BiCorr: Learning Bidirectional Latent Correspondences for Video Virtual Try-On

Anonymous authors
Paper under double-blind review

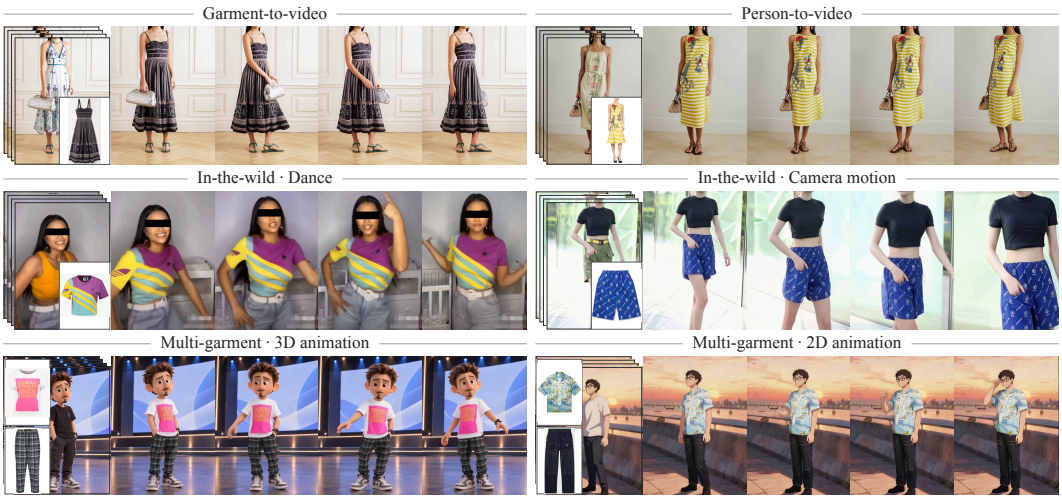

Figure 1: Video try-on results of BiCorr across reference types, motions, and visual styles. Eye regions of non-cartoon subjects are obscured throughout the paper for privacy.

Abstract

Video virtual try-on transfers a reference garment to a person in a video while preserving identity, motion, background, and garment appearance. Recent video diffusion Transformers have improved garment conditioning and temporal modeling, but fine-grained patterns and local garment structures can still be distorted, particularly under large motion and occlusion. Garment conditioning provides reference information but does not ensure that each person token retrieves relevant features from the correct garment region. To address this issue, we propose BiCorr, a framework for bidirectional latent correspondence learning between person-video and reference-garment tokens. We combine high-confidence direct matches with matches propagated along validated trajectories, and balance their coverage across time and space. These matches are mapped to the model's video and garment tokens and represented as distributional targets that account for mapping uncertainty while retaining distinct garment locations. The targets guide person tokens to retrieve features from corresponding garment regions, while a reverse objective associates garment tokens with their corresponding spatiotemporal locations throughout the video. All matching and correspondence losses are used only during training, requiring no additional matcher, tracker, or warping module at inference. Experiments show that BiCorr improves visual quality, garment fidelity, and temporal consistency.

1 Introduction

Video virtual try-on (VVT) aims to seamlessly transfer a reference garment onto a target individual in the video while preserving identity, motion, and background (Fang et al., 2024; He et al., 2025; Karras et al., 2024). The generated video must faithfully reproduce the garment's structure

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.

and appearance, including logos, text, and local patterns. However, maintaining this fidelity across frames is challenging due to non-rigid deformation, viewpoint changes, and self-occlusion. Recent VVT methods primarily address these difficulties by strengthening garment conditioning and temporal modeling. For garment conditioning, they use garment encoders and attention-based fusion to incorporate semantic, structural, and appearance cues (Fang et al., 2024; Li et al., 2025b). For temporal modeling, they introduce temporal modules (Fang et al., 2024; Karras et al., 2024), spatiotemporal attention (Zou et al., 2025) and keyframe guidance (He et al., 2026) to improve temporal consistency. Despite these advances, fine-grained garment details can still be distorted or misplaced. Such issues may arise when attention retrieves features from non-corresponding regions of the reference garment (Zhou et al., 2025; Kim et al., 2026). We therefore focus on explicitly learning person-garment correspondence, which links reference-garment regions to their locations on the person across frames.

