbf{z}_0$, the forward process injects Gaussian noise over N steps: $\mathbf{z}_n = \sqrt{\bar{\alpha}_n}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_n}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\bar{\alpha}_n = \prod_{s=1}^{n}(1-\beta_s)$ is the cumulative noise schedule. A denoising network ϵ_θ is trained to invert this process by minimizing the noise-prediction objective: $\mathcal{L} = \mathbb{E}_{\mathbf{z}_0,\boldsymbol{\epsilon},n}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{z}_n, n, \mathbf{c})\|_2^2\right]$, where $\mathbf{c}$ is an optional conditioning signal. The reverse process parameterization and sampling schedule are detailed in Appendix B. FusionDiff applies this framework not to pixels or VAE latents, but directly to *joint vision-language embeddings* in $\mathcal{Z}$, enabling generative modeling of the target embedding distribution conditioned on the composed query.

4 METHODOLOGY

We propose **FusionDiff**, which performs composed image retrieval directly within a unified joint vision-language space $\mathcal{Z}$ via a two-stage training strategy: (1) pre-training a Conditional Embedding Diffusion Model (CEDM) to learn a generative prior over $\mathcal{Z}$, and (2) adapting it to compositional editing via a lightweight Control-Adapter. An architectural overview is given in Figure 2.

4.1 PRE-TRAINING THE CONDITIONAL EMBEDDING DIFFUSION MODEL

The first stage aims to learn a generative prior over the manifold of valid joint vision-language embeddings in $\mathcal{Z}$. This prior captures the intrinsic structure of multimodal representations and provides a robust foundation for subsequent compositional editing. By pre-training on large-scale image-text pairs, the diffusion model learns the statistical regularities of the embedding space without requiring expensive triplet annotations.

Training Data and Feature Extraction. We train on a large corpus of image-text pairs $\mathcal{D}_{\text{pre}} = \{(I_i, T_i)\}_{i=1}^{M}$ assembled from publicly available datasets (CC12M (Changpinyo et al., 2021), SA-1B (Kirillov et al., 2023), JourneyDB (Sun et al., 2023)), without requiring any triplet annotations. For each pair (I, T), we compute the fused multimodal embedding using the frozen MLLM $f_{\mathcal{M}}$ via an instruction-augmented input:

$$\mathbf{z}_0 = f_{\mathcal{M}}(\texttt{[INST]}_{\text{fuse}}, I, T) \in \mathcal{Z}, \tag{2}$$

where $\texttt{[INST]}_{\text{fuse}}$ is the natural-language instruction prefix that directs the MLLM to produce a fused multimodal embedding. The full set of mode-specific prompt templates to be discussed later is provided in Appendix C. The embedding $\mathbf{z}_0$ jointly encodes the visual content of I and the semantic information in T, lying on the manifold of valid joint representations. The MLLM parameters remain frozen throughout training, ensuring that $\mathbf{z}_0$ consistently reflects the pre-trained embedding geometry.

Text-Conditioned Denoising Objective. We employ a frozen text encoder $\mathcal{E}_{\text{txt}}$ (Flan-T5-XXL (Chung et al., 2024)) to obtain the conditioning embedding $\mathbf{c}_T = \mathcal{E}_{\text{txt}}(T) \in \mathbb{R}^{d_c}$. The denoising network ϵ_θ receives the noisy embedding $\mathbf{z}_n$, timestep n (encoded via sinusoidal positional embeddings), and $\mathbf{c}_T$ injected through cross-attention layers. The pre-training objective minimizes the expected ℓ_2 reconstruction error between true and predicted noise:

$$\mathcal{L}_{\text{S1}} = \mathbb{E}_{(I,T)\sim\mathcal{D}_{\text{pre}},\, \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}),\, n\sim\mathcal{U}(1,N)}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{z}_n, n, \mathbf{c}_T)\|_2^2\right], \tag{3}$$

where $\mathbf{z}_n = \sqrt{\bar{\alpha}_n}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_n}\,\boldsymbol{\epsilon}$ is the noisy embedding at timestep n. Through this objective, ϵ_θ learns a text-conditioned generative prior over $\mathcal{Z}$: given a textual description, it can generate plausible joint vision-language embeddings that are semantically consistent with the text while respecting the geometric structure of the embedding manifold. This learned prior serves as the foundation for the compositional editing stage.

Classifier-Free Guidance. To enable control over the diversity-fidelity trade-off at inference time, we incorporate classifier-free guidance (CFG) (Ho & Salimans, 2022) during pre-training. With probability p_{drop}, the text condition is replaced by a learnable null embedding $\boldsymbol{\varnothing} \in \mathbb{R}^{d_c}$, jointly training both the conditional distribution $p(\mathbf{z}_0|T)$ and the unconditional distribution $p(\mathbf{z}_0)$ within a single network. At inference, the guided noise estimate is:

$$\tilde{\epsilon}_\theta(\mathbf{z}_n, n, \mathbf{c}_T) = \epsilon_\theta(\mathbf{z}_n, n, \boldsymbol{\varnothing}) + \gamma\big[\epsilon_\theta(\mathbf{z}_n, n, \mathbf{c}_T) - \epsilon_\theta(\mathbf{z}_n, n, \boldsymbol{\varnothing})\big], \tag{4}$$

where $\gamma \geq 0$ controls the strength of text guidance. Higher γ yields samples more tightly aligned with the conditioning text, enhancing fidelity at the potential cost of diversity. Setting $\gamma = 1$ recovers standard conditional sampling. In practice, we treat γ as a hyperparameter tuned on a validation set.

Architecture. We instantiate ϵ_θ as a Diffusion Transformer (DiT) (Peebles & Xie, 2023) with $L = 28$ blocks and hidden dimension $D = 1152$, operating directly on fusion embeddings $\mathbf{z}_0 \in \mathbb{R}^d$ without intermediate VAE decoding. To leverage the inductive biases of vision transformers, the embedding is first projected to a pseudo-spatial feature map via a learnable linear projector $\mathcal{P} : \mathbb{R}^d \to \mathbb{R}^{4\times32\times32}$:

$$\mathbf{H}_n = \mathcal{P}(\mathbf{z}_n) \in \mathbb{R}^{4\times32\times32}, \tag{5}$$

which is then patchified ($p = 2$) into $M = 256$ tokens for transformer processing. Timestep conditioning uses Adaptive Layer Normalization (AdaLN); text conditioning uses Multi-Head Cross-Attention (MHCA). Detailed network architecture is provided in Appendix A.

4.2 COMPOSITIONAL EDITING VIA CONTROL-ADAPTER

The second stage adapts the pre-trained CEDM to the composed image retrieval task. The core challenge is to learn a conditional distribution $p(\mathbf{z}_t|I_r, T_\Delta)$ that transforms a reference image I_r according to a modification instruction T_Δ, producing an embedding that matches the hypothetical target image—all without access to large-scale triplet annotations.

Training Data and Embeddings. We fine-tune on 200K synthetic triplets $\mathcal{D}_{\text{ft}} = \{(I_r, T_\Delta, I_t, T_t)\}$ from HQEdit (Hui et al., 2025), constructed via automated mining of visually similar image pairs with synthetically generated modification instructions. This is nearly two orders of magnitude fewer than prior diffusion-based methods, enabled by the strong generative prior preserved from Stage I.

The frozen MLLM encodes two embeddings per triplet. The *query embedding* fuses the reference image with the modification instruction, capturing both visual context and semantic intent:

$$\mathbf{z}_{r,\Delta} = f_{\mathcal{M}}(\texttt{[INST]}_{\text{query}}, I_r, T_\Delta) \in \mathcal{Z}. \tag{6}$$

The *target embedding* encodes the ground-truth target image with its description, serving as supervision for the compositional transformation:

$$\mathbf{z}_t = f_{\mathcal{M}}(\texttt{[INST]}_{\text{target}}, I_t, T_t) \in \mathcal{Z}. \tag{7}$$

Fine-tuning Objective. Let $\mathbf{c}_\Delta = \mathcal{E}_{\text{txt}}(T_\Delta) \in \mathbb{R}^{d_c}$ denote the encoded modification instruction. The fine-tuning objective trains the denoiser to reconstruct $\mathbf{z}_t$ from its noised version, conditioned on both the query embedding and the modification text:

$$\mathcal{L}_{\text{S2}} = \mathbb{E}_{(I_r,T_\Delta,I_t,T_t)\sim\mathcal{D}_{\text{ft}},\, \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}),\, n\sim\mathcal{U}(1,N)}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{z}_n, n, \mathbf{z}_{r,\Delta}, \mathbf{c}_\Delta)\|_2^2\right], \tag{8}$$

where $\mathbf{z}_n = \sqrt{\bar{\alpha}_n}\,\mathbf{z}_t + \sqrt{1-\bar{\alpha}_n}\,\boldsymbol{\epsilon}$. Critically, the ground-truth target description T_t is excluded from the denoiser's conditioning inputs during fine-tuning, forcing the model to learn the semantic transformation from (I_r, T_Δ) to $\mathbf{z}_t$.

Control-Adapter Architecture. Naïvely fine-tuning all backbone parameters risks catastrophic forgetting of the rich generative prior learned in Stage I, particularly given the limited fine-tuning data. We therefore propose the **Control-Adapter**—a lightweight, parameter-efficient module that injects query-specific information while keeping the backbone frozen.

For each of the L frozen backbone encoder blocks $\{F_l(\cdot;\Theta_l)\}$, we maintain a parallel trainable *control block* $F_l^c(\cdot;\Theta_l^c)$ initialized as a copy of F_l. Two families of linear projections, $\mathcal{Z}^{(1)} : \mathbb{R}^d \to \mathbb{R}^{d_h}$ and $\{\mathcal{Z}_l^{(2)} : \mathbb{R}^{d_h} \to \mathbb{R}^{d_h}\}_{l=1}^L$, are both initialized with *zero weights and biases*. The modulated output at each layer is computed as:

$$\mathbf{y}_{c,l} = F_l(\mathbf{h}_{l-1};\Theta_l) + \mathcal{Z}_l^{(2)}(F_l^c(\mathbf{h}_{c,l-1};\Theta_l^c))\,, \tag{9}$$

where the control branch input is defined recursively as:

$$\mathbf{h}_{c,l-1} = \begin{cases} \mathbf{h}_0 + \mathcal{Z}^{(1)}(\mathbf{z}_{r,\Delta}), & l = 1, \\ \mathbf{y}_{c,l-1}, & l > 1. \end{cases} \tag{10}$$

For the decoder, control signals are injected via residual addition with a zero-initialized learnable scalar λ:

$$\mathbf{h}_l^{\text{dec}} = \tilde{\mathbf{h}}_l^{\text{dec}} + \lambda \cdot \mathbf{y}_{c,L-l+1}. \tag{11}$$

The zero initialization of $\mathcal{Z}^{(1)}$, $\{\mathcal{Z}_l^{(2)}\}$, and λ guarantees a *warm start*: at the beginning of fine-tuning, the adapter contributes zero output and the model exactly recovers the pre-trained backbone behavior. As training progresses, the projections gradually learn non-zero weights, enabling query-conditioned modulation without destabilizing the prior. Only adapter parameters are updated during fine-tuning (approximately 30% of backbone size), offering two key advantages: preserving the strong generative prior learned from large-scale pre-training, and substantially reducing trainable parameters to prevent overfitting on the small triplet dataset. Detailed parameterization is provided in Appendix A.

4.3 Inference Pipeline

At inference, FusionDiff generates pseudo-target embeddings for a composed query and performs retrieval entirely within $\mathcal{Z}$. The pipeline consists of three stages.

