DetailAnywhere: Fashion Detail Generation via Cross-Modal Feature Alignment Distillation

Zijun Li[1,2*] **Yimin Zhou**[1*] **Jia Sun**[1†] **Honglie Wang**[1]
Pengcheng Wei[1] **Junlong Wu**[1] **Yongrui Heng**[1] **Jiyuan Wang**[1]
Huan Ouyang[1] **Boheng Zhang**[1] **Huaiqing Wang**[1] **Dewen Fan**[1]
Qianqian Gan[1] **Fan Yang**[1] **Tingting Gao**[1]
[1]Kuaishou Technology [2]AIM for Health Lab, Monash University



Figure 1: **Fashion Detail Generation.** Given a reference garment image with a bounding box indicating the target region, our model generates a high-fidelity, identity-consistent close-up of the indicated area.

Abstract

Diffusion-based generative AI has achieved remarkable success in e-commerce applications such as virtual try-on, poster generation, and product background synthesis. However, when making online purchasing decisions for apparel, consumers also desire the freedom to examine specific detail regions of interest, such as collars, cuffs, and fabric textures, yet existing methods have not explicitly studied this setting. We therefore formalize a new, non-template task: Fashion Detail Generation with focus conditioning, and release **FDBench**, the first benchmark comprising 40K+ human-verified reference-detail pairs across 41 different categories. This task poses a unique semantic gap challenge: the model must bridge the correspondence between a focus marker on a product reference image and a photorealistic close-up view of the indicated region, while faithfully preserving the garment's identity, without any precise prompt. To bridge this gap, we propose Cross-modal Feature Alignment Distillation (**CFAD**), which leverages a fine-tuned DINOv3 teacher to align both branches of a Multimodal Diffusion Transformer in a shared semantic space via dual-branch distillation. To further improve consistency between generated details and reference images, we introduce a consistency reward model that jointly scores image pairs along three quality axes and optimizes generation via reinforcement learning. Experiments show that **DetailAnywhere** achieves strong performance on FDBench and is preferred to the evaluated open-source baselines in human evaluation.

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Diffusion models [1–3] have become the dominant approach for high-fidelity image generation, and their impact on e-commerce visual content creation has been significant. Virtual try-on systems [4–6] allow consumers to preview garment-body combinations without physical fitting rooms. Automated poster generators [7, 8] produce marketing materials with accurate text rendering and harmonious layouts. Multimodal fashion editors [9, 10] support text- and sketch-guided garment manipulation. These tools have dramatically lowered production costs while raising visual fidelity, and many have been deployed on major e-commerce platforms.

Despite this rapid progress, a critical gap remains: existing methods are essentially non-interactive from the consumer's perspective, delivering pre-designed templates or globally manipulated images with no mechanism for users to freely explore the specific regions they find most informative. In practice, purchasing decisions for apparel frequently hinge on fine-grained garment details [11] that neither overall product images nor text descriptions can adequately convey [12, 13]. Consumers may wish to closely examine the stitching on a sweater cuff, the hardware finish of a zipper pull, or the weave pattern of a fabric—all of which directly shape perceived quality and drive purchasing confidence. Such detail display images have therefore become an essential component of product demonstration. However, acquiring them requires costly per-garment macro photography with specialized setups, imposing a heavy burden on small merchants and preventing scalable application.

Given the maturity of diffusion-based generation and the unified visual comprehension capabilities of recent multimodal models [14, 15], a natural question arises: **can we generate high-quality detail views end-to-end from a single reference image?** We formalize this as the Fashion Detail Generation task, and introduce **FDBench**, which contains 40k paired reference and detail images. Since providing a fine-grained text description that specifies both the product category and the target part (e.g., "the seam-stitching details at the connection area between the hood brim of the anorak and the upper rear back panel") is impractical for consumers, and the region of interest may lie at any location that lacks a standard part name, we define the focus marker as a bounding box that roughly delineates the area of interest. As illustrated in Fig. 1, the expected output is a local detail image of the indicated region that faithfully preserves identity consistency with the reference while presenting the garment's fine-grained details in a clear and visually appealing manner. Unlike reference-based super-resolution, this task permits changes in viewpoint and framing to synthesize a separately composed close-up.

However, generating detail views from a bounding-box-annotated reference presents two core challenges that distinguish it from standard image editing: **(1) Cross-branch semantic misalignment.** State-of-the-art editing models adopt the MMDiT architecture [3, 14, 15], which fuses a text branch and an image branch through joint attention. In Qwen-Image-Edit [14], the reference image is encoded into the text branch by a vision-language encoder (Qwen2.5-VL), yet this branch alone cannot interpret the spatial semantics of a bounding-box marker, failing to convey what the box indicates to the image branch and leading to incoherent or off-target generation. **(2) Detail consistency degradation.** Because our goal is to produce a close-up view of a specific local region, faithfully preserving the garment's identity (material, color, structural details) is essential. However, existing image editing models [10, 16] tend to "hallucinate" fine-grained details that do not exist in the reference, especially when the target region receives insufficient attention during generation.

Recent studies have shown that achieving cross-modal semantic alignment directly within large-scale generative models is notoriously difficult [17–19]. However, REPA [20] reveals that generative diffusion models implicitly learn internal representations aligned with those of small external visual encoders such as DINOv2 [21, 22]. This motivates a key insight: rather than forcing cross-modal alignment inside the large MMDiT, can we first complete the alignment in a small, trainable teacher model and then distill the feature distribution patterns back into the MMDiT? Based on this idea, we propose **CFAD**, a dual-branch cross-modal feature alignment distillation framework, shown in Fig. 2. We first train a view-bridging teacher model that maps both bounding-box-annotated references and detail close-ups into a shared semantic space, then distill this teacher's knowledge into both branches of the MMDiT [23]: the text branch receives semantic guidance from the teacher's encoding of the reference image, while the image branch receives structural guidance from the teacher's encoding of the ground-truth detail. To tackle consistency degradation, we further train a consistency reward model and apply reinforcement learning via negative-aware finetuning [24] to optimize generation along three complementary axes: aesthetic plausibility, identity consistency, and target fidelity.

In summary, our contributions are as follows:

1. We formalize the fashion detail generation task and construct FDBench, the first benchmark for this problem, comprising $\sim$40K manually verified reference-detail pairs for training, and 800 pairs for evaluation. Experiments on FDBench demonstrate strong performance across complementary automatic metrics and human preference evaluation.
2. We propose CFAD (Cross-modal Feature Alignment Distillation), which offloads the difficult cross-modal alignment within a large-scale generative model to a small, easily-trained discriminative teacher, then distills the aligned representations back into both branches of the MMDiT, bypassing direct branch-to-branch optimization.
3. We introduce a consistency reward model that scores reference-detail pairs along three task-specific quality dimensions and fine-tunes the generator via reinforcement learning, further improving identity consistency.

2 Related Work

2.1 Image Generation and Editing

Diffusion models [1, 2] have become the dominant paradigm for image synthesis. Early advances established key guidance mechanisms: classifier guidance [25] and classifier-free guidance (CFG) [26], which significantly improved controllability and sample fidelity. In parallel, CLIP [27] enabled language-aligned visual supervision and became a core signal for text-guided generation and editing. Latent diffusion models (LDMs) [2] then improved efficiency by operating in a compressed latent space, while transformer-based variants, including DiT [28] and MMDiT [3], further scaled quality by jointly processing text and image tokens through shared attention. Flow matching [29] provides an efficient training framework for these architectures. Subsequent works have further expanded diffusion-based modeling to a broad range of applications and explorations [30–36].

For image editing, InstructPix2Pix [10] enables natural-language-guided editing, while ControlNet [37] and IP-Adapter [16] introduce spatial and image-prompt conditioning. SDEdit [38] supports stochastic editing through forward-backward diffusion. Diffusion-based models and representations have also been increasingly applied to other domains such as 3D vision [39–43]. REPA [20] shows that diffusion transformers implicitly build discriminative representations similar to those of pretrained encoders like DINOv2 [21, 22], and that explicitly aligning the two accelerates training and improves quality. CFAD builds on this alignment principle, but uses a task-adapted teacher trained on reference and detail views and distills its features into both the image and conditioning branches. DDAE [44] and DIFT [45] further demonstrate that diffusion models can act as unified self-supervised learners whose internal features encode fine-grained spatial correspondence. Knowledge distillation [23] and cross-modal distillation [46] provide the broader foundation for such representation transfer. However, all the above methods target global edits or localized modifications within the same viewpoint. Our task requires generating a novel close-up view of a specified local region, which demands cross-view semantic understanding that existing editing pipelines do not address.

2.2 E-commerce Content Generation

AI-assisted content creation for e-commerce has been an active area of research. Existing studies can be grouped into several representative directions: virtual try-on [4, 47–50], fashion image editing with specific controls [9, 10, 16], product-background and scene synthesis for merchandising, and automatic poster generation for marketing content [7, 8].

Despite this breadth, most pipelines operate at the global image level and do not model fine-grained local detail generation under user-specified spatial focus. Our work addresses this gap by targeting identity-consistent close-up generation from a single reference image and a bounding box.

2.3 Reinforcement Learning for Image Editing

Reinforcement learning (RL) has been applied to align diffusion models with human preferences. DDPO [51] formulates denoising as a multi-step RL problem; DPOK [52] adds KL regulariza-

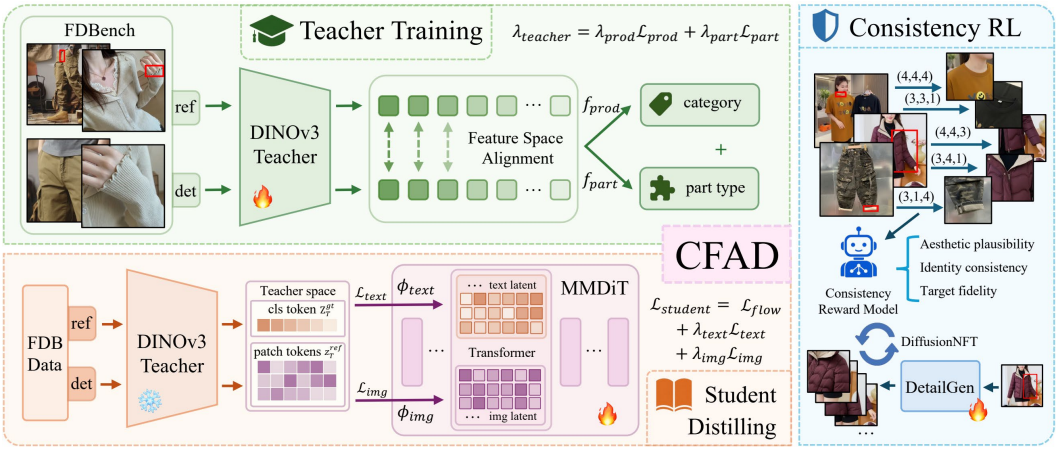

Figure 2: **Overview of DetailAnywhere.** Our framework consists of two stages: (1) CFAD trains a view-bridging teacher to align reference and detail representations, then distills this alignment into both branches of the generator; (2) A consistency reward model fine-tunes the generator via Negative-aware FineTuning to further improve identity preservation and detail fidelity.

tion for stability. GRPO [53] offers group-relative policy optimization without a critic network, while DiffusionNFT [24] introduces a forward-process RL formulation with continuous timestep support. To support RL training, several reward models have been developed: ImageReward [54] and HPSv2 [55] provide general-purpose preference scores for text-to-image generation, while EditReward [56] and EditScore [57] focus specifically on the quality of instruction-guided image edits. In the e-commerce domain, business-oriented signals such as CTR prediction [58, 59] and conversion-rate feedback have also been explored, reflecting a trend toward task-specific, domain-grounded rewards. RL-based or reward-driven optimization has also been adopted for multimodal visual reasoning, video reward modeling, and self-evolving MLLMs [60–64].

