SYMFASHION: LEARNING CROSS-TASK SYMMETRY FOR UNIFIED FASHION GENERATION

Anonymous authors
Paper under double-blind review

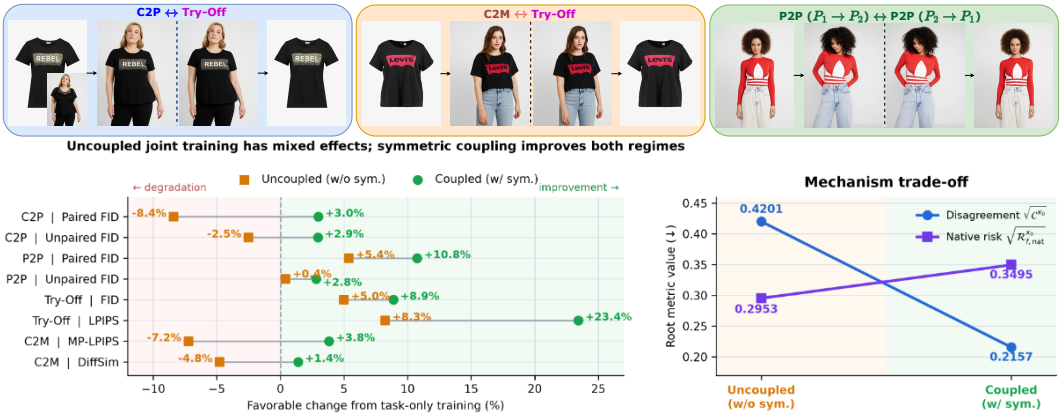

Figure 1: SymFashion couples C2P–Try-Off, C2M–Try-Off, and reverse P2P directions in one checkpoint. Uncoupled joint training has mixed effects; coupling improves all shown readouts and reduces cross-view disagreement despite higher native-target risk.

ABSTRACT

Can compatible forward and reverse fashion transformations provide useful supervision for one another, or does joint training merely share parameters? We address this question by proposing **SymFashion**, a unified fashion-generation model built on a single rectified-flow transformer that supports cloth-to-person try-on (C2P), person-to-person (P2P) garment transfer, virtual Try-Off, and pose-controllable cloth-to-model (C2M) generation. SymFashion maps compatible directions to a shared human–garment state and explicitly couples their role-aligned clean endpoints at the same timestep using permutation-aligned noise. Accounting for the latent mismatch introduced by the two encoding paths, our analysis recovers σ^2-weighted velocity agreement under exact latent alignment and supplies a descriptive cross-view root-risk bound. On VITON-HD and DressCode, a single checkpoint remains competitive across all four task families. Under a fixed total optimization budget, uncoupled joint training has mixed effects on the primary task-level readouts: it degrades C2P and C2M but improves P2P and Try-Off, whereas symmetric coupling improves both regimes. Direct diagnostics reveal substantially lower cross-view disagreement, especially at higher noise levels and in person-side garment regions, despite higher branchwise prediction risk. Thus, compatible transformations can support one another, but the benefit is more consistent across our evaluations when their shared semantic structure is explicitly coupled rather than left to uncoupled joint training.

1 INTRODUCTION

Putting related fashion transformations in one model does not, by itself, make them learn from one another. Cloth-to-person (C2P) virtual try-on dresses a target person while preserving identity and garment appearance (Cordier et al., 2001; Song et al., 2025a); diffusion models have improved its visual quality (Morelli et al., 2023; Choi et al., 2024; Hong et al., 2025; Kim et al., 2025). Broader fashion transformations include person-to-person garment transfer (P2P) (Raj et al., 2018; Zhu et al.,

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

2023; Chong et al., 2025), canonical garment extraction (Try-Off) (Velioglu et al., 2025b; Lobba et al., 2025; Velioglu et al., 2025a), and cloth-to-model (C2M) generation, i.e., garment-conditioned person generation optionally with pose control (Chen et al., 2024; Wang et al., 2024; Li et al., 2025).

Unified fashion models place multiple transformations in one architecture (Guo et al., 2025; Zhang et al., 2026; Yang et al., 2026), while bidirectional systems connect Try-On and Try-Off through shared objectives or sequential cycles (Lee & Kwak, 2025; Wu et al., 2026). Beyond consolidating task-specific systems, unification creates a shared optimization space for exploiting common human–garment structure, but does not make compatible predictions agree. We call two directions semantically compatible when paired views of the same underlying human–garment example admit a known role permutation that maps their complete target states to common person–garment coordinates. This raises our central question: can compatible directions supervise one another?

Compatible directions observe the same paired human–garment state through different conditions. Garment appearance links these views, while the complete person–garment state supplies the common target needed to compare their predictions. Joint training leaves this relation unused. Indeed, under a fixed total optimization budget, our uncoupled joint-training control improves the primary P2P and Try-Off readouts but degrades C2P and C2M relative to task-only training (Fig. 1, bottom-left plot).

We propose **SymFashion**, a unified fashion-generation model built on one shared rectified-flow transformer. Predefined task indicators and structured visual roles require neither user descriptions nor MLLM routing. C2P, P2P, C2M, and Try-Off share a two-slot latent canvas; compatible views use the same timestep and permutation-aligned noise. Their induced clean endpoints are returned to a common person–garment order and linked by role-aligned clean-endpoint consistency, hereafter *symmetric coupling*. This prediction-level interaction distinguishes SymFashion from ordinary sharing; symmetry means semantic-role consistency, not image-space inversion.

We derive a residual-aware clean-latent–velocity relation that becomes velocity disagreement weighted by the squared noise level under exact alignment. A cross-view root-risk bound separates estimation error from task disagreement without guaranteeing image quality.

The experiments ask whether one checkpoint competitively supports all four tasks (**RQ1**), whether coupling improves matched uncoupled training (**RQ2**), and what latent behavior it induces (**RQ3**).

Our contributions are:

- We formulate semantic compatibility between fashion transformations through a shared human–garment state and a known role permutation, and propose role-aligned clean-endpoint consistency to couple their garment-mask-weighted endpoint estimates.
- Across matched task configurations, symmetric coupling improves or ties every main-table task–dataset point estimate against its uncoupled counterpart.
- We analytically characterize the objective through a residual-aware clean-latent–velocity relation and a cross-view root-risk upper bound, then separate stronger cross-view coherence from branchwise prediction accuracy.
- We instantiate this principle in one rectified-flow transformer supporting all four tasks and characterize the coupling effect across datasets, noise levels, and spatial regions.

2 Related Work

Much of prior fashion-generation research improves fidelity within individual tasks or expands the set of supported transformations. SymFashion studies a complementary question: how semantically compatible directions should interact during training.

2.1 Virtual Try-On and Garment Transfer

Early virtual try-on methods separate geometric garment alignment from image synthesis (Han et al., 2018; Wang et al., 2018). Diffusion-based methods improve this pipeline through richer person–garment conditioning and feature integration (Morelli et al., 2023; Kim et al., 2024; Choi et al., 2024; Zeng et al., 2024; Chong et al., 2025; Kim et al., 2025; Hong et al., 2025). P2P considers the setting

in which the desired garment is observed on another person rather than provided as a canonical product image (Raj et al., 2018; Zhu et al., 2023), while compositional and multi-reference methods broaden the forward interface (Ning et al., 2024; Zhang et al., 2024). These works focus on forward person generation rather than coupling clean-endpoint estimates across directions.

2.2 Try-Off and Garment-Driven Human Generation

Try-Off infers a canonical garment from a dressed person (Velioglu et al., 2025b; Lobba et al., 2025; Velioglu et al., 2025a), whereas garment-driven methods generate a person from a garment, optionally under pose or scene control (Chen et al., 2024; Shen et al., 2025; Wang et al., 2024; Li et al., 2025). These directions share garment semantics but are not exact inverses: Try-Off must infer occluded content, while C2M admits multiple valid people and poses. We therefore do not assume exact image-space invertibility between these directions. SymFashion instead aligns their predicted person and garment roles across compatible views.

2.3 Unified and Bidirectional Fashion Models

Recent methods increasingly consolidate fashion transformations in one model, but differ in how their tasks interact. Any2AnyTryon broadens supported clothing configurations (Guo et al., 2025). Voost jointly models Try-On and Try-Off and optionally connects them through inference-time self-correction (Lee & Kwak, 2025). UniFit uses an MLLM to align natural-language instructions and reference images (Zhang et al., 2026); SymFashion instead uses predefined task indicators and structured visual roles, requiring neither a user-authored description nor MLLM-based routing at inference. Other work broadens forward VTON through category, editing, or mask-free interfaces (Li et al., 2026b; Du et al., 2025a;b; Feng et al., 2026).

Beyond coverage, CCVTON couples Try-Off and Try-On through a sequential image-space training cycle (Wu et al., 2026), while OrthoTryOn mitigates negative transfer by mapping task-specific bottleneck features into decorrelated coordinate frames (Yang et al., 2026). At the time of our evaluation, public checkpoints for CCVTON and UniFit were unavailable for reevaluation in our pipeline, so we discuss them rather than include them as quantitative baselines. These methods address coverage, routing, sequential correction, or task-feature interference. SymFashion instead couples parallel clean-endpoint estimates after semantic-role alignment within the same training step; Appendix Fig. B provides a schematic comparison.

2.4 Multi-Task Learning, Consistency, and Bidirectional Diffusion

Multi-task learning can improve generalization through shared representations (Caruana, 1997), but sharing alone does not require task predictions to agree. Cycle-consistent translation couples mappings by requiring their sequential composition to recover the input (Zhu et al., 2017); dual learning similarly couples primal and reverse translation models (He et al., 2016). The Bidirectional Diffusion Bridge Model models coupled distributions bidirectionally in one network (Kieu et al., 2025).

SymFashion differs along three axes. Its conditional tasks need not be bijective; correspondence is defined through explicit person and garment roles; and consistency is imposed between parallel clean-endpoint estimates using a shared timestep and permutation-aligned noise. It therefore neither relies on parameter sharing alone nor passes a sampled intermediate through a sequential cycle.

3 Method

SymFashion lets compatible task directions supervise one another through a shared two-slot latent representation, one conditional rectified-flow transformer, and role-aligned clean-endpoint consistency. Fig. 2 summarizes two stages. Panel (A) places the conditions for C2P, P2P, C2M, and Try-Off on the same canvas. Condition Garment supplies either a product garment or a full person image; unused conditions are zero-filled. For C2M, Dense Person is randomly enabled or zeroed during training to support pose-controlled and unconstrained generation. Panel (B) pairs compatible C2P–Try-Off views from one human–garment example. They share a timestep and permutation-aligned noise, and their complete endpoint estimates are returned to common $[P \mid G]$ order before comparison. Thus, tasks interact through predictions, not only shared parameters.

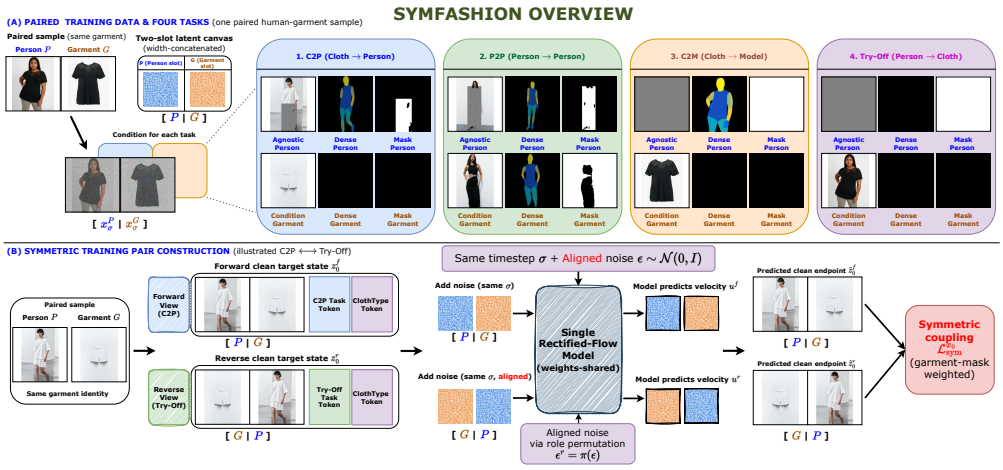

Figure 2: SymFashion training. (A) Each task constructs a conditioned two-slot canvas; Task and ClothType tokens specify the task and garment category. (B) Compatible views share a timestep and permutation-aligned noise; their induced clean endpoints are role-aligned before coupling.

3.1 A Shared Human–Garment State

A two-slot canvas is realized by two physical image tiles. A *semantic role* identifies the content assigned to a slot: P denotes the person-side role and G the garment-bearing role. A *physical tile position* is its left or right coordinate on the width-concatenated canvas. Depending on the task, either semantic role may be generated or provided as conditioning information, and G may correspond to a canonical product garment or to a reference person wearing the desired garment, as in P2P. Let π denote the known role permutation that reorders these semantic roles between physical tile positions when required; it does not select or change the task. Predictions are mapped back to a common P–G ordering before comparison. We arrange the corresponding VAE latents as

$$z_0 = [z_0^P \parallel z_0^G]. \tag{1}$$

Tab. 1 lists the compatible direction pairs used for symmetric training. Under this representation, C2P generates a dressed person from a product garment and a garment-agnostic person condition; P2P transfers the garment appearance from a reference person; Try-Off generates a canonical garment from a dressed person; and C2M generates a person from a garment with or without DensePose guidance (Güler et al., 2018). The four tasks are therefore represented as different conditional views over shared person and garment roles rather than as unrelated input–output mappings.

Table 1: Symmetric task pairs.

Forward Task	Reverse Task
C2P Try-On	Try-Off
C2M	Try-Off
P2P ($P_1 \rightarrow P_2$)	P2P ($P_2 \rightarrow P_1$)
Try-Off	C2P Try-On **or** C2M

Garment appearance links these task views, but comparing their complete predictions also requires paired targets in a common person–garment frame. We therefore call two directions compatible when, for each paired supervised example, a known role permutation maps their complete target states to a common semantic person and garment reference. Compatibility is a manually specified design criterion, not a relation learned from unpaired outputs; it requires neither identical conditioning interfaces, bijective tasks, nor sequential reconstruction of generated samples. Sharing an output space, garment category, or semantic-role layout alone is insufficient: the paired views must refer to the same underlying human–garment example.

3.2 Unified Rectified-Flow Transformer

SymFashion uses a single conditional rectified-flow transformer. Using the labels in Fig. 2, Agnostic Person and Condition Garment form the observed two-slot condition; Mask Person and Mask Garment form its spatial-mask condition; and Dense Person and Dense Garment form its geometry condition. We write these three packed components together with the predefined task and garment-category indicators as c^d. They are combined with the noised two-slot state through one shared conditional interface. An inapplicable cell is represented by zeros, so all four tasks retain the same input structure. The task indicator is selected by the requested interface and does not require a free-form user description. We use $d \in \{f, r\}$ to index the compatible forward and reverse views and

$k \in \{P, G\}$ to index semantic roles. Superscripts on z_0^d, x_σ^d, ϵ^d, and u^d identify the corresponding view. For a clean canvas z_0, noise $\epsilon \sim \mathcal{N}(0, I)$, and $\sigma \in [0, 1]$, rectified-flow training uses

$$x_\sigma = (1 - \sigma)z_0 + \sigma\epsilon, \qquad v^* = \epsilon - z_0. \tag{2}$$

The network $u_\theta(x_\sigma, \sigma, c)$ predicts v^*, yielding the clean endpoint estimate

$$\hat{z}_0(x_\sigma, \sigma, c) = x_\sigma - \sigma u_\theta(x_\sigma, \sigma, c). \tag{3}$$

During training, each branch predicts both semantic roles, even when only one is the task-facing output. This gives the two views complete estimates of the paired human–garment state for comparison after role alignment. Because garment appearance is their shared semantic link, we retain full-canvas comparison but give garment regions greater influence through role-aligned spatial weights,

$$w^P = 1 + 3m^P, \qquad w^G = 1 + 3m^G, \tag{4}$$

where m^P and m^G denote the garment-region masks associated with the person-side and garment-bearing roles, respectively. Garment regions on person images are obtained from Sapiens human parsing (Khirodkar et al., 2024), whereas foreground masks for flat product garments are obtained with SAM (Kirillov et al., 2023). The masks are resized to latent resolution and clamped to $[0, 1]$. Thus, the garment contribution has a fixed coefficient of 3: the weight is 1 outside the garment and 4 inside a binary garment region. We fix this coefficient to 3 for all tasks and datasets.