Query Encoding and Description Generation. Given a composed query (I_r, T_Δ), we first compute the query embedding $\mathbf{z}_{r,\Delta}$ via Eq. (6). We additionally leverage the MLLM's autoregressive generation capability to produce a target-oriented description of the hypothetical target image:

$$\tilde{T} = f_{\mathcal{M}}^{\text{gen}}(\texttt{[INST]}_{\text{desc}}, I_r, T_\Delta), \tag{12}$$

where $\texttt{[INST]}_{\text{desc}}$ prompts the model with instructions such as *"Describe the image that results from applying the following modification to the reference image."* and $f_{\mathcal{M}}^{\text{gen}}$ denotes the autoregressive text generation mode of the MLLM. The generated description $\tilde{T}$ incorporates visual attributes inferred from I_r and common-sense knowledge about the modification, and is encoded as $\mathbf{c}_{\tilde{T}} = \mathcal{E}_{\text{txt}}(\tilde{T})$ to provide semantic guidance for the reverse diffusion process.

Multi-Hypothesis Sampling. We generate S independent pseudo-target embeddings by running the reverse diffusion process with different random initializations. Starting from $\hat{\mathbf{z}}_N^{(s)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we iteratively apply the reverse diffusion update for timesteps $n = N, \ldots, 1$:

$$\hat{\mathbf{z}}_{n-1} = \frac{1}{\sqrt{\alpha_n}}\left(\hat{\mathbf{z}}_n - \frac{1-\alpha_n}{\sqrt{1-\bar{\alpha}_n}}\,\tilde{\epsilon}_\theta(\hat{\mathbf{z}}_n, n, \mathbf{z}_{r,\Delta}, \mathbf{c}_{\tilde{T}})\right) + \sigma_n \mathbf{w}_n, \tag{13}$$

where $\mathbf{w}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the stochastic noise term, $\sigma_n = \sqrt{\beta_n}$ is the noise standard deviation, and $\tilde{\epsilon}_\theta$ incorporates CFG as defined in Eq. (4). For accelerated inference, we use DPM-Solver++ (Lu et al., 2025), as detailed in Appendix B. The final pseudo-target embedding $\hat{\mathbf{z}}_t^{(s)} = \hat{\mathbf{z}}_0^{(s)} \in \mathcal{Z}$ is obtained after all N denoising steps. This multi-hypothesis strategy naturally captures the inherent ambiguity of compositional queries—multiple plausible interpretations of the target are explored in parallel.

Similarity-Based Ranking. Each gallery image I_g is pre-encoded as $\mathbf{z}_g = f_{\mathcal{M}}(\texttt{[INST]}_{\text{img}}, I_g) \in \mathcal{Z}$. We rank gallery images by max-pooled cosine similarity across all S hypotheses:

$$s_{\text{agg}}(I_g) = \max_{s \in \{1,\ldots,S\}} \frac{(\hat{\mathbf{z}}_t^{(s)})^\top \mathbf{z}_g}{\|\hat{\mathbf{z}}_t^{(s)}\|_2 \|\mathbf{z}_g\|_2}, \tag{14}$$

ensuring that a gallery image is relevant if it matches *any* plausible interpretation of the query. Both $\hat{\mathbf{z}}_t^{(s)}$ and $\mathbf{z}_g$ reside in the shared space, in which retrieval is conducted without the representation gap.

5 Experiments

We evaluate FusionDiff to address three core questions: **(Q1)** To what extent does FusionDiff establish a new state-of-the-art for ZS-CIR compared to existing cross-modal paradigms? **(Q2)** How do the conditional embedding diffusion model and the Control-Adapter synergistically contribute to accurate semantic alignment? **(Q3)** Does modeling directly within a unified vision-language space intrinsically bridge the representation gap inherent in traditional dual-encoder spaces?

5.1 Experimental Setup

Datasets & Baselines. We evaluate on three standard ZS-CIR benchmarks: **CIRR** (Liu et al., 2021) (21,552 real-world images, 4,148 test queries with fine-grained spatial/relational edits); **FashionIQ** (Wu et al., 2021) (77,684 fashion items with attribute-specific modifications); and **CIRCO** (Baldrati et al., 2023) (1,020 open-domain MS-COCO queries with multiple ground-truths). We benchmark against state-of-the-art inversion-based (Saito et al., 2023; Baldrati et al., 2023; Gu et al., 2024b), MLLM-based (Karthik et al., 2024; Yang et al., 2024; Tang et al., 2025), and generative-feedback paradigms (Gu et al., 2024a; Wang et al., 2025; Li et al., 2025).

Implementation Details. Following standard protocols, we report Recall@k, subset Recall$_s$@k, and mAP@k. We employ GME-Qwen2-VL-2B (Zhang et al., 2025) as the MLLM backbone and Flan-T5-XXL (Chung et al., 2024) for text conditioning. Stage 1 pre-trains on 33M image-text pairs

Table 1: **Results on CIRCO and CIRR.** Gray indicates training-free methods. **Bold**/underline denote best/second-best.

Dataset	Metric	*Reference*			*Inversion*		*Generative-Feedback*				*MLLM-Based*				*Ours*
		Image only	Text only	Image +Text	Pic2 Word	SEARLE -XL	LinCIR	LinCIR +CIG	LinCIR +IP-CIR	Compo Diff	CIReVL	LDRE	LDRE +IP-CIR	OSrCIR	**Fusion Diff**
CIRCO	mAP@5	1.34	2.56	2.65	8.72	11.68	19.71	20.64	25.70	15.33	26.77	31.12	32.75	30.47	**35.26**
	mAP@10	1.60	2.67	3.25	9.51	12.73	21.01	21.90	26.64	17.71	27.59	32.24	34.26	31.14	**36.58**
	mAP@25	2.12	2.98	4.14	10.64	14.33	23.13	24.04	29.09	19.45	29.96	34.95	36.86	35.03	**39.17**
	mAP@50	2.41	3.18	4.54	11.29	15.12	24.18	25.20	30.13	21.01	31.03	36.03	38.03	36.59	**40.37**
CIRR	R@1	6.89	21.81	11.71	23.90	24.24	35.25	34.43	35.37	26.71	34.65	36.15	39.25	37.26	**40.00**
	R@5	22.99	45.22	35.06	51.70	52.48	64.72	64.51	64.70	55.14	64.29	66.39	70.07	67.25	**70.27**
	R@10	33.68	57.42	48.94	65.30	66.29	76.05	76.12	76.15	74.52	75.06	77.25	80.00	77.33	**81.28**
	R@50	59.23	81.01	77.49	87.80	88.84	–	93.54	93.71	92.01	91.66	93.95	94.89	–	**95.33**
CIRR sub	R_s@1	21.04	62.24	32.77	53.76	53.76	63.35	62.24	62.58	64.54	67.95	68.82	**69.95**	69.22	69.83
	R_s@2	41.04	81.13	56.89	74.46	75.01	82.22	81.35	81.74	82.39	84.87	85.66	86.87	85.28	**86.90**
	R_s@3	60.31	90.70	74.96	87.07	88.19	91.98	91.28	91.35	91.81	93.21	93.76	94.22	93.55	**94.24**

Table 2: **Results on FashionIQ validation set.** Gray indicates training-free methods. **Bold**/underline denote best/second-best.

Category	Metric	*Reference*			*Inversion*		*Generative-Feedback*				*MLLM-Based*			*Ours*
		Image only	Text only	Image +Text	Pic2 Word	SEARLE	LinCIR	LinCIR +CIG	LinCIR +IP-CIR	Compo Diff	CIReVL	LDRE	OSrCIR	**Fusion Diff**
Shirt	R@10	6.92	19.87	13.44	26.20	26.89	46.76	47.35	48.04	41.31	34.01	35.94	38.65	**48.23**
	R@50	14.23	34.99	26.25	43.60	45.58	65.11	66.68	66.68	55.17	51.92	58.58	54.71	**68.45**
Dress	R@10	4.46	15.42	13.83	20.00	20.48	38.08	39.71	39.02	37.78	27.56	26.11	33.02	**41.15**
	R@50	12.19	35.05	30.88	40.20	43.13	60.88	60.93	61.03	49.10	50.04	51.12	54.78	**63.76**
Toptee	R@10	6.32	20.81	17.08	27.90	29.32	50.48	50.69	50.18	44.26	36.29	35.42	41.04	**51.30**
	R@50	13.77	40.49	31.67	47.40	49.97	71.09	71.39	71.14	56.41	56.63	56.67	61.83	**71.55**
Average	R@10	5.90	18.70	14.78	24.70	25.56	45.11	45.92	45.74	39.02	32.62	32.49	37.57	**46.89**
	R@50	13.37	36.84	29.60	43.70	46.23	65.69	66.34	66.28	51.71	52.86	55.46	57.11	**67.92**

(Changpinyo et al., 2021; Kirillov et al., 2023; Sun et al., 2023) with $N{=}1000$ diffusion steps. Stage 2 fine-tunes solely the Control-Adapter on 200K HQEdit triplets (Hui et al., 2025). Inference utilizes $N{=}14$ DPM-Solver++ steps, guidance scale $\gamma{=}2.5$, and adapter strength $\delta{=}1.0$. All experiments run on 8 NVIDIA A100 GPUs (Appendix J, K).

5.2 Main Results

CIRCO and CIRR (Table 1). FusionDiff achieves the best results on all CIRCO and CIRR metrics and on all but one CIRR_{sub} metric. On CIRCO, it improves upon the strongest baseline (LDRE+IP-CIR) by $+2.51$ absolute points on mAP@5 (a $+6.1\%$ relative gain at mAP@50). On CIRR, it achieves 40.00 and 81.28 at R@1 and R@10, respectively. In the subset evaluation (CIRR_{sub})—where distractors share visual similarity with the target—FusionDiff attains the best $\text{R}_s@\{2,3\}$ scores and is within 0.12 of the best $\text{R}_s@1$. This discriminative capability suggests that operating within the joint space preserves fine-grained visual details that cross-modal paradigms tend to discard.

FashionIQ (Table 2). FusionDiff achieves the best results in all three fashion categories, achieving average gains of $+0.97/+1.58$ over generative-feedback baselines at R@10/R@50. The largest improvement ($+1.44/+2.73$) occurs in **Dress**, the category demanding the most intricate attribute synthesis. MLLM-based methods underperform here, suggesting that subtle fashion attributes (e.g., fabric textures or exact silhouettes) are difficult to verbalize. This highlights the regime where our conditional embedding diffusion excels over language-bottlenecked approaches.

5.3 Ablation Studies

We conduct ablation studies to isolate the impact of our architectural components and validate the superiority of the joint vision-language space over standard dual-encoder alignments.

Component Necessity. Table 3 ablates the core diffusion mechanisms. Omitting diffusion entirely (Variant A: direct retrieval using raw fused query $\mathbf{z}_{r,\Delta}$) causes a 57.7% relative drop in CIRCO mAP@5, showing that diffusion-based semantic refinement is necessary. Applying Stage-1 text conditioning without the Control-Adapter's visual grounding (Variant B) performs *worse than no*

Table 3: **Ablation study.** (A) Direct retrieval without diffusion; (B) Stage-1 text conditioning only (no Control-Adapter); (C) Full FusionDiff.

Var.	Description	FashionIQ (Avg.) R@10	FashionIQ (Avg.) R@50	CIRCO mAP@5	CIRCO mAP@10	CIRCO mAP@25	CIRCO mAP@50
(A)	$\mathbf{z}_{r,\Delta}$-only: Direct retrieval without diffusion	40.57	61.22	14.91	16.44	18.57	19.50
(B)	Stage-1 only: Text conditioning without Control-Adapter	33.12	54.08	13.02	14.16	16.62	17.87
(C)	**FusionDiff**: Text conditioning + Control-Adapter	**46.89**	**67.92**	**35.26**	**36.58**	**39.17**	**40.37**

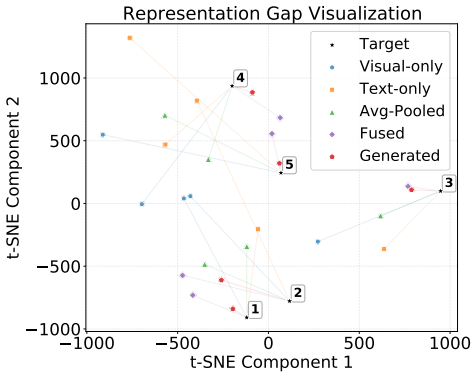


Figure 3: t-SNE Representation Gap Visualization on CIRR validation set.