However, none of these reward models captures the fine-grained consistency requirements of fashion detail generation. We therefore design a task-specific consistency reward model along three complementary dimensions and use it to guide generator fine-tuning via RL.

3 Method

3.1 Task Formulation

We define the fashion detail generation task as follows: given a reference garment image $\mathbf{x}_{\text{ref}} \in \mathbb{R}^{H\times W\times 3}$ depicting a full-body or half-body view of a clothing item, together with a bounding box b rendered directly on the image to indicate the specified target region, the goal is to generate a detail image $\mathbf{x}_{\text{det}} \in \mathbb{R}^{H'\times W'\times 3}$ that presents a high-fidelity close-up of that region.

Formally, we seek to learn a conditional generator G_θ such that:

$$\mathbf{x}_{\text{det}} = G_\theta(\mathbf{x}_{\text{ref}}^{\text{box}}), \quad \mathbf{x}_{\text{ref}}^{\text{box}} = \text{render}(\mathbf{x}_{\text{ref}}, b) \tag{1}$$

Here, $\text{render}(\mathbf{x}_{\text{ref}}, b)$ deterministically overlays a red rectangle at b on the reference, and its line width and opacity are specified in Appendix A.1. The generated detail must satisfy two first-class objectives: **(i)** High fidelity: photorealistic and visually pleasing rendering of the local region, free of artifacts; **(ii)** High consistency: faithful preservation of the garment's material, texture, color, and structural identity relative to the source.

3.2 CFAD: Cross-Modal Feature Alignment Distillation

The text-branch bottleneck and prompt enhancement. Leading image generation and editing models such as SD3 [3], FLUX.1 [65], and Qwen-Image-Edit [14] all adopt the MMDiT architecture, which fuses a text branch and an image branch through joint attention in each transformer block. When applied to our task, these models struggle to associate the bounding-box marker on

the reference with the target local region, producing incoherent or off-target detail images [17, 66]. The fundamental difficulty is that the text branch encodes a holistic understanding of the reference image but lacks region-level semantic grounding. It cannot convey which specific region the box indicates to the image branch. We observe empirically that employing a large vision-language model to localize and describe the boxed region, then injecting this description into the editing prompt (denoted "PE" in Table 1), substantially improves baseline performance, confirming that enriching the text branch with region-level semantics is a key factor in bridging the cross-branch gap. However, textual descriptions inherently cannot fully convey fine-grained structural cues and texture characteristics of the target region. Moreover, this approach introduces an expensive external VLM at inference time.

Key insight: feature-level alignment via a discriminative teacher. Recent work reveals that generative diffusion transformers already possess rich semantic representations internally. REPA [20] and follow-ups [67, 68] show that intermediate features of diffusion models closely resemble those of pretrained discriminative encoders such as DINOv2 [21, 22], and that explicit alignment accelerates training. Complementarily, DDAE [44] and DIFT [45] demonstrate that diffusion models can serve as unified representation learners whose internal activations encode fine-grained spatial correspondence. Together, these findings imply that the MMDiT's text and image branches already possess the *latent capacity* for rich semantic understanding. What they lack is not representational power, but an explicit mechanism to align their internal features into a shared space where joint attention can effectively bridge the cross-view correspondence.

This motivates a simple yet effective idea: rather than injecting region understanding through PE, we can implicitly align the text and image branches in a shared feature space, maximally leveraging the powerful representational capacity that the DiT already possesses. Specifically, we train a small discriminative teacher that accepts both the bounding-box-annotated reference and the corresponding detail close-up, and maps them into a shared semantic space via classification training. Following and extending the REPA paradigm [20], we then distill these aligned representations into both branches of the MMDiT simultaneously, which we call Cross-Modal Feature Alignment Distillation (CFAD). Because the two distillation targets reside in the same teacher space by construction, the text and image branches are implicitly guided toward mutual alignment through the teacher as an intermediary, entirely bypassing both the language bottleneck of PE and the intractable direct branch-to-branch optimization within the large generative model.

View-bridging teacher learning. We realize CFAD through two stages. In the first stage, we train a lightweight discriminative teacher that unifies the representation space of both views, providing the aligned targets for subsequent distillation.

Specifically, we fine-tune a DINOv3 [69] ViT-H/16+ backbone f_T to accept two types of input from each training pair: (i) the boxed reference $\mathbf{x}_{\text{ref}}^{\text{box}}$, and (ii) the corresponding detail close-up $\mathbf{x}_{\text{det}}$. Both views are passed through the same backbone f_T and trained with a shared classification objective over product categories and part types:

$$\mathcal{L}_{\text{teacher}} = \omega_{\text{prod}}\mathcal{L}_{\text{CE}}(f_{\text{prod}}(f_T(\mathbf{x})), y_{\text{prod}}) + \omega_{\text{part}}\mathcal{L}_{\text{CE}}(f_{\text{part}}(f_T(\mathbf{x})), y_{\text{part}}), \quad \mathbf{x} \in \{\mathbf{x}_{\text{ref}}^{\text{box}}, \mathbf{x}_{\text{det}}\}, \quad (2)$$

where f_{prod} and f_{part} are simple MLPs classification heads built on the [CLS] token, and $(y_{\text{prod}}, y_{\text{part}})$ is the shared label for each pair. y_{prod} and y_{part} denote the product-type and part-type labels, respectively. The teacher-loss weights ω_{prod} and ω_{part} are distinct from the generator distillation weights λ_{img} and λ_{text}; their values are specified in Section 4.2. The crucial design is shared label assignment, assigning the same product$\times$part label to both reference and detail views. This shared supervision encourages semantic compatibility between the two views within the discriminative teacher. After supervised fine-tuning on 80K samples (the total number of reference and detail images), the teacher achieves 84% joint accuracy, indicating that the teacher learns product and part semantics across the two input views.

Dual-branch distillation into the generator. In the second stage, we distill the aligned teacher representations into both branches of the generator at an intermediate layer l^*. We extract two types of targets from the trained teacher backbone: a semantic [CLS] token $\mathbf{z}_T \in \mathbb{R}^{d_t}$, and patch tokens $\mathbf{Z}_T \in \mathbb{R}^{N_p \times d_t}$ capturing fine-grained spatial structure, where d_t is the output hidden dimension of the teacher. Let $\mathbf{h}_{\text{text}}^{(l)} \in \mathbb{R}^{L \times D}$ and $\mathbf{h}_{\text{img}}^{(l)} \in \mathbb{R}^{S \times D}$ denote the text-branch and image-branch hidden states at block l, respectively.

For the image branch, we provide structural guidance by encoding the ground-truth detail $\mathbf{x}_{\text{det}}$ through the teacher to obtain patch tokens $\mathbf{Z}_T^{\text{gt}} \in \mathbb{R}^{S \times d_t}$. A lightweight projector $\phi_{\text{img}} : \mathbb{R}^D \to \mathbb{R}^{d_t}$ maps the image-branch hidden states to the teacher space, and the alignment loss maximizes patch-wise cosine similarity following REPA [20]:

$$\mathcal{L}_{\text{img}} = -\frac{1}{S} \sum_{i=1}^{S} \frac{\phi_{\text{img}}(\mathbf{H}_{\text{img},i}^{(l^*)}) \cdot \mathbf{Z}_{T,i}^{\text{gt}}}{\|\phi_{\text{img}}(\mathbf{H}_{\text{img},i}^{(l^*)})\| \cdot \|\mathbf{Z}_{T,i}^{\text{gt}}\|}. \tag{3}$$

where S is the number of image patch tokens and $\mathbf{H}_{\text{img},i}^{(l^*)}$ denotes the hidden state of the i-th patch at layer l^*. For the text branch, we provide semantic guidance by aligning it to the teacher's encoding of the boxed reference $\mathbf{z}_T^{\text{ref}}$. We mean-pool the text-branch hidden states $\mathbf{H}_{\text{text}}^{(l^*)}$ over non-padding tokens to obtain $\bar{\mathbf{h}}_{\text{text}}$, and minimize the cosine distance through projector $\phi_{\text{text}} : \mathbb{R}^D \to \mathbb{R}^{d_t}$:

$$\mathcal{L}_{\text{text}} = 1 - \frac{\phi_{\text{text}}(\bar{\mathbf{h}}_{\text{text}}) \cdot \mathbf{z}_T^{\text{ref}}}{\|\phi_{\text{text}}(\bar{\mathbf{h}}_{\text{text}})\| \cdot \|\mathbf{z}_T^{\text{ref}}\|}. \tag{4}$$

Let $\mathbf{z} = E(\mathbf{x}_{\text{det}})$ be the target-detail VAE latent, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathbf{z}_t = t\mathbf{z} + (1-t)\boldsymbol{\epsilon}$. With t sampled according to the base model's logit-normal training distribution, the flow-matching objective is

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{\mathbf{z},\boldsymbol{\epsilon},t}\left[\left\| v_\theta(\mathbf{z}_t, t \mid \mathbf{x}_{\text{ref}}^{\text{box}}) - (\mathbf{z} - \boldsymbol{\epsilon}) \right\|_2^2\right]. \tag{5}$$

The full training loss combines the flow-matching objective with both alignment losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{flow}} + \lambda_{\text{img}} \cdot \mathcal{L}_{\text{img}} + \lambda_{\text{text}} \cdot \mathcal{L}_{\text{text}}, \tag{6}$$

where λ_{img} and λ_{text} are scalar coefficients that balance the distillation strength relative to the primary flow-matching objective. Because $\mathbf{z}_T^{\text{ref}}$ and $\mathbf{Z}_T^{\text{gt}}$ reside in the same teacher feature space by construction, simultaneously distilling them into the two generative branches causes the branches to converge toward compatible representations, ultimately producing detail images with stronger identity consistency and more precise structural fidelity.