Correspondence-aware guidance has been shown to improve local garment fidelity in image-based virtual try-on (Kim et al., 2024; Wan et al., 2025; Zhou et al., 2025). In particular, supervising attention with external person-garment matches directly targets the retrieval issues described above (Kim et al., 2026; Takemoto & Koshinaka, 2027). However, extending this strategy to our video model presents two challenges: balancing correspondence reliability with spatiotemporal coverage, and adapting frame-level matches to compressed latent tokens. First, viewpoint changes and occlusion make direct matching between video frames and the reference garment unreliable. Strict filtering improves reliability but may concentrate supervision on easily matched frames and distinctive garment regions. Relaxing it can broaden coverage but risks admitting incorrect matches. Second, mapping frame-level matches to our model's spatially and temporally compressed latent grid (Team Wan et al., 2025) can introduce discretization errors alongside existing local matching uncertainty. Temporal compression may also map observations from different frames to the same person-video token, even when they correspond to distinct garment regions.

To address these challenges, we propose BiCorr, a framework for bidirectional latent correspondence learning in a garment-conditioned video diffusion Transformer (DiT). We combine high-confidence direct matches with validated matches propagated through point tracking. These estimated matches serve as pseudo-correspondences, which we map to the person-video and reference-garment token grids. After selecting tokens for supervision to balance spatiotemporal coverage, we construct distributional targets as probability distributions over token locations. Gaussian components accommodate local matching and discretization errors. Distinct garment locations retained for the same person-video token are represented by separate components. Neighborhoods around the component centers define support regions for attention alignment.

These resulting targets supervise person-garment correspondence in both directions. Person-to-garment (P→G) supervision guides each selected person-video token to retrieve features from corresponding garment regions. For each token, the same target supervises both garment cross-attention and garment-prefix self-attention, where garment tokens are prepended to the person-video sequence. In the reverse garment-to-person (G→P) direction, the mapped observations are grouped by garment token. Each reverse target supervises self-attention from a selected garment token to its corresponding person-video locations across frames. Matching, tracking, and correspondence supervision are used only for training, adding no overhead to the garment-conditioned generator at inference. For evaluation, we introduce local high-frequency spectral consistency (LHFSC) as a complementary measure of local high-frequency spectral agreement between generated frames and ground truth. We also curate BiCorr-Test, a video test set covering diverse garment appearances.

In summary, our contributions are as follows:

- We introduce bidirectional latent correspondence learning for garment-conditioned video DiTs. Joint supervision of P→G retrieval and G→P localization improves garment-detail fidelity without additional inference overhead.
- We formulate Gaussian-based distributional targets for temporally compressed video tokens. These targets accommodate local positional uncertainty and retain distinct garment matches assigned to the same person-video token.
- We construct coverage-balanced pseudo-correspondences by combining high-confidence direct matches with validated matches propagated through point tracking.
- We introduce LHFSC for paired evaluation and curate BiCorr-Test to assess fine-grained garment fidelity across diverse appearances.

2 Related Work

2.1 Video Virtual Try-On

Early VVT methods combine garment warping with temporal modeling (Dong et al., 2019; Zhong et al., 2021; Jiang et al., 2022). Diffusion methods extend image try-on through temporal guidance from pretrained models (He et al., 2025) or combine garment feature injection with temporal attention (Fang et al., 2024; Wang et al., 2024). Joint image-video training supplements video data with images (Karras et al., 2024). Video DiTs model spatiotemporal dependencies (Zheng et al., 2024; Zou et al., 2025). CatV2TON unifies image and video try-on through temporal concatenation of garment and person conditions (Chong et al., 2025b).

Detail-preserving designs use semantic, structural, and appearance cues (Li et al., 2025b), temporally smoothed garment-focused crops (Xu et al., 2024), or keyframe guidance (Zuo et al., 2025; He et al., 2026). UniVVT and InstructVVT use multimodal large language models for semantic conditioning without mask or pose inputs at inference (Cao et al., 2026; Shao et al., 2026). For evaluation, Eevee provides high-resolution full-shot and close-up videos and a garment consistency metric (Zeng et al., 2026). We focus on explicit person-garment correspondence supervision in garment-conditioned video DiTs.

