OrthoTryOn: Geometric Orthogonalization for Conflict-Free Unified Fashion Generation

Zhaotong Yang[1], Ying Tai[2] ⋆, Jiahui Zhan[3], Yu Zheng[1], Jianjun Qian[1], and Jian Yang[1,2] ⋆

[1] PCA Lab, School of Computer Science and Engineering, Nanjing University of Science and Technology
[2] PCA Lab, School of Intelligence Science and Technology, Nanjing University
[3] Shanghai Jiao Tong University

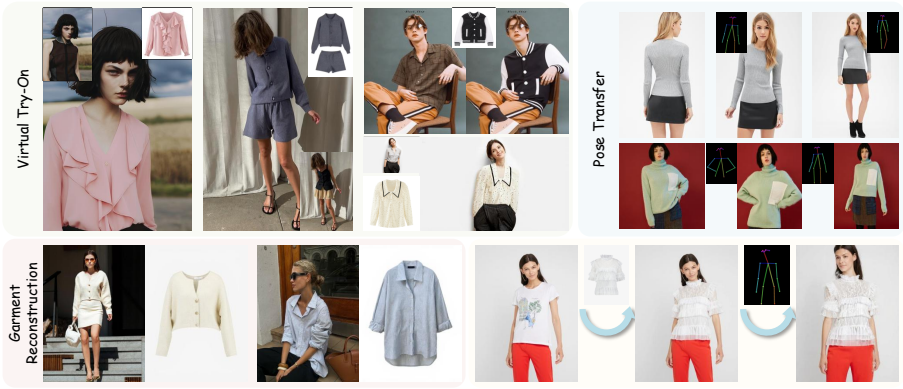


Fig. 1: We present OrthoTryOn, a unified generalist model capable of handling diverse fashion tasks within a single architecture, including virtual try-on, garment reconstruction, and pose transfer. The shared architecture also naturally supports sequential editing by chaining task-specific conditions.

Abstract. Unified fashion generation integrates tasks like virtual try-on and garment reconstruction into a single model to reduce task-specific adaptation costs. However, naive parameter sharing across semantically distinct tasks induces negative transfer through severe inter-task gradient conflict. We propose OrthoTryOn, a unified framework mitigating this interference within a shared Low-Rank Adaptation (LoRA) module. Its Orthogonal Subspace Projection (OSP) applies task-specific orthogonal rotations to bottleneck features, mapping them into decorrelated coordinate frames. To address residual semantic coupling at inference time, we further propose Fisher-guided Negative Guidance (FNG), a parameter-free strategy that utilizes diagonal Fisher information to quantify inter-task sensitivity overlap and explicitly repels generation trajectories from the most confusable task via Classifier-Free Guidance. Extensive experiments demonstrate that OrthoTryOn avoids the severe performance degradation typical of naive unified training and even surpasses independently trained task-specific models, achieving state-of-the-art results

⋆ Corresponding authors.

across multiple benchmarks while generalizing robustly across diverse diffusion backbones. Code is available at `https://github.com/NJU-PCALab/OrthoTryOn`.

Keywords: Virtual Try-On · Unified Fashion Generation · Diffusion Model

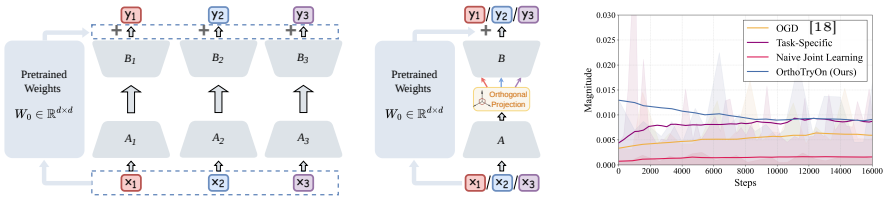
1 Introduction

Recent advances in diffusion-based image generation and editing have enabled increasingly controllable visual synthesis [8–11, 43, 62], creating new opportunities for digital fashion. Among various fashion-related tasks, Virtual Try-On (VTON) aims to synthesize realistic images of a person wearing a target garment. Despite achieving impressive visual fidelity, most existing methods [13,33,59] remain specialized solutions constrained by stringent inputs (*e.g.*, paired data or clean garment templates). Furthermore, current frameworks are typically designed to address a single predefined task. Consequently, supporting multiple fashion applications requires maintaining an ensemble of independently trained, task-specific Low-Rank Adaptation (LoRA) modules (as illustrated in Fig. 2(a)), which inherently lacks scalability and poses significant deployment burdens.

To break the limitations of single-task specificity, pioneering works such as Any2AnyTryon [20] and UniFit [64] have attempted to construct a unified generation paradigm. In such frameworks, adopting a single shared LoRA module for multi-task joint learning is a natural choice to maintain computational efficiency. However, this naive parameter sharing inevitably leads to noticeable performance degradation. Because different tasks possess distinct semantic objectives (*e.g.*, spatial alignment for VTON *vs*. structural preservation for reconstruction), forcing them into an identical parameter space makes it difficult for the model to capture task-specific subtle differences.

In this work, we point out that the fundamental cause of performance degradation in unified fashion generation lies in the *gradient conflict* during the multi-task joint optimization process. Through empirical analysis of gradient magnitudes within the shared LoRA parameters, we observe a sharp decay under naive multi-task training, as depicted in Fig. 2(c), suggesting destructive interference among task gradients in the low-rank parameter space. Such interference drives the model toward a compromised solution that is suboptimal for all tasks.

To address these challenges, we propose **OrthoTryOn**, a novel unified framework that structurally mitigates negative transfer within a single shared LoRA module. Specifically, we design the Orthogonal Subspace Projection (OSP) strategy (Fig. 2(b)), which introduces *task-specific orthogonal rotations* Q_i into the shared LoRA bottleneck. By rotating task-specific bottleneck features into decorrelated coordinate frames (without changing their magnitudes), OSP eliminates expected correlations between weight increments of different tasks and substantially reduces gradient-level interference, enabling stable joint training in a shared low-rank parameter space. In practice, each Q_i is sampled once and then frozen, introducing negligible overhead.



(a) Task-Specific LoRAs. (b) OrthoTryOn (Ours). (c) Gradient magnitude comparison.

Fig. 2: (a) Task-Specific LoRAs maintain independent adapters for each task. (b) Our OrthoTryOn within a shared LoRA, enforcing decorrelated coordinate frames to mitigate inter-task conflict. (c) Gradient magnitudes of different methods across training steps, showing that OrthoTryOn effectively mitigates the gradient decay caused by inter-task conflict.