3.3 Role-Aligned Clean-Endpoint Consistency

For a compatible forward–reverse pair with conditions c^f and c^r, we use the role permutation π defined above to align their semantic person and garment roles. We compare clean endpoints because they estimate the common semantic target state, while velocity agreement alone need not capture the implemented latent-construction difference. The forward canvas is encoded jointly after width concatenation, while the reverse canvas reorders the corresponding role latents. These two VAE construction paths need not commute exactly with role permutation. So, we define

$$\tilde{z}_0^r = \pi^{-1}(z_0^r), \qquad \delta_0 = z_0^f - \tilde{z}_0^r. \tag{5}$$

Equation 5 isolates δ_0, the deterministic mismatch induced by the implemented joint-canvas and role-wise VAE encoding paths after semantic realignment. It is not a separate model effect or a prediction error; retaining it simply makes the analysis match the implemented construction. From the same paired training state, both branches use the same timestep and a permutation-aligned copy of the same Gaussian noise:

$$\epsilon^r = \pi(\epsilon), \qquad x_\sigma^f - \pi^{-1}(x_\sigma^r) = (1 - \sigma)\delta_0. \tag{6}$$

Sharing the perturbation avoids adding independent-noise variation to the cross-view comparison, while δ_0 retains any clean-latent difference introduced by the two encoding paths. Let

$$\bar{\hat{z}}_0^r = \pi^{-1}(\hat{z}_0^r) \tag{7}$$

denote the reverse prediction mapped back to the forward semantic ordering. We then impose role-aligned clean-endpoint consistency:

$$\ell_{\text{sym}}^{x_0} = \frac{1}{2}\left[\|\hat{z}_0^f[P] - \bar{\hat{z}}_0^r[P]\|_{w^P}^2 + \|\hat{z}_0^f[G] - \bar{\hat{z}}_0^r[G]\|_{w^G}^2\right], \tag{8}$$

where $\|a\|_w^2 = |\Omega|^{-1}\sum_{i\in\Omega} w_i a_i^2$, with Ω denoting the spatial and channel elements of the corresponding latent role. The bracket notation $a[k]$ selects role k, while an overbar denotes a reverse-view quantity mapped back by π^{-1} to the forward semantic order. Both branches receive gradients from this term. Consequently, symmetric coupling directly links two role-aligned clean-endpoint estimates within the same training step, rather than feeding a generated intermediate through a second task. Each direction additionally receives its standard supervised flow-matching objective,

$$\mathcal{L}_{\text{flow}}^d = \mathbb{E}\left\|u_\theta(x_\sigma^d, \sigma, c^d) - (\epsilon^d - z_0^d)\right\|_2^2, \qquad d \in \{f, r\}, \tag{9}$$

The complete training objective is therefore

$$\mathcal{L} = \mathcal{L}_{\text{flow}}^f + \mathcal{L}_{\text{flow}}^r + \lambda_s \mathbb{E}[\ell_{\text{sym}}^{x_0}]. \tag{10}$$

Table 2: Representative C2P/P2P results; full results are in Appendix Tab. G. Paired P2P measures self-reference fidelity; unpaired FID/KID are the primary transfer comparison. KID is $\times 10^3$; blue parentheses show favorable coupling changes. †, *, and ^{R} denote retraining, reevaluation, and our Voost reproduction. **Bold** and underlined mark first and second overall; ties share a rank.

Setting	Method	VITON-HD						DressCode					
		Paired				Unpaired		Paired				Unpaired	
		LPIPS ↓	SSIM ↑	FID ↓	KID ↓	FID ↓	KID ↓	LPIPS ↓	SSIM ↑	FID ↓	KID ↓	FID ↓	KID ↓
Cloth-to-Person (C2P)													
Baselines	IDM-VTON* (Choi et al., 2024)	0.102	0.881	6.33	0.99	9.16	1.22	0.081	0.883	6.11	1.98	7.71	2.62
	CatVTON* (Chong et al., 2025)	0.108	0.834	8.85	0.41	11.36	1.09	0.083	0.875	5.12	1.02	6.54	1.83
	Any2Any* (Guo et al., 2025)	0.091	0.839	6.93	0.74	8.97	0.98	0.087	0.869	6.54	1.56	7.64	2.28
	OrthoTryOn* (Yang et al., 2026)	0.120	0.837	10.69	5.80	12.68	5.80	–	–	–	–	–	–
	Leffa* (Zhou et al., 2025)	0.081	0.857	6.31	1.21	9.03	0.69	0.073	0.877	4.81	1.22	6.61	1.78
	Voost^R (Lee & Kwak, 2025)	0.079	0.868	5.81	0.62	9.11	1.05	0.061	0.898	3.65	0.57	5.31	1.13
	OMFA* (Liu et al., 2025)	0.194	0.739	20.07	11.30	22.42	12.70	0.305	0.716	25.92	16.69	26.08	16.84
	RefTON* (Li et al., 2026a)	0.079	0.879	8.34	1.93	9.11	0.91	0.060	0.902	3.48	1.20	7.01	3.15
Full	Ours w/o sym.	**0.078**	0.881	6.19	0.61	9.39	1.10	0.062	0.901	2.64	0.23	4.93	1.00
	Ours w/ sym.	**0.078** (=)	**0.885** (+0.004)	**5.54** (+0.65)	**0.12** (+0.49)	**8.89** (+0.50)	0.65 (+0.45)	**0.054** (+0.008)	**0.907** (+0.006)	**2.41** (+0.23)	**0.11** (+0.12)	**4.51** (+0.42)	**0.51** (+0.49)
C2P + Try-Off	Ours w/o sym.	0.096	0.859	6.27	0.53	9.57	1.20	0.062	0.903	2.58	0.20	4.95	0.99
	Ours w/ sym.	0.092 (+0.004)	0.862 (+0.003)	5.68 (+0.59)	0.18 (+0.35)	8.95 (+0.62)	**0.50** (+0.70)	0.059 (+0.003)	0.905 (+0.002)	2.43 (+0.15)	0.12 (+0.08)	4.55 (+0.40)	0.70 (+0.29)
C2P Only	Ours	0.089	0.862	5.71	0.22	9.16	0.61	0.058	0.905	2.51	0.17	4.78	0.93
Person-to-Person (P2P)													
Baselines	StableVITON† (Kim et al., 2024)	0.098	0.876	7.01	0.92	10.35	1.96	–	–	–	–	–	–
	CAT-DM† (Zeng et al., 2024)	0.153	0.877	10.71	2.07	12.44	2.51	–	–	–	–	–	–
	ITA-MDT† (Hong et al., 2025)	0.078	0.881	8.02	3.88	11.91	4.61	0.061	0.913	3.31	0.74	5.41	1.34
	CatVTON* (Chong et al., 2025)	0.078	0.870	8.59	5.38	14.13	6.88	0.032	0.941	1.65	0.202	5.42	1.52
	OMFA* (Liu et al., 2025)	0.193	0.743	16.30	8.30	17.44	8.70	0.316	0.707	29.94	20.64	30.25	20.94
Full	Ours w/o sym.	0.014	0.974	2.11	0.50	9.60	1.31	0.014	0.975	0.96	0.087	4.59	0.85
	Ours w/ sym.	**0.012** (+0.002)	**0.984** (+0.010)	1.99 (+0.12)	0.20 (+0.30)	**9.37** (+0.23)	**1.27** (+0.04)	**0.011** (+0.003)	**0.988** (+0.013)	**0.90** (+0.06)	0.024 (+0.063)	**4.38** (+0.21)	**0.60** (+0.25)
P2P Only	Ours w/o sym.	0.014	0.979	2.23	0.56	9.64	1.42	0.016	0.976	1.05	0.089	4.74	0.90
	Ours w/ sym.	**0.012** (+0.002)	0.983 (+0.004)	**1.97** (+0.26)	**0.17** (+0.39)	9.63 (+0.01)	1.32 (+0.10)	**0.011** (+0.005)	0.984 (+0.008)	0.91 (+0.14)	**0.023** (+0.066)	4.59 (+0.15)	0.74 (+0.16)

$\lambda_s \geq 0$ is the global strength of cross-view coupling; setting $\lambda_s = 0$ retains both supervised directions but removes their explicit consistency constraint.

The two supervised flow terms anchor both views to their respective training targets, so agreement does not replace task supervision or introduce an additional ground-truth target. Instead, it regularizes finite-capacity joint optimization by coupling role-aligned endpoint estimates across compatible conditional views. Symmetric pairing is required only during training. At inference, SymFashion samples a single requested direction using its corresponding task conditions; the same C2M checkpoint operates with a supplied DensePose for target-pose control or with its geometry condition set to zero for unconstrained generation.

Analytical link. The clean-endpoint choice also determines how the two views are coupled across noise levels. Aligned noise implies $\hat{z}_0^f - \bar{\hat{z}}_0^r = (1-\sigma)\delta_0 - \sigma\Delta u$, where $\Delta u = u^f - \pi^{-1}(u^r)$. Under exact latent alignment ($\delta_0 = 0$), endpoint agreement reduces to σ^2-weighted velocity agreement; otherwise it additionally contains a residual and cross term. The full expansion and root-risk analysis are in Appendix G, while Tab. 4 compares the three prediction spaces. The analysis constrains cross-view geometry, not sampled-image quality, so lower disagreement need not imply lower single-view endpoint error; RQ3 tests this structured-bias interpretation.

4 Experiments

We answer the central question in three stages: RQ1 tests whether one checkpoint supports all compatible tasks, RQ2 isolates coupling from uncoupled joint training, and RQ3 probes the behavior coupling induces. Further evidence is organized by purpose in the appendix: Appendix A extends the mechanism diagnostics, while Appendices B and C examine the coupling components and reverse-canvas construction. Appendix D evaluates pose control, Appendix E collects supplemental comparisons, and Appendix F documents implementation and complexity. Full derivations, limitations, and representative failures follow in Appendices G–I.

4.1 Datasets, Settings, and Metrics

We evaluate the official VITON-HD and DressCode test splits under one pipeline. Matched variants fix the data, model, flow objectives, test IDs, and evaluator. Task composition and coupling are independent: *Full* trains all four compatible task families, while a task-pair setting, such as C2P + Try-Off, trains only the named compatible directions. Within either composition, *w/o sym.* and *w/*

Table 3: Try-Off/C2M results on VITON-HD and DressCode; full Try-Off metrics are in the appendix. KID is $\times 10^3$; * and R denote reevaluation and our Voost reproduction. Blue parentheses show favorable coupling changes. **Bold** and underlined mark the top two ranks; ties share rank.

(a) Virtual Try-Off

Setting	Method	VITON-HD				DressCode			
		LPIPS ↓	FID ↓	CLIP-FID ↓	KID ↓	LPIPS ↓	FID ↓	CLIP-FID ↓	KID ↓
Baselines	TryOffDiff* (Velioglu et al., 2025b)	0.413	21.08	6.78	7.46	–	–	–	–
	MGT* (Velioglu et al., 2025a)	0.471	24.39	7.31	10.40	0.305	11.35	6.33	3.40
	Inverse Virtual Try-On* (Lobba et al., 2025)	0.415	21.28	6.84	5.60	0.333	12.68	2.71	4.70
	Voost^R (Lee & Kwak, 2025)	0.310	10.06	1.64	2.48	0.295	5.34	0.99	0.52
	OMFA* (Liu et al., 2025)	0.299	10.18	3.15	2.53	0.438	30.09	5.87	13.79
	Any2Any* (Guo et al., 2025)	0.385	18.76	5.64	5.90	0.373	18.78	4.09	6.30
	OrthoTryOn* (Yang et al., 2026)	0.294	11.06	3.69	2.10	–	–	–	–
Full	Ours w/o sym.	0.278	8.55	1.49	0.93	0.267	5.02	0.828	0.49
	Ours w/ sym.	**0.232** (+0.046)	**8.20** (+0.35)	**1.38** (+0.11)	0.90 (+0.03)	0.248 (+0.019)	**4.41** (+0.61)	0.703 (+0.125)	0.30 (+0.19)
C2P + Try-Off	Ours w/o sym.	0.281	8.61	1.47	1.04	0.249	4.71	0.727	0.32
	Ours w/ sym.	0.263 (+0.018)	8.54 (+0.07)	1.41 (+0.06)	0.98 (+0.06)	**0.231** (+0.018)	4.46 (+0.25)	**0.701** (+0.026)	**0.27** (+0.05)
C2M + Try-Off	Ours w/o sym.	0.293	8.93	1.68	0.99	0.282	5.15	0.812	0.49
	Ours w/ sym.	0.269 (+0.024)	8.45 (+0.48)	1.51 (+0.17)	**0.84** (+0.15)	0.248 (+0.034)	4.74 (+0.41)	0.763 (+0.049)	0.41 (+0.08)
Try-Off Only	Ours	0.303	9.00	1.77	0.96	0.285	5.47	0.901	0.68

(b) C2M without target-pose conditioning

Setting	Method	VITON-HD				DressCode			
		MP-LPIPS ↓	CLIP-I ↑	DiffSim ↑	FFA ↑	MP-LPIPS ↓	CLIP-I ↑	DiffSim ↑	FFA ↑
External	Magic Clothing* (Chen et al., 2024)	0.283	0.789	0.400	0.480	0.289	0.794	0.337	0.442
	StableGarment* (Wang et al., 2024)	0.286	0.735	0.381	0.512	0.314	0.721	0.299	0.426
	Any2Any* (Guo et al., 2025)	0.351	0.804	0.395	0.438	0.345	0.805	0.479	0.467
Full	Ours w/o sym.	0.252	0.805	0.397	0.508	0.236	0.858	0.456	0.583
	Ours w/ sym.	**0.226** (+0.026)	**0.842** (+0.037)	**0.423** (+0.026)	**0.553** (+0.045)	**0.221** (+0.015)	**0.868** (+0.010)	**0.509** (+0.053)	**0.609** (+0.026)
C2M + Try-Off	Ours w/o sym.	0.256	0.822	0.398	0.505	0.249	0.849	0.469	0.539
	Ours w/ sym.	0.243 (+0.013)	0.830 (+0.008)	0.414 (+0.016)	0.527 (+0.022)	0.234 (+0.015)	0.861 (+0.012)	0.503 (+0.034)	0.593 (+0.054)
C2M Only	Ours	0.235	0.825	0.417	0.522	0.228	0.864	0.506	0.599

(c) Target-pose-conditioned C2M

Setting	Method	VITON-HD				DressCode			
		MP-LPIPS ↓	CLIP-I ↑	DiffSim ↑	FFA ↑	MP-LPIPS ↓	CLIP-I ↑	DiffSim ↑	FFA ↑
External	Magic Clothing + OpenPose* (Chen et al., 2024)	0.286	0.789	0.395	0.480	0.334	0.788	0.320	0.367
	StableGarment + OpenPose* (Wang et al., 2024)	0.287	0.733	0.329	0.489	0.361	0.744	0.236	0.357
Full	Ours w/o sym.	0.249	0.821	0.403	0.498	0.255	0.851	0.463	0.512
	Ours w/ sym.	**0.236** (+0.013)	**0.831** (+0.010)	**0.405** (+0.002)	**0.517** (+0.019)	**0.243** (+0.012)	**0.857** (+0.006)	**0.537** (+0.074)	**0.563** (+0.051)
C2M + Try-Off	Ours w/o sym.	0.251	0.814	0.397	0.497	0.272	0.839	0.444	0.484
	Ours w/ sym.	0.250 (+0.001)	0.817 (+0.003)	0.402 (+0.005)	0.503 (+0.006)	0.262 (+0.010)	0.847 (+0.008)	0.472 (+0.028)	0.513 (+0.029)
C2M Only	Ours	0.243	0.823	0.402	0.505	0.251	0.853	0.504	0.541

sym. use the same shared parameters and two supervised branches, with $\lambda_s = 0$ and 1, respectively. Full protocols appear in Appendix F.1.

4.2 RQ1: Can Compatible Tasks Share One Competitive Checkpoint?

RQ1 compares the same fixed *Full w/ sym.* checkpoint across all four tasks against external methods, without task-specific fine-tuning. Scores are compared only within each task–dataset protocol; matched variants are analyzed in RQ2. Under our common evaluator, this checkpoint achieves the best value among the evaluated external baselines across all twelve displayed C2P columns in Tab. 2. Its primary P2P unpaired FID/KID results are $9.37/1.27$ on VITON-HD and $4.38/0.60$ on DressCode, also the best among the evaluated baselines; paired P2P instead measures self-reference fidelity because the target's unmasked image supplies the garment condition.

Under the same evaluator, it also achieves the best value among the evaluated external baselines in every displayed Try-Off and C2M column in Tab. 3. Pose-controlled external methods use ControlNet-OpenPose (Chen et al., 2024; Wang et al., 2024), whereas SymFashion uses DensePose, making this a capability comparison; matched SymFashion rows keep the interface fixed. RQ1 establishes viability; matched RQ2 comparisons isolate coupling beyond task aggregation. Comprehensive outputs are shown in Appendix E.4, Figs. C–G.

4.3 RQ2: Does Symmetric Coupling Improve Uncoupled Joint Training?

RQ2 fixes the tasks, model, optimization, and both supervised flow branches. Its primary comparison changes only λ_s; controls intervene on the named coupling component. Tab. 4 progressively rules out weaker explanations on four VITON-HD readouts. The no-symmetry reference isolates uncoupled joint training. The *Incompatible pairing* row breaks sample correspondence while preserving the coupling mechanics and raises the four readouts to $72.52/80.06/140.09/0.614$, showing

Table 4: Coupling ablations on VITON-HD. C2P/P2P use unpaired protocols and C2M uses no pose. Incompatible pairing mismatches human–garment examples. **Bold** and underlined mark the top two ranks.