3.3 Consistency Reward Model and RL Fine-Tuning

While CFAD substantially improves cross-branch alignment during supervised training, the generated details may still exhibit subtle deficiencies that the supervised training objective does not explicitly penalize, such as color drift, implausible fabric textures, or identity deviation from the source garment. General-purpose preference models [54, 55, 70] optimize global aesthetics, and editing-specific rewards [56, 57, 71] target instruction-following on natural images; neither directly addresses the consistency requirements unique to fashion detail generation. We therefore introduce a task-specific *consistency reward model* R_ψ and use it to fine-tune the generator via reinforcement learning.

Multi-dimensional reward formulation. Given a boxed reference image and a candidate detail image $(\mathbf{x}_{\text{ref}}^{\text{box}}, \mathbf{x}_{\text{det}})$, the reward model produces three complementary scores: The three dimensions are: **aesthetic plausibility** (r_{as}), which measures visual naturalness and artifact-free quality; **identity consistency** (r_{id}), which measures whether material, color, pattern, and structure are preserved from the source garment; and **target fidelity** (r_{tf}), which measures whether the generated close-up correctly corresponds to the indicated local region. The composite reward is a weighted sum $r = \alpha_{\text{as}}\, r_{\text{as}} + \alpha_{\text{id}}\, r_{\text{id}} + \alpha_{\text{tf}}\, r_{\text{tf}}$. The three dimensions are predicted jointly for each image pair and aggregated into one scalar reward. We fine-tune a vision-language model on a curated dataset of about 60K real reference–detail image pairs with varying quality levels, each annotated with integer scores along the three dimensions. The details of data collection and annotation are provided in Appendix A.6.

RL fine-tuning. We adopt Negative-aware FineTuning [24], a forward-process RL framework that optimizes the generator using reward-weighted positive and negative reconstruction objectives. Unlike multi-step denoising RL (e.g., DDPO [51]), NFT operates on a small number of timesteps sampled from the forward process, enabling efficient online optimization. For each training sample, we generate a group of K candidates, score each with R_ψ, and apply group-relative advantage normalization [53] to obtain per-candidate advantages. These advantages are clipped and mapped to weights for the positive and negative NFT objectives.

4 Experiments

4.1 FDBench: Task and Data

Dataset Construction. We construct FDBench from a large-scale fashion e-commerce catalog, where each sample is a pair $(\mathbf{x}_{\text{ref}}^{\text{box}}, \mathbf{x}_{\text{det}})$ consisting of a boxed reference garment image and a corresponding detail close-up. The benchmark spans menswear, womenswear, and childrenswear with a total of 41 product categories. The final benchmark contains $\sim$40K training samples and a held-out 800-sample test set. To keep the main paper concise, we defer full data curation details, including annotation, collection channels, and filtering, to Appendix A.4.

4.2 Experimental Setup

Evaluation Metrics. We evaluate generated detail images along three complementary axes. SSIM and LPIPS [72] measure pixel-level and perceptual similarity between the generated detail and the ground-truth detail when paired ground truth exists. FID (Fréchet Inception Distance) measures distributional similarity between generated and real detail images over the test set. We further report CLIP Image Similarity between the generated detail and the ground-truth detail [27] as CIS, and CLIP Image Similarity between the generated detail and the cropped reference region as CCIS to quantify identity preservation. We additionally report **EditScore** [57] with sub-dimensions PF (Prompt Following), CS (Consistency), PQ (Perceptual Quality), and O (Overall, $\sqrt{\min(\text{PF}, \text{CS}) \times \text{PQ}}$); and **ImageEditJudge** [71] with sub-dimensions PD (Physical & Detail Fidelity), PC (Prompt Compliance), VS (Visual Seamlessness), and the average. For the prompt-related dimensions, all methods are evaluated against the same item-specific description produced by prompt enhancement (PE). This description summarizes the boxed region and serves as an evaluation reference, DetailAnywhere receives the boxed image and a fixed generic instruction, without this PE description or an auxiliary VLM at inference.

Implementation Details. We use Qwen-Image-Edit-2511 [14] as our base generator, a 20B-parameter MMDiT with hidden dimension $D = 3072$. The DINOv3 ViT-H/16+ view-bridging teacher ($d_t = 1280$) is fine-tuned on the 80K FDBench training images with product and part classification heads, achieving 87.2% part accuracy and 90.6% product accuracy (84.0% joint accuracy). The classification heads predict 41 product types and 20 part types, with teacher-loss weights $\omega_{\text{prod}} = 0.5$ and $\omega_{\text{part}} = 1.0$. For CFAD training, we set the alignment layer $l^* = 16$ following REPA [20], with loss weights $\lambda_{\text{img}} = 0.02$ and $\lambda_{\text{text}} = 0.05$. The consistency reward model R_ψ is built on Qwen3-VL-8B [73], fine-tuned for 2 epochs on $\sim$85K annotated real reference–detail image pairs with varying quality levels. RL fine-tuning applies DiffusionNFT [24] with LoRA (rank$= 64$, $\alpha = 128$) and composite reward weights $\alpha_{\text{as}} = 0.2$, $\alpha_{\text{id}} = 0.4$, $\alpha_{\text{tf}} = 0.4$. At inference time, all models generate at 1024×1024 resolution with 40 denoising steps; we use classifier-free guidance scale 4.0 for our model and select the recommended guidance scale for each baseline. Full training hyperparameters are provided in Appendix A.1.

4.3 Main Results

We compare **DetailAnywhere** (Ours) against a comprehensive set of state-of-the-art image editing models, divided into open-source and proprietary categories. Open-source baselines include Qwen-Image-Edit-2511 [14], FLUX.2 Dev [15], Step1X-Edit [74], ICEdit [75], FireRed-Image-Edit [76], and JoyAI-Image-Edit [77]. Proprietary baselines include GPT-Image-1 [78], Nano Banana 2 (Gemini 3.1 Flash Image Preview) [79], and Seedream 5.0 Lite [80]. For baselines without PE, the bounding box is rendered on the reference image and the instruction "Generate a detail view of the boxed region" is provided. PE baselines instead receive the unboxed reference and the generated region description, as detailed in Appendix A.2.

Table 1 reports results on the FDBench test set. The overall improvement combines adaptation to FDBench and the proposed training components. To separate these effects, Table 2 compares plain in-domain SFT with CFAD under the same base model and training schedule, without RL. CFAD improves SSIM from 0.347 to 0.366, FID from 74.1 to 70.9, and EditScore Overall from 8.48 to 8.79. DetailAnywhere (CFAD+NFT) achieves the best performance across all five image-quality metrics among open-source methods, substantially outperforming the strongest prompt-enhanced baseline (Qwen-Image-Edit PE) on both pixel-level fidelity (SSIM, FID) and identity preservation (CIS,



Figure 3: **Qualitative comparison.** Side-by-side visualization of detail generation results from JoyAI-Image-Edit [77], Qwen-Image-Edit (PE) [14], FireRed-Image-Edit (PE) [76], Seedream 5.0 Lite [80], and DetailAnywhere (Ours). Our method produces identity-consistent, photorealistic close-ups with faithful texture and structural preservation.

CCIS). Notably, baselines without prompt enhancement (Step1X-Edit, ICEdit, FLUX.2 Dev) exhibit extremely low Prompt Following and ImageEditJudge scores despite high Consistency, revealing that they fail to understand the task, and largely copy or minimally modify the reference rather than generating meaningful detail views. Prompt-enhanced baselines (FireRed PE, Qwen-Image-Edit PE) demonstrate significantly better task comprehension, yet still lag behind DetailAnywhere on identity preservation metrics, confirming that language-level region descriptions alone are insufficient to close the cross-view semantic gap. Proprietary models benefit from significantly larger model capacity and extensive training corpora, enabling strong task comprehension and high editing-quality scores across the board. Nevertheless, DetailAnywhere surpasses all three proprietary systems on pixel-level fidelity and identity preservation, and remains competitive on EditScore and ImageEditJudge, validating that our targeted cross-modal alignment strategy can effectively close the gap with large-scale closed-source models. Comparing the two DetailAnywhere variants, CFAD alone already leads all open-source baselines on ImageEditJudge, and RL fine-tuning with the consistency reward further improves fidelity metrics while maintaining strong editing scores, demonstrating that the dual-branch distillation provides a solid foundation that RL can effectively build upon.

We also compare against three reference-based super-resolution methods in Appendix A.9. Their strong crop consistency but lower target-detail similarity and editing scores distinguish crop restoration from fashion detail generation.

Qualitative analysis. Fig. 3 presents visual comparisons on representative FDBench examples. Our model consistently follows the visual marker and generates the correct garment region, whereas other methods occasionally misidentify the target area or are misled by PE-derived descriptions, as illustrated in the first row. Beyond localization accuracy, DetailAnywhere also excels in consistency and fidelity: in the second example, only our method faithfully reproduces the smooth button details. In the third example, our model correctly preserves the position of the brand logo, whereas Seedream 5.0 Lite [80] erroneously relocates it to an incorrect position, demonstrating the effectiveness of our consistency reward model in preserving fine-grained spatial details. More visualizations are provided in Appendix A.10. We further conduct a side-by-side human preference study as a practical indicator for e-commerce deployment; details are provided in Appendix A.3.

4.4 Ablation Studies

We conduct two sets of ablations to validate the two key components of DetailAnywhere: the CFAD distillation framework and the consistency reward model.

Table 1: **Main results on FDBench test set.** Comparison of **DetailAnywhere** against open-source and proprietary baselines. Best in **bold**, second-best underlined (marked within each group). Details of prompt enhancement (PE) are provided in Appendix A.2. SFT uses the same base model, FDBench data, and training schedule as CFAD, without RL.