Despite the statistical decorrelation introduced by OSP, residual semantic coupling may persist when the LoRA bottleneck dimension is small. To further enhance task discrimination at inference time, we introduce Fisher-guided Negative Guidance (FNG). FNG is a parameter-free strategy that quantifies inter-task sensitivity overlap using Fisher information and identifies the most confusable task as a hard negative condition within the Classifier-Free Guidance (CFG) framework [23]. OSP and FNG are highly synergistic: the former reduces gradient conflict during training, while the latter explicitly mitigates residual semantic leakage during generation.

Comprehensive experiments validate OrthoTryOn's effectiveness for multi-task fashion generation. Unlike conventional unified frameworks, OrthoTryOn not only avoids the performance degradation caused by negative transfer, but also consistently outperforms independently trained task-specific models by preserving cross-task knowledge sharing. Furthermore, it demonstrates strong cross-architecture generalizability as a universal plug-in. The key contributions of this work are summarized as follows:

- We propose OrthoTryOn with Orthogonal Subspace Projection (OSP), leveraging task-specific orthogonal rotations Q_i to construct decorrelated low-rank coordinate frames, enabling decorrelated joint optimization in expectation.
- We introduce Fisher-guided Negative Guidance (FNG), a parameter-free inference strategy that utilizes diagonal Fisher sensitivity to quantify inter-task overlap and handle residual coupling.
- OrthoTryOn achieves new state-of-the-art results across multiple benchmarks and consistently outperforms independently trained task-specific models, highlighting that properly structured parameter geometry can simultaneously suppress negative transfer and promote positive transfer in multi-task generation.

2 Related Work

Fashion Image Generation. Alongside recent progress in controllable image generation and editing [17, 28, 29, 37, 65, 67], Virtual Try-On (VTON) has

garnered significant attention due to its immense commercial potential [22, 33]. VITON [22] pioneered image-based synthesis in this domain, while subsequent flow-based approaches [12, 56, 59] improved clothing-body alignment via dense appearance flow estimation. Recently, OmniVTON [60] extended this paradigm to unconstrained real-world scenarios. However, mainstream VTON often relies on strict inputs, such as clean flat-lay garments. To bypass this, garment reconstruction [52,55] aims to extract standardized garments directly from human images. Concurrently, pose transfer [1,46,68] synthesizes novel poses while preserving identity and appearance, which is closely related to appearance-consistent correspondence under large deformations studied in visual tracking [16, 26].

To bridge these tasks, Any2AnyTryon [20] and UniFit [64] adopt a shared parameter space for unified fashion generation. However, forcing a single shared parameter space to jointly fit tasks with substantial semantic discrepancies inevitably induces severe inter-task gradient conflicts. OrthoTryOn mitigates this issue through task-specific orthogonal rotations in the shared LoRA bottleneck, enabling more effective utilization of large-scale multi-task data.

Task Decoupling in Deep Learning. Orthogonality has long been used to stabilize optimization via weight initialization or manifold constraints [2, 35, 42, 48, 54], while contrastive learning improves discriminability by separating semantically confusable representations [5, 58, 66]. More recently, projection-based methods such as PCGrad [61] and OGD [18] resolve conflicts by dynamically projecting gradients during backpropagation. However, these post-hoc manipulations require task-wise gradient isolation and additional backward passes, introducing non-negligible computational overhead. Moreover, explicitly discarding conflicting components may overly constrain optimization directions and suppress effective gradient magnitude.

In unified fashion generation, multiple tasks [20, 64] are typically accommodated within a single shared parameter space and distinguished only via conditional inputs. While computationally efficient, such tightly coupled parameter sharing inevitably induces severe inter-task gradient interference, leading to suboptimal convergence. In contrast, OrthoTryOn adopts a forward architectural orthogonalization strategy within a shared low-rank space by inserting task-specific orthogonal rotations, structurally decorrelating task updates in expectation and avoiding costly gradient manipulation, while Fisher-guided Negative Guidance further mitigates residual semantic leakage at inference.

Parameter-Efficient Fine-Tuning. As foundation models scale up rapidly, full-parameter fine-tuning becomes computationally prohibitive and prone to overfitting on downstream tasks. To alleviate this, Parameter-Efficient Fine-Tuning (PEFT) methods [24, 34, 36] have emerged, aiming to achieve comparable performance to full fine-tuning by updating only a minuscule fraction of parameters.

Among various PEFT techniques, Low-Rank Adaptation (LoRA) [25] is widely adopted due to its exceptional effectiveness and zero additional inference latency. For a pre-trained linear weight $W_0 \in \mathbb{R}^{d_{in} \times d_{out}}$ and an input feature x, LoRA freezes W_0 and introduces a trainable low-rank bypass. The forward propagation

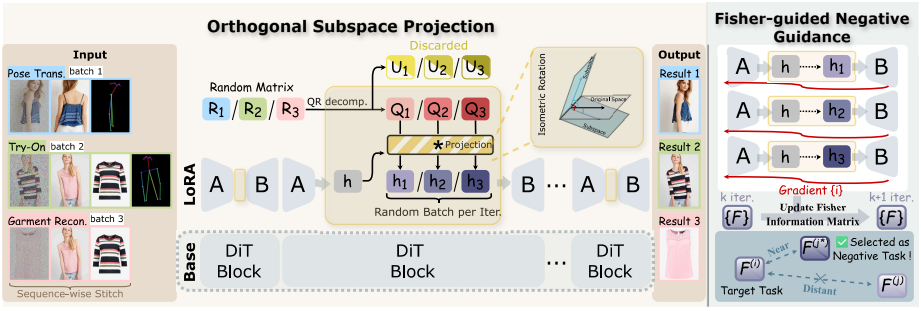


Fig. 3: Overview of OrthoTryOn. Orthogonal Subspace Projection utilizes task-specific orthogonal matrices Q_i in the shared LoRA module to rotate bottleneck features into decorrelated coordinate frames, achieving statistically decorrelated optimization in expectation. Based on parameter sensitivity tracked during training, Fisher-guided Negative Guidance explicitly suppresses the interfering task via CFG to prevent semantic leakage.

is formulated as:

$$y = xW_0 + \alpha xAB, \tag{1}$$

where $A \in \mathbb{R}^{d_{in} \times r}$ and $B \in \mathbb{R}^{r \times d_{out}}$ are the down- and up-projection matrices, respectively, with a bottleneck rank $r \ll \min(d_{in}, d_{out})$. The hyperparameter α scales the low-rank module and is omitted in subsequent derivations for simplicity.

Typically, A is initialized with a Gaussian distribution and B with zeros, ensuring the initial bypass output is zero to perfectly preserve pre-trained representations. While extensively injecting LoRA modules into Transformer layers (*e.g.*, attention and feed-forward networks) maximizes fitting capacity for multi-task generation, directly employing a shared LoRA space for joint learning inevitably triggers severe inter-task gradient conflicts, as revealed in Sec. 3.2.