Variant	Aligned noise η_n	Garment mask η_w	Two-sided grad. η_b	λ_s	C2P unpaired FID ↓	P2P unpaired FID ↓	Try-Off FID ↓	C2M MP-LPIPS ↓
No sym.	–	–	–	0.00	9.39	9.60	8.55	0.252
Independent noise	–	✓	✓	1.00	9.28	9.81	9.00	0.249
Uniform weights	✓	–	✓	1.00	9.30	9.61	9.19	0.238
One-sided gradient	✓	✓	–	1.00	9.39	9.52	9.02	0.232
Incompatible pairing	✓	✓	✓	1.00	72.52	80.06	140.09	0.614
Velocity sym.	✓	✓	✓	1.00	10.78	11.02	9.66	0.242
σ^2-velocity sym.	✓	✓	✓	1.00	10.04	10.29	8.96	0.229
Complete	✓	✓	✓	0.25	9.19	9.43	8.84	**0.226**
Complete	✓	✓	✓	0.50	9.04	9.43	8.81	0.229
Complete	✓	✓	✓	1.00	**8.89**	**9.37**	**8.20**	**0.226**
Complete	✓	✓	✓	2.00	10.77	11.04	9.79	0.239

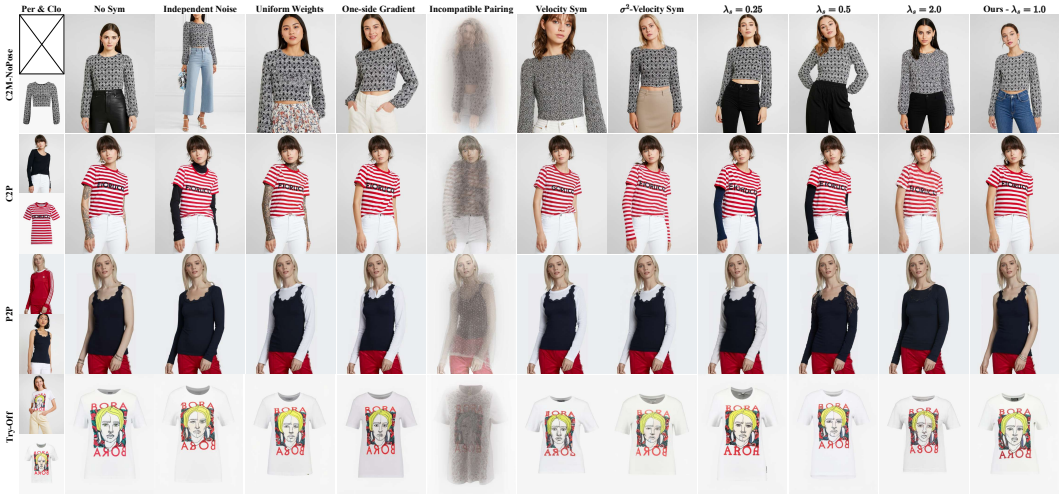

Figure 3: Qualitative controls for Tab. 4. The crosses mark unavailable inputs.

that arbitrary role-matched agreement is insufficient. σ^2-weighted velocity agreement improves over raw velocity matching but remains weaker than residual-aware x_0 coupling. At $\lambda_s = 1$, Complete is best or tied best on all four readouts; $\lambda_s = 2$ degrades every task. Within this tested construction, useful coupling therefore depends on semantic correspondence rather than merely adding a second-view consistency penalty. Aligned noise compares endpoints under the same perturbation, while two-sided gradients let both views adapt. Removing either component or garment-mask emphasis is worse than Complete on all four readouts; Appendix B gives the full analysis.

Fig. 3 shows the corresponding layering, texture, silhouette, and composition differences. The incompatible-pairing outputs are visibly blurred across all four rows, whereas the complete x_0 setting most consistently preserves the displayed garment details and structure.

As shown in Tabs. 2 and 3, under the fixed total optimization budget, uncoupled joint training degrades C2P ($5.71 \rightarrow 6.19$ FID) and C2M ($0.235 \rightarrow 0.252$ MP-LPIPS), whereas coupling improves them to 5.54 and 0.226. It already improves Try-Off ($9.00 \rightarrow 8.55$ FID) and P2P ($9.64 \rightarrow 9.60$ unpaired FID), with coupling further reaching 8.20 and 9.37. Thus, symmetric coupling improves both regimes: tasks degraded by uncoupled joint training and tasks already benefiting from it.

Coupling improves or ties every matched task–dataset readout in the main tables; these correlated metrics support a task-level trend rather than independent trials. Sampled-image metrics, however, cannot show whether the two compatible views actually become more coherent.

4.4 RQ3: What Behavior Does Symmetric Coupling Induce?

To test this, RQ3 measures three single-timestep latent quantities: $\mathcal{C}^{x_0}$ is aligned cross-view disagreement, $\mathcal{R}_r^{x_0}$ is reverse-view error against the common reference, and $\mathcal{R}_{f,\mathrm{nat}}^{x_0}$ is forward error against its own encoded target. Their combined quantity $B = \sqrt{\mathcal{R}_r^{x_0}} + \sqrt{\mathcal{C}^{x_0}}$ is a single-timestep latent diagnostic, not an image-quality metric. Appendix A details the balanced $n = 640$ protocol.

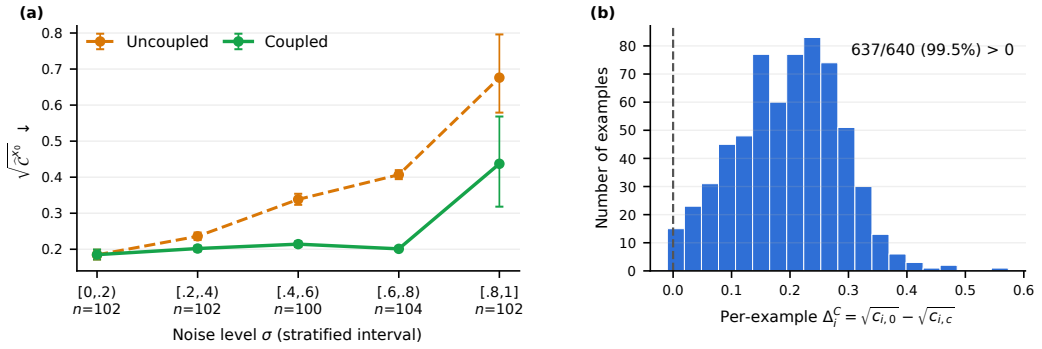

Figure 4: Coupling signature: (a) noise-resolved disagreement ($n = 510$; 95% CIs); (b) per-example natural-timestep change ($n = 640$; positive favors coupling).

Appendix Tab. A reports that coupling reduces $\sqrt{\mathcal{C}^{x_0}}$ by 48.7% and B by 23.8% but raises native forward risk by 18.3% (paired-bootstrap 95% CIs exclude zero), exposing a coherence–accuracy trade-off. Native risk measures single-timestep Euclidean error, whereas RQ1–RQ2 evaluate sampled images; the pattern is consistent with a structured semantic bias, not proof of causality.

In Fig. 4(a), the curves nearly coincide at the lowest noise but separate above $\sigma = 0.2$, consistent with increasing clean-endpoint sensitivity; panel (b) gives $\Delta_i^C > 0$ for $637/640$ examples, so the aggregate is not outlier-driven. Fig. 5 localizes the decrease to both roles and shows a larger reduction inside the person-side garment region (57.0% versus 28.3% outside), consistent with mask emphasis but not an effect confined to the weighted region. Appendix Tabs. B–C give task–dataset and regional breakdowns with spatial CIs, while Appendix Fig. A provides a descriptive depth-wise analysis. These figures answer complementary questions: Fig. 4 identifies when coupling acts and whether its effect is broadly shared, while Fig. 5 shows where the induced agreement is concentrated. Together with the task–dataset and depth breakdowns, this structure supports role-aligned coupling as a structured bias rather than improved single-view estimation.

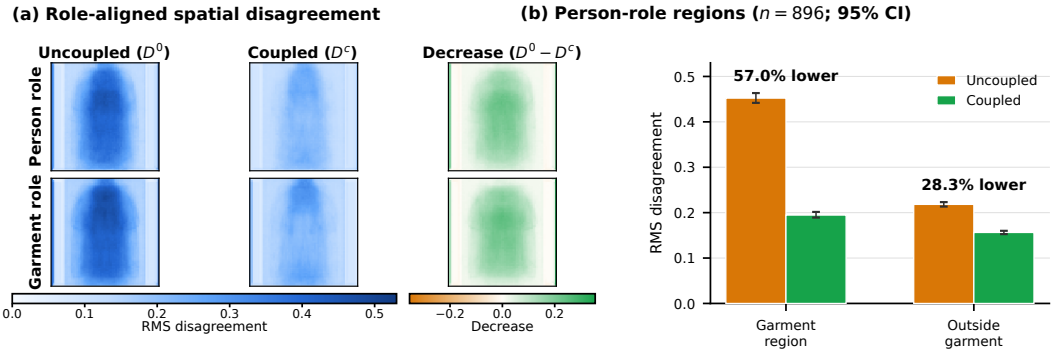

Figure 5: Spatial disagreement ($n = 896$): (a) role-wise values and their decrease; (b) person-role regions with 95% paired-bootstrap CIs. Values are channel normalized.

5 Conclusion

Across four fashion tasks, SymFashion turns compatible directions into structured supervision beyond parameter sharing. One checkpoint remains competitive on C2P, P2P, Try-Off, and C2M. Under matched fixed-budget training, uncoupled sharing has mixed effects, whereas role-aligned clean-endpoint coupling improves or ties every main-table task–dataset point estimate. Diagnostics show stronger cross-view coherence across tasks, garment regions, and transformer depth, with a higher-noise trend despite higher branchwise risk. This describes the induced behavior; sampled-image results establish its task-level benefit. Thus, compatibility-based supervision, not task aggregation alone, offers a broader design principle for unified conditional generation. Future work could examine whether this principle extends to other domains with related semantic structure and whether compatible correspondences can be learned rather than specified a priori.

ADDITIONAL STATEMENTS

REPRODUCIBILITY STATEMENT

SymFashion uses FLUX.1-Fill-dev (Black Forest Labs, 2024) as its single shared rectified-flow transformer and fine-tunes it at 512×512. We fine-tune only its transformer attention layers and input/output projections; all remaining pretrained parameters are frozen. This scope is fixed across all matched comparisons. The training mixture combines VITON-HD, DressCode, DeepFashion, and the curated LRVS-Fashion subset. For P2P only, StableVITON, CAT-DM, and ITA-MDT are retrained with their original architectures on the same task-compatible mixture. Our Voost reproduction uses this same data mixture but is trained only on its native Try-On and Try-Off tasks; all remaining baselines use released checkpoints.

All SymFashion configurations use the same pretrained FLUX backbone and flow-matching objective augmented only by the proposed symmetric loss. Optimization uses AdamW with learning rate 10^{-5}, bf16 precision, ZeRO-2, global batch size 32, and 80,000 optimizer steps on eight NVIDIA B200 GPUs. We use one global $\lambda_s = 1.0$ for every task and dataset, without task- or dataset-specific tuning; Table 4 reports its sensitivity. The matched non-symmetric model sets $\lambda_s = 0$. The garment-mask spatial weighting uses $w^k = 1 + 3m^k$ globally for every task. We fix this coefficient to 3 for all tasks and datasets and do not tune it separately; all other matched settings are unchanged. Both matched variants execute the two supervised directional transformer passes during training, so the symmetric term adds an endpoint comparison rather than a third model pass. Relative to a single-direction setup, this bidirectional protocol requires two transformer evaluations per training example; inference uses only the requested direction. Inference uses 30 flow-matching steps with guidance scale 30; all matched comparisons use the same test IDs and inference conditions.

ETHICS STATEMENT

This work uses established fashion and human-image datasets and does not collect new personal data or involve human-subject experimentation. We will document the licenses and permitted uses of all training and evaluation sources and will release only artifacts compatible with those terms. Because the task generates images of people, outputs may reproduce identity, body-shape, skin-tone, pose, or garment-category biases present in the data. The model is intended for research and virtual-fashion evaluation; it should not be used for identity impersonation, non-consensual image manipulation, or decisions about people.

AI USE STATEMENT

In this work, we used generative AI tools only for language assistance and manuscript polishing. We reviewed all AI-assisted revisions and take responsibility for the final content of this work, including its text, claims, and artifacts.

REFERENCES

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pp. 3686–3693, 2014. 30

Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*, 2020. 30

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 26, 30

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 10, 25

Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997. 3

Weifeng Chen, Tao Gu, Yuhao Xu, and Arlene Chen. Magic clothing: Controllable garment-driven image synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6939–6948, 2024. 2, 3, 7, 22, 30

Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 14131–14140, 2021. 26

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*, pp. 206–235. Springer, 2024. 1, 2, 6, 22, 24

Zheng Chong, Xiao Dong, Haoxiang Li, Wenqing Zhang, Hanqing Zhao, Dongmei Jiang, Xiaodan Liang, et al. Catvton: Concatenation is all you need for virtual try-on with diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 66586–66601, 2025. 2, 6, 22, 24

Frédéric Cordier, WonSook Lee, HyeWon Seo, and Nadia Magnenat-Thalmann. From 2d photos of yourself to virtual try-on dress on the web. In *People and computers XV—interaction without frontiers: Joint proceedings of HCI 2001 and IHM 2001*, pp. 31–46. Springer, 2001. 1

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 30

Chenghu Du, Junyin Wang, Kai Liu, Shengwu Xiong, and Yi Rong. Mask does not matter: a unified latent diffusion-enhanced framework for mask-free virtual try-on. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 936–944, 2025a. 3

Chenghu Du, Shengwu Xiong, and Yi Rong. All parts matter: A unified mask-free virtual try-on framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19525–19534, 2025b. 3

Yutong Feng, Linlin Zhang, Hengyuan Cao, Yiming Chen, Xiaoduan Feng, Jian Cao, Yuxiong Wu, and Bin Wang. Omnitry: Virtual try-on anything without masks. *Advances in Neural Information Processing Systems*, 38:132640–132667, 2026. 3, 22, 24

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7297–7306, 2018. 4

Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Jiaming Liu, and Chuang Zhang. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19085–19096, 2025. 2, 3, 6, 7, 22, 24

Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7543–7552, 2018. 2

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016. 3

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 26, 30

Ji Woo Hong, Tri Ton, Trung X Pham, Gwanhyeong Koo, Sunjae Yoon, and Chang D Yoo. Ita-mdt: Image-timestep-adaptive masked diffusion transformer framework for image-based virtual try-on. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28284–28294, 2025. 1, 2, 6, 22, 24, 29

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pp. 206–228. Springer, 2024. 5, 28

Duc Kieu, Kien Do, Toan Nguyen, Dang Nguyen, and Thin Nguyen. Bidirectional diffusion bridge models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025. 3

Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8176–8185, 2024. 2, 6, 22, 24, 29

Jeongho Kim, Hoiyeong Jin, Sunghyun Park, and Jaegul Choo. Promptdresser: Improving the quality and controllability of virtual try-on via generative textual prompt and prompt-aware mask. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16026–16036, 2025. 1, 2, 22, 24

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023. 5, 28

Klemen Kotar, Stephen Tian, Hong-Xing Yu, Dan Yamins, and Jiajun Wu. Are these the same apple? comparing images based on object intrinsics. *Advances in Neural Information Processing Systems*, 36:40853–40871, 2023. 30

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fr\'echet inception distance. *arXiv preprint arXiv:2203.06026*, 2022. 30

Seungyong Lee and Jeong-gi Kwak. Voost: A unified and scalable diffusion transformer for bidirectional virtual try-on and try-off. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025. 2, 3, 6, 7, 22, 24

Simon Lepage, Jérémie Mary, and David Picard. Lrvs-fashion: Extending visual search with referring instructions. *arXiv:2306.02928*, 2023. 26

Liuzhuozheng Li, Yue Gong, Shanyuan Liu, Zanyi Wang, Dengyang Jiang, Leibucha Wu, Bo Cheng, Yuhang Ma, Dawei Leng, and Yuhui Yin. Refton: Reference person shot assist virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14915–14925, 2026a. 6, 22, 24

Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271, 2020. 29

Qi Li, Shuwen Qiu, Kee Kiat Koo, Julien Han, and Karim Bouyarmane. Dit-vton: Diffusion transformer framework for unified multi-category virtual try-on and virtual try-all with integrated image editing. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 202–211. IEEE, 2026b. 3, 22

Xinghui Li, Qichao Sun, Pengze Zhang, Fulong Ye, Zhichao Liao, Wanquan Feng, Songtao Zhao, and Qian He. Anydressing: Customizable multi-garment virtual dressing via latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23723–23733. IEEE, 2025. 2, 3

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and L Zitnick. Microsoft coco: Common objects in context. eccv. 2014. In *Lecture Notes in Computer Science*, volume 8693. Springer, 2014. 30

Jinxi Liu, Zijian He, Guangrun Wang, Guanbin Li, and Liang Lin. One model for all: Unified try-on and try-off in any pose via llm-inspired bidirectional tweedie diffusion. *arXiv preprint arXiv:2508.04559*, 2025. 6, 7, 22, 24