	Image Quality					EditScore ↑				ImageEditJudge ↑			
Method	SSIM ↑	LPIPS ↓	FID ↓	CIS ↑	CCIS ↑	PF	CS	PQ	O	PD	PC	VS	Avg
Proprietary													
GPT-Image-1 [78]	0.274	0.650	82.6	0.797	0.726	9.16	8.82	**9.85**	8.97	4.26	4.28	4.26	4.27
Nano Banana 2 [79]	**0.319**	**0.619**	**68.2**	**0.856**	0.732	**9.78**	**9.23**	9.66	**9.34**	4.35	4.43	4.33	4.37
Seedream 5.0 Lite [80]	0.275	0.638	80.8	0.841	**0.759**	9.61	9.17	9.63	9.14	**4.46**	**4.61**	**4.46**	**4.51**
Open-source													
FLUX.2 Dev [15]	0.339	0.699	118.7	0.726	0.632	0.12	**9.85**	8.37	0.21	1.41	1.41	1.41	1.41
Step1X-Edit [74]	0.345	0.693	120.5	0.697	0.622	1.62	9.55	7.77	2.32	2.97	2.97	2.96	2.97
ICEdit [75]	0.349	0.691	118.4	0.690	0.631	1.25	9.31	7.58	2.18	2.96	2.99	2.95	2.97
FireRed-Image-Edit [76]	0.321	0.701	118.8	0.718	0.646	5.06	7.39	9.11	5.36	3.52	3.68	3.52	3.57
FireRed-Image-Edit (PE)	0.274	0.660	85.0	0.791	0.719	**9.64**	8.39	9.47	8.71	3.88	3.98	3.88	3.91
Qwen-Image-Edit [14]	0.309	0.686	98.6	0.757	0.665	4.37	6.10	9.43	4.41	3.82	3.99	3.80	3.86
Qwen-Image-Edit (PE)	0.254	0.648	82.6	0.804	0.739	9.51	8.77	9.42	8.85	4.09	4.12	4.09	4.10
JoyAI-Image-Edit [77]	0.324	0.680	93.9	0.754	0.706	7.72	8.80	8.62	7.23	3.51	3.65	3.51	3.55
Qwen-Image-Edit SFT	0.347	0.623	74.1	0.844	0.722	9.19	8.36	9.34	8.48	4.09	4.20	4.09	4.13
DetailAnywhere (CFAD)	0.366	0.608	70.9	0.861	0.741	9.41	8.61	9.48	8.79	4.29	4.31	4.29	4.30
DetailAnywhere (CFAD+NFT)	**0.394**	**0.606**	**66.9**	**0.863**	**0.748**	9.57	8.82	**9.59**	**8.95**	**4.31**	**4.34**	**4.30**	**4.32**

Table 2: **Ablation on CFAD.** Effect of removing each distillation branch. All variants share the same training schedule and base model, without RL fine-tuning.

					EditScore ↑			
Variant	SSIM ↑	FID ↓	CIS ↑	CCIS ↑	PF	CS	PQ	O
w/o CFAD	0.347	74.1	0.844	0.722	9.19	8.36	9.34	8.48
Vanilla REPA	0.342	74.5	0.846	0.729	9.11	8.29	9.30	8.44
w/ img-branch	0.358	71.4	0.852	**0.743**	9.21	8.56	9.42	8.56
Full CFAD (Ours)	**0.366**	**70.9**	**0.861**	0.741	**9.41**	**8.61**	**9.48**	**8.79**

Ablation on CFAD. To isolate the contribution of dual-branch alignment, we compare four variants in Table 2: vanilla supervised fine-tuning without any distillation (w/o CFAD), an image-branch-only variant that keeps $\mathcal{L}_{\text{img}}$ but removes text-branch distillation (w/ img-branch), and the full CFAD model with both alignment paths enabled. We additionally include vanilla REPA, using a frozen DINOv3 teacher and image-branch-only alignment. All variants share the same base model and training schedule without RL fine-tuning.

Full CFAD achieves the best results on the majority of metrics. Adding only image-branch distillation improves pixel-level fidelity (SSIM, FID) and Consistency Score, as the student successfully mimics the teacher's visual encoding of the image [20, 67]. However, identity preservation metrics fluctuate slightly, because distilling the image branch alone does not help it coordinate with the conditioning branch to understand what to generate. Once the text-branch alignment is introduced, the two branches of the model converge in a shared semantic space, leading to not only the best image-level quality but also a substantial boost in Prompt Following, which reflects that the generator now more correctly interprets the editing instruction. These results support complementary contributions from image-branch and conditioning-branch distillation in this setting. A feature distribution analysis in Appendix A.7 further corroborates this conclusion: CFAD induces clear part-type-aware clustering in the MMDiT's hidden space, providing direct internal evidence for the effectiveness of dual-branch alignment distillation. Cross-view alignment results are provided in Appendix A.8.

Ablation on Reward Model. Starting from the same CFAD-trained generator, we compare four RL settings in Table 3: no RL update, RL with EditReward [56], RL with HPSv3 [70], and RL with our task-specific consistency reward from Section 3.3. All runs use the same DiffusionNFT optimizer, group size, and number of update steps. EditReward optimizes general editing quality and improves Perceptual Quality, but substantially degrades pixel-level fidelity and identity preservation, indicating that a generic editing reward misaligns with the consistency requirements of our task. HPSv3 focuses on global aesthetics and maintains SSIM, yet worsens distributional quality and fails to improve identity metrics. In contrast, our consistency reward simultaneously improves all fidelity

Table 3: **Ablation on RL reward.** All variants share the same CFAD-trained initialization and the same DiffusionNFT training schedule; only the reward function differs.

					EditScore ↑			
Reward	SSIM ↑	FID ↓	CIS ↑	CCIS ↑	PF	CS	PQ	O
No RL	0.366	70.9	0.861	0.741	9.41	8.61	9.48	8.79
EditReward [56]	0.323	70.8	0.845	0.722	9.46	8.58	**9.71**	8.83
HPSv3 [70]	0.365	75.1	0.851	0.736	9.17	8.65	9.66	8.62
Consistency Reward (Ours)	**0.394**	**66.9**	**0.863**	**0.748**	**9.57**	**8.82**	9.59	**8.95**

and identity metrics while also boosting EditScore Overall, confirming that a task-specific, multi-dimensional reward grounded in fashion detail consistency is critical for effective RL fine-tuning.

5 Conclusion

We formalized fashion detail generation and presented FDBench, the first benchmark comprising 40K+ human-verified reference–detail pairs. To address the cross-branch semantic gap in MMDiT architectures, we proposed CFAD, which offloads cross-modal alignment to a lightweight view-bridging teacher and distills the aligned representations into both generative branches simultaneously. A task-specific consistency reward model further improves identity preservation via RL fine-tuning. Together, these components form **DetailAnywhere**, which significantly outperforms all open-source baselines and matches state-of-the-art proprietary models across automatic metrics and human evaluations.

Limitations and future work. Our evaluation is limited to the fashion domain; generalization to other product categories (e.g., electronics, furniture) remains to be validated. The view-bridging teacher and the reward model both require paired, annotated data that may not be readily available in all domains. When the target region is severely occluded, the model may hallucinate plausible but incorrect details. A bounding box containing multiple candidate parts can also lead to generation of the wrong part. We plan to extend FDBench to additional domains, explore self-supervised alternatives to the view-bridging teacher, and investigate multi-scale alignment strategies that provide layer-adaptive supervision across the MMDiT depth.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[2] 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.

[3] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorber, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

[4] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[5] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[6] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *European Conference on Computer Vision (ECCV)*, 2024.

[7] Yifan Gao, Zihang Lin, Chuanbin Liu, Min Zhou, Tiezheng Ge, Bo Zheng, and Hongtao Xie. Postermaker: Towards high-quality product poster generation with accurate text rendering. *arXiv preprint arXiv:2504.06632*, 2025.

[8] Sixiang Chen, Jianyu Lai, Jialin Gao, Hengyu Shi, Zhongying Liu, Tian Ye, Junfeng Luo, Xiaoming Wei, and Lei Zhu. Posteromni: Generalized artistic poster creation via task distillation and unified reward feedback. *arXiv preprint arXiv:2602.12127*, 2026.

[9] Alberto Baldrati, Davide Morelli, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[10] 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.

[11] Nannan Li, Kevin J Shih, and Bryan A Plummer. Enhancing virtual try-on with synthetic pairs and error-aware noise scheduling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21238–21247, 2025.

[12] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] Meiqi Sun, Mingyu Li, and Junxiong Zhu. E-comiq-zh: A human-aligned dataset and benchmark for fine-grained evaluation of e-commerce posters with chain-of-thought. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[14] Qwen Team. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

[15] Black Forest Labs. FLUX.2: Frontier visual intelligence. `https://bfl.ai/blog/flux-2`, 2025.

[16] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[17] Liang Chen, Shuai Bai, Wenhao Chai, Weichu Xie, Haozhe Zhao, Leon Vinci, Junyang Lin, and Baobao Chang. Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6146–6156, 2025.

[18] Jaa-Yeon Lee, Byunghee Cha, Jeongsol Kim, and Jong Chul Ye. Aligning text to image in diffusion models is easier than you think. *arXiv preprint arXiv:2503.08250*, 2025.

[19] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023.

[20] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations (ICLR)*, 2025.

[21] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

[22] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[23] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning Workshop*, 2015.

[24] Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025.

[25] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[26] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[28] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[29] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *International Conference on Learning Representations (ICLR)*, 2023.

[30] Jiyuan Wang, Chunyu Lin, Lei Sun, Zhi Cao, Yuyang Yin, Lang Nie, Zhenlong Yuan, Xiangxiang Chu, Yunchao Wei, Kang Liao, et al. Geometry-guided reinforcement learning for multi-view consistent 3d scene editing. *arXiv preprint arXiv:2603.03143*, 2026.

[31] Honglie Wang, Yan-Ming Zhang, Wangzi Yao, Fei Yin, and Cheng-Lin Liu. Learning to generate stylized handwritten text via a unified representation of style, content, and noise. In *The Fourteenth International Conference on Learning Representations*, 2026.

[32] Honglie Wang, Minsi Ren, Yan-Ming Zhang, Fei Yin, and Cheng-Lin Liu. Template-guided cascaded diffusion for stylized handwritten chinese text-line generation. In *International Conference on Document Analysis and Recognition*, pages 149–166. Springer, 2025.

[33] Nanxu Gong, Zijun Li, Sixun Dong, Haoyue Bai, Wangyang Ying, Xinyuan Wang, and Yanjie Fu. Sculpting features from noise: Reward-guided hierarchical diffusion for task-optimal feature transformation. *Advances in Neural Information Processing Systems*, 38:23452–23474, 2026.

[34] JiYuan Wang, Chunyu Lin, Lei Sun, Rongying Liu, Lang Nie, Mingxing Li, Kang Liao, Xiangxiang Chu, and Yao Zhao. From editor to dense geometry estimator. *arXiv preprint arXiv:2509.04338*, 2025.

[35] Hongyu Yan, Zijun Li, Kunming Luo, Li Lu, and Ping Tan. Symmcompletion: High-fidelity and high-consistency point cloud completion with symmetry guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9094–9102, 2025.