3 Methods

3.1 Overview of Universal Fashion Generation

We present a universal generative paradigm for multi-task fashion image editing. Rather than devising task-specific subnetworks, we seamlessly concatenate visual conditions along the sequence dimension: virtual try-on requires a reference model, a target garment, and a pose skeleton map; garment reconstruction utilizes a reference model and a reference garment (providing background attributes); and pose transfer relies on a reference model and a target pose skeleton map. Guided by task-specific text prompts, this concatenated sequence is directly fed into a unified Diffusion Transformer backbone.

While this paradigm unifies tasks at the input level, employing standard LoRA for joint multi-task learning inevitably encounters a parameter optimization bottleneck due to gradient conflicts (elaborated in Sec. 3.2). To overcome

this, we propose the OrthoTryOn framework, comprising Orthogonal Subspace Projection (OSP) and Fisher-guided Negative Guidance (FNG), which enables architecture-agnostic multi-task decoupling and synergistic optimization within the shared LoRA space.

3.2 Motivation

To investigate the performance degradation in naive joint multi-task learning, we analyze the gradient dynamics during optimization. Our empirical analysis suggests that inter-task gradient interference is a major source of the model's performance deterioration. Specifically, we track and quantify the L_2 norm of the backpropagated gradients for the virtual try-on task under two settings: single-task training and joint multi-task learning. As illustrated in Fig. 2(c), the gradient norm under single-task fine-tuning stabilizes at approximately 8.5×10^{-3}, whereas under joint multi-task optimization, it is reduced to about one-fifth of the single-task baseline. This pronounced attenuation is consistent with destructive interactions among task updates in the shared parameter space. Because our training framework employs uniform task sampling, updates from semantically distinct tasks are alternated throughout training and may partially counteract one another. Consequently, the shared parameters are driven toward a compromise that can be suboptimal for individual tasks.

3.3 Orthogonal Subspace Projection

To resolve the gradient cancellation bottleneck, we propose Orthogonal Subspace Projection (OSP), which rotates task-specific features into decorrelated coordinate frames within the shared low-rank bottleneck to minimize inter-task interference.

Formally, suppose we jointly optimize N fashion generation tasks within a shared LoRA parameter space θ, with $\mathcal{L}_i$ denoting the loss of the i-th task. An ideal multi-task optimization would minimize the joint objective while keeping cross-task gradients orthogonal:

$$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}_i(\theta) \quad \text{s.t.} \quad \langle \nabla_\theta \mathcal{L}_i, \nabla_\theta \mathcal{L}_j \rangle = 0, \forall i \neq j, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. Rather than explicitly enforcing this constraint, which would require costly per-task gradient isolation and projection, we pursue a forward reparameterization that reduces expected cross-task gradient correlation through architectural design.

Task-specific orthogonal rotations in LoRA. Standard LoRA computes the weight increment $\Delta W = AB$ via a low-rank down-projection $A \in \mathbb{R}^{d_{in} \times r}$ and up-projection $B \in \mathbb{R}^{r \times d_{out}}$. In OSP, for each task $i \in \{1, \ldots, N\}$, we interpose a task-specific orthogonal matrix $Q_i \in \mathbb{R}^{r \times r}$ between A and B, where $Q_i^\top Q_i = I$. The forward pass for an input feature x becomes:

$$y = xW_0 + xAQ_iB. \tag{3}$$

Intuitively, Q_i performs an *isometric rotation* in the bottleneck subspace, assigning each task a distinct coordinate frame without changing feature magnitudes.
Weight increment decorrelation. The weight increment of task i is $\Delta W_i = AQ_iB$. For distinct tasks $i \neq j$, their Frobenius inner product is:

$$\langle \Delta W_i, \Delta W_j \rangle_F = \mathrm{tr}(B^\top Q_i^\top A^\top A Q_j B). \tag{4}$$

When Q_j is sampled independently from the Haar measure on $\mathcal{O}(r)$, symmetry implies $\mathbb{E}[Q_j] = \mathbf{0}$ (for $r \geq 2$). Therefore,

$$\mathbb{E}_{Q_j}[\langle \Delta W_i, \Delta W_j \rangle_F] = \mathrm{tr}\big(B^\top Q_i^\top A^\top A \cdot \mathbb{E}[Q_j] \cdot B\big) = 0. \tag{5}$$

Thus, OSP achieves *exact* decorrelation of cross-task weight increments in expectation.
Gradient interference analysis. Let $G_i = \frac{\partial \mathcal{L}_i}{\partial (AQ_iB)}$ denote the gradient of the loss with respect to the weight increment. By the chain rule,

$$\nabla_B \mathcal{L}_i = Q_i^\top A^\top G_i, \qquad \nabla_A \mathcal{L}_i = G_i B^\top Q_i^\top. \tag{6}$$

We analyze cross-task interference under a *single-step* setting: at any given iteration, A and B are fixed from the previous update, so G_i depends on Q_i (through Eq. 3) but is *functionally independent of* Q_j for $j \neq i$. Taking the cross-task gradient inner product on B as a representative case:

$$\langle \nabla_B \mathcal{L}_i, \nabla_B \mathcal{L}_j \rangle = \mathrm{tr}(G_i^\top A Q_i Q_j^\top A^\top G_j), \tag{7}$$

and analogously $\langle \nabla_A \mathcal{L}_i, \nabla_A \mathcal{L}_j \rangle = \mathrm{tr}(Q_i B G_i^\top G_j B^\top Q_j^\top)$ for parameter A. We state the unified result below.

Property 1. Assume the loss function $\mathcal{L}$ is twice differentiable. At any optimization step conditional on fixed A and B, let the task-specific rotations $Q_i, Q_j \in \mathcal{O}(r)$ be independently sampled from the Haar measure. Then the expected cross-task gradient inner product satisfies, for both parameters A and B:

$$\big|\mathbb{E}_{Q_i,Q_j}[\langle \nabla \mathcal{L}_i, \nabla \mathcal{L}_j \rangle \mid A, B]\big| \leq \mathcal{O}(1/r) \cdot C(A, B, G, \mathcal{H}), \tag{8}$$

where C is a constant determined by the parameter norms, per-step gradient scales, and bounded Hessian approximations, independent of the rank r.

Unlike the weight-increment case, this gradient bound relaxes to $\mathcal{O}(1/r)$ because G_j functionally depends on Q_j via the forward pass. While Haar symmetry ($\mathbb{E}[Q_j] = \mathbf{0}$) eliminates the first-order interference, the remaining second-order coupling stems from the Hessian operator. Applying the Haar-orthogonal conjugation identity to this term reveals the exact $1/r$ decay rate. Detailed proofs are provided in the supplementary material.
Sampling and freezing orthogonal rotations. Each Q_i is generated once per LoRA layer and per task, and remains frozen throughout training. In practice, we sample a Gaussian matrix and apply QR decomposition, retaining the orthogonal factor as Q_i. Since $Q_i^\top Q_i = I$, OSP is isometric by construction ($\|hQ_i\|_2 = \|h\|_2$), avoiding spectral scaling and anisotropic distortion that may arise from unnormalized random projections. The storage footprint is negligible: only a fixed seed is required.