2.2 Correspondence Learning and Attention Supervision

Image virtual try-on methods promote person-garment alignment through correspondence-based conditioning (Wan et al., 2025) and auxiliary supervision (Kim et al., 2024; Zhou et al., 2025). External matches provide explicit attention supervision: CORAL aligns attention-derived coordinates with DINOv3 matches and minimizes attention entropy (Kim et al., 2026). SIFT-VTON instead supervises cross-attention distributions with normalized spatial histograms of filtered SIFT matches (Takemoto & Koshinaka, 2027).

In VVT, auxiliary constraints include cross-frame self-attention regularization (Li et al., 2025a), region-based attention guidance (Li et al., 2025c; Wang et al., 2025), and optical-flow-guided feature consistency (Chen et al., 2026). PEMF-VTO uses frame-garment matches and tracked points to condition dedicated attention modules, with optional point guidance at inference (Chang et al., 2026).

Beyond try-on, DiffTrack analyzes temporal correspondence in the query-key similarities of video DiTs (Nam et al., 2025). Point trajectories provide feature-level supervision for video diffusion models (Jeong et al., 2025; Shi et al., 2025), while CAMEO supervises cross-view attention with geometric correspondences (Kwon et al., 2026). MVTrack4Gen combines attention supervision across views and latent frames with an auxiliary point-tracking head (Lee et al., 2026).

Building on correspondence-based attention supervision, BiCorr maps frame-level person-garment matches to compressed video latent grids and groups them to supervise attention in both directions. Its Gaussian-based targets accommodate local positional uncertainty and retain distinct garment locations associated with the same person-video token.

3 Method

BiCorr adapts a pretrained video DiT for virtual try-on through bidirectional correspondence supervision. As shown in Figure 2, frame-level pseudo-correspondences are mapped to token grids to construct distributional targets for training.

3.1 Garment-Conditioned Video DiT

The DiT is conditioned on a garment-agnostic person video with the original garment masked out, the corresponding binary mask sequence, and the person's pose sequence (Li et al., 2025b). Text and global garment features provide additional conditioning through semantic cross-attention. During training, a video variational autoencoder (VAE) encodes the person video and reference garment image into latent representations. The video latent is mixed with Gaussian noise, concatenated channel-wise with representations of the garment-agnostic video, mask, and pose, and patchified