Table 4: **CLIP space vs. joint VL space for embedding diffusion.** Only the underlying embedding space differs across variants.

Method	CIRR R@1	CIRR R@10	FashionIQ (Avg.) R@10	FashionIQ (Avg.) R@50
CLIP-Direct	26.14	63.57	31.22	53.89
CLIP-CEDM	29.83	68.41	36.47	58.63
$\mathbf{z}_{r,\Delta}$-Direct	36.52	77.91	40.57	61.22
FusionDiff	**40.00**	**81.28**	**46.89**	**67.92**

diffusion at all. Without structural visual cues, the conditional embedding diffusion model correctly matches the text but hallucinates or loses the reference image's foundational identity. Thus, the Control-Adapter is essential for anchoring generative edits to the reference image.

Joint VL Space vs. CLIP Space. Table 4 compares identical CEDM architectures in CLIP ViT-L/14 (d=768) and our joint VL space. CLIP variants use the ℓ_2-normalized weighted average of image and text embeddings for direct retrieval (CLIP-Direct) or diffusion input (CLIP-CEDM); Alternative CLIP fusion strategies are detailed in Appendix G. Unlike CLIP's independent modality encoding (Radford et al., 2021), the MLLM enables deep token-level cross-modal interactions, allowing $\mathbf{z}_{r,\Delta}$ to capture compositional intent. CEDM improves CIRR R@1 in both CLIP (+3.69) and joint VL (+3.48) spaces, with joint-space CEDM surpassing CLIP-CEDM (+10.17). These gains suggest that joint embeddings better capture compositional edits, enabling diffusion to refine query–target alignment.

5.4 Representation Gap Analysis

Figure 3 presents a t-SNE visualization of embeddings extracted from CIRR validation queries using our MLLM backbone. Visual-only, text-only, and average-pooled representations show pronounced separation from ground-truth target embeddings. In contrast, MLLM-derived fused embeddings $\mathbf{z}_{r,\Delta}$ substantially reduce this separation, while FusionDiff-generated embeddings exhibit closer alignment with the ground-truth targets. This progression validates that early deep multimodal fusion narrows the query–target representation gap, with CEDM refinement further improving alignment and alleviating the representation gap plaguing traditional ZS-CIR methodologies. Additional quantitative analysis is reported in Appendix H.

5.5 Qualitative Results and Sensitivity

Embedding Reconstructions. By fine-tuning a SANA 1.5 (Xie et al., 2025) decoder to visualize our generated fusion embeddings as spatial images, we observe that FusionDiff correctly resolves complex spatial relations on CIRR (Figure 4) and effectively models subtle physical manipulations on FashionIQ (Figure 5) while preserving base garment textures.

Hyperparameter Sensitivity. Figure 6 demonstrates FusionDiff's robustness, displaying a stable performance plateau around guidance scale γ=2.5 and adapter strength δ=1.0. Inference accuracy

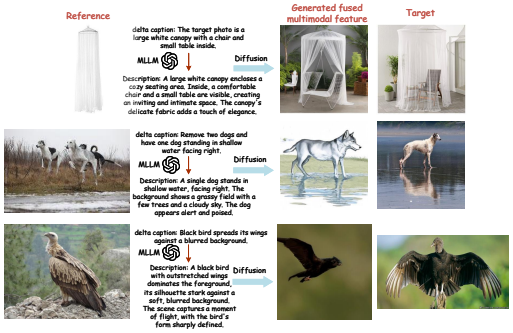
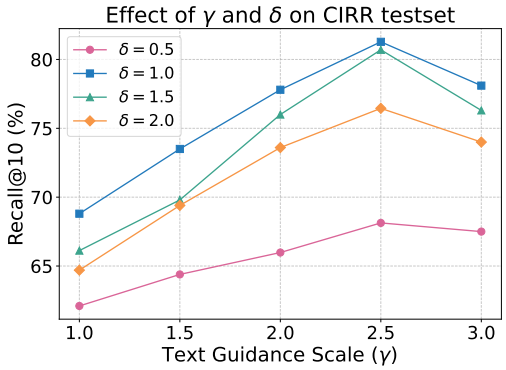


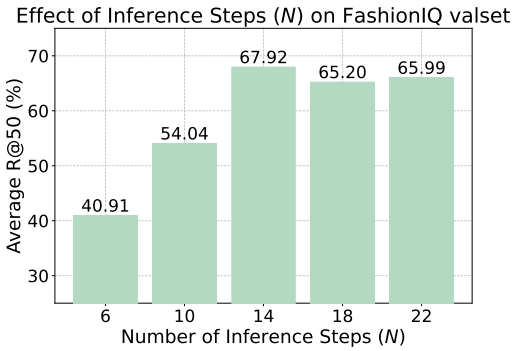
Figure 4: **Qualitative results on CIRR.** Reference image, modification, MLLM-generated description, decoded fusion-embedding visualization, and the retrieved target.



Figure 5: **Qualitative results on FashionIQ.** FusionDiff handles fine-grained attribute changes (color, pattern, structural cuts).



(a) Guidance scale γ & strength δ.



(b) Denoising steps N.

Figure 6: **Hyperparameter sensitivity.** (a) CIRR R@10 vs. (γ, δ); (b) FashionIQ R@50 vs. N.

on FashionIQ R@50 saturates at $N{=}14$ denoising steps. Extending steps yields no measurable gain while increasing latency, making $N{=}14$ the optimal tradeoff for real-time deployment ($\sim$0.32 s per query on a single A100). Further results are provided in Appendix N.

5.6 Data Efficiency

FusionDiff is data-efficient in triplet supervision: it requires only 200K Stage-2 triplets, nearly two orders of magnitude fewer than CompoDiff (18M). Although Stage 1 uses 33M image–text pairs, such pairs are abundant and public, whereas composed-retrieval triplets are costly to construct. We attribute this efficiency to the Control-Adapter, whose frozen backbone preserves the Stage-1 generative prior. FusionDiff remains competitive (46.01 R@10 on FashionIQ) with only 150K triplets (Appendix I).

6 Conclusion and Limitations

We present **FusionDiff**, which addresses the ZS-CIR representation gap via a *conditional embedding diffusion model* operating directly within a unified vision-language space. Driven by MLLM-generated semantics and a parameter-efficient *Control-Adapter*, FusionDiff achieves state-of-the-art results on CIRCO, CIRR, and FashionIQ using only 200K synthetic triplets—nearly two orders of magnitude less data than prior methods. One limitation is that iterative denoising incurs higher inference latency, and MLLM hallucinations can propagate to retrieval, necessitating uncertainty-aware sampling in future work.

AI USE STATEMENT

In this work, we used generative AI tools to formulate mathematical equations and to implement methods. We also used them to polish the paper for readability. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including any text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have made the following efforts to ensure reproducibility. The FusionDiff framework, including the conditional embedding diffusion model, the Control-Adapter, and the inference pipeline, is fully specified in Section 4, with architecture details and complete formulations in Appendices A and B, the prompt templates in Appendix C, and pseudocode for training and inference in Appendix L. All experiments use public data: Stage 1 pre-training uses public image–text corpora, Stage 2 uses the public HQEdit triplets, and evaluation uses the public CIRR, FashionIQ, and CIRCO benchmarks (Section 5); Appendix D describes the zero-shot protocol and data settings. Appendix J reports the training strategy and all hyperparameters of both stages, and Appendix K the compute resources, training cost, and inference speed. The MLLM backbone, the text encoder, and all datasets are obtained from their public sources. We will release the source code upon acceptance.

REFERENCES

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 21434–21442. IEEE, 2022. doi: 10.1109/CVPR52688.2022.02080. URL https://doi.org/10.1109/CVPR52688.2022.02080.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 15292–15301. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01407. URL https://doi.org/10.1109/ICCV51070.2023.01407.

Tong Bao, Che Liu, Derong Xu, Zhi Zheng, and Tong Xu. MLLM-I2W: harnessing multimodal large language model for zero-shot composed image retrieval. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pp. 1839–1849. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.125/.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 18-22, 2023*, pp. 18392–18402. IEEE Computer Society, 2023. doi: 10.1109/CVPR52729.2023.01764. URL https://doi.org/10.1109/CVPR52729.2023.01764.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 3558–3568. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00356. URL https://openaccess.thecvf.com/content/CVPR2021/html/Changpinyo_Conceptual_12M_Pushing_Web-Scale_Image-Text_Pre-Training_To_Recognize_Long-Tail_Visual_CVPR_2021_paper.html.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James T. Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=eAKmQPe3m1.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, Daniel Salz, Xi Xiong, Daniel Vlasic, Filip Pavetic, Keran Rong, Tianli Yu, Daniel Keysers, Xiaohua Zhai, and Radu Soricut. Pali-3 vision language models: Smaller, faster, stronger. *CoRR*, abs/2310.09199, 2023. doi: 10.48550/ARXIV.2310.09199. URL https://doi.org/10.48550/arXiv.2310.09199.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25:70:1–70:53, 2024. URL https://jmlr.org/papers/v25/23-0870.html.

Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. ARTEMIS: attention-based retrieval with text-explicit matching and implicit similarity. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=CVfLvQq9gLo.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 15180–15190. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01457. URL https://doi.org/10.1109/CVPR52729.2023.01457.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *Trans. Mach. Learn. Res.*, 2024, 2024a. URL https://openreview.net/forum?id=mKtlzW0bWc.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only efficient training of zero-shot composed image retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13225–13234. IEEE, 2024b. doi: 10.1109/CVPR52733.2024.01256. URL https://doi.org/10.1109/CVPR52733.2024.01256.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. doi: 10.48550/ARXIV.2207.12598. URL https://doi.org/10.48550/arXiv.2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Cihang Xie, and Yuyin Zhou. Hq-edit: A high-quality dataset for instruction-based image editing. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=mZptYYttFj.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-V: universal embeddings with multimodal large language models. *CoRR*, abs/2407.12580, 2024. doi: 10.48550/ARXIV.2407.12580. URL https://doi.org/10.48550/arXiv.2407.12580.

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=EDPxCjXzSb.

Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. Dual compositional learning in interactive image retrieval. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 1771–1779. AAAI Press, 2021. doi: 10.1609/AAAI.V35I2.16271. URL https://doi.org/10.1609/aaai.v35i2.16271.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 3992–4003. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00371. URL https://doi.org/10.1109/ICCV51070.2023.00371.

Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 802–812. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00086. URL https://openaccess.thecvf.com/content/CVPR2021/html/Lee_CoSMo_Content-Style_Modulation_for_Image_Retrieval_With_Text_Feedback_CVPR_2021_paper.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

You Li, Fan Ma, and Yi Yang. Imagine and seek: Improving composed image retrieval with an imagined proxy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 3984–3993. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00377. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_Imagine_and_Seek_Improving_Composed_Image_Retrieval_with_an_Imagined_CVPR_2025_paper.html.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal LLMS. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=i45NQb2iKO.

Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. Universal vision-language dense retrieval: Learning A unified representation space for multi-modal retrieval. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=PQOlkgsBsik.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 2105–2114. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00213. URL https://doi.org/10.1109/ICCV48922.2021.00213.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Mach. Intell. Res.*, 22(4): 730–751, 2025. doi: 10.1007/S11633-025-1562-4. URL https://doi.org/10.1007/s11633-025-1562-4.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *10th International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=aBsCjcPu_tE.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 4172–4182. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00387. URL https://doi.org/10.1109/ICCV51070.2023.00387.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 19305–19314. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01850. URL https://doi.org/10.1109/CVPR52729.2023.01850.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. Journeydb: A benchmark for generative image understanding. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/9bc59aff4685e39e1a8175d5303248a1-Abstract-Datasets_and_Benchmarks.html.