3.4 Fisher-guided Negative Guidance

While OSP reduces expected cross-task gradient correlation, the $\mathcal{O}(1/r)$ suppression factor indicates that non-negligible residual coupling may persist, particularly within highly compressed low-rank bottlenecks (*e.g.*, $r = 4$). When tasks share similar visual semantics, this residual overlap in parameter sensitivities can induce coupled conditional velocity field predictions during inference. Intuitively, relying on these overlapping parameter subsets causes the learned vector fields to exhibit correlated directional biases, increasing the risk of semantic entanglement at generation time. To handle this residual semantic leakage, we introduce Fisher-guided Negative Guidance (FNG), a plug-and-play inference strategy designed to operate synergistically with OSP.

The core philosophy of FNG is to proactively identify the most severe "interfering task" and suppress it during the decoding phase. Instead of relying on feature-space similarity, we approximate inter-task sensitivity overlap using parameter-space statistics accumulated during training. Specifically, following standard practices [4, 30, 49], we implement an empirical-Fisher-style proxy for per-parameter sensitivity. Since computing the full Fisher Information Matrix is intractable for large-scale models, we use its diagonal approximation to characterize the sensitivity of the i-th task to specific parameters.

To capture stable task sensitivities and mitigate early-stage gradient noise, we maintain an exponential moving average of squared gradients for each task to estimate its diagonal empirical-Fisher-style proxy $F^{(i)}$. At any given iteration step k, the vector is efficiently updated as:

$$F^{(i)}|_k = \beta \cdot F^{(i)}|_{k-1} + (1-\beta) \cdot \mathbb{E}\left[(\nabla_\theta \mathcal{L}_i)^2\right], \tag{9}$$

where $\beta \in [0, 1)$ is the momentum coefficient, the expectation is taken over the current mini-batch, θ denotes the trainable LoRA parameters across all adapted layers, and $F^{(i)}$ is updated only when task i is sampled. This tracked statistic reflects how strongly each parameter contributes to optimizing a specific task. After training, fully converged sensitivity vectors are used only offline to identify each task's most interfering task j^*. The high-dimensional Fisher vectors are then discarded, leaving only a discrete $i \mapsto j^*$ mapping for inference, which requires negligible storage and no trainable parameters.

During inference for task i, we identify the task with the highest Fisher similarity j^* via cosine similarity between Fisher vectors, computed as $j^* = \arg\max_{j \neq i} S(i, j)$, where:

$$S(i,j) = \frac{\langle F^{(i)}, F^{(j)} \rangle}{\|F^{(i)}\|_2 \|F^{(j)}\|_2}. \tag{10}$$

While multi-negative guidance is theoretically possible, adopting the single most severe interferer provides an optimal balance between disambiguation efficacy and computational cost. We then modify the conditional velocity prediction by replacing the unconditional null-prompt in standard Classifier-Free Guidance

Fig. 4: Qualitative comparison of virtual try-on on VITON-HD dataset [12].

with the explicit condition of the interfering task:

$$\hat{v} = v(x_t, t, c_{j^*}) + s \cdot \Big(v(x_t, t, c_i) - v(x_t, t, c_{j^*}) \Big), \tag{11}$$

where c_i and c_{j^*} represent the condition inputs for the target and the hard negative tasks, respectively, and s is the guidance scale. In the learned conditional velocity field, this operation explicitly pushes the generation trajectory away from the most heavily coupled semantic sub-manifold. Geometrically, this modifies the local vector field by introducing a repulsive component along the most correlated task direction, while preserving attraction toward the desired conditional manifold. Benefiting from this design, FNG effectively reduces semantic leakage without introducing any trainable parameters, mitigating task confusion in joint multi-task generation.

4 Experiments

4.1 Experimental Settings

Implementation Details. We adopt LongCat-Image-Edit [51] as our backbone. All experiments are implemented in PyTorch 2.6.0 using four NVIDIA RTX A6000 GPUs. During joint training, we employ uniform task sampling and optimize the model for $10,000$ iterations using the AdamW optimizer [40] (batch size 16, base learning rate 1×10^{-4} with $1,000$ warmup steps). The LoRA rank is fixed at 128. Input resolutions are set to 512×384 for virtual try-on and garment reconstruction, and 512×352 for pose transfer. For the dynamic diagonal Fisher Information Matrix estimation in FNG, the EMA momentum coefficient β is set to 0.99. During inference, we use 50 sampling steps with an FNG scale of 2.0 for VTON and 1.5 for the remaining tasks.

Datasets. We jointly train our model on VITON-HD [12] and DeepFashion [39]. For virtual try-on and garment reconstruction, following Any2AnyTryon [20], we construct a shared subset from VITON-HD comprising 11,647 training and 2,032 testing quadruplets, from which the requisite condition-target tuples for each specific task are seamlessly extracted. For pose transfer, following the protocol in [69], we partition DeepFashion into 101,966 training and 8,570 testing

Table 1: Quantitative comparison across virtual try-on, garment reconstruction, and pose transfer. Each task is evaluated using four core metrics. The best and second-best results are highlighted in **bold** and underline, respectively. Missing values for single-task experts are denoted with -.

Method	Capability	Virtual Try-On				Garment Recon.				Pose Transfer			
		LPIPS↓	SSIM↑	FID↓	KID↓	LPIPS↓	DISTS↓	FID↓	CLIP-I↑	LPIPS↓	SSIM↑	FID↓	CLIP-I↑
GP-VTON [56]	VTON Only	0.068	0.872	11.708	3.990	-	-	-	-	-	-	-	-
OOTDiffusion [57]	VTON Only	0.132	0.784	15.136	5.774	-	-	-	-	-	-	-	-
IDM-VTON [13]	VTON Only	0.082	0.816	10.745	2.229	-	-	-	-	-	-	-	-
CatVTON [14]	VTON Only	**0.057**	0.870	9.015	1.091	-	-	-	-	-	-	-	-
FitDiT [27]	VTON Only	0.106	0.830	10.340	1.648	-	-	-	-	-	-	-	-
OmniVTON [60]	VTON Only	0.145	0.832	9.621	1.323	-	-	-	-	-	-	-	-
TryOffDiff [52]	Recon. Only	-	-	-	-	0.212	0.227	15.273	0.884	-	-	-	-
TryOffAnyone [55]	Recon. Only	-	-	-	-	0.245	0.211	11.488	0.904	-	-	-	-
CoCosNet-v2 [68]	Pose Only	-	-	-	-	-	-	-	-	0.227	0.724	13.325	-
NTED [46]	Pose Only	-	-	-	-	-	-	-	-	0.200	0.736	7.645	0.916
PoCoLD [21]	Pose Only	-	-	-	-	-	-	-	-	0.192	0.743	8.416	-
CFLD [41]	Pose Only	-	-	-	-	-	-	-	-	0.182	0.748	7.149	0.922
MCLD [38]	Pose Only	-	-	-	-	-	-	-	-	0.176	**0.756**	7.079	0.926
AnyDoor [7]	Unified	0.113	0.808	13.403	5.793	0.474	0.340	65.130	0.781	0.636	0.346	61.643	0.716
Any2AnyTryon [20]	Unified	0.077	0.846	10.143	2.668	0.250	0.218	10.771	0.907	0.175	0.705	12.250	0.916
FLUX.2-klein [32]	Unified	0.114	0.836	14.651	7.126	0.356	0.272	30.491	0.864	0.256	0.634	8.598	0.881
LongCat-Image-Edit [51]	Unified	0.101	0.840	10.587	1.528	0.292	0.237	16.157	0.880	0.205	0.694	18.703	0.899
OrthoTryOn (Ours)	Unified	0.064	**0.876**	**8.312**	**0.532**	**0.192**	**0.191**	**9.563**	**0.931**	**0.146**	0.728	**6.364**	**0.936**