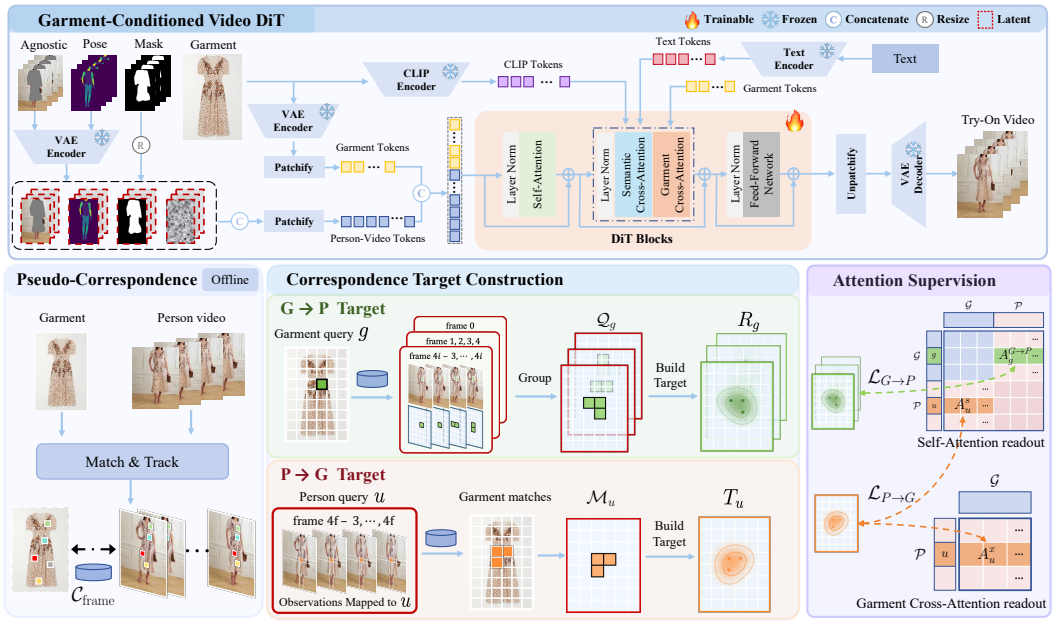

Figure 2: Overview of BiCorr. Lower panels show training-only target construction and supervision. Person-video and garment queries are selected separately for balanced coverage. G→P groups matches for garment query g across latent frames i. For P→G, $f = \tau(u)$ is the latent-frame index of person-video query u. T_u supervises Self- and Cross-P→G; R_g supervises G→P in self-attention. Matrix rows index queries; columns index keys. Dashed arrows indicate auxiliary supervision.

into person-video tokens. The clean garment latent is separately patchified into reference-garment tokens. The reference-garment tokens are prepended to the person-video sequence for joint self-attention and also serve as keys and values in garment cross-attention.

3.2 Pseudo-Correspondence Construction and Selection

We estimate correspondences offline by matching DINOv3 patch descriptors between person-video frames and the reference garment image (Siméoni et al., 2026). Direct matching is restricted to garment regions in each frame and the reference image. We filter direct matches based on forward-reverse consistency, matching ambiguity, and frame quality. Selected high-confidence matches initialize CoTracker3 (Karaev et al., 2025), which tracks person-side locations across frames while their reference-garment locations remain fixed. Propagated matches are validated by visibility, appearance consistency, membership in the person-side garment mask, and agreement with tracks reinitialized from intermediate frames.

We combine direct and validated propagated matches, retaining at most one garment match per person location per frame. The N retained observations form $\mathcal{C}_{\text{frame}} = \{(t_i, p_i, g_i)\}_{i=1}^{N}$, where t_i indexes frames and p_i, g_i index person and garment descriptor grids. During training, observations are mapped to token grids (Sec. 3.3). We select person-video and garment queries under separate budgets to balance temporal and spatial coverage. Match confidence guides both filtering and query selection (Appendix B.1).

3.3 Distributional Correspondence Targets for Video Latent Tokens

Let $\mathcal{P}$ and $\mathcal{G}$ denote the index sets of the person-video and reference-garment token grids, with sizes $F_P \times H_P \times W_P$ and $H_G \times W_G$, respectively. Here F_P counts latent frames, and the remaining dimensions are spatial. For each sampled clip, we select its observations from $\mathcal{C}_{\text{frame}}$ and reindex t_i within the clip. Spatial maps $\mathcal{R}_P, \mathcal{R}_G$ convert descriptor-grid indices to spatial token indices, while τ_{VAE} maps frame indices to latent-frame indices:

$$u_i = \big(\tau_{\text{VAE}}(t_i), \mathcal{R}_P(p_i)\big) \in \mathcal{P}, \qquad \bar{g}_i = \mathcal{R}_G(g_i) \in \mathcal{G}. \tag{1}$$

Observations with the same mapped temporal and spatial indices share a person-video query u. Temporal compression may therefore associate u with different garment locations across frames. For P→G supervision, $\mathcal{M}_u \subseteq \mathcal{G}$ retains up to K distinct garment locations with the highest observation counts. For G→P supervision, $\mathcal{Q}_g = \{u_i \mid \bar{g}_i = g\}$ collects distinct person-video locations associated with each garment token.