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1096–1104, 2016. 26

Davide Lobba, Fulvio Sanguigni, Bin Ren, Marcella Cornia, Rita Cucchiara, and Nicu Sebe. Inverse virtual try-on: Generating multi-category product-style images from clothed individuals. *arXiv preprint arXiv:2505.21062*, 2025. 2, 3, 7, 22, 24

Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *European Conference on Computer Vision*, pp. 345–362. Springer, 2022. 26

Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 8580–8589, 2023. 1, 2

Shuliang Ning, Duomin Wang, Yipeng Qin, Zirong Jin, Baoyuan Wang, and Xiaoguang Han. Picture: Photorealistic virtual try-on from unconstrained designs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6976–6985, 2024. 3

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. *arXiv preprint arXiv:2304.07193*, 2023. 30

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*, pp. 8748–8763. PmLR, 2021. 30

Amit Raj, Patsorn Sangkloy, Huiwen Chang, Jingwan Lu, Duygu Ceylan, and James Hays. Swapnet: Garment transfer in single view images. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 666–682, 2018. 1, 3

Mehul P Sampat, Zhou Wang, Shalini Gupta, Alan Conrad Bovik, and Mia K Markey. Complex wavelet structural similarity: A new image similarity index. *IEEE transactions on image processing*, 18(11):2385–2401, 2009. 30

Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang. Imagdressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6795–6804, 2025. 3

Dan Song, Xuanpu Zhang, Juan Zhou, Weizhi Nie, Ruofeng Tong, Mohan Kankanhalli, and An-An Liu. Image-based virtual try-on: A survey. *International Journal of Computer Vision*, 133(5): 2692–2720, 2025a. 1

Yiren Song, Xiaokang Liu, and Mike Zheng Shou. Diffsim: Taming diffusion models for evaluating visual similarity. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16904–16915. IEEE, 2025b. 30

Riza Velioglu, Petra Bevandic, Robin Chan, and Barbara Hammer. Mgt: Extending virtual try-off to multi-garment scenarios. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 6098–6107. IEEE, 2025a. 2, 3, 7, 22, 24

Riza Velioglu, Petra Bevandic, Robin Chan, and Barbara Hammer. Tryoffdiff: Virtual-try-off via high-fidelity garment reconstruction using diffusion models. In *BMVC*, 2025b. https://doi.org/nt3n. 2, 3, 7, 22, 24

Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 589–604, 2018. 2

Rui Wang, Hailong Guo, Jiaming Liu, Huaxia Li, Haibo Zhao, Xu Tang, Yao Hu, Hao Tang, and Peipei Li. Stablegarment: Garment-centric generation via stable diffusion. *arXiv preprint arXiv:2403.10783*, 2024. 2, 3, 7, 22

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The thrity-seventh asilomar conference on signals, systems & computers, 2003*, volume 2, pp. 1398–1402. Ieee, 2003. 30

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 26, 30

Jia Wu, Yijing Dai, Tingfeng Cao, Meiling Wu, Tao Luo, Jian Dong Zhang, Guangming Lu, and Xiaoyi Zeng. High-fidelity virtual try-on beyond paired data scarcity via diffusion-based cycle-consistent learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35821–35830, 2026. 2, 3, 22

Yuhao Xu, Tao Gu, Weifeng Chen, and Arlene Chen. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8996–9004, 2025. 22, 24

Zhaotong Yang, Yuhui Li, Shengfeng He, Xinzhe Li, Yangyang Xu, Junyu Dong, and Yong Du. Omnivton: Training-free universal virtual try-on. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16702–16711, 2025. 22, 24

Zhaotong Yang, Ying Tai, Jiahui Zhan, Yu Zheng, Jianjun Qian, and Jian Yang. Orthotryon: Geometric orthogonalization for conflict-free unified fashion generation. *arXiv preprint arXiv:2606.27880*, 2026. 2, 3, 6, 7, 22, 24

Jianhao Zeng, Dan Song, Weizhi Nie, Hongshuo Tian, Tongtong Wang, and An-An Liu. Cat-dm: Controllable accelerated virtual try-on with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8372–8382, 2024. 2, 6, 22, 24, 29

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018. 26, 30

Wei Zhang, Yeying Jin, Xin Li, Yan Zhang, Xiaofeng Cong, Cong Wang, Fengcai Qiao, and Zhichao Lian. Unifit: Towards universal virtual try-on with mllm-guided semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12816–12824, 2026. 2, 3, 22

Xujie Zhang, Ente Lin, Xiu Li, Yuxuan Luo, Michael Kampffmeyer, Xin Dong, and Xiaodan Liang. Mmtryon: Multi-modal multi-reference control for high-quality fashion generation. *arXiv preprint arXiv:2405.00448*, 2024. 3

Zijian Zhou, Shikun Liu, Xiao Han, Haozhe Liu, Kam Woh Ng, Tian Xie, Yuren Cong, Hang Li, Mengmeng Xu, Juan-Manuel Pérez-Rúa, et al. Learning flow fields in attention for controllable person image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2491–2501. IEEE, 2025. 6, 22, 24

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2223–2232, 2017. 3

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4606–4615, 2023. 1, 3

Appendix

Appendix Contents

A Mechanism diagnostics
B Symmetry ablations
C Reverse-canvas construction
D Pose-control evaluation
E Supplemental comparisons
F Experimental details
G Analytical details
H Limitations
I Failure analysis

A Mechanism Risk Diagnostics

Table A reports pooled aggregate and noise-resolved clean-latent diagnostics for the full C2P, P2P, Try-Off, and C2M task mixture over the VITON-HD–DressCode evaluation pool. The aggregate panel uses the natural training timestep distribution and 640 examples, balanced at 80 examples in each task–dataset cell. The lower panel is a separate stratified-sigma evaluation with 510 matched examples. Table B reports the cell-level natural-timestep results; the sigma panel remains a pooled diagnostic rather than a claim about every task separately.

Table A: Cross-view risk diagnostics under natural and stratified timesteps.

Balanced aggregate, natural timesteps ($n = 640$)

Variant	Space	$\sqrt{R_f}\downarrow$	$\sqrt{R_f^{\mathrm{nat}}}\downarrow$	$\sqrt{R_r}\downarrow$	$\sqrt{C}\downarrow$	$B\downarrow$
No sym.	None	0.3260	0.2953	0.3441	0.4201	0.7642
Sym.	Clean latent	0.3616	0.3495	0.3668	0.2157	0.5825

Stratified-sigma clean-latent improvement over no sym.

σ interval	n	$\Delta\sqrt{C} = \sqrt{C_0} - \sqrt{C_s}\uparrow$	$\Delta B = B_0 - B_s\uparrow$
$[0.0, 0.2)$	102	-0.0005	$+0.0045$
$[0.2, 0.4)$	102	$+0.0339$	$+0.0408$
$[0.4, 0.6)$	100	$+0.1242$	$+0.1229$
$[0.6, 0.8)$	104	$+0.2057$	$+0.1936$
$[0.8, 1.0]$	102	$+0.2390$	$+0.2055$

The unequal totals reflect three separate sampling designs rather than missing outputs. The natural-timestep aggregate uses 80 examples in each of eight task–dataset cells ($n = 640$). The sigma-stratified evaluator runs until every bin contains at least 100 examples and retains the final complete batch, yielding $n = 510$ with 100–104 examples per bin. The spatial audit retains seven complete 16-example batches per cell, yielding 112 examples per cell ($n = 896$). Within each protocol, the symmetric and non-symmetric checkpoints use identical IDs, timesteps, noise, and sample counts; numerical comparisons are made only within a protocol.

The diagnostic is computed on held-out test examples using the same sample identifiers, timestep draws, and aligned noise for the two paired directions. For each branch, the predicted velocity is converted to a clean endpoint with Eq. 3; the reverse estimate is then mapped back through the role permutation before comparison. We report the square-root risks so that the displayed quantities have the same norm scale as the triangle inequality underlying Eq. 30. The cross-view term measures disagreement between the two aligned predictions, whereas B combines this disagreement with reverse-view estimation error. The evaluation examples are disjoint from the training data. The mechanism diagnostics are reported post hoc and are not optimized during training. The balanced aggregate gives every task–dataset cell equal weight.

Concretely, for each matched example i, we compute the weighted squared quantities below. Each role norm is averaged over its packed latent elements, as in Eq. 8, and the two role values are averaged with equal weight. Here $z^\star_{0,i} = [z^\star_{0,P,i} \,\|\, z^\star_{0,G,i}]$ is the common reference formed from the evaluator's person-side and garment-side VAE encodings; by construction it is also the aligned reverse clean target, while the jointly encoded forward canvas may differ by $\delta_{0,i}$ in Eq. 5: For C2M, this is the reference person in that paired supervised example, not a unique canonical person associated with the garment.

$$
\begin{aligned}
r_{f,i} &= \|\hat{z}^f_{0,i} - z^\star_{0,i}\|^2_W, & \widehat{\mathcal{R}}^{x_0}_f &= \frac{1}{n}\sum_i r_{f,i},\\
r_{r,i} &= \|\bar{\hat{z}}^r_{0,i} - z^\star_{0,i}\|^2_W, & \widehat{\mathcal{R}}^{x_0}_r &= \frac{1}{n}\sum_i r_{r,i},\\
c_i &= \|\hat{z}^f_{0,i} - \bar{\hat{z}}^r_{0,i}\|^2_W, & \widehat{\mathcal{C}}^{x_0} &= \frac{1}{n}\sum_i c_i.
\end{aligned}
\tag{11}
$$

We additionally separate the forward-path construction residual by evaluating the prediction against the jointly encoded forward clean canvas $z_{0,i}^{f,\text{joint}}$ used to form its noisy input:

$$r_{f,\text{nat},i} = \|\hat{z}_{0,i}^{f} - z_{0,i}^{f,\text{joint}}\|_W^2, \qquad \widehat{\mathcal{R}}_{f,\text{nat}}^{x_0} = \frac{1}{n}\sum_i r_{f,\text{nat},i}. \tag{12}$$

Consequently, $\mathcal{R}_f^{x_0}$ and $\mathcal{R}_r^{x_0}$ are mechanism diagnostics relative to one common evaluator reference, not a symmetric leaderboard of branch quality. The forward diagnostic intentionally includes the construction residual δ_0, whereas $\mathcal{R}_{f,\text{nat}}^{x_0}$ directly measures forward prediction error relative to the native clean target used by that path. The table reports $\sqrt{\widehat{\mathcal{R}}_f^{x_0}}$, $\sqrt{\widehat{\mathcal{R}}_{f,\text{nat}}^{x_0}}$, $\sqrt{\widehat{\mathcal{R}}_r^{x_0}}$, $\sqrt{\widehat{\mathcal{C}}^{x_0}}$, and $B = \sqrt{\widehat{\mathcal{R}}_r^{x_0}} + \sqrt{\widehat{\mathcal{C}}^{x_0}}$. For a sigma interval I_b, the same sample averages are recomputed over $\mathcal{S}_b = \{i : \sigma_i \in I_b\}$; the reported deltas are $\Delta\sqrt{C} = \sqrt{\widehat{C}_0} - \sqrt{\widehat{C}_s}$ and $\Delta B = B_0 - B_s$, where subscripts 0 and s denote the non-symmetric and symmetric checkpoints, respectively; positive values are favorable.

We separately audit the latent construction residual on a balanced held-out pool of 640 examples, with 80 examples in each task–dataset cell. This gives $\sqrt{\mathbb{E}\|\delta_0\|_W^2} = 0.1675$. Directly evaluating Eq. 25 gives a mean absolute discrepancy of 0.00383 and a maximum discrepancy of 0.0692 under bf16 execution. These values quantify finite-precision implementation error rather than prediction quality; the residual-aware equality holds exactly under the stated construction. Since δ_0 is defined before noising, we report one aggregate value rather than interpreting it as a timestep-dependent model effect.

For uncertainty, we perform 10,000 paired bootstrap replicates. The aggregate analysis resamples examples with replacement within each task–dataset cell, preserving all eight cell sizes and using the same sampled indices for the symmetric and non-symmetric checkpoints. In each replicate, we first recompute the means of r_f, $r_{f,\text{nat}}$, r_r, and c, then apply the square roots and form B; the reported 95% intervals are the 2.5th and 97.5th percentiles of the paired checkpoint differences. The sigma-resolved analysis applies the same paired procedure independently within each fixed interval.

For the balanced $n = 640$ evaluation, symmetric coupling reduces $\sqrt{\mathcal{C}^{x_0}}$ from 0.4201 to 0.2157 and B from 0.7642 to 0.5825. In contrast, $\sqrt{\mathcal{R}_f^{x_0}}$ increases from 0.3260 to 0.3616 and $\sqrt{\mathcal{R}_r^{x_0}}$ from 0.3441 to 0.3668. Under 10,000 task–dataset-stratified paired-bootstrap resamples, the favorable difference is 0.2044 for $\sqrt{C}$ (95% CI $[0.1979, 0.2110]$) and 0.1817 for B (95% CI $[0.1758, 0.1875]$). The corresponding differences for $\sqrt{R_f}$ and $\sqrt{R_r}$ are negative, confirming that agreement improves without a reduction in the two common-reference branchwise risks. The native forward risk also increases from 0.2953 to 0.3495, with favorable difference -0.0542 (95% CI $[-0.0572, -0.0513]$). Thus, the forward-risk increase is not solely an artifact of using the common role-wise reference. At the same time, the inequality slack $S = B - \sqrt{R_f}$ decreases from 0.4382 to 0.2209; the observed inequality slack is smaller, but this does not imply a lower bounded forward risk, which increases in this comparison.

Table B: Cell-wise mechanism changes (80 examples/cell). Deltas are no symmetry minus symmetry; percentages are relative reductions.

Task	Dataset	$\Delta\sqrt{C}\uparrow$	$\Delta B\uparrow$
C2P	VITON-HD	$+0.2587$ (54.8%)	$+0.2415$ (27.6%)
C2P	DressCode	$+0.2114$ (51.0%)	$+0.2071$ (26.6%)
P2P	VITON-HD	$+0.1587$ (37.6%)	$+0.1377$ (17.8%)
P2P	DressCode	$+0.1341$ (37.0%)	$+0.1209$ (18.2%)
Try-Off	VITON-HD	$+0.2314$ (52.8%)	$+0.1812$ (23.7%)
Try-Off	DressCode	$+0.2059$ (52.4%)	$+0.1550$ (23.1%)
C2M	VITON-HD	$+0.2283$ (51.5%)	$+0.2098$ (25.9%)
C2M	DressCode	$+0.2062$ (50.8%)	$+0.1938$ (25.7%)

Table B shows the same qualitative trade-off in every task–dataset cell. The relative reduction in $\sqrt{C}$ ranges from 37.0% to 54.8%, while B decreases by 17.8%–27.6%. C2P on VITON-HD gives the largest changes; the smallest $\sqrt{C}$ and B reductions occur for P2P on DressCode and VITON-HD, respectively. Both deltas remain favorable in every cell. Native forward risk increases in all eight cells, with favorable differences from -0.0865 to -0.0253. The pooled result is therefore not driven by one task or dataset, while the differing magnitudes caution against claiming a uniform effect size.

The stratified-sigma panel reports the favorable change relative to the non-symmetric reference within each noise interval. Its bin counts are $102, 102, 100, 104, 102$. The first bin shows a negligible adverse change in $\sqrt{\mathcal{C}^{x_0}}$ (-0.0005), while the remaining bins improve by 0.0339, 0.1242, 0.2057, and 0.2390; ΔB is positive in all five bins. Under 10,000 paired-bootstrap resamples, the

95% interval for $\Delta\sqrt{C}$ contains zero only in the first bin and excludes zero in the remaining four; the interval for ΔB excludes zero in all five bins. We therefore describe the result as a higher-noise reduction trend, not as uniform improvement over the entire noise schedule.

Spatial localization. Table C asks where cross-view agreement changes on the person-side role. Let $m_i^{\mathrm{gar}}(p) \in [0,1]$ denote the Sapiens garment mask on the person tile, resized to the latent grid; soft values occur only along resized boundaries. We use "in" for its garment-weighted region and "out" for the complement. For each example i and person-tile position p, let $d_i^P(p) = D_z^{-1}\|\hat{z}^f_{0,P,i}(p) - \bar{\tilde{z}}^r_{0,P,i}(p)\|_2^2$, where $D_z = 16$ is the latent-channel count. We compute equal-sample averages inside and outside the garment region as

$$\widehat{\mathcal{C}}^{x_0}_{P,\mathrm{in}} = \frac{1}{n}\sum_i \frac{\sum_p m_i^{\mathrm{gar}}(p) d_i^P(p)}{\sum_p m_i^{\mathrm{gar}}(p)}, \qquad \widehat{\mathcal{C}}^{x_0}_{P,\mathrm{out}} = \frac{1}{n}\sum_i \frac{\sum_p (1 - m_i^{\mathrm{gar}}(p)) d_i^P(p)}{\sum_p (1 - m_i^{\mathrm{gar}}(p))}. \tag{13}$$

The channel normalization puts the table on the same per-latent-element scale as Fig. 4; it remains a person-tile regional audit rather than an additive decomposition of the two-role, W-weighted aggregate. Lower values mean stronger spatial agreement. The Δ row is Uncoupled minus Coupled, so positive values favor coupling. The balanced evaluation uses 896 matched examples, with 112 examples in each of the eight task–dataset cells. Symmetric coupling reduces $\sqrt{\widehat{\mathcal{C}}^{x_0}_{P,\mathrm{in}}}$ from 0.452 to 0.195 and $\sqrt{\widehat{\mathcal{C}}^{x_0}_{P,\mathrm{out}}}$ from 0.218 to 0.156. These are relative reductions of 57.0% and 28.3%, respectively. Under 10,000 task–dataset-stratified paired-bootstrap resamples, the favorable changes are 0.258 (95% CI $[0.250, 0.265]$) inside the garment and 0.062 (95% CI $[0.059, 0.064]$) elsewhere. Thus, agreement improves across the canvas but is more strongly concentrated in the garment-sensitive region; the effect is not exclusively localized there.