[36] Yichong Xia, Yimin Zhou, Jinpeng Wang, Baoyi An, Haoqian Wang, Yaowei Wang, and Bin Chen. Diffpc: Diffusion-based high perceptual fidelity image compression with semantic refinement. In *The Thirteenth International Conference on Learning Representations*, 2025.

[37] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[38] 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 *International Conference on Learning Representations (ICLR)*, 2022.

[39] Zijun Li, Hongyu Yan, Shijie Li, Kunming Luo, Li Lu, Xulei Yang, and Weisi Lin. Diffpcn: Latent diffusion model based on multi-view depth images for point cloud completion. *arXiv preprint arXiv:2509.23723*, 2025.

[40] Jiyuan Wang, Chunyu Lin, Cheng Guan, Lang Nie, Jing He, Haodong Li, Kang Liao, and Yao Zhao. Jasmine: Harnessing diffusion prior for self-supervised depth estimation. *arXiv preprint arXiv:2503.15905*, 2025.

[41] Hongyu Yan, Kunming Luo, Weiyu Li, Yixun Liang, Shengming Li, Jingwei Huang, Chunchao Guo, and Ping Tan. Posemaster: Generating 3d characters in arbitrary poses from a single image. *arXiv preprint arXiv:2506.21076*, 2025.

[42] Jiyuan Wang, Chunyu Lin, Lang Nie, Kang Liao, Shuwei Shao, and Yao Zhao. Digging into contrastive learning for robust depth estimation with diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM 24, page 41294137. ACM, October 2024. doi: 10.1145/3664647.3681168. URL `http://dx.doi.org/10.1145/3664647.3681168`.

[43] Jiyuan Wang, Chunyu Lin, Lang Nie, Shujun Huang, Yao Zhao, Xing Pan, and Rui Ai. Weatherdepth: Curriculum contrastive learning for self-supervised depth estimation under adverse weather conditions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, page 49764982. IEEE, May 2024. doi: 10.1109/icra57147.2024.10611100. URL `http://dx.doi.org/10.1109/icra57147.2024.10611100`.

[44] Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[45] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[46] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[47] Jeongho Kim, Gyojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[48] Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, Dongmei Jiang, and Xiaodan Liang. Catvton: Concatenation is all you need for virtual try-on with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

[49] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *ACM Multimedia*, 2023.

[50] Xu Yang, Changxing Ding, Zhibin Hong, Junhao Huang, Jin Tao, and Xiangmin Xu. Texture-preserving diffusion models for high-fidelity virtual try-on. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[51] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024.

[52] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[53] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[54] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[55] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

[56] Keming Wu, Sicong Jiang, Max Ku, Ping Nie, Minghao Liu, and Wenhu Chen. Editreward: A human-aligned reward model for instruction-guided image editing. *arXiv preprint arXiv:2509.26346*, 2025.

[57] Xin Luo, Jiahao Wang, Chenyuan Wu, Shitao Xiao, Xiyan Jiang, Defu Lian, Jiajun Zhang, Dong Liu, and Zheng Liu. Editscore: Unlocking online RL for image editing via high-fidelity reward modeling. *arXiv preprint arXiv:2509.23909*, 2025.

[58] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. Deepfm: A factorization-machine based neural network for ctr prediction. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[59] Jiahao Fan, Yuxin Qin, Wei Feng, Yanyin Chen, Yaoyu Li, Ao Ma, Yixiu Li, Li Zhuang, Haoyi Bian, Zheng Zhang, Jingjing Lv, Junjie Shen, and Ching Law. Autopp: Towards automated product poster generation and optimization, 2025. URL `https://arxiv.org/abs/2512.21921`.

[60] Chaoya Jiang, Yongrui Heng, Wei Ye, Han Yang, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. Vlm-r^3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. *Advances in Neural Information Processing Systems*, 38:63841–63869, 2025.

[61] Junlong Wu, Yi Cheng, Hang Liu, and Houde Liu. Arc: Robots adaptive risk-aware robust control via distributional reinforcement learning. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10656–10663. IEEE, 2025.

[62] Jiyuan Wang, Huan Ouyang, Jiuzhou Lin, Chunyu Lin, Dewen Fan, Boheng Zhang, Haonan Fan, Fei Zuo, Jia Sun, Huaiqing Wang, et al. Cac: Advancing video reward models via hierarchical spatiotemporal concentrating. *arXiv preprint arXiv:2605.11723*, 2026.

[63] Yimin Zhou, Yichong Xia, Sicheng Pan, Bin Chen, Yaowei Li, Jiawei Li, Mingyao Hong, Zhi Wang, and Yaowei Wang. High quality underwater image compression with adaptive color correction. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8587–8591. IEEE, 2026.

[64] Yongrui Heng, Chaoya Jiang, Han Yang, Shikun Zhang, and Wei Ye. Eve: Verifiable self-evolution of mllms via executable visual transformations. *arXiv preprint arXiv:2604.18320*, 2026.

[65] Black Forest Labs. FLUX. `https://github.com/black-forest-labs/flux`, 2024.

[66] Jeeyung Kim, Erfan Esmaeili, and Qiang Qiu. Text embedding is not all you need: Attention control for text-to-image semantic alignment with text self-attention maps. *arXiv preprint arXiv:2411.15236*, 2024.

[67] Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18262–18272, 2025.

[68] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

[69] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[70] Yuhang Ma et al. HPSv3: Towards wide-spectrum human preference score. *arXiv preprint arXiv:2508.03789*, 2025.

[71] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025.

[72] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[73] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[74] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

[75] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. In *NeurIPS*, 2025.

[76] Super Intelligence Team. FireRed-Image-Edit-1.0 technical report. *arXiv preprint arXiv:2602.13344*, 2026.

[77] Lin Song, Wenbo Li, Guoqing Ma, Wei Tang, Bo Wang, Yuan Zhang, Yijun Yang, Yicheng Xiao, Jianhui Liu, et al. Awaking spatial intelligence in unified multimodal understanding and generation. *arXiv preprint arXiv:2605.04128*, 2026.

[78] OpenAI. GPT-Image-1: Image generation in the OpenAI API. `https://openai.com/index/image-generation-api/`, 2025.

[79] Google DeepMind. Nano Banana 2 (Gemini 3.1 Flash Image Preview). `https://ai.google.dev/gemini-api/docs/image-generation`, 2026. Model ID: `gemini-3.1-flash-image-preview`; accessed 2026-05-07.

[80] ByteDance Seed. Seedream 5.0 Lite. `https://seed.bytedance.com/en/seedream5_0_lite`, 2026.

[81] ByteDance Seed. Seed 2.0: A multimodal understanding and agent model series. https://seed.bytedance.com/en/seed2, 2026.

[82] Google DeepMind. Gemini 3.1 Pro. https://deepmind.google/models/gemini/pro/, 2025.

[83] Byeonghun Lee, Hyunmin Cho, Hong Gyu Choi, Soo Min Kang, Iljun Ahn, and Kyong Hwan Jin. Reference-based super-resolution via image-based retrieval-augmented generation diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10764–10774, 2025.

[84] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5791–5800, 2020.

[85] Yuan Wang, Yuhao Wan, Siming Zheng, Bo Li, Qibin Hou, and Peng-Tao Jiang. Trust but verify: Adaptive conditioning for reference-based diffusion super-resolution via implicit reference correlation modeling. In *International Conference on Learning Representations (ICLR)*, 2026.

A Supplementary Materials

A.1 Additional Implementation Details

Teacher training. The DINOv3 teacher uses a ViT-H/16+ backbone with output dimension $d_t = 1280$. We fully fine-tune the backbone with AdamW using a differential learning rate schedule: 2×10^{-5} for the backbone and 1×10^{-4} for the classification heads. Training uses bf16 mixed precision, batch size 32 per GPU on 4 GPUs, weight decay 0.01, warmup ratio 0.05, label smoothing 0.05, and a maximum of 200 epochs with early stopping on joint accuracy. The red bounding box rendered on reference images uses a relative line width of 0.06 and an opacity of 0.5. The two classification losses in Eq. (2) are weighted as $\omega_{\text{prod}} = 0.5$ and $\omega_{\text{part}} = 1.0$, reflecting that part-level discrimination is the harder subtask.

CFAD distillation. For CFAD training of the base generator, we use DeepSpeed ZeRO-3 with bf16 mixed precision, learning rate 1×10^{-5} with linear warmup over 500 steps, and train for 15 epochs on FDBench ($\sim$40K samples) using 2 GPU nodes with batch size 1 per GPU and gradient checkpointing. Each distillation projector (ϕ_{img} and ϕ_{text}) is a two-layer MLP: LayerNorm(D) $\rightarrow$Linear$(D, 2048)$ $\rightarrow$GELU$\rightarrow$Linear$(2048, d_t)$, where D=3072 and d_t=1280.

Reward model. The consistency reward model R_ψ training uses the same Qwen3-VL-8B [73] backbone with learning rate 2×10^{-5} and batch size 64. The annotation dataset comprises $\sim$85K reference-detail pairs with integer scores (1–4) along three dimensions.

RL fine-tuning. RL fine-tuning uses DiffusionNFT with LoRA (rank$= 64$, $\alpha = 128$) on 2 GPU nodes. We set NFT $\beta = 0.1$, advantage clipping range 5.0, learning rate 8×10^{-5} with AdamW (weight decay 10^{-4}), and EMA decay 0.9. Each epoch samples K=16 candidates per case with per-device batch size 1, 48 unique cases, and 2 gradient steps per epoch. We sample 2 training timesteps per iteration with time shift 3.0 and generate with 15 inference steps at guidance scale 4.

A.2 Prompt Enhancement Details

Prompt enhancement (PE) is designed to compensate for the relatively weak region-level understanding of the condition branch in off-the-shelf editing models. For each boxed reference image, we first use Qwen3-VL-30B-A3B as an external image understanding model to localize the marked region and summarize its garment semantics. The VLM is prompted to return a strict JSON object with four fields:

VLM Understanding Prompt

```
The image shows a full-view garment, with a red bounding box marking a
region.

Please carefully inspect the image and answer the following four questions:
1. position: the location of the red box in the image. Choose the most
suitable option from: upper-left, top, upper-right, left, center, right,
lower-left, bottom, lower-right.
2. clothing_type: the type of garment, described concisely in no more than
10 words.
3. detail_part: the garment part corresponding to the red-box region,
described concisely in no more than 20 words.
4. textural_features: distinctive texture, accessory, decoration, or style
information of the garment, described concisely in no more than 20 words.

Return strict JSON only, without any additional text:
{"position": "xxx", "clothing_type": "xxx", "detail_part": "xxx",
"textural_features": "xxx"}
```

We then parse the JSON response and assemble the final editing prompt as follows:

Assembled Editing Prompt

```
Generate a high-fidelity close-up detail image of the target garment region.
The target region is located at {position} in the reference image.  The
garment is {clothing_type}.  The region corresponds to {detail_part}.
Preserve the garment identity, including color, material, structure, and
the following visual characteristics:  {textural_features}.  Produce a detail
view of this region, with clear texture and faithful local structure.
```

For PE baselines, we remove the red bounding box from the input image and provide the unboxed reference image together with the assembled prompt to the corresponding editing model. All results marked with "(PE)" in Table 1 are obtained using this procedure.