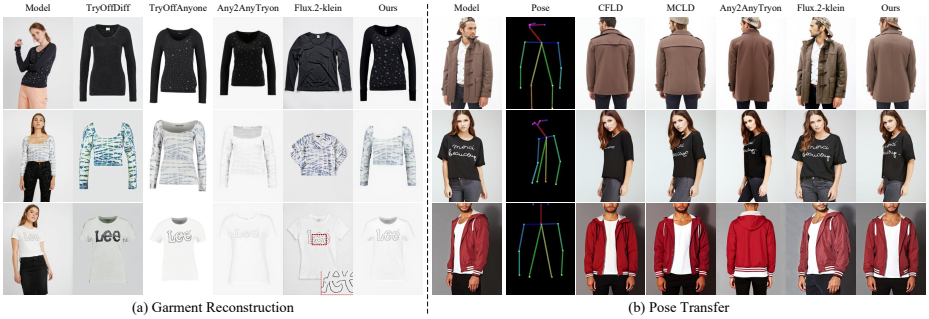


Fig. 5: Qualitative comparison of garment reconstruction on VITON-HD dataset [12] and pose transfer on DeepFashion dataset [39]. Please zoom in for better view.

pairs depicting the same identity under different poses. Text conditions are generated via Qwen2.5-VL-7B-Instruct [3], and human skeletons are extracted using HRNet [50].

4.2 Comparison with State-of-the-Art Methods

To evaluate OrthoTryOn, we benchmark against four unified foundation models (AnyDoor [7], Any2AnyTryon [20], LongCat-Image-Edit, and FLUX.2-klein [32]) alongside numerous task-specific expert models. While the first three unified baselines are retrained on our multi-task dataset using their official implementations, we directly employ the pre-trained weights of FLUX.2-klein for zero-shot evaluation, taking full advantage of its inherent multi-conditional image generation capabilities without additional fine-tuning.

Virtual Try-On. We employ LPIPS [63] and SSIM [53] to evaluate perceptual quality and structural consistency, alongside FID [44] and KID [6] under an unpaired setting to simulate real-world scenarios. Beyond general baselines, we compare OrthoTryOn against six task-specific SOTA experts: GP-VTON [56], OOTDiffusion [57], IDM-VTON [13], CatVTON [14], FitDiT [27], and OmniV-TON [60]. Quantitative results in Tab. 1 indicate that OrthoTryOn significantly outperforms all general baselines. Notably, despite being a unified framework, it even surpasses the majority of single-task experts. This superiority is largely attributed to OSP's effective disentanglement: by suppressing inter-task gradient conflicts at a rate that scales inversely with the bottleneck dimension, our model mitigates destructive interference during joint optimization, enabling stable and efficient utilization of large-scale multi-task data. Visual comparisons in Fig. 4 demonstrate our method's exceptional garment fidelity, effectively preventing texture distortions and reducing residual artifacts from the original clothing. By operating under a shared parameter budget, OrthoTryOn highlights that structured parameter geometry can prevent the capacity dilution typically observed in naive multi-task learning.

Garment Reconstruction. We evaluate garment reconstruction using FID, LPIPS, CLIP-I [45], and DISTS [15]. As detailed in Tab. 1, compared to general baselines and two task-specific experts (TryOffDiff [52], TryOffAnyone [55]), OrthoTryOn achieves superior performance across all metrics. Notably, it excels in LPIPS and DISTS, indicating a highly accurate preservation of both global perceptual realism and local structural details. Despite the semantic gap between object-level garment reconstruction and human-centric generation, OSP successfully shields garment texture features from inter-task interference during joint optimization. As depicted in Fig. 5 (a), unlike baselines that struggle with severe texture distortions and incorrect garment categorization, our method delivers accurate reconstruction, strictly preserving intricate textural patterns and complex topological structures.

Pose Transfer. For pose transfer, we employ FID, LPIPS, SSIM, and CLIP-I. We compare against five task-specific SOTAs: CoCosNet-v2 [68] and PoCoLD [21] (reporting official results), alongside NTED [46], CFLD [41], and MCLD [38] (evaluated using generated images released by the authors). According to the quantitative benchmarks, OrthoTryOn significantly outperforms all general baselines and surpasses experts across most metrics. Despite a slight SSIM trade-off (likely attributable to the use of dense spatial priors like DensePose [19] in some expert models, whereas our framework relies on sparse skeletal conditions), our method achieves a 0.715 FID improvement, indicating better alignment with real-world distributions. As shown in Fig. 5 (b), OrthoTryOn infers high-fidelity target views from single-view inputs and strictly maintains fine-grained garment textures without artifacts or blurring, even under drastic pose variations.

4.3 Ablation Study

We conduct comprehensive ablations to validate our core components. Quantitative and qualitative results are summarized in Tab. 2 and Fig. 6.

Table 2: Ablation study across three fashion generation tasks. Base (Joint-Learning): naive multi-task learning using a single shared LoRA. Base (Task-Specific): trained exclusively on individual tasks. OSP-R: inserting non-orthogonal random matrices in the LoRA bottleneck. FNG*: guidance using the task with the lowest Fisher similarity.