Table C: Channel-normalized person-tile clean-latent disagreement inside and outside the garment mask ($n = 896$). Lower is better; Δ is Uncoupled minus Coupled.

Variant	$\sqrt{\widehat{\mathcal{C}}^{x_0}_{P,\mathrm{in}}}\downarrow$	$\sqrt{\widehat{\mathcal{C}}^{x_0}_{P,\mathrm{out}}}\downarrow$
Uncoupled	0.452	0.218
Coupled	0.195	0.156
Δ (95% CI)	$+0.258$ $[0.250, 0.265]$	$+0.062$ $[0.059, 0.064]$

Figure 5 visualizes the same matched audit in the main paper. The aggregate decrease is concentrated over person and garment support, whereas the person-role bar chart reproduces the regional quantities in Table C. The maps are explanatory rather than an additional quantitative estimate; the table carries the exact regional comparison and uncertainty.

Representation depth. We additionally measure how role-aligned agreement evolves through the 57 trainable attention blocks. For token t in semantic role $k \in \{P, G\}$, write $\widetilde{h} = h/(\|h\|_2 + 10^{-8})$ and define

$$\begin{aligned} d_{\ell,i,t} &= 1 - \left\langle \widetilde{h}^f_{\ell,i,t}, \widetilde{\bar{h}}^{,r}_{\ell,i,t} \right\rangle, \\ \Delta D_{\ell,b} &= D^0_{\ell,b} - D^c_{\ell,b}. \end{aligned} \tag{14}$$

where $\bar{h}^r$ is semantically realigned, $D^q_{\ell,b}$ averages d over tokens, examples in noise bin b, and the two roles with equal weight, and $q \in \{0, c\}$ indexes the uncoupled and coupled checkpoints. Thus, $\Delta D > 0$ denotes lower normalized hidden-state disagreement with coupling. This separate matched evaluation contains 528 examples, with 106, 106, 103, 108, and 105 examples in ascending noise bins.

Figure A shows all 285 block–noise cells, with $\Delta D > 0$ in 256. Every cell above the lowest-noise interval is positive; the $[0, 0.2)$ interval is mixed, with 28 positive and 29 negative cells. The two-sided BH–FDR test retains 277 cells (253 positive and 24 negative), while the remaining eight all lie in the lowest-noise interval. Averaged over noise levels, the early, middle, and late thirds of the transformer improve by 0.0200 (95% CI $[0.0193, 0.0208]$), 0.0097 ($[0.0093, 0.0101]$), and 0.0137 ($[0.0131, 0.0144]$), respectively, under 10,000 paired-bootstrap replicates. For the late group, the reduction rises from 0.0049 in the lowest-noise bin to 0.0224 in the highest. The mean hidden-norm ratios between checkpoints remain within $[0.920, 1.025]$ across blocks, arguing against norm collapse as the explanation. All three task-pair families have a positive mean reduction.

We test instance specificity at 12 prespecified blocks $\{0, 6, 12, 18, 19, 24, 30, 36, 42, 48, 54, 56\}$. Ten cyclic derangements within each task-pair-family–dataset–noise-bin cell replace the aligned reverse

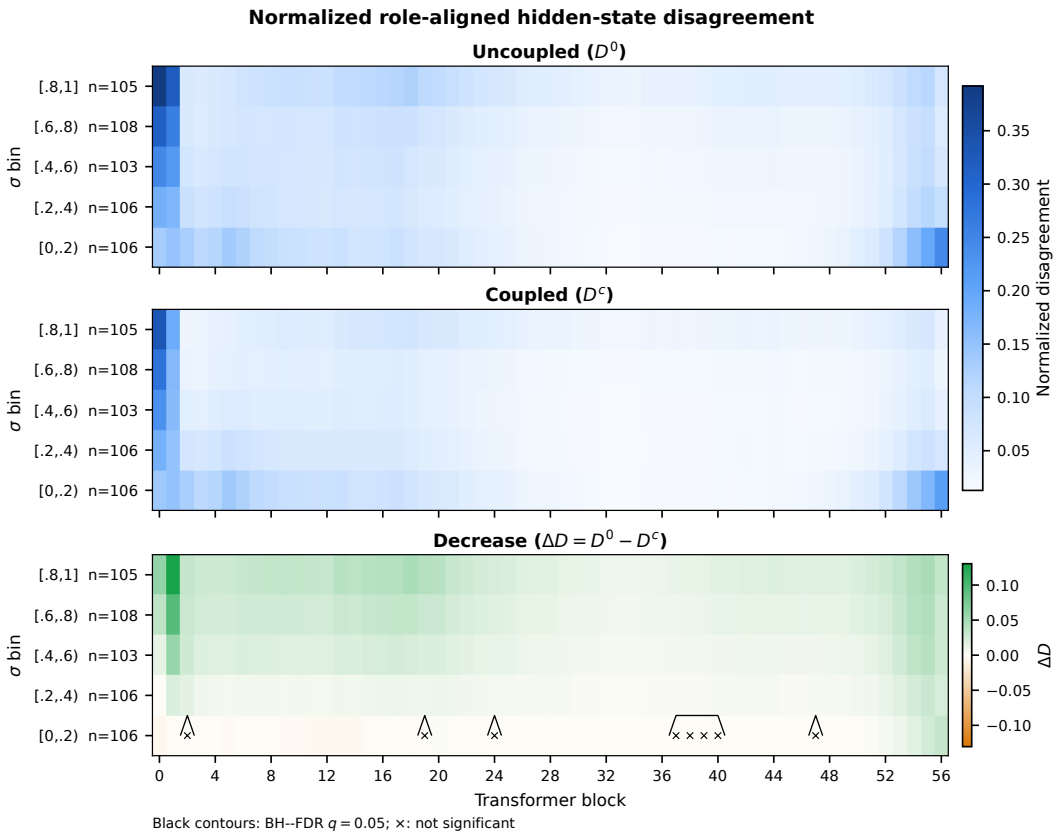

Figure A: Role-aligned hidden-state disagreement over transformer depth and noise level. The first two panels report absolute normalized disagreement for the uncoupled and coupled checkpoints on a shared scale; the last reports $\Delta D = D^0 - D^c$, where positive values indicate stronger alignment with coupling. Black contours mark cells passing a two-sided paired-bootstrap BH–FDR at $q = 0.05$; crosses mark the remaining cells. This normalized diagnostic measures representation alignment, not output quality or causal layer importance.

example. Let $S^q = D^{q,\mathrm{shuf}} - D^{q,\mathrm{comp}}$ and $\Delta S = S^c - S^0$; hence, $S > 0$ means matched views are closer than shuffled views, and $\Delta S > 0$ means coupling widens this gap. Both checkpoints give $S > 0$ in all 60 probe-block–noise cells; ΔS is positive in 57, with mean 0.01845. Its early, middle, and late averages are 0.0340 (95% CI $[0.0335, 0.0346]$), 0.0109 ($[0.0106, 0.0113]$), and 0.0104 ($[0.0096, 0.0112]$). This sparse control supports instance specificity at coarse depths, not at every intervening block.

Takeaway. This section establishes the mechanism signature of symmetric coupling: lower cross-view disagreement recurs across task–dataset cells, noise levels, spatial regions, and representation depth, with a wider matched-versus-shuffled gap. The aggregate reduction itself is expected from the objective and is not independent evidence for image quality. Because native-target forward risk increases even as B falls, the result identifies a coherence–accuracy trade-off rather than better clean-latent estimation; RQ1–RQ2 separately establish sampled-image quality.

B Symmetry Component and Prediction-Space Ablations

Table 4 reports component controls, prediction-space controls, and the λ_s sensitivity sweep. The component rows remove aligned noise, fixed garment-mask emphasis, or two-sided gradient flow while holding the remaining settings fixed. The *Velocity sym.* and σ^2*-velocity sym.* rows replace the clean-endpoint comparison by velocity or σ^2-weighted velocity agreement. C2P and P2P use their VITON-HD unpaired protocols; Try-Off uses FID and C2M uses no-pose MP-LPIPS. These task-

level readouts follow the primary evaluation goal of each direction: unpaired distribution quality for try-on and garment transfer, output-distribution quality for canonical-garment generation in Try-Off, and preservation of the conditioning garment identity for C2M.

Mathematical encoding. Let $\boldsymbol{\eta} = (\eta_n, \eta_w, \eta_b) \in \{0,1\}^3$ switch noise alignment, fixed garment-mask emphasis, and bidirectional symmetric gradients, respectively. Both views retain the same timestep σ, while

$$\epsilon^r_{\eta_n} = \begin{cases} \pi(\epsilon^f), & \eta_n = 1, \\ \epsilon',\ \epsilon' \sim \mathcal{N}(0, I),\ \epsilon' \perp \epsilon^f, & \eta_n = 0, \end{cases} \quad w^k_{\eta_w} = 1 + 3\eta_w m^k, \quad k \in \{P, G\}. \tag{15}$$

Define $T_1(y) = y$ and $T_0(y) = \mathrm{sg}(y)$, where sg preserves the forward value but stops its gradient. Thus, when $\eta_b = 0$, the aligned reverse prediction is treated as a stop-gradient target and the symmetric term updates only the forward prediction. The reverse branch remains trainable through $\mathcal{L}^r_{\text{flow}}$; when $\eta_b = 1$, the symmetric term propagates through both predictions. For prediction space $\rho \in \{x_0, v, \sigma^2 v\}$, define

$$q^d_{x_0} = \hat{z}^d_0, \qquad q^d_v = u^d, \qquad q^d_{\sigma^2 v} = \sigma u^d. \tag{16}$$

As before, $\bar{q}^r_\rho = \pi^{-1}(q^r_\rho)$ denotes semantic realignment. Thus, squared disagreement between $q^f_{\sigma^2 v}$ and its aligned reverse counterpart is exactly $\sigma^2 \|u^f - \bar{u}^r\|^2_W$. The controlled symmetric term and complete objective can be written as

$$\ell^\rho_{\text{sym}}(\boldsymbol{\eta}) = \frac{1}{2} \sum_{k \in \{P,G\}} \left\| q^f_\rho[k] - T_{\eta_b}\big(\bar{q}^r_\rho[k; \epsilon^r_{\eta_n}]\big) \right\|^2_{w^k_{\eta_w}}, \tag{17}$$

$$\mathcal{L}(\rho, \boldsymbol{\eta}, \lambda_s) = \mathcal{L}^f_{\text{flow}} + \mathcal{L}^r_{\text{flow}} + \lambda_s \mathbb{E}\big[\ell^\rho_{\text{sym}}(\boldsymbol{\eta})\big]. \tag{18}$$

Thus, the table settings are No sym. $= (x_0; -, -, -; 0)$, Independent noise $= (x_0; 0, 1, 1; 1)$, Uniform weights $= (x_0; 1, 0, 1; 1)$, One-sided gradient $= (x_0; 1, 1, 0; 1)$, Velocity sym. $= (v; 1, 1, 1; 1)$, σ^2-velocity sym. $= (\sigma^2 v; 1, 1, 1; 1)$, and Complete $= (x_0; 1, 1, 1; \lambda_s)$, where each tuple denotes $(\rho; \eta_n, \eta_w, \eta_b; \lambda_s)$. In Table 4, ✓and –correspond to switch values one and zero, while "–" denotes an inactive switch when $\lambda_s = 0$.

All rows keep the data mixture, optimizer, backbone, task configuration, and two supervised flow branches fixed. The no-symmetry reference disables only cross-view coupling, whereas each component control changes one ingredient of that coupling. These interventions therefore test the proposed construction without changing task supervision or model capacity.

Compatibility negative control. The 'Incompatible pairing' row tests whether correspondence between compatible views matters beyond adding an arbitrary consistency penalty. For a local minibatch of size b, let p be a fixed-point-free cyclic permutation, so $p(i) \neq i$. Before noising the reverse view, we jointly permute its clean canvas, observed condition latent, spatial mask, DensePose condition, and task/category conditioning. We leave the forward batch, timestep σ_i, aligned reverse noise $\pi(\epsilon_i)$, and forward-example weights w^P_i, w^G_i unchanged, then reconstruct the reverse noised state and its supervised velocity target. The control therefore uses

$$\mathcal{L}^{x_0}_{\text{sym,inc}} = \frac{1}{2b} \sum_{i=1}^{b} \sum_{k \in \{P,G\}} \left\| \hat{z}^f_{0,i}[k] - \bar{\hat{z}}^r_{0,p(i)}[k; \sigma_i, \pi(\epsilon_i)] \right\|^2_{w^k_i}. \tag{19}$$

Semantic roles are not crossed: person remains compared with person and garment with garment, but the two sides describe different human–garment examples. The control therefore preserves the task interfaces, timestep, aligned-noise construction, role ordering, and symmetric objective while breaking the human–garment correspondence that defines a compatible pair. Because every reverse tensor moves under the same permutation and its flow target is recomputed after noising, both branches retain valid supervised flow objectives. Aligned noise, garment-mask weighting, two-sided gradients, x_0 prediction space, and $\lambda_s = 1$ otherwise match the Complete configuration exactly. If a minibatch contains no derangement with distinct example keys, its symmetric term is omitted while both supervised flow terms remain active. This is not a garment-only causal intervention: both the human identity and its garment correspondence differ across the two views. Instead, it tests whether

generic role-matched agreement is sufficient, or whether useful cross-task supervision depends on semantically corresponding human–garment views.

The negative control fails sharply on every readout: C2P and P2P unpaired FID increase to 72.52 and 80.06, Try-Off FID to 140.09, and C2M MP-LPIPS to 0.614, compared with 8.89, 9.37, 8.20, and 0.226 for the Complete configuration. Because the two supervised flow branches remain valid and the intervention changes the cross-view correspondence, this result shows that the symmetric term is not a generic benefit when applied between arbitrary role-matched examples. Within this tested construction, useful coupling depends on semantically corresponding views.

The *Velocity sym.* and σ^2*-velocity sym.* controls keep aligned noise, fixed garment-mask emphasis, two-sided gradients, and $\lambda_s = 1$ fixed. They therefore test whether the clean-endpoint formulation contributes beyond directly matching velocity, and whether the exact-alignment special case $\sigma^2 \ell^v_{\mathrm{sym}}$ captures the residual-aware objective used by the full model.

The prediction-space results form a consistent hierarchy. Relative to raw velocity agreement, σ^2 weighting lowers C2P, P2P, and Try-Off FID by 0.74, 0.73, and 0.70, respectively, and lowers C2M MP-LPIPS by 0.013. Nevertheless, the residual-aware x_0 objective at $\lambda_s = 1$ further improves these four readouts by 1.15, 0.92, 0.76, and 0.003. Both velocity-space controls underperform no symmetry on the three distributional FID measurements, although they improve C2M MP-LPIPS. In contrast, x_0 agreement improves all four measurements over no symmetry. These results support the prediction-space choice: σ^2 weighting mitigates direct velocity matching, but does not capture the nonzero alignment residual and cross term in Eq. 26. This is task-level evidence for the residual-aware formulation, not a proof of the identity's causal contribution.

The component removals produce mixed task-dependent responses relative to no symmetry, but each remains worse than the complete construction on all four measurements. Relative to independent noise, the Complete configuration lowers C2P, P2P, and Try-Off FID by 0.39, 0.44, and 0.80, respectively, and C2M MP-LPIPS by 0.023. The corresponding improvements over uniform weighting are 0.41, 0.24, 0.99, and 0.012, and those over one-sided gradients are 0.50, 0.15, 0.82, and 0.006. The degradation profile varies across tasks, with Try-Off being the most consistently sensitive direction. The complete construction nevertheless performs best across all four measurements.

Try-Off provides the clearest stress test: independent noise (9.00), uniform weighting (9.19), and one-sided gradients (9.02) all underperform the no-symmetry reference (8.55), whereas the complete construction improves FID to 8.20. Thus, an incompletely specified consistency term can be harmful in this direction; consistent cross-task gains depend on how the views are coupled, not merely on adding a consistency penalty.