A.3 Human Preference Evaluation

To complement automatic metrics with practical quality assessment, we conduct a side-by-side (SBS) human preference study on a 800-sample subset of the FDBench test set. For each test sample, three unaffiliated professional quality-inspection annotators are independently shown the reference image with bounding box together with two generated detail images (Ours vs. a baseline) in randomized order, and each annotator selects one of four options: A is better, B is better, both are equally good, or both are equally bad. We shuffled the left–right order and anonymized method identities so that annotators could not identify our result. The final label for each sample is determined by majority vote: if at least two of the three annotators agree, the majority option is taken; if all three annotators choose different options, the sample is discarded as inconclusive. We compare DetailAnywhere against two representative open-source baselines (Qwen-Image-Edit PE and FireRed PE) and the strongest proprietary baseline (Nano Banana 2). We additionally compare the final CFAD+RL model against its pre-RL CFAD checkpoint to isolate the practical effect of RL. Results are summarized in Table 4.

Table 4: Human preference evaluation (side-by-side). For each pair, annotators label one of four outcomes: Ours wins, Both good, Baseline wins, or Both bad. Percentages are computed over valid samples.

Baseline	Ours Wins	Both Good	Baseline Wins	Both Bad
Qwen-Image-Edit (PE)	**66.1%**	2.4%	14.9%	16.6%
FireRed-Image-Edit (PE)	**74.9%**	2.0%	4.7%	18.3%
Nano Banana 2	40.7%	14.5%	**41.1%**	3.7%
CFAD (no RL)	44.2%	22.0%	26.3%	7.4%

DetailAnywhere is strongly preferred over both open-source PE baselines, winning 66.1% and 74.9% of comparisons against Qwen-Image-Edit PE and FireRed PE, respectively, while the baselines are preferred in only 14.9% and 4.7% of cases. Against Nano Banana 2, the win rates are nearly tied (40.7% vs. 41.1%). Against pre-RL CFAD, CFAD+RL wins 44.2% of valid comparisons and loses 26.3%, a 17.9-percentage-point margin. Together with the controlled reward ablation in Table 3, this supports a perceptible benefit beyond automatic metrics. These results indicate competitive perceived quality relative to the evaluated proprietary baseline and a clear preference over the two open-source PE baselines. The preference for CFAD+RL over pre-RL CFAD further supports that consistency reward optimization improves the perceptual quality valued by human evaluators.

A.4 FDBench Construction Details

FDBench is constructed through a six-stage pipeline applied to real e-commerce photography, followed by multi-round quality filtering. The overall flow is summarized below, with all VLM prompts reproduced verbatim.

Stage 1: Reference Image Curation. We collect ∼85K product-reference image pairs from a large-scale fashion e-commerce platform, selecting only items whose product pages contain at least one dedicated detail photograph. We first discard images with resolution below 720p, then score the remaining images for aesthetic quality using Doubao-Seed-2.0-pro [81]. Following the evaluation

dimensions of HPSv3 [70], we prompt the VLM to assess each image along five axes: composition, color harmony, lighting, texture clarity, and structural integrity. The scoring prompt is shown below.

Aesthetic Scoring Prompt

```
You are an expert image quality assessor. Evaluate the given product
photograph along the following five dimensions, each scored from 1 (worst)
to 5 (best):

1. Composition: Is the garment well-framed, with a clear subject and
balanced layout?
2. Color Harmony: Are the colors natural, well-balanced, and free of color
cast?
3. Lighting: Is the lighting even and professional, without harsh shadows
or overexposure?
4. Texture Clarity: Are fabric textures, stitching, and material details
clearly visible?
5. Structural Integrity: Is the garment properly displayed without wrinkles,
distortion, or occlusion?

Return strict JSON only:
{"composition": N, "color_harmony": N, "lighting": N, "texture_clarity":
N, "structural_integrity": N, "overall": N}
```

We retain the top 75% of images by overall aesthetic score, yielding 60K reference images, each with at least one corresponding real detail photograph.

Stage 2: Detail–Reference Grounding. For each existing real detail image, we use Gemini 3.1 Pro [82] to localize the corresponding region on the reference image via bounding-box grounding, producing a spatial annotation for the reference–detail pair.

Detail Grounding Prompt

```
Image 1 is the full product image; Image 2 is a detail reference image.
Locate the region in Image 1 that best matches Image 2.

Return exactly one bounding box covering only the most matching local
region--do not expand to the entire garment.

Output a JSON array: [{"box_2d": [ymin, xmin, ymax, xmax]}].
box_2d uses Image 1 coordinates, range 0-1000.
Do not return markdown, explanations, or any other text.
```

Stage 3: Part-Type Clustering. We deploy a locally hosted Qwen3-VL-30B-A3B [73] to classify each reference–detail pair into one of 20 predefined garment part types. The 20 part labels are selected from the top-30 business-relevant detail categories by manual curation, chosen to cover the vast majority of garment structures: *Collar, Cuff, Fabric/Material, Coat Hem, Pocket, Trouser Cuff, Placket, Bottom Hem, Shoulder Seam, Button, Waistband, Stitching, Skirt Hem, Embroidery, Trouser Waist, Zipper, Quilting, Waistline, Cap Brim, Fur/Wool Texture*. Note that this taxonomy is intentionally soft: a detail image only needs to contain the relevant part rather than correspond to it exclusively. The part-type label serves solely for dataset balancing and analysis, as it is not used as a generation condition during training or inference.

Part-Type Clustering Prompt

```
You are given two images of a garment. Image 1 is the full product image
with a red bounding box marking a region; Image 2 is a close-up detail of
that region.

Classify the detail region into exactly one of the following 20 part types:
[Collar, Cuff, Fabric/Material, Coat Hem, Pocket, Trouser Cuff, Placket,
Bottom Hem, Shoulder Seam, Button, Waistband, Stitching, Skirt Hem,
```

```
Embroidery, Trouser Waist, Zipper, Quilting, Waistline, Cap Brim, Fur/Wool
Texture]

If the detail does not include any of the above categories, return "None".

Return strict JSON only: {"part_type": "xxx"}
```

Pairs classified as `None` (i.e., unclassifiable) are discarded, leaving 52K grounded reference–detail pairs covering approximately 22K unique reference images.

Stage 4: Automated Quality Filtering. We apply a final round of VLM-based quality filtering using Gemini 3.1 Pro, evaluating each pair along three binary dimensions:

Quality Filtering Prompt

```
You will see two images.
Image 1: the original product image (full garment).
Image 2: the candidate detail image.

Product category: {category}-{sub_category}
Detail part type: {detail_part}

Evaluate strictly along three dimensions:
1. part_match: Does Image 2 actually show a close-up of the specified
detail part?
  - correct: Image 2 clearly shows the specified part in detail.
  - wrong: Image 2 does not show the specified part, or the part is
ambiguous.
2. product_consistency: Are Image 1 and Image 2 the same product, with
consistent details (color, fabric, texture, style)?
  - correct: Same product, details are consistent.
  - wrong: Clearly not the same product, or details do not match.
3. structure_quality: Is the structure of Image 2 reasonable? E.g.,
garment is complete, no distortion, natural appearance.
  - correct: Structure is reasonable and natural.
  - wrong: Obvious structural problems.

Return strict JSON only:
{"part_match": "correct/wrong", "product_consistency": "correct/wrong",
"structure_quality": "correct/wrong", "reason": "brief explanation"}
```

Pairs that fail any of the three checks are removed, reducing the dataset to approximately 46K pairs.

Stage 5: Manual Screening. A final round of comprehensive human review is conducted following the same three-dimensional criteria as Stage 4, with additional attention to fine-grained identity preservation: material fidelity, accessory retention (e.g., buttons, zippers, embroidery), and color consistency between the reference and detail images. After manual screening, the dataset contains **40893** verified reference–detail pairs.

Stage 6: Train/Test Split. We perform stratified random sampling by category proportion to construct an **800-pair test set** and a **40093-pair training set**. The split is conducted at the product-SKU level to prevent data leakage between training and evaluation.

A.5 FDBench Composition and Split Details

The final FDBench dataset contains **40,893** reference–detail pairs spanning three garment categories and 41 product types. Table 5 summarizes the category-level distribution.

Across product categories, the dataset covers 41 distinct product types, defined as category×subcategory combinations (e.g., Womenswear×T-shirt and Menswear×T-shirt are counted as separate types). The top-5 product types are: Female Short Jacket (3,317), Female Base Layer Top (3,040),

Table 5: FDBench category distribution.

Category	Count	Percentage
Womenswear	23,389	57.2%
Menswear	11,577	28.3%
Childrenswear	5,927	14.5%
Total	**40,893**	**100%**

Table 6: Test set product-type distribution (800 pairs, 32 sub-category types present; the parent category Womenswear/Menswear/Childrenswear is collapsed). The distribution is smoothed relative to the training set to ensure adequate coverage of rare types.

Product Type	Count	Product Type	Count
Short Jacket	59	Skirt	24
Base Layer Top	58	Regular Coat	24
T-shirt	53	Jeans	24
Casual Trousers	51	Cardigan	21
Wool Coat	40	Hoodie/Fleece	19
Pullover Knit	40	Sweater	19
Dress	37	Sweatshirt	13
Down Jacket	37	Middle-aged Top	11
Jacket	32	Casual Set	8
Shirt	29	Misc. Set	6
Padded Jacket	29	Leggings	6
Fur Coat	27	Sportswear Set	6
Casual Pants	27	Padded Coat	6
Denim Trousers	27	Tank Top	6
Vest	27	Leather Jacket	5
Knitwear	24	Leisure Set	5
Total		**800**	

Female T-shirt (2,420), Female Down Jacket (2,309), and Female Casual Trousers (2,135), while the smallest categories include Female Sweatpants (203) and Male Sportswear Sets (217).

To construct the test set, we apply *smoothed* stratified sampling rather than strict proportional sampling: while the overall category ratio (Womenswear/Menswear/Childrenswear) is approximately preserved, the per-product-type distribution is deliberately flattened so that product types with low frequency in the training set still receive a sufficient number of test samples for reliable evaluation. This design ensures that the model's generalization to rare product types can be properly assessed. Table 6 lists the product-type composition of the 800-pair test set.