Setting	Virtual Try-On				Garment Recon.				Pose Transfer			
	LPIPS↓	SSIM↑	FID↓	KID↓	LPIPS↓	DISTS↓	FID↓	CLIP-I↑	LPIPS↓	SSIM↑	FID↓	CLIP-I↑
(a) Base (Joint-Learning)	0.101	0.840	10.587	1.528	0.292	0.237	16.157	0.880	0.205	0.694	18.703	0.899
(b) Base (Task-Specific)	0.095	0.840	9.487	0.941	0.211	**0.190**	10.428	0.925	0.188	0.685	7.445	0.926
(c) Base + OSP-R	0.114	0.812	9.318	1.014	0.491	0.347	20.412	0.790	0.388	0.602	15.423	0.873
(d) Base + OSP	0.084	0.867	8.386	0.576	0.203	0.194	9.787	0.929	0.161	0.716	6.632	0.928
(e) Base + OSP + FNG*	**0.064**	**0.876**	8.387	0.640	0.194	0.193	9.871	0.930	0.154	0.723	7.318	0.930
OrthoTryOn (Ours)	**0.064**	**0.876**	**8.312**	**0.532**	**0.192**	0.191	**9.563**	**0.931**	**0.146**	**0.728**	**6.364**	**0.936**

Table 3: Quantitative evaluation of cross-architecture generalizability. For Any2AnyTryon, we only apply OSP without FNG, since its distilled FLUX.1-dev backbone already incorporates strong built-in CFG priors.

Setting	Virtual Try-On				Garment Recon.				Pose Transfer			
	LPIPS↓	SSIM↑	FID↓	KID↓	LPIPS↓	DISTS↓	FID↓	CLIP-I↑	LPIPS↓	SSIM↑	FID↓	CLIP-I↑
Any2AnyTryon [20]	0.077	0.846	10.143	2.668	0.250	0.218	10.771	0.907	0.175	0.705	12.250	0.916
Any2AnyTryon + OSP	**0.058**	**0.868**	**9.180**	**1.800**	**0.224**	**0.207**	**10.127**	**0.913**	**0.170**	**0.707**	**11.609**	**0.916**
AnyDoor [7]	**0.113**	0.808	13.403	5.793	0.474	0.340	65.130	0.781	0.636	0.346	61.643	0.716
AnyDoor + OSP	0.114	0.808	13.249	5.768	0.373	0.278	21.604	0.852	0.593	0.359	52.489	0.772
AnyDoor + OSP + FNG	**0.113**	**0.811**	**12.806**	**5.369**	**0.346**	**0.263**	**18.810**	**0.860**	**0.592**	**0.364**	**48.375**	**0.783**

Effectiveness of OSP. Naive joint learning (variant (a)) suffers from severe artifacts and texture blur across all tasks, underperforming even the task-specific expert models (variant (b)). This degradation primarily stems from gradient conflicts and negative transfer among heterogeneous tasks. By introducing OSP (variant (d)), the model overcomes this bottleneck, outperforming both naive joint learning and expert models across most metrics. This superiority arises because OSP assigns each task a distinct low-rank coordinate frame while preserving feature norms, thereby reducing destructive cross-task interference yet retaining shared human and physical priors under stable joint optimization. To further verify that orthogonality is the key to stable decoupling, we replace Q_i with non-orthogonal random matrices R_i (variant (c)). This variant severely impairs generative capability: compared to variant (d), its FID scores on garment reconstruction and pose transfer deteriorate drastically by 10.625 and 8.791, respectively. Without the isometric constraint, random mixing introduces uncontrolled spectral scaling and anisotropic distortion in the bottleneck, which destabilizes optimization and amplifies negative transfer, ultimately leading to significant performance degradation.

Effectiveness of FNG. Although OSP effectively mitigates interference during joint learning, residual semantic coupling may persist under highly compressed low-rank bottlenecks, which can trigger semantic leakage during inference. As highlighted by the red boxes in Fig. 6, variant (d) relying solely on OSP still exhibits visual artifacts when generating local details. To address this, we intro-

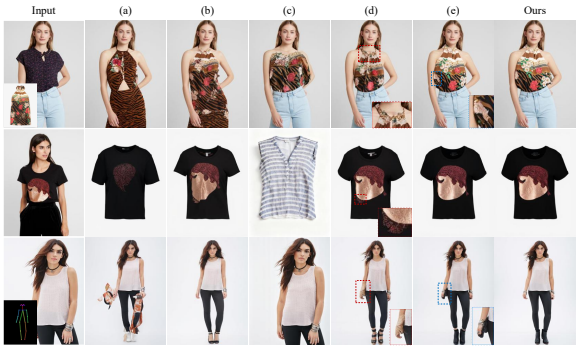


Fig. 6: Qualitative ablation study on different variants. Rows from top to bottom: virtual try-on, garment reconstruction, and pose transfer.

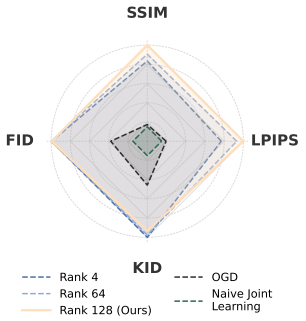


Fig. 7: Performance comparison of different virtual try-on variants across multiple metrics.

duce Fisher-guided Negative Guidance (FNG). We first evaluate a suboptimal variant FNG* (variant (e)), which forces the task with the lowest Fisher similarity (*i.e.*, maximum semantic discrepancy) as the negative prompt. As shown in Tab. 2, while this strategy enhances structural alignment and eliminates the red-box artifacts, it introduces new visual flaws (blue boxes) and suffers from FID degradation. This likely occurs because the most distant task has minimal distributional overlap with the target task; thus, forced repulsion fails to accurately isolate genuine interfering features and instead injects biased perturbations. In contrast, OrthoTryOn selects the task with the highest Fisher similarity as the hard negative task. Utilizing the fully converged Fisher statistics tracked via EMA during training, this operation precisely targets highly confusable semantics. The final results indicate that the complete FNG module eradicates all visual flaws and achieves state-of-the-art performance across all metrics (*e.g.*, boosting pose transfer FID to 6.364), demonstrating the necessity and effectiveness of repelling highly correlated tasks based on parameter sensitivity.

Robustness of Orthogonal Decoupling. To validate the robustness of OSP, we evaluate OrthoTryOn across varying LoRA ranks $r \in \{4, 64, 128\}$. As depicted in Fig. 7, while reducing the rank to 4 yields a marginal decline in SSIM and LPIPS due to constrained parameter capacity, the overall generative fidelity (FID and KID) remains highly stable and significantly outperforms the naive joint learning baseline. This indicates that orthogonal coordinate frames remain beneficial even under extreme low-rank constraints. Furthermore, we benchmark against Orthogonal Gradient Descent (OGD), a representative post-hoc gradient projection method. Although OGD partially mitigates negative transfer compared to naive joint learning, discarding conflicting gradient components may hinder convergence by removing potentially useful optimization signals. Consequently, it is clearly worse than OrthoTryOn across all evaluation metrics.