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 Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pp. 5180–5188. AAAI Press, 2024. doi: 10.1609/AAAI.V38I6.28324. URL https://doi.org/10.1609/aaai.v38i6.28324.

Yuanmin Tang, Jue Zhang, Xiaoting Qin, Jing Yu, Gaopeng Gou, Gang Xiong, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Wu. Reason-before-retrieve: One-stage reflective chain-of-thoughts for training-free zero-shot composed image retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 14400–14410. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01343. URL https://openaccess.thecvf.com/content/CVPR2025/html/Tang_Reason-before-Retrieve_One-Stage_Reflective_Chain-of-Thoughts_for_Training-Free_Zero-Shot_Composed_Image_Retrieval_CVPR_2025_paper.html.

Lan Wang, Wei Ao, Vishnu Naresh Boddeti, and Ser-Nam Lim. Generative zero-shot composed image retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 29690–29700. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02764. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Generative_Zero-Shot_Composed_Image_Retrieval_CVPR_2025_paper.html.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogério Feris. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 11307–11317. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01115. URL https://openaccess.thecvf.com/content/CVPR2021/html/Wu_Fashion_IQ_A_New_Dataset_Towards_Retrieving_Images_by_Natural_CVPR_2021_paper.html.

Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, Bingchen Liu, Daquan Zhou, and Song Han. SANA 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/xie25b.html.

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 Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (eds.), *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pp. 80–90. ACM, 2024. doi: 10.1145/3626772.3657740. URL https://doi.org/10.1145/3626772.3657740.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 3813–3824. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00355. URL https://doi.org/10.1109/ICCV51070.2023.00355.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 9274–9285. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00866. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Bridging_Modalities_Improving_Universal_Multimodal_Retrieval_by_Multimodal_Large_Language_CVPR_2025_paper.html.

A Architecture Details

A.1 Transformer Block Structure

As illustrated in Figure 7, the core denoising network ϵ_θ is built upon a Diffusion Transformer (DiT) architecture. Each of the $L = 28$ DiT (Peebles & Xie, 2023) blocks comprises Multi-Head Self-Attention (MHSA), Multi-Head Cross-Attention (MHCA) for text conditioning, and a Feed-Forward Network (FFN), composed as:

$$\mathbf{X}'_\ell = \mathbf{X}_{\ell-1} + \mathrm{MHSA}\big(\mathrm{AdaLN}(\mathbf{X}_{\ell-1}, n)\big), \tag{15}$$

$$\mathbf{X}''_\ell = \mathbf{X}'_\ell + \mathrm{MHCA}\big(\mathrm{AdaLN}(\mathbf{X}'_\ell, n), \mathbf{C}_{\text{text}}\big), \tag{16}$$

$$\mathbf{X}_\ell = \mathbf{X}''_\ell + \mathrm{FFN}\big(\mathrm{AdaLN}(\mathbf{X}''_\ell, n)\big), \tag{17}$$

where $\mathbf{C}_{\text{text}}$ is the text condition encoded by $\mathcal{E}_{\text{txt}}$ followed by an MLP, and AdaLN modulates hidden states with timestep-dependent scale and shift parameters:

$$\mathrm{AdaLN}(\mathbf{X}, n) = \gamma(\boldsymbol{\tau}) \odot \mathrm{LN}(\mathbf{X}) + \beta(\boldsymbol{\tau}), \tag{18}$$

where $\boldsymbol{\tau} \in \mathbb{R}^D$ is the timestep embedding produced by a sinusoidal encoding followed by an MLP, with $\gamma(\cdot)$ and $\beta(\cdot)$ being learned affine projections. This design allows the denoising network to adapt its normalization statistics at each layer based on the current noise level, providing fine-grained timestep-aware modulation without requiring separate normalization modules. Zero-initialized gating is applied to each residual branch for stable training dynamics, ensuring that newly added capacity starts with zero contribution before learning non-trivial transformations.

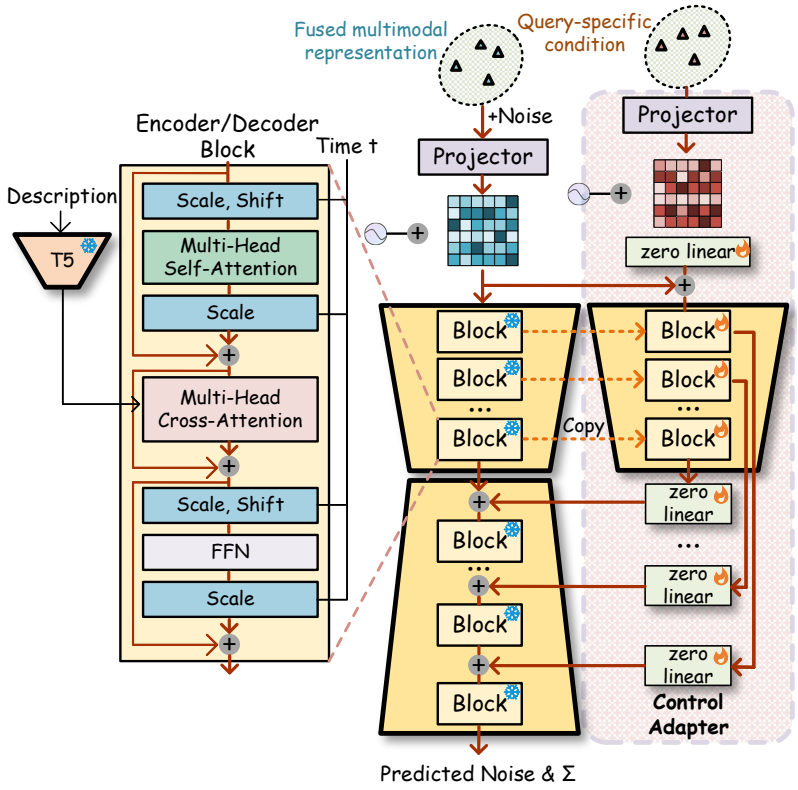


Figure 7: Detailed network architecture of the conditional embedding diffusion model of FusionDiff.

A.2 Feature Projection and Output

Since the DiT architecture was originally designed for 2D image latents, operating it directly on a 1D embedding vector would forgo the inductive biases of spatial self-attention patterns. We address

this with a learnable linear projector $\mathcal{P}$ that reshapes the fusion embedding $\mathbf{z}_n \in \mathbb{R}^{d=1536}$ into a pseudo-spatial feature map $\mathbf{H}_n \in \mathbb{R}^{4\times32\times32}$, which is then patchified into non-overlapping $p \times p$ patches (with $p=2$), yielding a sequence of $M=256$ tokens. These patches are linearly projected to a hidden dimension $D = 1152$ and augmented with sinusoidal positional embeddings:

$$\mathbf{X}_0 = \mathrm{Linear}(\mathrm{Patchify}(\mathbf{H}_n)) + \mathbf{E}_{\mathrm{pos}} \in \mathbb{R}^{M\times D}. \tag{19}$$

After processing through L transformer blocks, the final hidden states $\mathbf{X}_L$ are reshaped and projected to recover the predicted noise in the original embedding space via a learnable linear projector $\mathcal{P}_{\mathrm{out}}$:

$$\epsilon_\theta(\mathbf{z}_n, n, \mathbf{c}_T) = \mathcal{P}_{\mathrm{out}}(\mathbf{X}_L) \in \mathbb{R}^d. \tag{20}$$

The model is optimized using the diffusion objective in Eq. (3), learning to denoise in the joint vision-language embedding space.

A.3 Control-Adapter Parameterization

During Stage II fine-tuning, only the following parameters are updated while the backbone $\{\Theta_l\}_{l=1}^L$ remains frozen:

$$\Theta_{\text{train}} = \{\Theta_1^c, \ldots, \Theta_L^c\} \cup \{\mathbf{W}^{(1)}, \mathbf{b}^{(1)}\} \cup \{\mathbf{W}_l^{(2)}, \mathbf{b}_l^{(2)}\}_{l=1}^L \cup \{\lambda\}, \tag{21}$$

where $\mathbf{W}^{(1)}, \mathbf{b}^{(1)}$ are the weights and biases of the input projection $\mathcal{Z}^{(1)} : \mathbb{R}^d \to \mathbb{R}^{d_h}$, and $\mathbf{W}_l^{(2)}, \mathbf{b}_l^{(2)}$ are those of the per-layer output projections $\{\mathcal{Z}_l^{(2)} : \mathbb{R}^{d_h} \to \mathbb{R}^{d_h}\}_{l=1}^L$. All projection weights and biases, as well as the decoder injection scalar λ, are initialized to zero:

$$\mathbf{W}^{(1)}=\mathbf{0}, \quad \mathbf{b}^{(1)}=\mathbf{0}, \quad \mathbf{W}_l^{(2)}=\mathbf{0}, \quad \mathbf{b}_l^{(2)}=\mathbf{0}, \quad \lambda=0, \quad \forall l. \tag{22}$$

The control blocks $\{F_l^c\}$ are initialized as exact copies of the corresponding backbone blocks $\{F_l\}$. This zero-initialization scheme guarantees a *warm start*: at the onset of fine-tuning, the Control-Adapter contributes exactly zero to each residual branch, so the model behavior is identical to the pre-trained backbone. As training progresses, the projections gradually develop non-zero weights, enabling query-conditioned modulation without destabilizing the conditional embedding diffusion model learned in Stage I. In practice, the Control-Adapter adds approximately 30% additional parameters relative to the backbone encoder. Crucially, because the backbone is frozen, the rich generative prior over $\mathcal{Z}$ is fully preserved, and the adapter specializes only the query-conditioning pathway—explaining the strong generalization observed despite fine-tuning on only 200K triplets.

B Detailed Formulations

B.1 MLLM Architecture

An MLLM $f_{\mathcal{M}}$ comprises three components: a visual encoder $\mathcal{V}$, a modality alignment module $\mathcal{A}$, and a language model backbone $\mathcal{L}$. The visual encoder maps an input image $I \in \mathbb{R}^{H\times W\times 3}$ to patch embeddings:

$$\mathbf{V} = \mathcal{V}(I) = [\mathbf{v}_1, \ldots, \mathbf{v}_P] \in \mathbb{R}^{P\times d_v}. \tag{23}$$

The modality alignment module then projects these visual patch embeddings into the language model's token space:

$$\tilde{\mathbf{V}} = \mathcal{A}(\mathbf{V}) \in \mathbb{R}^{P\times d_l}, \tag{24}$$

bridging the representational gap between vision and language feature spaces. For multimodal input (I, T), text is tokenized into $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_K] \in \mathbb{R}^{K\times d_l}$. The unified sequence $\mathbf{X} = [\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_P, \mathbf{t}_1, \ldots, \mathbf{t}_K]$ is processed by the language model backbone $\mathcal{L}$ to produce contextualised hidden states $\mathbf{H} = \mathcal{L}(\mathbf{X})$, where every text token can attend to every image patch and vice versa through full bidirectional attention. This joint attention is the key architectural advantage over dual-encoder models such as CLIP (Radford et al., 2021), which encode image and text independently and therefore cannot model token-level cross-modal interactions. The joint embedding is extracted from the final sequence position and ℓ_2-normalized as in Eq. (1) of the main paper.

Table 5: Prompt templates used for different MLLM encoding modes.