The sweep shows task-dependent sensitivity. As λ_s increases from 0.25 to 0.50 and 1.00, C2P FID decreases monotonically from 9.19 to 9.04 and 8.89. P2P is stable at 9.43 for the two lower weights and then improves to 9.37, while C2M remains within 0.226–0.229. Try-Off is more sensitive: the two lower weights yield 8.84 and 8.81, both above the no-symmetry value of 8.55, whereas $\lambda_s = 1.00$ improves it to 8.20. At $\lambda_s = 2.00$, however, C2P, P2P, and Try-Off FID increase to 10.77, 11.04, and 9.79, and C2M MP-LPIPS rises to 0.239. Thus, $\lambda_s = 1.00$ gives the best or tied-best value on every measurement, while overweighting consistency reverses much of the gain. The response is therefore neither uniformly monotonic nor insensitive to coupling strength. We use $\lambda_s = 1.00$.

Taken together with the matched rows in the main tables, the task-composition results distinguish two regimes under the fixed total optimization budget. Uncoupled joint training degrades C2P and C2M but improves the selected P2P and Try-Off readouts; symmetric coupling improves all four relative to the corresponding uncoupled model. This separates useful structured interaction from joint training alone without claiming a uniform effect across tasks.

Within this ablation protocol, the complete construction reduces C2P unpaired FID from 9.39 to 8.89, P2P unpaired FID from 9.60 to 9.37, and Try-Off FID from 8.55 to 8.20; C2M MP-LPIPS decreases from 0.252 to 0.226. These metrics are computed from the same trained checkpoints and evaluation data; their shared favorable direction complements the controlled component comparisons above.

Takeaway. This section establishes that the complete coupling construction, rather than an arbitrary consistency penalty, produces the strongest four task-level readouts. Semantic correspondence is essential, aligned noise, garment-mask emphasis, and two-sided gradients work best together, and clean-endpoint coupling outperforms the named velocity-space alternatives. The λ_s rows characterize sensitivity to coupling strength rather than repeated trials.

C Reverse-Canvas Construction

The two semantic roles can be arranged in two ways when constructing the reverse branch. Let $\mathcal{J}(X_P, X_G) = [X_P \,\|\, X_G]$ denote width concatenation in the forward order, and let $\pi\mathcal{J}(X_P, X_G) = \mathcal{J}(X_G, X_P)$ exchange the two tiles. Here X_k^r denotes the reverse-view tensor for semantic role k, and X may represent a clean or noised latent, a masked condition, a spatial mask, or spatial guidance. Reverse-canvas construction must apply one ordering consistently to every such tensor; otherwise a latent token, its mask, and its conditioning signal would refer to different semantic locations.

Physical swap (main method). SymFashion constructs the reverse view as

$$\mathcal{X}_r^{\text{swap}} = \pi\mathcal{J}(X_P^r, X_G^r) = [X_G^r \,\|\, X_P^r]. \tag{20}$$

Thus the role requested by the reverse task is moved into the designated output position. The same permutation is applied to the clean latent, aligned noise, noised state, masked condition, mask, and spatial guidance. In particular, if the forward noise is $\epsilon_f = [\epsilon_P \,\|\, \epsilon_G]$, the aligned reverse noise is $\epsilon_r = [\epsilon_G \,\|\, \epsilon_P]$. This preserves tokenwise correspondence between the two views after semantic realignment.

Fixed-slot ablation. The alternative leaves the semantic roles at their forward coordinates,

$$\mathcal{X}_r^{\text{fixed}} = \mathcal{J}(X_P^r, X_G^r) = [X_P^r \,\|\, X_G^r], \tag{21}$$

and expresses the reversed direction through the reverse task condition and the location of the active mask. The corresponding latent, noise, condition, mask, and guidance tensors all retain this identity ordering. This tests whether task and mask conditioning alone can disambiguate direction, or whether explicitly placing the requested role in a consistent output position is beneficial.

The implemented fixed-slot variant is hybrid for P2P: P2P retains physical swapping because its reverse mask can represent different target extents across samples. Enforcing a fixed position in that case would vary the position–mask meaning within the same branch. Consequently, its P2P score measures the effect of training the shared checkpoint with the hybrid construction; the direct positional intervention applies to C2P and C2M.

Table D: Reverse-canvas construction on VITON-HD. C2P/P2P use unpaired FID; C2M uses no-pose MP-LPIPS. Both variants retain full task coverage.

Variant	C2P unpaired FID ↓	P2P unpaired FID ↓	Try-Off FID ↓	C2M MP-LPIPS ↓
SymFashion	8.89	9.37	8.20	0.226
SymFashion (hybrid fixed-slot)	9.44	9.56	8.94	0.231

Semantic realignment. The symmetric objective never compares raw left/right positions. After the reverse prediction, we first recover the semantic estimates

$$(\hat{z}_{0,P}^r, \hat{z}_{0,G}^r) = \begin{cases} (\text{right}(\hat{z}_0^r), \text{left}(\hat{z}_0^r)), & \text{swap}, \\ (\text{left}(\hat{z}_0^r), \text{right}(\hat{z}_0^r)), & \text{fixed slot}, \end{cases} \tag{22}$$

and then apply the same role-aligned consistency objective defined in Sec. 3. Hence the ablation changes only how the reverse view is presented to the transformer, not which semantic quantities are coupled. The C2P reverse branch requests the garment role and therefore realizes the Try-Off direction within the same construction.

Table D shows a consistent positive trend for physical role swapping. Relative to the hybrid fixed-slot variant, it lowers C2P and P2P unpaired FID from 9.44 and 9.56 to 8.89 and 9.37, Try-Off FID from 8.94 to 8.20, and C2M MP-LPIPS from 0.231 to 0.226. This supports explicit output-position alignment across the four reported measurements, while the hybrid P2P qualification above still applies.

Takeaway. This ablation establishes that presenting compatible directions with explicit output-position alignment improves all four reported task readouts over the hybrid fixed-slot construction. It supports physical role swapping as part of the coupling design, while the retained P2P swap prevents interpreting its P2P entry as a direct fixed-slot intervention.

D POSE-CONTROL EVALUATION

Pose control requires a separate controlled evaluation because the external methods and SymFashion use different pose representations. For each source garment, we keep the model and all non-pose inference conditions fixed and evaluate three settings: no pose condition, a shuffled pose drawn from another test identity, and the intended target pose. An independent pose estimator evaluates every attempted output, and detection failures are reported rather than excluded from the evaluation.

Table E: Target-pose evaluation.

Dataset	Pose condition (reference)	PCK@.05 ↑	PCK@.10 ↑	OKS ↑	Detection ↑
VITON-HD	No pose (vs. target)	0.258	0.610	0.249	0.989
VITON-HD	Shuffled (vs. target) ↓	0.232	0.584	0.235	0.999
VITON-HD	Shuffled (vs. donor) ↑	0.970	0.994	0.854	0.999
VITON-HD	Correct pose (vs. target)	0.977	0.996	0.864	0.998
DressCode	No pose (vs. target)	0.418	0.660	0.288	0.798
DressCode	Shuffled (vs. target) ↓	0.422	0.720	0.312	0.952
DressCode	Shuffled (vs. donor) ↑	0.933	0.968	0.749	0.953
DressCode	Correct pose (vs. target)	0.947	0.973	0.790	0.957

"Target" = the sample's own true photo (its intended pose); "donor" = the true photo of the sample whose DensePose was substituted in under Shuffled. Pose control is evidenced by Shuffled scoring low vs. target and high vs. donor, both approaching the Correct-pose row.

Table E evaluates whether the generated person follows the supplied geometry rather than merely remaining visually plausible under pose conditioning. Comparing the no-pose, shuffled-pose, and correct-pose settings provides a controlled intervention on the geometric condition. Since the external baselines use OpenPose whereas SymFashion uses DensePose, we treat this internal intervention as the primary evidence for pose-condition usage and retain the external results as a capability-level comparison rather than a strictly matched pose-control ablation.

The no-pose row is the reference condition, the shuffled-pose rows test whether the model responds to a geometric signal that is unrelated to the target, and the correct-pose row supplies the intended target geometry. The shuffled condition is scored twice against two different references, since a single score cannot distinguish "the model ignores the pose input" from "the model follows it": *vs. target* compares the output to the sample's own true photo (its originally intended pose, which was withheld under this condition), and *vs. donor* compares the same output to the true photo of whichever other test identity's DensePose was substituted in. If the model genuinely uses the pose condition, the output should track the donor pose, so the vs. target row should score low (close to or below No pose) while the vs. donor row should score high (approaching Correct pose); if the model instead ignores DensePose and just extrapolates from the person photo, the two shuffled rows would land close to each other instead. PCK and OKS measure keypoint-level alignment, while Detection records whether an output can be scored reliably. Garment-preservation metrics remain in the main C2M comparison rather than being repeated here. Since the external methods use OpenPose and SymFashion uses DensePose,

Table F: Task coverage of prior methods and SymFashion. Checkmarks denote capabilities evaluated by the cited method; evaluation protocols may differ across papers. ‡ denotes a reported capability without a public checkpoint available for our reevaluation.

Method	C2P	P2P	C2M	Try-Off
Virtual try-on methods				
IDM-VTON (Choi et al., 2024)	✓	–	–	–
OOTD (Xu et al., 2025)	✓	–	–	–
CatVTON (Chong et al., 2025)	✓	✓	–	–
StableVITON (Kim et al., 2024)	✓	–	–	–
CAT-DM (Zeng et al., 2024)	✓	–	–	–
ITA-MDT (Hong et al., 2025)	✓	–	–	–
OmniVTON (Yang et al., 2025)	✓	✓	–	–
OmniTry (Feng et al., 2026)	✓	–	–	–
PromptDresser (Kim et al., 2025)	✓	–	–	–
Leffa (Zhou et al., 2025)	✓	–	–	–
RefTON (Li et al., 2026a)	✓	✓	–	–
DiT-VTON (Li et al., 2026b)	✓	–	–	–
Unified and bidirectional fashion methods				
Voost (Lee & Kwak, 2025)	✓	–	–	✓
CCVTON‡ (Wu et al., 2026)	✓	–	–	✓
UniFit‡ (Zhang et al., 2026)	✓	✓	✓	✓
OrthoTryOn (Yang et al., 2026)	✓	–	–	✓
OMFA (Liu et al., 2025)	✓	✓	–	✓
Any2Any (Guo et al., 2025)	✓	–	✓	✓
Garment-driven person generation methods				
Magic Clothing (Chen et al., 2024)	–	–	✓	–
StableGarment (Wang et al., 2024)	–	–	✓	–
Try-Off and inverse try-on methods				
TryOffDiff (Velioglu et al., 2025b)	–	–	–	✓
MGT (Velioglu et al., 2025a)	–	–	–	✓
Inverse VTO (Lobba et al., 2025)	–	–	–	✓
SymFashion (Ours)	✓	✓	✓	✓

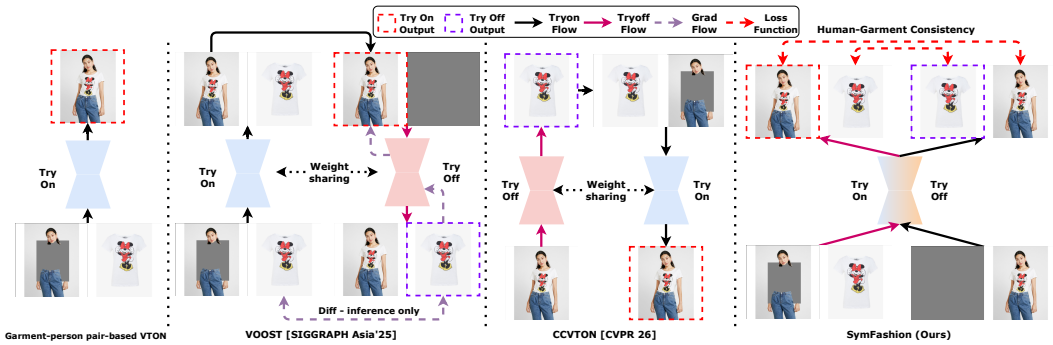

Figure B: Cross-task interaction in Voost, CCVTON, and SymFashion. SymFashion couples parallel role-aligned clean endpoints rather than composing generated images sequentially.

the within-SymFashion intervention is the controlled result; cross-method numbers should be read as capability-level evidence.

The measured values support this interpretation on both datasets. On VITON-HD, the shuffled output scores only 0.232 PCK@0.05 and 0.235 OKS against the target pose, but 0.970 PCK@0.05 and 0.854 OKS against the donor pose; the corresponding correct-pose row reaches 0.977 and 0.864. On DressCode, the same donor-versus-target separation is observed: the shuffled output scores 0.933 PCK@0.05 and 0.749 OKS against the donor, compared with 0.422 and 0.312 against the original target, while the correct-pose row reaches 0.947 and 0.790. The pattern is therefore driven by the supplied pose condition rather than by a generic increase in pose-estimation success. The no-pose and shuffled-vs-target rows are not expected to be identical across datasets; they provide an absolute reference, whereas the donor-versus-target separation is the direct intervention test.

Takeaway. This intervention establishes that changing only the DensePose condition moves the generated geometry toward the supplied donor or target pose on both datasets. It therefore verifies that SymFashion uses its pose input; garment fidelity and image quality remain separate questions, and the result is not a matched comparison against OpenPose-based external methods.

E Supplemental Comparisons

E.1 Relationship to Bidirectional Fashion Systems

Figure B separates three forms of interaction between forward and reverse fashion transformations. Voost shares parameters across Try-On and Try-Off objectives and can optionally apply an inference-time self-correction loop. CCVTON composes the two directions sequentially in image space during training. SymFashion instead evaluates compatible views in parallel and directly couples their role-aligned clean-endpoint estimates in one training step. The comparison concerns where and how task interaction is imposed, rather than claiming identical task coverage or training protocols.

E.2 Task Coverage

Table F summarizes the task coverage of SymFashion relative to prior fashion-generation methods. SymFashion supports C2P, P2P, C2M, and Try-Off with a single set of model weights. The table is a capability summary rather than a performance ranking, since task definitions, conditioning interfaces, training data, and evaluation protocols differ across prior work.

The checkmarks denote transformation interfaces reported and evaluated by the cited work; they do not imply equal training conditions or equal output quality. Nor do they imply identical inference interfaces: UniFit uses MLLM-aligned natural-language instructions, whereas SymFashion uses predefined task indicators and structured visual conditions without a user-provided description. A double dagger marks a reported capability for which no public checkpoint was available for reevaluation in our pipeline.

Table G: Full C2P/P2P results on VITON-HD and DressCode. Paired P2P measures self-reference fidelity; unpaired FID/KID are the primary transfer comparison. KID is $\times 10^3$; blue parentheses show favorable coupling changes, with (=) denoting a tie. Symbols follow Tab. 2; **Bold** and underlined mark first and second overall; ties share a rank.