4.4 Cross-Architecture Generalizability

To validate architectural generalizability, we adapt OrthoTryOn to Any2AnyTryon (FLUX.1-dev [31]) and AnyDoor (Stable Diffusion 2.1 [47]). As shown in Tab. 3, our components yield consistent gains across distinct generative paradigms.

For Any2AnyTryon, integrating OSP yields consistent improvements across all three evaluated tasks. In the Virtual Try-On setting, OSP significantly enhances fine-grained detail preservation and overall image realism, evidenced by a marked drop in LPIPS from 0.077 to 0.058 and a reduction in FID from 10.143 to 9.180. Similar gains are observed in Garment Reconstruction and Pose Transfer (*e.g.*, Garment Recon. CLIP-I increases to 0.913 and Pose Transfer FID decreases to 11.609), indicating effective mitigation of inter-task gradient interference within the shared LoRA space. As noted, FNG is omitted in this setup due to its partial functional overlap with the distilled FLUX.1-dev backbone's built-in CFG priors. Nevertheless, OSP alone effectively stabilizes multi-task optimization by structurally reducing gradient conflicts in the shared low-rank space.

For AnyDoor, both OSP and FNG are seamlessly integrated. OrthoTryOn enables a substantial improvement in garment reconstruction, with FID dropping from 65.130 to 18.810. While the improvements in VTON are relatively modest, this is primarily because AnyDoor's pre-training already incorporates the VITON-HD dataset, leaving limited space for further optimization. Furthermore, although the absolute metrics for pose transfer remain suboptimal due to the inherent limitations of AnyDoor's local inpainting paradigm in handling large-scale spatial deformations, OrthoTryOn still achieves a substantial improvement in relative performance. This demonstrates that our framework effectively mitigates inter-task interference, allowing the backbone to better realize its multi-task capacity.

5 Conclusion

In this paper, we present OrthoTryOn, a highly effective framework for unified fashion generation that overcomes the negative transfer inherent in shared LoRA adaptation. By introducing Orthogonal Subspace Projection (OSP), we structurally enforce decorrelated coordinate frames via task-specific orthogonal rotations, which significantly suppresses gradient interference and enables stable joint optimization in expectation. Complementarily, Fisher-guided Negative Guidance (FNG) leverages empirical parameter sensitivities to explicitly mitigate residual semantic leakage during inference, without introducing any trainable parameters. Extensive experiments demonstrate that OrthoTryOn not only avoids the performance degradation typically observed in unified training, but also surpasses independently trained task-specific models, highlighting the importance of structured parameter geometry in unlocking effective multi-task generation. Moreover, OrthoTryOn generalizes robustly across diverse diffusion backbones, establishing itself as a universal plug-and-play adaptation mechanism.

Limitation. The $\mathcal{O}(1/r)$ interference bound of OSP implies that, under extremely small LoRA ranks coupled with many heterogeneous tasks, residual gradient coupling may become non-negligible. Although FNG partially compensates at inference time, it cannot fully recover training-stage information loss. In practice, moderately increasing the rank provides sufficient degrees of freedom to achieve better performance, as validated by our rank ablation study.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. U24A20330, 62361166670, and 62406135; the Natural Science Foundation of Jiangsu Province under Grant No. BK20241198; and the Gusu Innovation and Entrepreneur Leading Talents Program under Grant No. ZXL2024362.

References

1. An, X., Zhao, L., Gong, C., Wang, N., Wang, D., Yang, J.: Sharpose: Sparse high-resolution representation for human pose estimation. In: AAAI. vol. 38, pp. 691–699 (2024)
2. Arjovsky, M., Shah, A., Bengio, Y.: Unitary evolution recurrent neural networks. In: ICML. pp. 1120–1128. PMLR (2016)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Bao, J., Qian, J., Yan, M., Yang, J.: Progressive representation learning for multimodal sentiment analysis with incomplete modalities. arXiv preprint arXiv:2603.09111 (2026)
5. Bi, J., Wu, Q., Qian, J., Luo, L., Yang, J.: Dual manifold regularization steered robust representation learning for point cloud analysis. In: AAAI. vol. 39, pp. 1844–1852 (2025)
6. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. In: ICLR (2018)
7. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. In: CVPR. pp. 6593–6602 (2024)
8. Chen, Z., Gao, R., Xiang, T.Z., Lin, F.: Diffusion model for camouflaged object detection. In: ECAI (2023)
9. Chen, Z., Li, Y., Wang, H., Chen, Z., Jiang, Z., Li, J., Wang, Q., Yang, J., Tai, Y.: Ragd: Regional-aware diffusion model for text-to-image generation. In: ICCV. pp. 19331–19341 (2025)
10. Chen, Z., Zhu, J., Chen, X., Zhang, J., Chen, J., Zeng, Z., Zhang, W., Wang, C., Yang, J., Tai, Y.: L2p: Unlocking latent potential for pixel generation. arXiv preprint arXiv:2605.12013 (2026)
11. Chen, Z., Zhu, J., Chen, X., Zhang, J., Hu, X., Zhao, H., Wang, C., Yang, J., Tai, Y.: Dip: Taming diffusion models in pixel space. In: CVPR. pp. 36136–36146 (2026)
12. Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: CVPR. pp. 14131–14140 (2021)

13. Choi, Y., Kwak, S., Lee, K., Choi, H., Shin, J.: Improving diffusion models for authentic virtual try-on in the wild. In: ECCV. pp. 206–235. Springer (2024)
14. Chong, Z., Dong, X., Li, H., Zhang, S., Zhang, W., Zhang, X., Zhao, H., Jiang, D., Liang, X.: Catvton: Concatenation is all you need for virtual try-on with diffusion models. In: ICLR (2025)
15. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE TPAMI **44**(5), 2567–2581 (2020)
16. Ding, T., Yang, H., Shi, L., Li, J., Hu, X., Yang, J., Tai, Y.: Adaptive depth lightweight rgb-t tracking with holistic token routing. In: CVPR. pp. 20942–20952 (2026)
17. Du, Y., Zhan, J., Li, X., Dong, J., Chen, S., Yang, M.H., He, S.: One-for-all: towards universal domain translation with a single stylegan. IEEE TPAMI **47**(4), 2865–2881 (2025)
18. Farajtabar, M., Azizan, N., Mott, A., Li, A.: Orthogonal gradient descent for continual learning. In: AISTATS. pp. 3762–3773. PMLR (2020)
19. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: CVPR. pp. 7297–7306 (2018)
20. Guo, H., Zeng, B., Song, Y., Zhang, W., Zhang, C., Liu, J.: Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. In: ICCV. pp. 19085–19096 (2025)
21. Han, X., Zhu, X., Deng, J., Song, Y.Z., Xiang, T.: Controllable person image synthesis with pose-constrained latent diffusion. In: ICCV. pp. 22768–22777 (2023)
22. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: CVPR. pp. 7543–7552 (2018)
23. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
24. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: ICML. pp. 2790–2799. PMLR (2019)
25. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)
26. Hu, X., Tai, Y., Zhao, X., Zhao, C., Zhang, Z., Li, J., Zhong, B., Yang, J.: Exploiting multimodal spatial-temporal patterns for video object tracking. In: AAAI. vol. 39, pp. 3581–3589 (2025)
27. Jiang, B., Hu, X., Luo, D., He, Q., Xu, C., Peng, J., Zhang, J., Wang, C., Wu, Y., Fu, Y.: Fitdit: Advancing the authentic garment details for high-fidelity virtual try-on. arXiv preprint arXiv:2411.10499 (2024)
28. Jin, J., Shen, Y., Fu, Z., Yang, J.: Customized generation reimagined: Fidelity and editability harmonized. In: ECCV. pp. 410–426. Springer (2024)
29. Jin, J., Yu, Z., Shen, Y., Fu, Z., Yang, J.: Latexblend: Scaling multi-concept customized generation with latent textual blending. In: CVPR. pp. 23585–23594 (2025)
30. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. PNAS **114**(13), 3521–3526 (2017)
31. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
32. Labs, B.F.: Flux.2. `https://github.com/black-forest-labs/flux2` (2025)
33. Lee, S., Gu, G., Park, S., Choi, S., Choo, J.: High-resolution virtual try-on with misalignment and occlusion-handled conditions. In: ECCV. pp. 204–219. Springer (2022)

34. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: EMNLP. pp. 3045–3059 (2021)
35. Lezcano-Casado, M., Martınez-Rubio, D.: Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In: ICML. pp. 3794–3803. PMLR (2019)
36. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: ACL. pp. 4582–4597 (2021)
37. Li, X., Zhan, J., He, S., Xu, Y., Dong, J., Zhang, H., Du, Y.: Personamagic: Stage-regulated high-fidelity face customization with tandem equilibrium. In: AAAI. vol. 39, pp. 4995–5003 (2025)
38. Liu, J., Zhang, J., Rota, P., Sebe, N.: Multi-focal conditioned latent diffusion for person image synthesis. In: CVPR. pp. 16019–16028 (2025)
39. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: CVPR. pp. 1096–1104 (2016)
40. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
41. Lu, Y., Zhang, M., Ma, A.J., Xie, X., Lai, J.: Coarse-to-fine latent diffusion for pose-guided person image synthesis. In: CVPR. pp. 6420–6429 (2024)
42. Mishkin, D., Matas, J.: All you need is a good init. In: ICLR (2016)
43. Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., Li, X., Yang, J., Tai, Y.: Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In: ICLR. vol. 2025, pp. 1045–1064 (2025)
44. Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation. In: CVPR. pp. 11410–11420 (2022)
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
46. Ren, Y., Fan, X., Li, G., Liu, S., Li, T.H.: Neural texture extraction and distribution for controllable person image synthesis. In: CVPR. pp. 13535–13544 (2022)
47. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
48. Saxe, A.M., McClelland, J.L., Ganguli, S.: Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In: ICLR (2014)
49. Schwarz, J., Czarnecki, W., Luketina, J., Grabska-Barwinska, A., Teh, Y.W., Pascanu, R., Hadsell, R.: Progress & compress: A scalable framework for continual learning. In: ICML. pp. 4528–4537. PMLR (2018)
50. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR. pp. 5693–5703 (2019)
51. Team, M.L., Ma, H., Tan, H., Huang, J., Wu, J., He, J.Y., Gao, L., Xiao, S., Wei, X., Ma, X., et al.: Longcat-image technical report. arXiv preprint arXiv:2512.07584 (2025)
52. Velioglu, R., Bevandic, P., Chan, R., Hammer, B.: Tryoffdiff: Virtual-try-off via high-fidelity garment reconstruction using diffusion models. In: BMVC (2025)
53. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004)
54. Wisdom, S., Powers, T., Hershey, J., Le Roux, J., Atlas, L.: Full-capacity unitary recurrent neural networks. NeurIPS **29** (2016)
55. Xarchakos, I., Koukopoulos, T.: Tryoffanyone: Tiled cloth generation from a dressed person. arXiv preprint arXiv:2412.08573 (2024)
56. Xie, Z., Huang, Z., Dong, X., Zhao, F., Dong, H., Zhang, X., Zhu, F., Liang, X.: Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In: CVPR. pp. 23550–23559 (2023)

57. Xu, Y., Gu, T., Chen, W., Chen, A.: Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In: AAAI. vol. 39, pp. 8996–9004 (2025)
58. Yang, Z., Chen, Y., Li, Y., He, S., Xu, Y., Dong, J., Yang, J., Du, Y.: Deshadow-mamba: Deshadowing as 1d sequential similarity. arXiv preprint arXiv:2510.24260 (2025)
59. Yang, Z., Jiang, Z., Li, X., Zhou, H., Dong, J., Zhang, H., Du, Y.: D^4-vton: Dynamic semantics disentangling for differential diffusion based virtual try-on. In: ECCV. pp. 36–52. Springer (2024)
60. Yang, Z., Li, Y., He, S., Li, X., Xu, Y., Dong, J., Du, Y.: Omnivton: Training-free universal virtual try-on. In: ICCV. pp. 16702–16711 (2025)
61. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. NeurIPS **33**, 5824–5836 (2020)
62. Zhan, J., Sun, X., Zhu, X., Ji, Y., Liu, R., Zhang, L., Zhang, J.: Towards source-aware object swapping with initial noise perturbation. In: CVPR. pp. 4400–4409 (2026)
63. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)
64. Zhang, W., Jin, Y., Li, X., Zhang, Y., Cong, X., Wang, C., Qiao, F., et al.: Unifit: Towards universal virtual try-on with mllm-guided semantic alignment. In: AAAI (2026)
65. Zheng, Y., Fang, W., Hu, X., Fan, J., Weng, J., Li, J., Zhang, K., Yang, J.: Dvdpec: Driving-video dehazing via position embedding-based codebook. IEEE TCSVT (2026)
66. Zheng, Y., Zhan, J., He, S., Dong, J., Du, Y.: Curricular contrastive regularization for physics-aware single image dehazing. In: CVPR. pp. 5785–5794 (2023)
67. Zheng, Y., Zhang, K., Zhu, W., Liu, Q., Hu, X., Li, J., Yang, J.: Dvar: Dynamic visual autoregressive modeling for image super-resolution. In: CVPR. pp. 23378–23387 (2026)
68. Zhou, X., Zhang, B., Zhang, T., Zhang, P., Bao, J., Chen, D., Zhang, Z., Wen, F.: Cocosnet v2: Full-resolution correspondence learning for image translation. In: CVPR. pp. 11465–11475 (2021)
69. Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: Progressive pose attention transfer for person image generation. In: CVPR. pp. 2347–2356 (2019)