Instruction	Prompt Template
$[\text{INST}]_{\text{fuse}}$	Represent the following image and its description as a unified embedding for retrieval: <image> <text>
$[\text{INST}]_{\text{query}}$	Represent the intended target for retrieval given by the source image and modification: <image> Modification: <text>
$[\text{INST}]_{\text{target}}$	Represent the following image and its description as a unified embedding for retrieval: <image> <text>
$[\text{INST}]_{\text{desc}}$	Describe the image that results from applying the following modification to the reference image: <image> Modification: <text>
$[\text{INST}]_{\text{img}}$	Represent the following image for retrieval: <image>

B.2 Conditional Embedding Diffusion Model: Reverse Process and DPM-Solver++ Sampling

The Conditional Embedding Diffusion Model (CEDM) explicitly models the conditional distribution $p_\theta(\mathbf{z}_0 \mid \mathbf{c})$ of target fusion embeddings given a conditioning signal $\mathbf{c}$ (e.g., modification text or query embedding). The reverse process factorizes as

$$p_\theta(\mathbf{z}_{0:N}) = p(\mathbf{z}_N) \prod_{n=1}^{N} p_\theta(\mathbf{z}_{n-1}|\mathbf{z}_n), \tag{25}$$

with each transition modeled as a Gaussian:

$$p_\theta(\mathbf{z}_{n-1}|\mathbf{z}_n) = \mathcal{N}\big(\mathbf{z}_{n-1};\, \boldsymbol{\mu}_\theta(\mathbf{z}_n, n),\, \sigma_n^2 \mathbf{I}\big)\,. \tag{26}$$

The mean is parameterized via the predicted noise ϵ_θ following the standard DDPM (Ho et al., 2020) reparameterization:

$$\boldsymbol{\mu}_\theta(\mathbf{z}_n, n) = \frac{1}{\sqrt{\alpha_n}} \left(\mathbf{z}_n - \frac{1-\alpha_n}{\sqrt{1-\bar{\alpha}_n}}\, \epsilon_\theta(\mathbf{z}_n, n, \mathbf{c}) \right). \tag{27}$$

Intuitively, the CEDM learns to iteratively refine a noisy embedding toward the manifold of valid target fusion embeddings that are consistent with the conditioning signal $\mathbf{c}$—analogous to how pixel-space diffusion models denoise toward natural images, but operating entirely within $\mathcal{Z}$.

DPM-Solver++ Sampling. For accelerated inference, we employ DPM-Solver++ (Lu et al., 2025), a fast, high-order ODE solver specifically designed for accelerating diffusion model sampling. DPM-Solver++ reformulates the reverse SDE as a probability-flow ODE and applies a tailored exponential integrator that achieves substantially lower truncation error than first-order methods (e.g., DDPM) at the same number of function evaluations. Concretely, using a second-order DPM-Solver++ with $N{=}14$ function evaluations, the per-query inference latency is approximately $0.32\,\text{s}$ on a single NVIDIA A100 GPU (excluding MLLM encoding time), with negligible accuracy degradation compared to the full 1000-step DDPM schedule. This represents a roughly $70\times$ reduction in diffusion sampling steps while maintaining retrieval performance. Including MLLM encoding, the total latency is approximately $0.67\,\text{s}$ per query, which remains practical for interactive retrieval.

C Prompt Templates

Table 5 lists the instruction prefixes used for different encoding modes. Each prefix is implemented as a natural-language prompt template and is prepended to the corresponding image, text, or multimodal input.

D Zero-Shot Protocol and Data Settings

In all experiments, we strictly adhere to the zero-shot composed image retrieval (ZS-CIR) protocol. Our model is never trained on any triplets, images, or annotations from the evaluation benchmarks

Table 6: Direct embedding retrieval baselines across benchmarks. FIQ denotes FashionIQ average performance. Although stronger VLM embeddings improve over CLIP-Direct, direct retrieval remains substantially below FusionDiff.

Method	CIRCO mAP@50	CIRR R@1	CIRR R@10	FIQ R@10	FIQ R@50
CLIP-Direct (Radford et al., 2021)	5.56	26.14	63.57	31.22	53.89
BLIP-2 Direct (Li et al., 2023)	8.30	27.09	65.17	31.94	55.32
E5-V-8B Direct (Jiang et al., 2024)	17.13	33.02	74.40	37.15	58.87
$\mathbf{z}_{r,\Delta}$-Direct	19.50	36.52	77.91	40.57	61.22
FusionDiff	**40.37**	**40.00**	**81.28**	**46.89**	**67.92**

Table 7: Comparison with naive supervised fine-tuning baselines. FIQ denotes FashionIQ average R@10. Direct supervised fine-tuning provides limited gains, whereas full fine-tuning of the diffusion model can degrade performance, indicating the importance of preserving the pretrained generative prior while adapting the conditioning pathway.

Method	CIRCO mAP@50	CIRR R@1	FIQ R@10
$\mathbf{z}_{r,\Delta}$-Direct	19.50	36.52	40.57
Naive SFT on MLLM	22.13	37.01	41.72
LoRA SFT on MLLM	22.30	36.65	40.94
CEDM + Naive Full FT	25.42	30.33	39.31
FusionDiff	**40.37**	**40.00**	**46.89**

(CIRR (Liu et al., 2021), FashionIQ (Wu et al., 2021), and CIRCO (Baldrati et al., 2023)). All training data are derived from public datasets with domains disjoint from the test sets, ensuring that evaluation images and captions remain completely unseen during the training phase. This strict separation guarantees that all reported improvements reflect genuine zero-shot generalization rather than data contamination.

E Direct Embedding and Supervised Fine-Tuning Baselines

Direct embedding retrieval. In Section 5.3, we evaluated whether pretrained vision-language embeddings can be used directly for composed image retrieval without diffusion-based refinement. To broaden this comparison, we additionally evaluate direct retrieval with BLIP-2 (Li et al., 2023), and E5-V-8B (Jiang et al., 2024). For each model, we construct a composed query embedding from the reference image and modification text, and retrieve gallery images by nearest-neighbor search in the corresponding embedding space.

As shown in Table 6, while the MLLM's fused embeddings already substantially outperform CLIP-Direct (e.g., +10.38 R@1 on CIRR), they still fall short of FusionDiff ($36.52 \rightarrow 40.00$ R@1 on CIRR; $40.57 \rightarrow 46.89$ Avg. R@10 on FashionIQ). As discussed in Section 2, recent unified-embedding models model the static joint distribution of vision and language but do not address the CIR-specific challenge of capturing the semantic transformation from a composed query to its target—precisely the distributional gap our conditional embedding diffusion bridges.

Naive supervised fine-tuning. We also compare FusionDiff with direct supervised fine-tuning alternatives. These baselines test whether the performance gain can be obtained by simply adapting the MLLM or by fully fine-tuning the diffusion model on triplet supervision.

As reported in Table 7, naive SFT and LoRA SFT on the MLLM underperform because supervised contrastive fine-tuning with 200K triplets learns a deterministic point-to-point mapping. However, CIR is inherently a one-to-many distributional problem (one query may correspond to multiple valid targets), which deterministic mappings fundamentally cannot capture. CEDM + Naive Full FT, which uses an MLP as the adapter, suffers from catastrophic forgetting: naively fine-tuning all diffusion parameters on only 200K triplets destroys the rich generative prior learned from 33M pairs in Stage 1, leading to performance below the no-training baseline on CIRR R@1 (30.33 vs. 36.52).

Table 8: Cross-backbone portability of FusionDiff. FIQ denotes FashionIQ average R@10. FusionDiff achieves comparable performance with both tested backbones.

Backbone	Params	CIRR R@1	CIRR R@10	CIRCO mAP@50	FIQ R@10
GME-Qwen2-VL (Zhang et al., 2025)	2B	**40.00**	**81.28**	**40.37**	**46.89**
E5-V (Jiang et al., 2024)	8B	38.72	79.85	38.94	45.31

Our Control-Adapter elegantly resolves this tension: it freezes the generative prior while adapting only the query-conditioning pathway ($\sim$30% of the backbone parameters), achieving the best of both worlds.

F Cross-Backbone Portability

We evaluate whether FusionDiff depends on a specific MLLM embedding backbone by replacing GME-Qwen2-VL (Zhang et al., 2025) with E5-V (Jiang et al., 2024), an alternative joint vision-language embedding model. For each backbone, we keep the overall CEDM and Control-Adapter design unchanged, while adapting the input dimensionality of the pseudo-spatial projector to match the corresponding embedding dimension. The diffusion model is trained in the embedding space of each backbone.

Table 8 shows that FusionDiff achieves comparable performance with both tested backbones, with a CIRR R@1 difference of 1.28 percentage points, supporting the applicability of the CEDM and Control-Adapter design to both embedding spaces. Both backbones outperform CLIP-CEDM in Table 4 by 8.89–10.17 percentage points in CIRR R@1, suggesting that the observed advantage of joint-space diffusion is not limited to GME-Qwen2-VL.

G CLIP-Based Fusion Details

CLIP is a dual-encoder model which encodes modalities independently and lacks deep token-level cross-modal interactions. For the CLIP-based variants, we first extract the image and text embeddings separately:

$$\mathbf{z}_I^{\mathrm{CLIP}} = \mathrm{CLIP}_{\mathrm{img}}(I_r), \qquad \mathbf{z}_T^{\mathrm{CLIP}} = \mathrm{CLIP}_{\mathrm{txt}}(T_\Delta). \tag{28}$$

The fused feature is obtained by a weighted average followed by ℓ_2 normalization:

$$\mathbf{z}_{\mathrm{fused}}^{\mathrm{CLIP}} = \frac{\alpha \mathbf{z}_I^{\mathrm{CLIP}} + (1-\alpha)\mathbf{z}_T^{\mathrm{CLIP}}}{\|\alpha \mathbf{z}_I^{\mathrm{CLIP}} + (1-\alpha)\mathbf{z}_T^{\mathrm{CLIP}}\|_2}, \quad \alpha = 0.5. \tag{29}$$

For CLIP-Direct, $\mathbf{z}_{\mathrm{fused}}^{\mathrm{CLIP}}$ is used directly as the retrieval query. For CLIP-CEDM, the same fused representation is used as the conditioning input to a CEDM trained in the CLIP embedding space with dimensionality $d = 768$. To verify that the performance gap between CLIP-CEDM and FusionDiff is not merely caused by this simple averaging strategy, we also evaluate alternative fusion mechanisms within CLIP-CEDM, including concatenation followed by an MLP and a cross-attention fusion module. As shown in Table 9, these CLIP-CEDM variants improve over the weighted-average baseline but remain substantially below FusionDiff in the MLLM joint embedding space.

H Quantitative Analysis of Representation Gap

To supplement the qualitative t-SNE visualization in Section 5.4, we quantify the representation gap on the CIRR validation set. For each query representation, we compute its mean cosine similarity and mean ℓ_2 distance to the corresponding target embedding. We also report centered kernel alignment (CKA) similarity to measure the global structural alignment between each representation set and the target embedding set.

As shown in Table 10, the representation gap monotonically decreases from single-modality representations to fused representations and further to FusionDiff-generated embeddings. This confirms

Table 9: Comparison of CLIP-space fusion strategies within CLIP-CEDM and FusionDiff in the MLLM joint embedding space. CLIP-space fusion improves when using learned fusion modules, but the gap to FusionDiff remains substantial.

Fusion Method	CIRR R@1	CIRR R@10	FashionIQ R@10	FashionIQ R@50
Weighted Average ($\alpha = 0.5$)	29.83	68.41	36.47	58.63
Concatenation + MLP	30.12	69.05	37.01	59.14
Cross-attention Fusion	31.47	70.23	37.89	59.72
FusionDiff (MLLM joint space)	**40.00**	**81.28**	**46.89**	**67.92**

Table 10: Quantitative representation gap analysis on the CIRR validation set. FusionDiff-generated embeddings are substantially closer to the target embeddings than single-modality, average-pooled, and raw fused representations.

Representation	Mean Cosine Sim. ↑	Mean ℓ_2 Dist. ↓	CKA Sim. ↑
Visual-only	0.312	1.172	0.287
Text-only	0.358	1.131	0.341
Avg-Pooled	0.401	1.094	0.392
Fused ($\mathbf{z}_{r,\Delta}$)	0.627	0.864	0.618
FusionDiff Generated	**0.781**	**0.661**	**0.773**

that the proposed conditional embedding diffusion model does not merely improve retrieval accuracy, but also moves the query representation closer to the target embedding distribution in the joint vision-language space.