The split is performed at the product-SKU level to ensure no reference image appears in both sets, preventing data leakage during evaluation.

A.6 Consistency Reward Dataset and Model Validation

Dataset construction. We construct the consistency reward dataset from the $\sim$60K reference–detail pairs produced at Stage 2 of the FDBench pipeline (Section A.4), *before* manual screening. Each pair is re-annotated by Gemini 3.1 Pro using a detailed scoring prompt that evaluates three independent dimensions on a 1–4 integer scale: **aesthetic plausibility** (fabric authenticity, craft rationality, visual integrity, commercial usability), **product identity consistency** (fabric, craft, trim, and identity-exclusive feature matching against the reference), and **target part fidelity** (part accuracy, structure fidelity, detail completeness, focus precision). Each dimension has strict rubrics with explicit positive/negative indicators; the full scoring prompt is reproduced below.

Consistency Reward Scoring Prompt

```
You are a senior professional judge specializing in fashion product visual
quality inspection and cross-image detail consistency verification, with
professional knowledge of clothing technology, fabric texture, hardware
accessories, and garment structure.

### Input Information
You will receive 2 fixed inputs, and your evaluation must be strictly based
on the following inputs:
1. Image A (Reference Image): Full-view product image of the target
clothing, with a red bounding box marking the target detail region.
2. Image B (Candidate Detail Image): The close-up image to be evaluated,
which is supposed to show a detail view of the boxed region in Image A.

### Evaluation Rules
Evaluate Image B against Image A from the following 3 independent dimensions.
For each dimension, you must strictly use only the integer scores from 1 to 4
according to the exclusive scoring criteria for clothing products.

--
#### Dimension 1: aesthetic_plausibility
This dimension focuses on the visual authenticity, physical rationality, and
commercial usability of the candidate image as a real clothing product detail
shot.

Key Criteria:
• Fabric Authenticity: Matches the physical texture and properties of the
specified fabric.
• Craft Rationality: Garment structure, stitching and hardware comply with
standard clothing production rules.
• Visual Integrity: No artifacts, blurring or distortion that impairs detail
recognition.
• Commercial Usability: Meets professional standards for commercial clothing
detail imagery.

Negative Indicators:
• Severe fabric/structure distortion that is physically impossible.
• Obvious craft defects (twisted stitching, through-modeled hardware) or
major visual artifacts.
• Content is unrecognizable as a clothing detail.

Scoring Rubric (1-4 Scale):
• 4 (Very Good): Fully meets professional detail shot standards. Fabric,
craft and visuals are completely authentic and flawless.
• 3 (Relatively Good): Overall usable and reasonable, with only
imperceptible minor flaws that do not affect core detail recognition.
• 2 (Relatively Poor): Has obvious defects in fabric, craft or visual
quality that impair usability and detail recognition.
• 1 (Very Poor): Completely unusable. The image is severely distorted,
structurally impossible, corrupted or unrecognizable.

#### Dimension 2: Product Identity Consistency
This dimension focuses on whether the candidate image belongs to the exact
same clothing product as the reference image, verified by matching unique
product-specific features.

Key Criteria:
• Fabric Consistency: Exact match of fabric material, texture, color and
finish to the reference.
• Craft Consistency: Full alignment of stitching, thread details and seam
structure with the reference.
• Trim Consistency: Precise match of hardware, buttons, embroidery and
decorative parts.
• Identity Exclusivity: Matches unique, non-generic features that
distinguish the garment from similar products.

Negative Indicators:
```

• Core product features (fabric, craft, trims) are completely mismatched with the reference.
• Only generic color/category similarity, with no matching of unique product-specific details.
• The candidate image clearly belongs to a different clothing product.

Scoring Rubric (1-4 Scale):
• 4 (Very Good): All unique product features are 100% consistent with the reference, fully confirming the same garment with no deviations.
• 3 (Relatively Good): All core product features are highly matched, with only minor, imperceptible deviations that do not affect same-product judgment.
• 2 (Relatively Poor): Only generic features overlap; core unique features are weakly matched or altered, cannot confirm the same product.
• 1 (Very Poor): Completely unrelated product. All core features contradict the reference, or content is unrecognizable.

Dimension 3: target_part_fidelity
This dimension focuses on how accurately, completely and clearly the candidate image depicts the specified target clothing detail part, while preserving its correct local structure from the reference.

Key Criteria:
• Part Accuracy: Content exactly matches the requested target part, with no incorrect parts included.
• Structure Fidelity: Local structure and spatial relationship of the part fully align with the reference.
• Detail Completeness: All core process and decorative details of the target part are fully presented.
• Focus Precision: The target part is the clear focus of the image, with no incorrect framing.

Negative Indicators:
• The target part is missing, unrecognizable, or replaced by a wrong part.
• Local structure of the target part is severely distorted or inconsistent with the reference.
• Core details are omitted, or the focus is completely off the requested part.

Scoring Rubric (1-4 Scale):
• 4 (Very Good): Accurately and completely depicts the target part, with structure, details and framing perfectly consistent with the reference.
• 3 (Relatively Good): Clearly presents the target part with all core details matched, with only minor framing/focus deviations that do not affect recognition.
• 2 (Relatively Poor): The target part is present but has obvious deviations in structure, missing details, or incorrect framing.
• 1 (Very Poor): Completely fails to show the target part. The part is wrong, missing, unrecognizable, or the image is corrupted.

Mandatory Scoring Constraints (must be strictly followed)
1. Independent Judgment: The three dimensions must be scored independently, and the score of one dimension must not interfere with the others.
2. High Score Restriction: Generic similarity is not enough for a score of 3 and above. High scores must have clear and verifiable matching evidence of exclusive clothing details.
3. Strict Principle for Uncertainty: When there is any uncertainty about feature matching, part accuracy, or rationality, you must choose the lower score gear, and overestimation is strictly prohibited.
4. Score Range Limit: Only integer scores from 1 to 4 are allowed.

Output Requirements
Return only a single standard JSON object without any additional text, notes, or explanations, strictly following the schema below. The "reason" must be a specific description corresponding to the scoring criteria, pointing out the exact clothing detail features/mismatches/defects, and vague descriptions such as "has defects" or "good consistency" are prohibited.

```
{"aesthetic_plausibility": {"score": 1-4, "reason": "specific reason
corresponding to the scoring criteria"}, "product_identity_consistency":
{"score": 1-4, "reason": "specific reason corresponding to the scoring
criteria"}, "target_part_fidelity": {"score": 1-4, "reason": "specific
reason corresponding to the scoring criteria"}}
```

We retain only the three integer scores from each response, discarding the free-text reasons, to form the final $\sim$**60K consistency reward dataset**.

Model training. We fine-tune Qwen3-VL-8B [73] as the reward backbone via supervised fine-tuning for 2 epochs. The model takes a boxed reference image (with the red bounding box rendered) and the candidate detail image as input, without any text description from prompt enhancement. The input prompt is fixed as:

Reward Model Input Prompt

```
<image> is the reference image with a red bounding box marking the target
region. <image> is a candidate detail image of the boxed region. Please
score the image pair along aesthetic_plausibility, Product Identity
Consistency, and target_part_fidelity. For each dimension, you must strictly
use only the integer scores from 1 to 4 according to the exclusive scoring
criteria for clothing products.
```

Validation. We evaluate the trained reward model on a held-out test set of 200 reference–detail pairs (100 positive, 100 negative). Positive samples are pairs that passed the Stage 5 manual screening in the FDBench pipeline; negative samples are pairs that were rejected during the same screening round. Table 7 reports the mean normalized score (0–1 scale) for positive and negative samples, their separation (gap), effect size (Cohen's d), and statistical significance (Welch's t-test).

Table 7: Consistency reward model validation on 200 held-out pairs (100 positive, 100 negative). Scores are normalized to $[0, 1]$. ***: $p < 0.001$.

Dimension	Positive μ	Negative μ	Gap	Cohen's d
Aesthetic Plausibility	0.923	0.907	0.017	0.08
Identity Consistency***	0.760	0.547	0.213	0.54
Target Fidelity***	0.833	0.617	0.217	0.60

The reward model achieves statistically significant separation on the two consistency-critical dimensions: identity consistency ($p = 0.0002$, $d = 0.54$) and target fidelity ($p < 0.0001$, $d = 0.60$), confirming its ability to discriminate between high-quality and defective detail generations. The aesthetic plausibility dimension shows a small, non-significant gap ($d = 0.08$), which is expected: rejected pairs in Stage 6 were primarily filtered for identity or fidelity failures rather than aesthetic defects, so both positive and negative samples tend to have comparable visual quality.

A.7 Feature Distribution Analysis of CFAD

To directly verify that CFAD aligns the internal representations of the MMDiT, we extract hidden features from four layers (10, 16, 20, 30) of the image branch and text branch for both the CFAD-trained model and a vanilla SFT baseline (trained with the same schedule but without any distillation loss). We select 90 test samples (30 per category) from three representative part types—Cuff, Collar, and Pocket—and perform a single forward pass per sample at a fixed intermediate timestep ($t \approx 0.5$) to capture the mean-pooled hidden states.

We measure intra-class cosine similarity (average pairwise similarity among samples of the same part type) and inter-class cosine similarity (average similarity between samples of different part types), and report their difference as the *separation* score. Higher separation indicates that the features form tighter, more discriminative clusters by part type. Results are summarized in Table 8.

Table 8: **Feature clustering analysis across MMDiT layers.** Intra-/inter-class cosine similarity and separation score for the image branch and text branch at layers 10, 16, and 30. CFAD substantially improves separation in both branches, with layer 16 (the alignment target l^*) showing the largest gain.

Layer	Model	Img-Branch Sep.	Text-Branch Sep.
10	w/o CFAD	0.00030	0.00086
	w/ CFAD	0.00191	0.00624
16	w/o CFAD	0.00056	0.00062
	w/ CFAD	0.00521	0.03638
20	w/o CFAD	0.00017	0.00039
	w/ CFAD	0.00143	0.01469
30	w/o CFAD	0.00006	0.00034
	w/ CFAD	0.00033	0.00382

Without CFAD, all layers exhibit negligible separation ($<$0.001) in both branches: the features are essentially collapsed into an undifferentiated cluster regardless of which garment region the bounding box indicates. To better illustrate the layer-wise effect, Fig. 4 plots the *separation ratio* (w/ CFAD divided by w/o CFAD) across the four layers. Both branches peak sharply at the alignment target layer 16: the text branch achieves a $58.7\times$ ratio, and the image branch a $9.3\times$ ratio. Neighboring layers also benefit, with layer 20 still showing $37.7\times$ for text and $8.4\times$ for image, but the gain attenuates rapidly toward earlier and later layers. This peaked pattern is consistent with the REPA design [20]: distillation is applied at a single intermediate layer, and the alignment signal propagates partially to adjacent layers but is strongest at the injection point. The disproportionately large text-branch improvement confirms that text-branch distillation is critical for injecting region-level semantic understanding, which is precisely the bottleneck identified in Section 3.2. These results provide direct evidence that CFAD successfully induces part-type-aware clustering in the MMDiT's internal feature space, corroborating the quantitative gains observed in the main ablation study (Table 2).