Setting	Method	VITON-HD						DressCode					
		Paired				Unpaired		Paired				Unpaired	
		LPIPS ↓	SSIM ↑	FID ↓	KID ↓	FID ↓	KID ↓	LPIPS ↓	SSIM ↑	FID ↓	KID ↓	FID ↓	KID ↓
Cloth-to-Person (C2P)													
Baselines	CAT-DM* (Zeng et al., 2024)	0.104	0.855	7.70	0.83	10.26	–	–	–	–	–	–	–
	IDM-VTON* (Choi et al., 2024)	0.102	0.881	6.33	0.99	9.16	1.22	0.081	0.883	6.11	1.98	7.71	2.62
	OOTD* (Xu et al., 2025)	0.081	0.878	8.81	0.82	9.31	0.97	0.076	0.888	5.92	1.44	7.21	2.31
	CatVTON* (Chong et al., 2025)	0.108	0.834	8.85	0.41	11.36	1.09	0.083	0.875	5.12	1.02	6.54	1.83
	OmniVTON* (Yang et al., 2025)	0.145	0.832	7.76	–	9.62	–	0.119	0.865	5.33	–	6.45	–
	Any2Any* (Guo et al., 2025)	0.091	0.839	6.93	0.74	8.97	0.98	0.087	0.869	6.54	1.56	7.64	2.28
	OrthoTryOn* (Yang et al., 2026)	0.120	0.837	10.69	5.80	12.68	5.80	–	–	–	–	–	–
	OmniTry* (Feng et al., 2026)	0.186	0.769	17.81	10.08	20.97	12.01	0.117	0.851	8.41	4.63	10.87	5.81
	ITA-MDT* (Hong et al., 2025)	0.083	0.876	7.77	0.87	9.83	0.96	0.062	0.892	3.55	1.64	5.52	1.89
	PromptDresser* (Kim et al., 2025)	0.151	0.819	11.15	1.16	13.06	0.77	0.087	0.874	4.89	1.19	7.48	2.08
	Leffa* (Zhou et al., 2025)	0.081	0.857	6.31	1.21	9.03	0.69	0.073	0.877	4.81	1.22	6.61	1.78
	Voost^R (Lee & Kwak, 2025)	0.079	0.868	5.81	0.62	9.11	1.05	0.061	0.898	3.65	0.57	5.31	1.13
	OMFA* (Liu et al., 2025)	0.194	0.739	20.07	11.30	22.42	12.70	0.305	0.716	25.92	16.69	26.08	16.84
	RefTON* (Li et al., 2026a)	0.079	0.879	8.34	1.93	9.11	0.91	0.060	0.902	3.48	1.20	7.01	3.15
Full	Ours w/o sym.	**0.078**	0.881	6.19	0.61	9.39	1.10	0.062	0.901	2.64	0.23	4.93	1.00
	Ours w/ sym.	**0.078** (=)	**0.885** (+0.004)	**5.54** (+0.65)	**0.12** (+0.49)	**8.89** (+0.50)	0.65 (+0.45)	**0.054** (+0.008)	**0.907** (+0.006)	**2.41** (+0.23)	**0.11** (+0.12)	**4.51** (+0.42)	**0.51** (+0.49)
C2P + Try-Off	Ours w/o sym.	0.096	0.859	6.27	0.53	9.57	1.20	0.062	0.903	2.58	0.20	4.95	0.99
	Ours w/ sym.	0.092 (+0.004)	0.862 (+0.003)	5.68 (+0.59)	0.18 (+0.35)	8.95 (+0.62)	**0.50** (+0.70)	0.059 (+0.003)	0.905 (+0.002)	2.43 (+0.15)	0.12 (+0.08)	4.55 (+0.40)	0.70 (+0.29)
C2P Only	Ours	0.089	0.862	5.71	0.22	9.16	0.61	0.058	0.905	2.51	0.17	4.78	0.93
Person-to-Person (P2P)													
Baselines	StableVITON† (Kim et al., 2024)	0.098	0.876	7.01	0.92	10.35	1.96	–	–	–	–	–	–
	CAT-DM† (Zeng et al., 2024)	0.153	0.877	10.71	2.07	12.44	2.51	–	–	–	–	–	–
	ITA-MDT† (Hong et al., 2025)	0.078	0.881	8.02	3.88	11.91	4.61	0.061	0.913	3.31	0.74	5.41	1.34
	CatVTON* (Chong et al., 2025)	0.078	0.870	8.59	5.38	14.13	6.88	0.032	0.941	1.65	0.202	5.42	1.52
	OMFA* (Liu et al., 2025)	0.193	0.743	16.30	8.30	17.44	8.70	0.316	0.707	29.94	20.64	30.25	20.94
Full	Ours w/o sym.	0.014	0.974	2.11	0.50	9.60	1.31	0.014	0.975	0.96	0.087	4.59	0.85
	Ours w/ sym.	**0.012** (+0.002)	**0.984** (+0.010)	1.99 (+0.12)	0.20 (+0.30)	**9.37** (+0.23)	**1.27** (+0.04)	**0.011** (+0.003)	**0.988** (+0.013)	**0.90** (+0.06)	0.024 (+0.063)	**4.38** (+0.21)	**0.60** (+0.25)
P2P Only	Ours w/o sym.	0.014	0.979	2.23	0.56	9.64	1.42	0.016	0.976	1.05	0.089	4.74	0.90
	Ours w/ sym.	**0.012** (+0.002)	0.983 (+0.004)	**1.97** (+0.26)	**0.17** (+0.39)	9.63 (+0.01)	1.32 (+0.10)	**0.011** (+0.005)	0.984 (+0.008)	0.91 (+0.14)	**0.023** (+0.066)	4.59 (+0.15)	0.74 (+0.16)

Table H: Try-Off results on VITON-HD and DressCode. KID values are multiplied by 10^3. * denotes a released checkpoint reevaluated with our pipeline, and ^{R} our Try-On/Try-Off-only Voost reproduction; **bold** and underlined entries mark the best and second-best overall values, with ties sharing a rank. Blue parentheses show favorable coupling changes.

Setting	Method	VITON-HD								DressCode							
		SSIM ↑	MS-SSIM ↑	CW-SSIM ↑	LPIPS ↓	FID ↓	CLIP-FID ↓	KID ↓	DISTS ↓	SSIM ↑	MS-SSIM ↑	CW-SSIM ↑	LPIPS ↓	FID ↓	CLIP-FID ↓	KID ↓	DISTS ↓
Baselines	TryOffDiff* (Velioglu et al., 2025b)	0.665	0.513	0.681	0.413	21.08	6.78	7.46	0.249	–	–	–	–	–	–	–	–
	MGT* (Velioglu et al., 2025a)	0.667	0.501	0.678	0.471	24.39	7.31	10.40	0.376	0.759	0.642	0.524	0.305	11.35	6.33	3.40	0.218
	Inverse Virtual Try-On* (Lobba et al., 2025)	0.648	0.507	0.673	0.415	21.28	6.84	5.60	0.265	0.759	0.668	0.398	0.333	12.68	2.71	4.70	0.236
	Voost^R (Lee & Kwak, 2025)	0.689	0.620	0.715	0.310	10.06	1.64	2.48	0.202	0.704	0.649	0.743	0.295	5.34	0.99	0.52	0.202
	OMFA* (Liu et al., 2025)	0.661	0.612	0.718	0.299	10.18	3.15	2.53	0.198	0.606	0.461	0.652	0.438	30.09	5.87	13.79	0.278
	Any2Any* (Guo et al., 2025)	0.672	0.554	0.695	0.385	18.76	5.64	5.90	0.253	0.662	0.538	0.691	0.373	18.78	4.09	6.30	0.256
	OrthoTryOn* (Yang et al., 2026)	0.690	0.628	0.714	0.294	11.06	3.69	2.10	0.193	–	–	–	–	–	–	–	–
Full	Ours w/o sym.	0.699	0.642	0.735	0.278	8.55	1.49	0.93	0.194	0.725	0.666	0.726	0.267	5.02	0.828	0.49	0.192
	Ours w/ sym.	**0.715** (+0.016)	**0.659** (+0.017)	**0.775** (+0.040)	**0.232** (+0.046)	**8.20** (+0.35)	**1.38** (+0.11)	0.90 (+0.03)	**0.184** (+0.010)	**0.761** (+0.036)	**0.705** (+0.039)	**0.781** (+0.055)	0.248 (+0.019)	**4.41** (+0.61)	0.703 (+0.125)	0.30 (+0.19)	0.187 (+0.005)
C2P + Try-Off	Ours w/o sym.	0.704	0.644	0.733	0.281	8.61	1.47	1.04	0.191	0.738	0.686	0.770	0.249	4.71	0.727	0.32	0.186
	Ours w/ sym.	0.711 (+0.007)	0.658 (+0.014)	0.744 (+0.011)	0.263 (+0.018)	8.54 (+0.07)	1.41 (+0.06)	0.98 (+0.06)	0.187 (+0.004)	0.751 (+0.013)	0.702 (+0.016)	**0.781** (+0.011)	**0.231** (+0.018)	4.46 (+0.25)	**0.701** (+0.026)	**0.27** (+0.05)	**0.178** (+0.008)
C2M + Try-Off	Ours w/o sym.	0.692	0.635	0.721	0.293	8.93	1.68	0.99	0.196	0.717	0.660	0.747	0.282	5.15	0.812	0.49	0.199
	Ours w/ sym.	0.709 (+0.017)	0.655 (+0.020)	0.738 (+0.017)	0.269 (+0.024)	8.45 (+0.48)	1.51 (+0.17)	**0.84** (+0.15)	0.189 (+0.007)	0.736 (+0.019)	0.688 (+0.028)	0.766 (+0.019)	0.248 (+0.034)	4.74 (+0.41)	0.763 (+0.049)	0.41 (+0.08)	0.183 (+0.016)
Try-Off Only	Ours	0.684	0.623	0.714	0.303	9.00	1.77	0.96	0.199	0.711	0.652	0.742	0.285	5.47	0.901	0.68	0.200

Takeaway. This comparison establishes the coverage part of RQ1: one SymFashion checkpoint exposes all four human–garment transformation interfaces without requiring a user description. Because prior interfaces and protocols differ, the table documents capability rather than performance; the matched tables carry the quantitative claims.

E.3 Full Quantitative Results

Table G restores all C2P baselines omitted from the representative main-paper view and retains the complete P2P comparison. No method or value is changed between the compact and full versions.

Table H provides the complete Try-Off metric breakdown behind the four representative columns in the main paper. LPIPS, FID, CLIP-FID, KID, and DISTS are lower-is-better, whereas SSIM, MS-SSIM, and CW-SSIM are higher-is-better. KID is displayed after multiplication by 10^3. A superscript star identifies a released checkpoint reevaluated with our evaluator, and ^{R} marks our Try-On/Try-Off-only Voost reproduction. Missing cells denote unavailable outputs or incompatible evaluation interfaces rather than zero scores.

Try-Off results in Table H extend the evaluation to the reverse generation direction. The full checkpoint achieves strong overall performance across the reported VITON-HD and DressCode metrics,

with the relative ranking depending on the metric and dataset. This indicates that sharing the checkpoint with the forward tasks does not preclude competitive reverse-direction generation.

Takeaway. The full C2P/P2P table preserves the rankings reported in the compact main view. Across all expanded Try-Off metrics, the complete checkpoint also exceeds the evaluated external baselines, while task-pair variants attain some best overall values. Thus, unified training remains competitive beyond the selected main-paper readouts; missing baseline cells denote unavailable measurements, not negative results.

E.4 Qualitative Comparison

Figures C–G compare the different transformation and conditioning settings. Within each row, all methods receive the same available task inputs. The C2P panel provides the target person and product garment; the P2P panel provides the target person and a source person wearing the desired garment; and the Try-Off panel includes the paired canonical garment reference. The two C2M panels separately evaluate generation without and with target-pose conditioning.

Qualitative findings. Across the displayed examples, the full symmetric checkpoint generally shows a stronger balance of garment preservation, output realism, and condition adherence than its matched non-symmetric counterpart. In C2P (Fig. C), symmetric coupling more consistently preserves fine stripe spacing, neckline geometry, printed lettering, sleeve and torso proportions, and the silhouettes of dresses and lower-body garments while retaining the target person. By comparison, the matched variants without symmetric coupling occasionally alter garment structure, for example by changing a sleeveless neckline, introducing an unsupported outer layer, or modifying the hem and panel structure of long and asymmetric garments. These examples suggest that the benefit is not restricted to upper-body appearance transfer but also extends to dress and lower-body structure.

In P2P (Fig. D), the symmetric variants more consistently retain chest graphics, layered outerwear, patterned garments, and overall garment structure while preserving the target-person layout. Relative to the corresponding non-symmetric variants, they also show fewer examples of spurious sleeves, garment-layer changes, and target-outfit drift.

For Try-Off (Fig. E), the symmetric model more closely preserves the structure of the paired canonical garments in the displayed cases, including logo-shirt silhouettes, vertical stripe layouts, and ruffled-dress structure. Removing symmetric coupling can instead distort lettering, introduce unsupported garment elements, or rearrange layered details. Crossed cells indicate cases for which the corresponding baseline output is unavailable.

The C2M results show a similar tendency. Without target-pose conditioning (Fig. F), the full symmetric checkpoint generates plausible people while retaining high-frequency prints, knit textures, asymmetric cuts, and multicolor patterns. Under target-pose conditioning (Fig. G), it provides a consistent joint match to the supplied garment and body geometry across the displayed examples. Across the matched task configurations, symmetric coupling tends to reduce crop and silhouette drift, garment-category changes, and the loss of recognizable garment details relative to the corresponding non-symmetric models.

The paired *w/ sym* and *w/o sym* columns therefore provide a visual counterpart to RQ2: the observed improvements concern not only image quality, but also more stable preservation of the shared human–garment state across task views. These examples complement, rather than replace, the aggregate quantitative evaluation in the main paper. The panels are illustrative rather than a random-sample estimate, so all visual observations are restricted to the displayed examples; the quantitative tables carry the dataset-level claims.

F Experimental and Implementation Details

Implementation backbone. We instantiate SymFashion with FLUX.1-Fill-dev (Black Forest Labs, 2024) as the single shared rectified-flow transformer for all four task interfaces. We fine-tune its transformer attention layers and input/output projections while keeping the remaining pretrained parameters frozen.



Figure C: Qualitative comparison for cloth-to-person try-on (C2P). Each row fixes the target person and product garment. The examples span upper-body garments, long dresses, and structured lower-body garments, highlighting fine-detail preservation, identity retention, localized editing, and silhouette fidelity across external methods and SymFashion configurations. Crossed cells denote unavailable baseline outputs.

F.1 Training and Evaluation Protocol

We train on the official VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022) training splits, augmented with task-compatible samples from DeepFashion (Liu et al., 2016) and a curated LAION-based LRVS-Fashion subset (Lepage et al., 2023). Evaluation is restricted to the official VITON-HD and DressCode test splits. C2P and P2P use paired LPIPS (Zhang et al., 2018), SSIM (Wang et al., 2004), FID and KID (Heusel et al., 2017; Bińkowski et al., 2018); unpaired evaluation uses FID and KID. Try-Off uses perceptual, structural, and distributional garment-reference fidelity metrics, while C2M uses garment-preservation and semantic-similarity measures. KID values are multiplied by 10^3 for readability; the expanded Try-Off metric set is reported below.

For P2P, the target person is supplied in agnostic form, while the garment-bearing condition is a full person image accompanied by its predicted garment-location mask. The paired protocol uses the target's own original unmasked photograph as this condition and evaluates against the same photograph. We therefore interpret paired LPIPS, SSIM, FID, and KID as aligned self-reference fidelity rather than identity-disjoint garment transfer.

In the unpaired protocol, the garment-bearing condition is drawn from a different source-image ID in the official unpaired list. Because no pixel-aligned photograph of the target person wearing that source garment exists, this protocol is evaluated with FID and KID and provides the primary evidence for P2P transfer. Both protocols use the same checkpoint, resolution, sampling schedule, and evaluator.

Construction and sampling of compatible pairs. Each training item supplies two role-aligned views of the same underlying person–garment example; the reverse view is not formed by drawing an unrelated second item. C2P and C2M items deterministically use Try-Off as the reverse direction. For worn-garment P2P items, the two person images exchange source and target roles; a flat-reference fallback, when present, is labeled as Try-Off rather than treated as a P2P reverse. Dedicated Try-Off items draw uniformly from three compatible reverse configurations: one C2P-style inpainting configuration and two C2M-style full-generation configurations with and without pose guidance. The full loader contains 136,311 C2P, 89,799 P2P, 136,311 Try-Off, and 136,311 C2M



Figure D: Qualitative comparison for person-to-person garment transfer (P2P). Each row fixes the target person and a source person wearing the transferred garment, exposing both garment-extraction errors and target-person editing artifacts. Crossed cells denote unavailable outputs for the corresponding baseline.



Figure E: Qualitative comparison for virtual Try-Off. The first column shows the dressed-person input and paired canonical garment reference. The remaining columns compare garment silhouette, printed details, and layered structure; crossed cells denote unavailable baseline outputs.

items (498,732 total), corresponding to approximate example-level proportions of 27.3%, 18.0%, 27.3%, and 27.3%. It concatenates these task datasets and shuffles examples globally, so exposure follows these proportions rather than explicit task balancing. Both views use the same sampled timestep within an optimizer step. For C2M examples, the *Dense Person* condition is independently enabled or replaced by zeros with equal probability. This trains the same checkpoint for the pose-controlled and no-pose interfaces shown in Fig. 2; they are not separately trained C2M models.

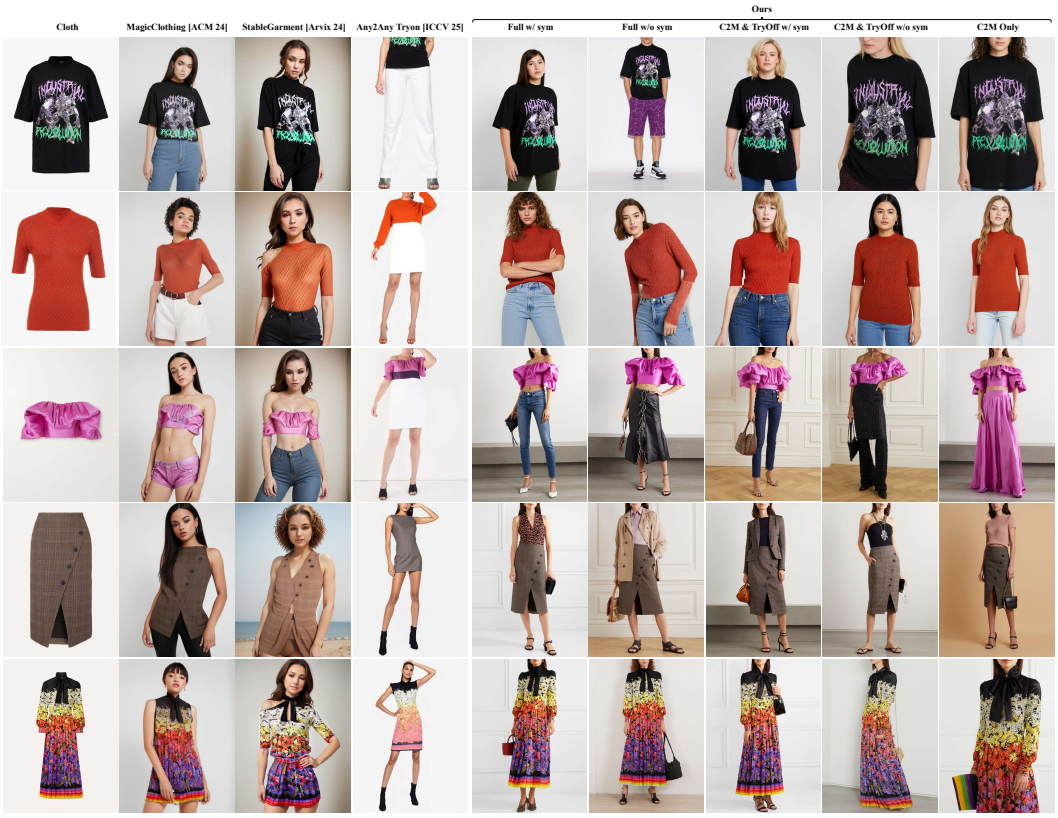

Figure F: Qualitative comparison for cloth-to-model generation (C2M) without a target-pose condition. Each row fixes the product garment across external methods and SymFashion configurations, while person appearance and pose are otherwise generated by each method.