I Effect of Stage-2 Training Set Size

To evaluate the data efficiency of FusionDiff, we vary the number of synthetic triplets used in Stage-2 fine-tuning while keeping the Stage-1 image-text pretraining and all model hyperparameters fixed. The default setting uses 200K triplets, corresponding to the natural size of the HQEdit training split used in our experiments.

As shown in Table 11, performance improves consistently as the number of Stage-2 triplets increases. Nevertheless, FusionDiff already achieves strong performance with 100K triplets and reaches highly competitive results with 150K triplets. This suggests that the model does not rely solely on large-scale triplet supervision; instead, the Stage-1 image-text pretraining provides a useful generative prior that improves data efficiency during task-specific fine-tuning.

J Training Strategy and Hyperparameters

Stage 1: Conditional Embedding Diffusion Model Pre-training. We instantiate the diffusion backbone using the DiT-XL/2 architecture configuration. Furthermore, we enforce full-precision attention computation to ensure numerical precision during the diffusion process, which is particularly important for the embedding-space denoising objective where small floating-point errors can accumulate across timesteps.

The model is optimized using AdamW (Loshchilov & Hutter, 2019) with a global batch size of 4096. We employ the following hyperparameter settings:

- Learning rate: $\eta = 2 \times 10^{-5}$;
- Weight decay: $\lambda = 3 \times 10^{-2}$;
- Optimizer stability term: $\varepsilon = 10^{-10}$.

To stabilize training dynamics under this large effective batch size, we apply global gradient clipping with an ℓ_2-norm threshold of 0.01. The model is trained for a total of 4 epochs on the Stage 1

Table 11: Effect of the number of Stage-2 synthetic triplets on retrieval performance.

Stage-2 Triplets	CIRCO	CIRR		FashionIQ Avg.	
	mAP@50	R@1	R@10	R@10	R@50
50K	33.59	35.25	76.20	41.93	60.27
100K	36.12	37.04	78.01	44.20	63.35
150K	38.80	38.63	80.59	46.01	66.11
200K	**40.37**	**40.00**	**81.28**	**46.89**	**67.92**

corpus. We utilize a learning rate warmup strategy for the initial $1{,}000$ optimization steps, followed by a cosine annealing schedule for the remainder of the training process. This warmup is necessary because the large batch size and randomly initialized parameters can produce large initial gradients; the cosine schedule then smoothly decays the learning rate to prevent overshooting near convergence.

Stage 2: Control-Adapter Fine-tuning. In the second stage, we freeze all parameters of the pre-trained conditional embedding diffusion model to fully preserve the robust generative prior over $\mathcal{Z}$ learned from large-scale pre-training. We instantiate the Control-Adapter by duplicating the weights of the DiT encoder blocks; the zero-initialization of the linear projection layers (connecting the adapter to the backbone) ensures that the initial forward pass is identical to the frozen CEDM, providing a stable starting point for fine-tuning. The model is fine-tuned on the synthetic HQEdit dataset containing approximately 200K triplets (Hui et al., 2025).

We continue to use the AdamW optimizer (Loshchilov & Hutter, 2019) for fine-tuning. Given the reduced dataset size and the parameter-efficient nature of the Control-Adapter, we adopt the following specific settings:

- Learning rate: $\eta = 2 \times 10^{-5}$;
- Weight decay: $\lambda = 2 \times 10^{-2}$.

Training is distributed across 8 NVIDIA A100 GPUs. To prevent overfitting to the synthetic edit patterns, we maintain the global gradient clipping threshold at 0.01. The Control-Adapter is trained for 2 epochs on the HQEdit dataset. We employ a constant learning rate schedule without warmup, as the Control-Adapter is initialized from pre-trained weights and the zero-initialization strategy already provides a smooth start—the adapter begins with zero output and only gradually diverges from the frozen CEDM as training progresses.

K Training Cost and Inference Speed

We provide a detailed analysis of the computational resources, training costs, and inference latency associated with FusionDiff. Table 12 summarizes the full configuration and efficiency metrics. It is important to note that Stage 1 is a pure pre-training phase used solely to build the conditional embedding diffusion model over the joint vision–language space $\mathcal{Z}$; retrieval inference is performed only after Stage 2 fine-tuning, utilizing the frozen CEDM together with the trained Control-Adapter.

Training Efficiency. All models were trained on a node equipped with $8\times$ NVIDIA A100 (80GB) GPUs. Pre-training is the most computationally intensive phase, requiring approximately 80 GPU-hours to align the DiT-XL/2 backbone (total depth $2L{=}28$ blocks) with the joint vision–language embedding space using the 33M-sample corpus. In contrast, Stage 2 is highly efficient: by freezing the backbone and training only the lightweight Control-Adapter on 200K synthetic triplets, we reduce the fine-tuning time to approximately 26 minutes (0.43 GPU-hours). This two-order-of-magnitude reduction in Stage 2 cost compared to Stage 1 demonstrates that, once the conditional embedding diffusion model is established over $\mathcal{Z}$, adapting FusionDiff to the composed image retrieval task is computationally inexpensive.

Accelerated Inference via DPM-Solver++. Standard diffusion sampling (e.g., DDPM (Ho et al., 2020)) typically necessitates hundreds to thousands of iterative denoising steps, introducing pro-

Table 12: Hyperparameters, training details, and compute resources for FusionDiff. Metrics are reported separately for Stage 1 (conditional embedding diffusion model pre-training) and Stage 2 (feature editing fine-tuning). Stage 1 is training-only; inference latency is reported only for the Stage 2 model, which uses DPM-Solver++ for accelerated sampling.

Parameter / Metric	**Stage 1: Pre-training**	**Stage 2: Feature Editing**
Architecture (DiT-XL/2)		
Hidden dimension (d)	1152	1152
Attention heads	16	16
Encoder/Decoder Blocks (L)	14	14
Patch size (p)	2	2
Trainable Parameters	954M	318M
Training Configuration		
Global Batch Size	4096	1024
Optimizer	AdamW	AdamW
Learning Rate	2×10^{-5}	2×10^{-5}
Weight Decay	3×10^{-2}	2×10^{-2}
Data Scale	33M (Image-Text Pairs)	200K (Synthetic Triplets)
Total Training Steps	$\sim$32,800	$\sim$400
Approx. Training Time	$\sim$80 GPU-hours (4 epochs)	$\sim$26 min (2 epochs)
Inference & Efficiency		
Sampler	—	DPM-Solver++
Inference Steps (N)	—	14
Guidance Scale (γ)	—	2.5
Control Scale (δ)	—	1.0
Inference Latency (Diffusion)	—	$\sim$0.32 s / query

hibitive latency for retrieval applications. To enable practical deployment, we employ DPM-Solver++ (Lu et al., 2025), a fast, high-order ODE solver for diffusion models. DPM-Solver++ exploits the semi-linear structure of the diffusion ODE and applies exponential integrators with analytically computed linear parts, achieving much lower discretization error per step than standard Euler-type solvers. Specifically, we utilize a second-order DPM-Solver++ with $N{=}14$ inference steps—a more than $70\times$ reduction compared to the 1000-step training schedule. This optimization reduces the per-query inference latency to approximately 0.32 s (excluding MLLM encoding time) on a single NVIDIA A100, with negligible degradation in retrieval accuracy relative to the full-step baseline. Including MLLM encoding, the total latency is approximately 0.67 s per query, which remains practical for interactive retrieval.

L Algorithm

We present the complete algorithmic framework of FusionDiff, encompassing four key components: (1) **VL-Embed** (Algorithm 1): construction of the joint vision–language manifold using the MLLM; (2) **Stage 1 Pre-training** (Algorithm 2): learning the conditional embedding diffusion model over fused multimodal embeddings; (3) **Stage 2 Fine-tuning** (Algorithm 3): freezing the conditional embedding diffusion model and training the Control-Adapter for query-specific compositional editing; and (4) **FusionDiff Inference** (Algorithm 4): the complete zero-shot retrieval pipeline that samples pseudo-target embeddings and ranks gallery images.

M Additional Qualitative Results

To provide a more comprehensive assessment of FusionDiff, we present additional qualitative comparisons against two representative zero-shot baselines: CIReVL (Karthik et al., 2024), a training-free LLM-reasoning method, and LinCIR+CIG (Wang et al., 2025), a generative method that synthesizes pseudo-target images in pixel space. We visualize Recall@1 retrieved results on both CIRR and FashionIQ. These examples substantiate our claim that modeling the conditional

Algorithm 1 VL-Embed: Joint Vision-Language Space Construction

Require: Image-Text Pair Dataset $\mathcal{D}_{\text{pair}}$, frozen MLLM f_M
Ensure: Latent Dataset $\mathcal{D}_{\text{latent}} = \{(\mathbf{z}_0, \mathbf{c}_T)\}$
$\mathcal{D}_{\text{latent}} \leftarrow \emptyset$
for each pair $(I, T) \in \mathcal{D}_{\text{pair}}$ **do**
  // Step 1: Construct multimodal input with fusion instruction
  $x_{\text{pair}} \leftarrow \texttt{[INST]}_{\text{fuse}} \oplus I \oplus T$
  // Step 2: Extract fused joint embedding
  $\mathbf{z}_0 \leftarrow f_M(x_{\text{pair}})$ ▷ Joint VL embedding in $\mathbb{R}^{d_{\text{VL}}}$
  $\mathbf{z}_0 \leftarrow \mathbf{z}_0 / \|\mathbf{z}_0\|_2$ ▷ ℓ_2-normalize onto the embedding manifold
  // Step 3: Extract text conditioning embedding
  $\mathbf{c}_T \leftarrow \mathcal{E}_{\text{text}}(T)$ ▷ Text encoder output for diffusion conditioning
  $\mathcal{D}_{\text{latent}} \leftarrow \mathcal{D}_{\text{latent}} \cup \{(\mathbf{z}_0, \mathbf{c}_T)\}$
end for
return $\mathcal{D}_{\text{latent}}$

Algorithm 2 Stage 1: Conditional Embedding Diffusion Model Pre-training

Require: Latent Dataset $\mathcal{D}_{\text{latent}}$, DiT denoiser ϵ_θ
Ensure: Pre-trained CEDM parameters Θ_{prior}
Randomly initialize Θ_{prior}
for epoch $= 1$ to N_{epochs} **do**
  for each $(\mathbf{z}_0, \mathbf{c}_T) \in \mathcal{D}_{\text{latent}}$ **do**
    $n \sim \mathcal{U}(1, N)$ ▷ Sample diffusion timestep
    $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\mathbf{z}_n \leftarrow \sqrt{\bar{\alpha}_n}\,\mathbf{z}_0 + \sqrt{1 - \bar{\alpha}_n}\,\boldsymbol{\epsilon}$ ▷ Forward diffusion
    // Classifier-Free Guidance (CFG) training
    $u \sim \mathcal{U}(0, 1)$
    if $u < p_{\text{cfg}}$ **then** ▷ Drop text condition with probability $p_{\text{cfg}} = 0.1$
      $\mathbf{c}_{\text{cond}} \leftarrow \varnothing$ ▷ Learnable null embedding
    else
      $\mathbf{c}_{\text{cond}} \leftarrow \mathbf{c}_T$
    end if
    $\boldsymbol{\epsilon}_{\text{pred}} \leftarrow \epsilon_\theta(\mathbf{z}_n, n, \mathbf{c}_{\text{cond}})$
    $\mathcal{L}_{\text{S1}} \leftarrow \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\text{pred}}\|_2^2$ ▷ Noise-prediction loss
    Update Θ_{prior} via $\nabla_{\Theta_{\text{prior}}} \mathcal{L}_{\text{S1}}$
  end for
end for
return Θ_{prior}

Algorithm 3 Stage 2: Control-Adapter Fine-tuning for Compositional Editing

Require: Triplet Dataset $\mathcal{D}_{\text{edit}}$, pre-trained conditional embedding diffusion model $\epsilon_\theta(\cdot;\Theta_{\text{prior}})$
Ensure: Trained Control-Adapter parameters Θ_{ctrl}