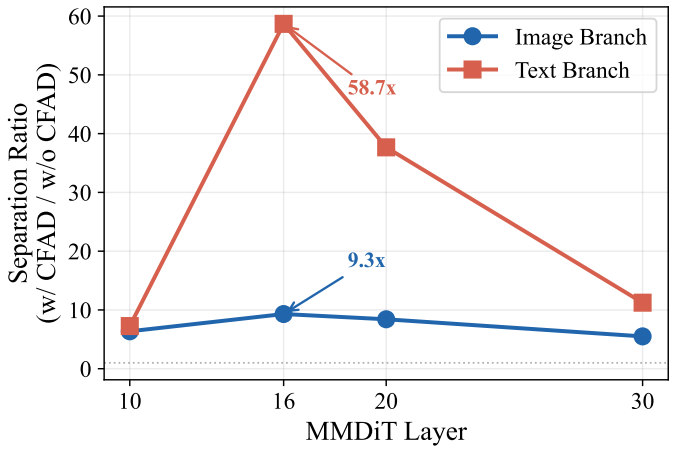

Figure 4: **Separation ratio across MMDiT layers.** Ratio of the separation score (w/ CFAD / w/o CFAD) for image and text branches. Both branches peak at the alignment target layer 16 (l^*), with the text branch reaching $58.7\times$.

A.8 Cross-View Alignment Analysis

We measure correspondence between boxed references and their paired details in the teacher and generator representation spaces on all 800 test pairs. Table 9 reports matched and nonmatching cosine similarities and cross-view retrieval. For each query, Own is its cosine similarity to the matched view, and Random averages similarities to the 799 nonmatches. We compute the confidence interval of the mean Own-minus-Random gap with 10,000 bootstrap resamples. Retrieval ranks the paired match among 800 candidates; the chance expectation for R@1 is 0.125%.

In teacher space, we compare frozen DINOv3 and the fine-tuned view-bridging teacher using the same image preprocessing and evaluate reference-to-detail retrieval. In generator space, we mean-pool the boxed-reference conditioning-branch hidden states over non-padding tokens at layer $l^* = 16$ and $t = 0.5$, with the noisy ground-truth detail latent in the image branch. The reference representation is passed through the text projector into the teacher's 1024-dimensional global embedding space and compared with the teacher detail embedding. CFAD uses its learned projector; SFT uses an architecturally identical, untrained projector. Generator retrieval is reported in the detail-to-reference direction. Thus, these generator-space results evaluate the projected conditioning representation together with its projector.

Table 9: Cross-view alignment on 800 FDBench test pairs. Own and Random are matched and mean nonmatching cosine similarities. Gap is Own minus Random, with a 95% bootstrap CI. Teacher retrieval is reference-to-detail; generator retrieval is detail-to-reference.

Space	Model	Own	Random	Gap [95% CI]	R@1	R@10
Teacher	Frozen DINOv3	0.524	0.134	0.389 [0.368, 0.412]	**44.0%**	76.0%
Teacher	View-bridging teacher	**0.786**	0.219	**0.567** [0.540, 0.594]	39.7%	**81.0%**
Generator	SFT	−0.010	−0.010	0.000 [−0.003, 0.003]	0.3%	4.0%
Generator	CFAD	**0.419**	0.375	**0.044** [0.032, 0.056]	**5.0%**	**29.3%**

Teacher fine-tuning increases the Own-minus-Random gap from 0.389 to 0.567. Because this gap accounts for similarity to nonmatching details, its increase indicates stronger selectivity for the corresponding reference–detail pair. This supports the role of the view-bridging teacher in providing a task-adapted cross-view target for distillation. In the generator's projected representation, CFAD yields a positive gap of 0.044, with a 95% confidence interval of [0.032, 0.056], and raises R@10 from 4.0% to 29.3% relative to SFT. These complementary observations support the intended two-stage design: teacher adaptation strengthens cross-view correspondence, and dual-branch distillation transfers this signal into the generator, with the conditioning-branch result directly supporting its learned association between the boxed reference and the target detail.

A.9 Comparison with Reference-Based Super-Resolution

We evaluate iRAG [83], TTSR [84], and Ada-RefSR [85] on FDBench. The bounding-box crop is the low-resolution input, the full garment image is the reference, and the paired detail image is the evaluation target. Table 10 reports image-quality and EditScore results on all 800 test samples, with HPSv3 evaluated on the same set. CIS compares each generated image with the ground-truth detail, whereas CCIS compares it with the reference crop.

RefSR methods preserve the input crop and consequently obtain high CCIS, but score lower on target-detail similarity, editing quality, and visual preference. For example, Ada-RefSR obtains CIS 0.753 and EditScore Overall 6.72, compared with 0.863 and 8.95 for DetailAnywhere (CFAD+NFT). These results support the distinction between restoring an input crop and synthesizing an identity-consistent close-up with appropriate viewpoint and framing.

Table 10: Comparison with reference-based super-resolution on FDBench. PF, CS, PQ, and O denote EditScore dimensions. All scores are reported using the same evaluation definitions as the main comparison; HPSv3 measures visual preference.

Method	SSIM ↑	LPIPS ↓	FID ↓	CIS ↑	CCIS ↑	PF ↑	CS ↑	PQ ↑	O ↑	HPSv3 ↑
iRAG	0.361	0.679	98.2	0.732	0.841	6.31	5.82	5.25	4.83	−2.48
TTSR	0.371	0.685	87.0	0.723	**0.923**	6.96	6.82	4.20	4.74	−2.97
Ada-RefSR	0.366	0.657	75.8	0.753	0.879	8.31	7.74	6.80	6.72	−0.77
Qwen-Image-Edit (PE)	0.254	0.648	82.6	0.804	0.739	9.51	8.77	9.42	8.85	4.96
FireRed (PE)	0.274	0.660	85.0	0.791	0.719	**9.64**	8.39	9.47	8.71	5.18
DetailAnywhere (CFAD)	0.366	0.608	70.9	0.861	0.741	9.41	8.61	9.48	8.79	4.93
DetailAnywhere (CFAD+NFT)	**0.394**	**0.606**	**66.9**	**0.863**	0.748	9.57	**8.82**	**9.59**	**8.95**	**5.41**

A.10 Additional Visualizations

In addition to the main paper visualization, we provide more side-by-side comparisons of generated detail images from different methods on representative FDBench examples in Fig. 5.

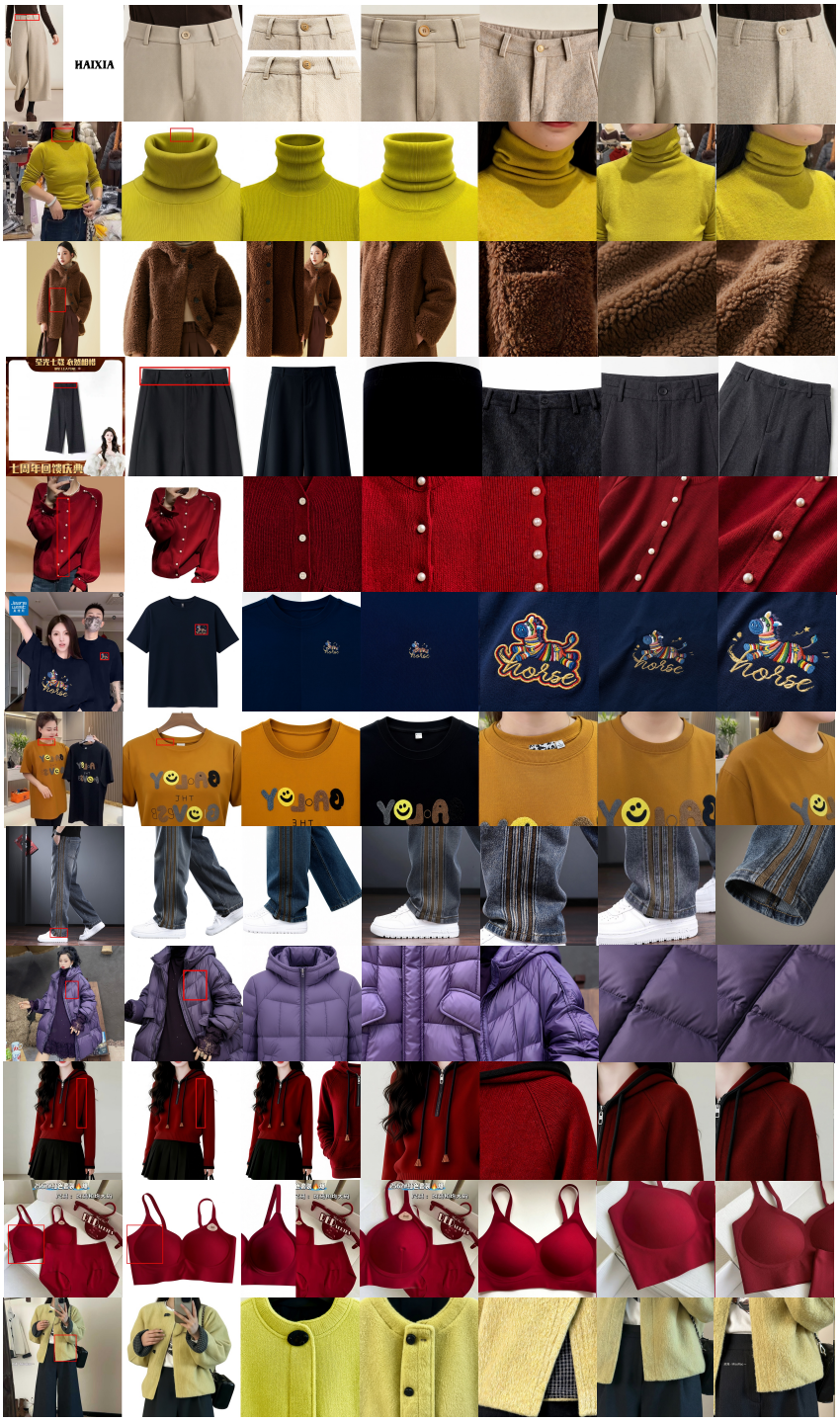

Figure 5: **Additional Fashion Detail Visualizations.** Additional visualizations.