Task and ClothType text conditions. The *Task Token* and *ClothType Token* in Fig. 2 are fixed categorical text conditions, not user-written descriptions. The exact task strings are `TryOn` for C2P and P2P, `TryOff` for Try-Off, and `A [person] wear the [garment]` for C2M. C2P and P2P deliberately share the `TryOn` token: the product-garment versus worn-person reference, together with the *Dense Garment* and *Mask Garment* conditions in Panel (A), distinguishes these two interfaces. Whether C2M uses pose guidance does not change its task string; that choice is represented by *Dense Person*.

The ClothType condition is selected from four fixed strings: `inner shirt or inner t-shirt or top, or bra` for upper-body garments, `shorts, or pants, or skirt, or trousers, or underwear` for lower-body garments, `dresses, or jumpsuits, or overalls` for full-body garments, and `jackets, or blazers, or outer shirt, or long padding jackets` for outerwear. It is obtained from the dataset category during training and evaluation, and from the requested garment category for user-provided inputs. Compatible forward and reverse views retain the same ClothType condition but receive the Task Token of their respective requested direction; for example, a C2P–Try-Off pair uses `TryOn` and `TryOff`, respectively.

Source and use of garment masks. The spatial weights in Eq. 4 use two mask sources matched to the input representation. For a person image, including a worn-person garment condition, the garment region is obtained from Sapiens human parsing (Khirodkar et al., 2024). For a flat product image, the garment foreground is obtained with SAM (Kirillov et al., 2023). Person-image masks are adjusted to exclude overlapping hair and bag regions. All masks are binarized at 512×512, resized to latent resolution, and clamped to $[0, 1]$; resizing can therefore produce soft boundary values. The resulting w^P, w^G tensors weight only the symmetric training objective and are neither model inputs nor required when sampling. The ordinary task-specific inpainting-mask preprocessing is unchanged at inference.

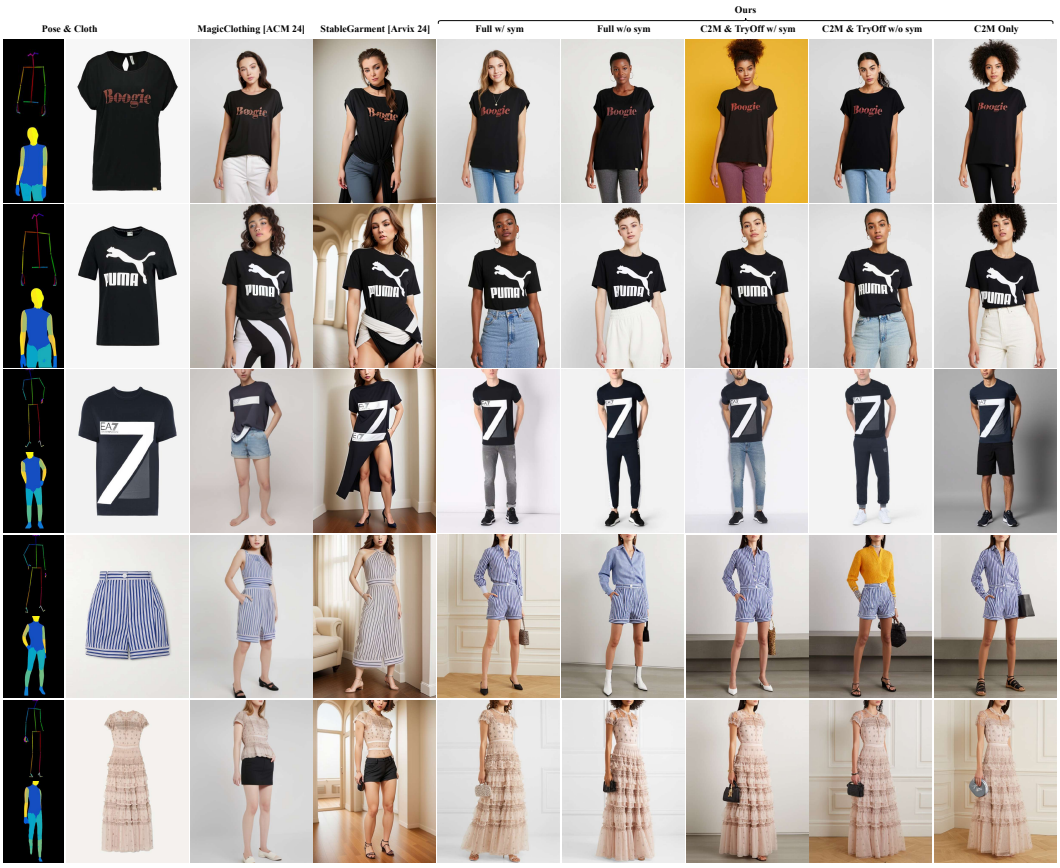

Figure G: Qualitative comparison for pose-controlled C2M. Each row provides a target pose and product garment and compares pose adherence, garment-detail preservation, and human plausibility across external methods and SymFashion configurations. External methods use OpenPose, whereas SymFashion uses its supported DensePose interface.

Main/task-composition symmetric variants use Eq. 8; the prediction-space controls in Table 4 instead use Eq. 17. Matched non-symmetric counterparts use the same forward and reverse flow objectives, backbone, task definitions, data splits, and evaluation pipeline, differing only by $\lambda_s = 0$. For controlled P2P comparisons, StableVITON (Kim et al., 2024), CAT-DM (Zeng et al., 2024), and ITA-MDT (Hong et al., 2025) are retrained using their original architectures on the same task-compatible data mixture, with garment inputs extracted using SCHP (Li et al., 2020). Our Voost reproduction uses this same data mixture but is trained only on its native Try-On and Try-Off tasks; all remaining baseline results use released checkpoints. All available metrics are recomputed with the same preprocessing, test IDs, and evaluator, while missing entries indicate unavailable outputs or incompatible settings. Before full-reference scoring, every generated and matched reference image is decoded as RGB and resized to 512×512 with the same Lanczos interpolation. FID and KID are computed from the corresponding saved 512×512 generated and real-reference sets; no method-specific metric preprocessing is applied. Within each matched SymFashion comparison, the test manifest, output resolution, sampling schedule, and guidance scale are held fixed; the training-time coupling weight is the only intervention between the corresponding symmetric and non-symmetric rows.

F.2 Metric and Model Complexity Details

Model complexity. Table I reports the exact parameter counts of the shared transformer and the canonical trainable checkpoint. The BF16 storage entry is the raw memory required by all transformer parameters, rather than peak device memory. The architecture uses a width-concatenated two-slot canvas: VAE downsampling and 2×2 latent packing produce $2{,}048$ image tokens, which are processed jointly with 154 conditioning tokens.

We measure compute at batch size one with an operator-level counter, using the paper's 512×512 per-tile resolution and counting one multiply–add as two FLOPs. The per-step count includes the transformer and its input/output projections; it excludes VAE encoding/decoding and input preprocessing and therefore should not be read as end-to-end runtime. The 30-step value is the per-step count multiplied by the fixed inference schedule. Symmetric coupling is used only during training and introduces no additional module or denoising branch at inference, so the reported sampling cost is shared by all four tasks and by the matched symmetric and non-symmetric variants.

Metric implementation. All image pairs are matched by sample identifier. Full-reference metrics include SSIM (Wang et al., 2004), MS-SSIM (Wang et al., 2003), CW-SSIM (Sampat et al., 2009), LPIPS (Zhang et al., 2018), and DISTS (Ding et al., 2020). Distribution metrics include FID (Heusel et al., 2017), CLIP-FID (Kynkäänniemi et al., 2022), and KID (Bińkowski et al., 2018). They compare the generated and matched real-reference sets separately for each dataset; KID is multiplied by 10^3 only for display. CLIP-I uses OpenAI CLIP ViT-B/32 image embeddings (Radford et al., 2021) and reports cosine similarity to the matched garment reference. MP-LPIPS (Chen et al., 2024), FFA (Kotar et al., 2023), and DiffSim (Song et al., 2025b) follow their corresponding definitions in the common evaluator; FFA uses DINOv2-base features (Oquab et al., 2023).

Table I: SymFashion complexity at 512^2 resolution.

Quantity	Value
Two-slot denoising canvas	512×1024
Image / conditioning tokens	2,048 / 154
Hidden width	3,072
Double / single-stream blocks	19 / 38
Total parameters	11.908B
Trainable parameters	2.530B (21.25%)
Frozen parameters	9.377B (78.75%)
BF16 parameter storage	22.18 GiB
Trainable checkpoint size	4.71 GiB
MACs / denoising step	15.92T
FLOPs / denoising step	31.84T
FLOPs / 30-step sample	0.955P

For pose adherence, an independent MediaPipe BlazePose estimator (Bazarevsky et al., 2020) detects 33 keypoints in both the generated image and its reference person image. PCK@t (Andriluka et al., 2014) is the fraction of jointly visible keypoints whose Euclidean error is below t times the reference-person bounding-box height, with $t \in \{0.05, 0.10\}$. Detection success is the fraction of attempted outputs for which both images yield a valid pose; detection failures remain in this denominator. For COCO-style OKS (Lin et al., 2014), the MediaPipe landmarks are mapped to the COCO-17 order (nose, eyes, ears, shoulders, elbows, wrists, hips, knees, and ankles) using indices $(0, 2, 5, 7, 8, 11, 12, 13, 14, 15, 16, 23, 24, 25, 26, 27, 28)$. We use the standard COCO per-joint sigmas, a visibility threshold of 0.5, and the area of the reference visible-keypoint bounding box for normalization. OKS is averaged over outputs with valid mapped keypoints, whereas Detection retains every attempted output in its denominator. The pose table retains only PCK@0.05, PCK@0.10, OKS, and Detection; garment and realism metrics remain in the main C2M comparison.

G Analytical Details

For a two-role latent, define

$$\|a\|_W^2 = \frac{1}{2}\left(\|a[P]\|_{w^P}^2 + \|a[G]\|_{w^G}^2\right), \tag{23}$$

and let $\langle a, b\rangle_W$ denote its corresponding weighted inner product. Substituting Eq. 3 in both branches gives

$$\hat{z}_0^f - \bar{\hat{z}}_0^r = \left(x_\sigma^f - \pi^{-1}(x_\sigma^r)\right) - \sigma\left(u^f - \bar{u}^r\right). \tag{24}$$

Using Eq. 6 yields the residual-aware identity

$$\hat{z}_0^f - \bar{\hat{z}}_0^r = (1-\sigma)\delta_0 - \sigma\Delta u, \qquad \Delta u = u^f - \bar{u}^r. \tag{25}$$

Taking the role-weighted squared norm and expanding the cross term gives

$$\begin{aligned} \ell_{\mathrm{sym}}^{x_0} &= \|(1-\sigma)\delta_0 - \sigma\Delta u\|_W^2 \\ &= (1-\sigma)^2\|\delta_0\|_W^2 + \sigma^2\|\Delta u\|_W^2 \\ &\quad - 2\sigma(1-\sigma)\langle\delta_0, \Delta u\rangle_W. \end{aligned} \tag{26}$$

The triangle inequality also gives

$$\sqrt{\ell_{\mathrm{sym}}^{x_0}} \le (1-\sigma)\|\delta_0\|_W + \sigma\|\Delta u\|_W. \tag{27}$$

If $\delta_0 = 0$, the residual term vanishes and

$$\ell_{\mathrm{sym}}^{x_0} = \sigma^2 \ell_{\mathrm{sym}}^{v}, \tag{28}$$

where ℓ_{sym}^{v} replaces the aligned endpoint differences in Eq. 8 with aligned velocity differences. Thus, the exact σ^2 relation is an aligned-latent special case, whereas Eq. 26 matches the implemented joint-canvas and role-wise VAE construction.

Let $z_0^\star$ denote the common role-wise reference latent used by the mechanism evaluator and define

$$\begin{aligned} \mathcal{R}_f^{x_0} &= \mathbb{E}\|\hat{z}_0^f - z_0^\star\|_W^2, \\ \mathcal{R}_r^{x_0} &= \mathbb{E}\|\bar{\hat{z}}_0^r - z_0^\star\|_W^2, \\ \mathcal{C}^{x_0} &= \mathbb{E}\|\hat{z}_0^f - \bar{\hat{z}}_0^r\|_W^2 = \mathbb{E}\ell_{\mathrm{sym}}^{x_0}. \end{aligned} \tag{29}$$

Writing $\hat{z}_0^f - z_0^\star = (\hat{z}_0^f - \bar{\hat{z}}_0^r) + (\bar{\hat{z}}_0^r - z_0^\star)$ and applying Minkowski's inequality in the role-weighted L_2 space over evaluation samples gives

$$\sqrt{\mathcal{R}_f^{x_0}} \le \sqrt{\mathcal{R}_r^{x_0}} + \sqrt{\mathcal{C}^{x_0}}, \tag{30}$$

This is a one-way upper bound. Its right-hand side $B = \sqrt{\mathcal{R}_r^{x_0}} + \sqrt{\mathcal{C}^{x_0}}$ is used only as a descriptive diagnostic: a smaller B across models need not imply a smaller $\mathcal{R}_f^{x_0}$, and agreement can preserve a shared error.

For completeness, let $\bar{v}_r^* = \epsilon - z_0^\star$ denote the aligned reverse velocity target. Equation 3 gives $\bar{\hat{z}}_0^r - z_0^\star = -\sigma(\bar{u}^r - \bar{v}_r^*)$. With $\kappa = \sup_{a \neq 0} \|a\|_W^2 / \|a\|_2^2$, finite for the bounded garment-mask weights, define the reverse flow risk on the same mechanism-evaluation distribution as

$$\mathcal{R}_{\mathrm{flow},r}^{\mathrm{eval}} = \mathbb{E}_{\mathrm{eval}}\|\bar{u}^r - \bar{v}_r^*\|_2^2.$$

Since $\sigma^2 \le 1$, we have $\mathcal{R}_r^{x_0} \le \kappa \mathcal{R}_{\mathrm{flow},r}^{\mathrm{eval}}$. Hence,

$$\sqrt{\mathcal{R}_f^{x_0}} \le \sqrt{\kappa \mathcal{R}_{\mathrm{flow},r}^{\mathrm{eval}}} + \sqrt{\mathbb{E}_{\mathrm{eval}} \ell_{\mathrm{sym}}^{x_0}}. \tag{31}$$

Here $z_0^\star$ is the aligned reverse target in the evaluator, and $\mathbb{E}_{\mathrm{eval}} \ell_{\mathrm{sym}}^{x_0} = \mathcal{C}^{x_0}$. The two evaluation risks have the same pointwise forms as the reverse supervised flow and per-example symmetric training terms, respectively; we do not equate their expectations with the training losses without an additional generalization assumption. The bound is a one-way consequence under the stated assumptions; it does not exclude shared errors, forward-path encoding residuals, or finite optimization effects.

H LIMITATIONS

Evaluation is limited to VITON-HD and DressCode under a common 512×512 protocol; performance at native high resolution and on in-the-wild imagery remains untested. The external results are system-level comparisons under this common evaluator rather than reproductions of native-resolution published scores. Paired P2P measures aligned self-reference fidelity, whereas the current unpaired transfer evaluation uses distributional metrics and does not directly quantify garment identity. Pose-controlled comparisons also use each method's supported interface: OpenPose for the external methods and DensePose for SymFashion. The internal pose intervention therefore provides the strictly controlled evidence. Finally, compatible task mixtures share a fixed total optimization budget, causing per-task exposure to vary with mixture composition, and the observed coherence gain does not imply lower branchwise prediction risk.



Figure H: Representative C2P, Try-Off, and P2P failure cases.

I FAILURE ANALYSIS

Figure H shows three remaining limitations. C2P hallucinates an unsupported bracelet-like accessory. In Try-Off, tuck-in hides the lower shirt, so its exact stripe count is underdetermined and extrapolated incorrectly. P2P transfers a back-view cue as a front-facing neckline, exposing unresolved front–back ambiguity. Thus, symmetry does not eliminate unsupported details or ambiguities from hidden regions and viewpoints. These examples show representative failure types; their prevalence is not estimated here.