// Initialize Control-Adapter with warm start
Freeze Θ_{prior} ▷ Preserve the conditional embedding diffusion model
Initialize control blocks $\{F_l^c\}$ as copies of backbone blocks $\{F_l\}$
Zero-initialize all projection weights $\mathbf{W}^{(1)}$, $\{\mathbf{W}_l^{(2)}\}$, biases, and scalar λ ▷ Warm start: adapter outputs zero initially
for iteration $= 1$ to N_{iter} **do**
  Sample $(I_r, T_\Delta, I_t, T_t) \sim \mathcal{D}_{\text{edit}}$
  $\mathbf{z}_{r,\Delta} \leftarrow f_M(\texttt{[INST]}_{\text{query}}, I_r, T_\Delta)$ ▷ Query embedding: fuses reference image and modification
  $\mathbf{z}_t \leftarrow f_M(\texttt{[INST]}_{\text{target}}, I_t, T_t)$ ▷ Target embedding: ground-truth supervision
  $\mathbf{c}_\Delta \leftarrow \mathcal{E}_{\text{text}}(T_\Delta)$ ▷ Text conditioning for the modification instruction
  $n \sim \mathcal{U}(1, N), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{z}_n \leftarrow \sqrt{\bar{\alpha}_n}\,\mathbf{z}_t + \sqrt{1-\bar{\alpha}_n}\,\boldsymbol{\epsilon}$ ▷ Corrupt target embedding
  // Control-Adapter forward pass
  Compute per-layer control signals $\{y_{c,l}\}_{l=1}^{L}$ using $\mathbf{z}_{r,\Delta}$ and $\mathbf{z}_n$
  $\boldsymbol{\epsilon}_{\text{pred}} \leftarrow \epsilon_\theta(\mathbf{z}_n, n, \mathbf{c}_\Delta, \{y_{c,l}\})$
  $\mathcal{L}_{\text{S2}} \leftarrow \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\text{pred}}\|_2^2$ ▷ Noise-prediction loss conditioned on query
  Update Θ_{ctrl} via $\nabla_{\Theta_{\text{ctrl}}}\mathcal{L}_{\text{S2}}$
end for
return Θ_{ctrl}

Algorithm 4 FusionDiff Inference: Zero-Shot Composed Image Retrieval

Require: Composed query (I_r, T_Δ), image gallery $\mathcal{D}_{\text{gal}}$, number of hypotheses S
Ensure: Ranked list of gallery images

// Step 1: Query encoding and description generation
$\mathbf{z}_{r,\Delta} \leftarrow f_M(\texttt{[INST]}_{\text{query}}, I_r, T_\Delta)$ ▷ Joint query embedding
$\tilde{T} \leftarrow f_M^{\text{gen}}(\texttt{[INST]}_{\text{desc}}, I_r, T_\Delta)$ ▷ MLLM generates target-oriented description
$\mathbf{c}_{\tilde{T}} \leftarrow \mathcal{E}_{\text{text}}(\tilde{T})$ ▷ Text conditioning for reverse diffusion
// Step 2: Multi-hypothesis sampling via DPM-Solver++
for $s = 1$ to S **do**
  $\hat{\mathbf{z}}_N^{(s)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ Independent noise initialization for each hypothesis
  for each DPM-Solver++ step $n = N, \ldots, 1$ **do**
    $\boldsymbol{\epsilon}_{\text{cond}} \leftarrow \epsilon_\theta(\hat{\mathbf{z}}_n^{(s)}, n, \mathbf{c}_{\tilde{T}}, \mathbf{z}_{r,\Delta})$ ▷ Conditional prediction
    $\boldsymbol{\epsilon}_{\text{uncond}} \leftarrow \epsilon_\theta(\hat{\mathbf{z}}_n^{(s)}, n, \boldsymbol{\varnothing}, \boldsymbol{\varnothing})$ ▷ Unconditional prediction
    $\tilde{\boldsymbol{\epsilon}} \leftarrow \boldsymbol{\epsilon}_{\text{uncond}} + \gamma\big[\boldsymbol{\epsilon}_{\text{cond}} - \boldsymbol{\epsilon}_{\text{uncond}}\big]$ ▷ Classifier-free guidance
    $\hat{\mathbf{z}}_{n-1}^{(s)} \leftarrow \text{DPM-Solver++Step}(\hat{\mathbf{z}}_n^{(s)}, \tilde{\boldsymbol{\epsilon}}, n)$ ▷ High-order ODE solver update
  end for
  $\hat{\mathbf{z}}_t^{(s)} \leftarrow \hat{\mathbf{z}}_0^{(s)}$ ▷ Pseudo-target embedding for hypothesis s
end for
// Step 3: Max-pooled similarity-based gallery ranking
for each gallery image $I_g \in \mathcal{D}_{\text{gal}}$ **do**
  $\mathbf{z}_g \leftarrow f_M(\texttt{[INST]}_{\text{img}}, I_g)$ ▷ Gallery embedding (pre-computed offline)
  $s_{\text{agg}}(I_g) \leftarrow \max_{s\in\{1,\ldots,S\}} \frac{(\hat{\mathbf{z}}_t^{(s)})^\top \mathbf{z}_g}{\|\hat{\mathbf{z}}_t^{(s)}\|_2\,\|\mathbf{z}_g\|_2}$
end for
return Gallery images ranked in descending order of s_{agg}

distribution directly in the joint vision–language space yields superior alignment compared to text-centric reasoning or pixel-space generation.

M.1 Additional Qualitative Results on CIRR

Figure 8 illustrates qualitative results on the CIRR validation set. Queries in CIRR often involve complex spatial modifications, viewpoint changes, or state transitions that are difficult to capture through text alone.

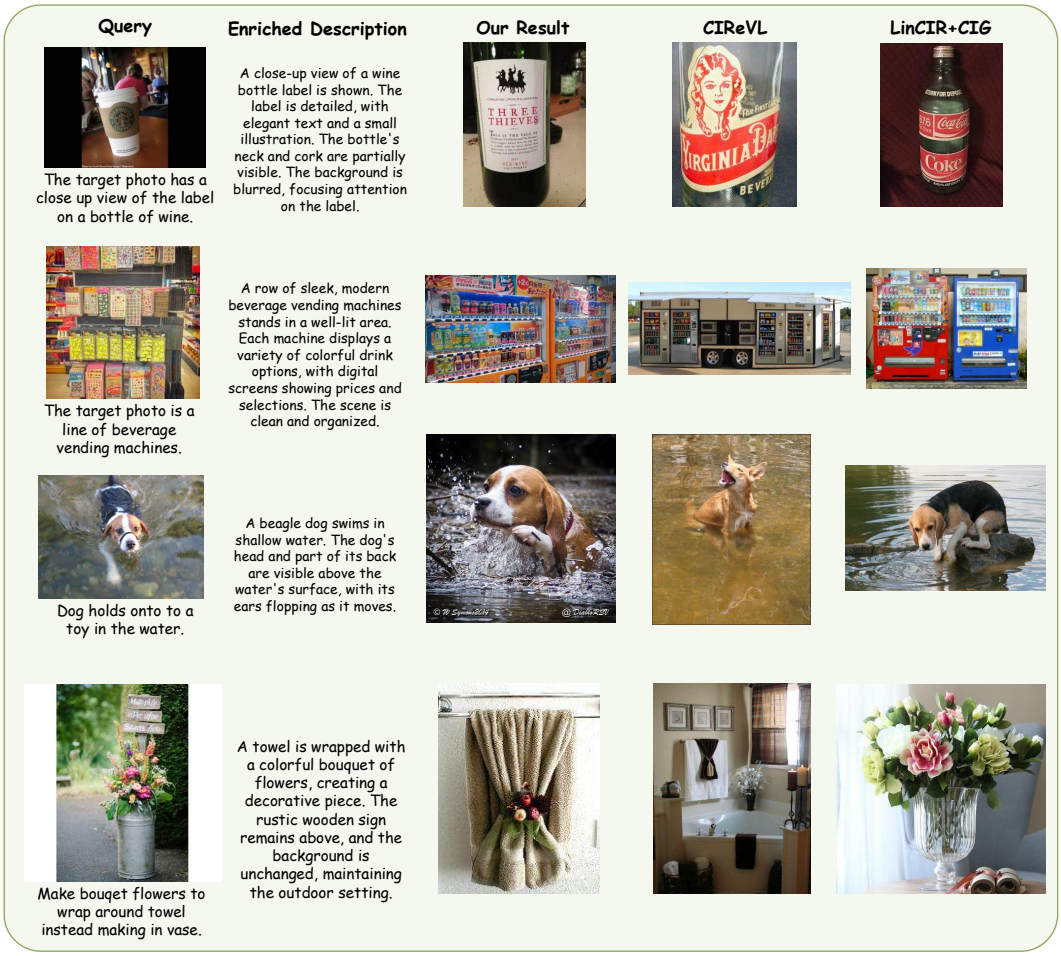


Figure 8: Additional qualitative comparison on the **CIRR** validation set (Recall@1). We compare FusionDiff against CIReVL (Karthik et al., 2024) and LinCIR+CIG (Wang et al., 2025). FusionDiff accurately captures fine-grained visual details and spatial relationships where baselines suffer from semantic drift or hallucination.

Analysis of Semantic Consistency. In the first row (Wine Bottle), the query demands a specific visual viewpoint change ("close up view of the label"). LinCIR+CIG suffers from severe hallucination, retrieving a carbonated soft drink bottle that is visually distinct from the queried wine bottle category. This failure likely stems from noise inherent in pixel-space diffusion: when the textual condition ("bottle") is broad, the generative model produces a plausible but incorrect image category, and the subsequent retrieval step propagates this error. CIReVL, relying heavily on textual reasoning, retrieves a bottle with a completely different label structure and color palette compared to the ground truth—the text description alone cannot fully specify the fine-grained visual identity of the reference. By contrast, FusionDiff successfully retrieves the correct target, demonstrating that joint-space diffusion preserves the fine-grained visual identity and label layout of the reference while faithfully executing the viewpoint change.

Analysis of Complex Compositional Queries. The fourth row (Flowers and Towel) presents a challenging compositional query: "wrap around towel instead making in vase." CIReVL fails to capture the "wrapping" action, retrieving a standard bathroom interior that ignores the flowers entirely. LinCIR+CIG retrieves a generic vase, ignoring the modification instruction entirely—the pixel-space synthesis focuses on the dominant visual category ("vase") and loses the semantic relationship specified by the edit. FusionDiff is the only method that retrieves the correct image where the bouquet is tied around the towel. This highlights the effectiveness of the **Control-Adapter**, which injects the structural layout of the query embedding directly into the conditional embedding diffusion model, enabling precise compositional reasoning that pure language models and pixel-space generators miss.

M.2 Additional Qualitative Results on FashionIQ

Figure 9 showcases results on the FashionIQ validation set, focusing on fine-grained attribute manipulation such as texture, pattern, and structural edits.

Pattern and Symbol Alignment. The second row (Tie-dye Shirt) provides a compelling example of fine-grained pattern recognition. The edit instruction explicitly asks for "a heart in the center." LinCIR+CIG retrieves a shirt with a peace sign—a common failure mode in generative retrieval where the synthesized pseudo-visual features capture the general style (tie-dye) but hallucinate the wrong symbol due to noise in the pixel-space diffusion model, which cannot reliably distinguish between abstract graphic motifs. CIReVL retrieves a shirt with text ("Peace Love Softball"), failing to align with the visual concept of a "heart"—language descriptions do not uniquely specify visual symbols. FusionDiff accurately retrieves the shirt featuring a heart pattern, confirming that operating in the joint VL space avoids the modality gap where text descriptions fail to fully encapsulate complex visual patterns.

Attribute and Silhouette Fidelity. In the first row (Grey Dress), the query specifies "one shoulder strap." LinCIR+CIG retrieves a dress with a completely different silhouette (flared skirt, standard straps), failing to preserve either the structural edit or the reference's fitted cut. CIReVL retrieves a dress that partially satisfies the "one shoulder" attribute but misses the specific texture (sequins) and silhouette fit of the reference. FusionDiff retrieves the exact match, preserving the tight-fitting silhouette and the metallic texture of the reference while satisfying the structural one-shoulder edit. This case illustrates a key advantage of embedding-level diffusion: the conditional embedding diffusion model jointly encodes structural and textural attributes within a single continuous embedding, avoiding the information bottleneck introduced by intermediate language descriptions.

Summary. Across both datasets, the comparisons reveal that FusionDiff substantially reduces two types of errors prevalent in baseline methods: (1) *hallucination* of irrelevant objects or symbols in generative methods (LinCIR+CIG), caused by noise in pixel-space synthesis; and (2) *semantic drift* in text-centric methods (CIReVL), where fine-grained visual constraints are lost during text-to-vision cross-modal matching. By grounding retrieval within the joint vision–language space through the conditional embedding diffusion model and the Control-Adapter, FusionDiff simultaneously avoids both failure modes, which strongly corroborates the quantitative gains reported in the main paper.

N Accuracy–Latency Trade-off for Denoising Steps

We evaluate the effect of the number of denoising steps N on retrieval accuracy and inference latency. All settings use the same trained model and differ only in the number of reverse denoising steps used at inference time. Latency is measured as diffusion-side inference time per query.

As shown in Table 13, performance degrades gracefully when using fewer denoising steps. At $N = 10$, FusionDiff already achieves 39.31 CIRR R@1, while reducing diffusion latency to approximately 0.22 s per query. Increasing N beyond 14 brings negligible improvements on CIRR but increases latency. We therefore use $N = 14$ as the default setting, which offers a favorable accuracy–latency trade-off.

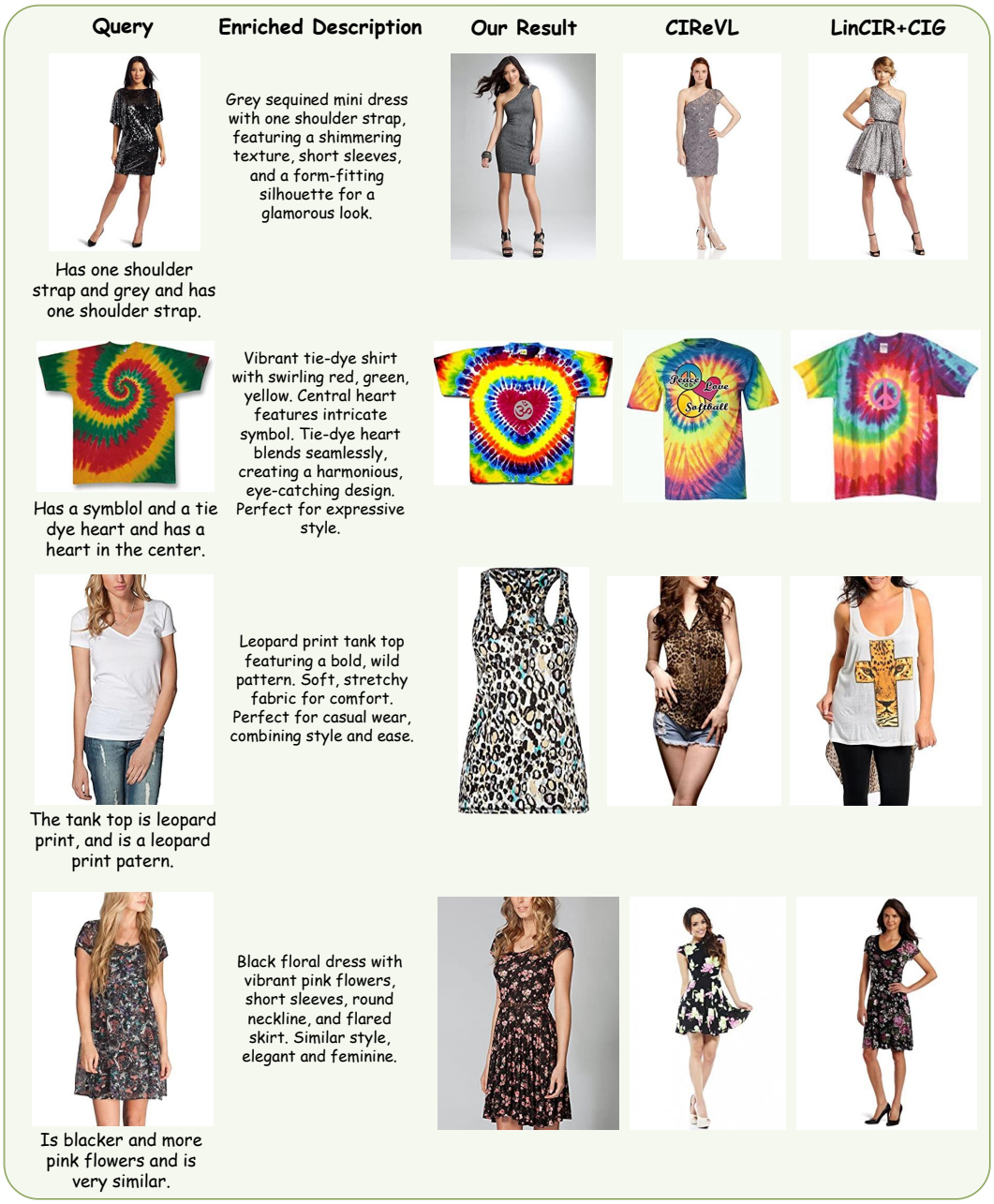


Figure 9: Additional qualitative comparison on the **FashionIQ** validation set (Recall@1). FusionDiff demonstrates superior capability in following fine-grained attribute editing instructions (e.g., specific logos, patterns, and cuts) compared to CIReVL (Karthik et al., 2024) and LinCIR+CIG (Wang et al., 2025).

O Failure Mode Analysis

We conduct a systematic failure analysis on the CIRR validation set. A failure is defined as a query for which the target image is not retrieved within the top-ranked candidates under the evaluated metric. We manually categorize these cases according to the dominant type of compositional modification.

Table 14 shows that failure rates vary across modification types, ranging from 3.7% to 8.4%. Viewpoint, framing, and focus changes exhibit the highest failure rate (8.4%), followed by abstract or stylistic modifications (8.3%) and counting or quantity changes (8.2%). Fine-grained color and

Table 13: Accuracy–latency trade-off under different numbers of denoising steps. FIQ denotes FashionIQ average R@50. Performance improves rapidly as N increases and saturates around $N = 14$, while latency grows approximately linearly with the number of steps.

Steps N	CIRR R@1	CIRR R@10	FIQ R@50	Latency (s/query)
4	34.12	74.33	57.68	∼0.10
6	36.87	77.41	61.91	∼0.13
8	38.54	79.62	64.22	∼0.18
10	39.31	80.45	66.04	∼0.22
14	**40.00**	**81.28**	**67.92**	∼0.32
18	40.08	81.31	67.20	∼0.44
22	40.05	81.26	67.39	∼0.57

Table 14: Failure mode analysis on the CIRR validation set. Each query is assigned to one dominant modification category according to its edit instruction. The failure rate is computed as the number of failed queries in a category divided by the total number of validation queries in that category.

Modification Category	Total	Failed	Failure Rate	Example Failed Instruction
Object/entity substitution	1260	59	4.7%	"Change the penguin to a dog"
Complex spatial relations and object placement	606	30	5.0%	"Put a mountain on the left and blue sky in the background."
Counting / quantity changes	378	31	8.2%	"Reduce the number of dogs to one, add a human"
Attribute negation / object removal	231	9	3.9%	"remove the snow in the picture"
Pose / action / state changes	984	36	3.7%	"make them sit on the green grass"
Viewpoint / framing / focus	310	26	8.4%	"more focused on its head"
Fine-grained color / appearance attributes	364	19	5.2%	"Paint the body of the animal red."
Abstract / stylistic modifications	48	4	8.3%	"Blur the green grass behind the brown dog and person."
Overall	4181	214	5.1%	–

appearance changes (5.2%) and complex spatial relations (5.0%) are close to the overall failure rate of 5.1%, whereas object/entity substitution (4.7%), attribute negation or object removal (3.9%), and pose, action, or state changes (3.7%) exhibit lower failure rates. Overall, the results indicate that FusionDiff effectively handles a broad range of concrete semantic and attribute-level modifications, while leaving room for improvement in geometric reframing, numerical reasoning, and abstract transformations.

P Broader Impacts

Our proposed framework, **FusionDiff**, advances the state-of-the-art in Zero-Shot Composed Image Retrieval (ZS-CIR). The ability to accurately retrieve images based on fine-grained multimodal queries carries several broader societal implications, encompassing both positive applications and potential risks.

Potential Positive Impacts. FusionDiff significantly enhances user-controllable visual search, which has direct positive implications for e-commerce (e.g., fashion recommendation), interior design, and digital content creation. By allowing users to naturally express visual modifications through language, it democratizes access to complex search mechanisms without requiring specialized technical skills. Furthermore, our *Control-Adapter* demonstrates strong data efficiency, achieving strong performance using only 200K synthetic triplets. This reduces the reliance on large-scale, labor-intensive human annotations, lowering the barrier to entry for resource-constrained researchers and organizations to develop specialized retrieval systems.

Potential Negative Impacts and Ethical Considerations. Despite these benefits, our framework relies on foundational pre-trained models—specifically, Multimodal Large Language Models (MLLMs)

and Diffusion Transformers (DiTs) trained on massive, internet-scraped datasets (e.g., CC12M, SA-1B).

- *Inherited Biases:* These pre-training corpora inevitably contain societal, gender, and cultural biases. Because FusionDiff deeply fuses vision and language into a joint space, there is a risk that the model might propagate or even amplify these biases during retrieval. For instance, in open-domain or fashion queries, the model might associate certain clothing styles or attributes with specific demographics, leading to skewed or stereotypical retrieval results.
- *Environmental Footprint:* While our Stage-2 fine-tuning is computationally lightweight (taking less than an hour on 8 A100 GPUs), the Stage-1 pre-training of the Conditional Embedding Diffusion Model and the reliance on heavy MLLM backbones still carry a notable carbon footprint.
- *Potential for Misuse:* Advanced compositional retrieval technologies could hypothetically be repurposed for unauthorized surveillance or non-consensual tracking if applied to sensitive biometric or facial datasets (e.g., "find this person but wearing a hat").

Mitigation Strategies. To mitigate these risks, we strictly evaluate our models on standard, public benchmarks that do not involve sensitive personal data. We encourage future deployments of ZS-CIR systems to incorporate fairness-aware evaluation protocols to detect and debias multimodal embeddings before commercial application. Furthermore, the embedding-level diffusion proposed in our work bypasses the need for high-resolution pixel-space generation, inherently limiting the capability to generate deepfakes or visually harmful content.

Q Statistical Significance Analysis

To assess whether the performance improvements of our proposed framework, **FusionDiff**, are statistically significant rather than attributable to variation across evaluation queries, we conducted statistical significance tests on the validation splits of the corresponding benchmarks.

Specifically, we performed paired-sample t-tests with a significance level of $\alpha = 0.05$, comparing the per-query retrieval scores of FusionDiff against the strongest generative-feedback baselines, such as LinCIR+CIG and LDRE+IP-CIR. We used Recall@10 for CIRR and FashionIQ and per-query AP@50 for CIRCO, whose average corresponds to the reported mAP@50 metric.

The resulting p-values were below 0.05 across all validation splits, with $p < 0.01$ for highly compositional settings. These statistically significant results provide further evidence that FusionDiff delivers robust and consistent improvements over the compared generative-feedback baselines.