We place spatial Gaussians at these matched locations to account for matching and remapping uncertainty. For video samples, multiple locations receive separate components with equal coefficients; a single location receives additional weight at the matched token. Let $\kappa_G(g; g_0)$ and $\kappa_P(v; q)$ denote these unnormalized components centered at $g_0 \in \mathcal{M}_u$ and $q \in \mathcal{Q}_g$, respectively. For nonempty center sets, the bidirectional targets are

$$\begin{aligned} T_u(g) &= Z_u^{-1} \sum\nolimits_{g_0 \in \mathcal{M}_u} \kappa_G(g; g_0), && g \in \mathcal{G}, \\ R_g(v) &= Z_g^{-1} \sum\nolimits_{q \in \mathcal{Q}_g} \kappa_P(v; q) \mathbf{1}[\tau(v) = \tau(q)], && v \in \mathcal{P}, \end{aligned} \tag{2}$$

where Z_u, Z_g sum the unnormalized mixtures over $\mathcal{G}$ and the full $\mathcal{P}$, respectively. Here $\tau(v)$ is the latent-frame index of v, and $\mathbf{1}[\cdot]$ denotes the indicator function.

Support regions $S_u \subseteq \mathcal{G}$ and $S_g^P \subseteq \mathcal{P}$ unite local spatial neighborhoods around the component centers, and each G→P component and its neighborhood lie within the center's latent frame. These regions are used for correspondence supervision (Sec. 3.4). Appendices B.2 and B.3 specify remapping, components, and image-specific normalization for $F_P = 1$.

3.4 Bidirectional Correspondence Learning

The target T_u supervises P→G correspondence in garment-prefix self-attention ($a = s$) and garment cross-attention ($a = x$), while R_g supervises G→P correspondence in self-attention. For block ℓ and head h, let $Q^a_{\ell h}, K^a_{\ell h}$ denote the query and key matrices of pathway a, with brackets indicating the token selection by domain. With H heads of dimension d_h, the auxiliary distributions are

$$\begin{aligned} A_\ell^{a,P\to G} &= H^{-1} \sum\nolimits_{h=1}^{H} \mathrm{softmax}_{\mathcal{G}} \left(d_h^{-1/2} Q^a_{\ell h}[\mathcal{P}] K^a_{\ell h}[\mathcal{G}]^\top \right), \quad a \in \{s, x\}, \\ A_\ell^{G\to P} &= H^{-1} \sum\nolimits_{h=1}^{H} \mathrm{softmax}_{\mathcal{P}} \left(d_h^{-1/2} Q^s_{\ell h}[\mathcal{G}] K^s_{\ell h}[\mathcal{P}]^\top \right). \end{aligned} \tag{3}$$

Each softmax normalizes over the indicated key domain, and the resulting distributions are averaged across heads. We then aggregate them across blocks using shared, learned noise-dependent weights that are nonnegative and sum to one. Rows of the aggregated matrices give A^s_u, A^x_u for person-video query u and $A_g^{G\to P}$ for garment query g (Appendix B.4). Self-P→G supervision constrains relative probabilities within $\mathcal{G}$ without directly constraining the total self-attention mass assigned to garment tokens.

For prediction A, target T or R, and local support region S, let $A(S^c) = \sum_{i \notin S} A(i)$ denote outside probability mass. Superscript S denotes restriction and renormalization within S. The per-query losses are

$$\begin{aligned} \mathcal{D}(A, T, S) &= \mathrm{CE}(T, A) + \lambda_{\mathrm{ent}} \gamma_{\mathrm{ent}}(T, S)\, \mathrm{Ent}(A) + \lambda_{\mathrm{out}} A(S^c), \\ \mathcal{D}_G(A, R, S) &= \mathrm{CE}(R, A) + \lambda_{\mathrm{shape}} D_{\mathrm{KL}}(R^S \| A^S) + \lambda_{\mathrm{region}} A(S^c). \end{aligned} \tag{4}$$

Cross-entropy aligns each prediction with its full target distribution, while the outside-region penalties discourage attention to distant locations. For P→G, entropy minimization encourages concentrated attention. The gate $\gamma_{\mathrm{ent}}(T, S) \in [0, 1]$ reduces this penalty when the target entropy is high relative to the support size. For G→P, the KL term additionally aligns relative probabilities within S. All loss weights are nonnegative. Appendix B.4 provides the gate and numerical definitions.

For nonempty sets of selected queries with valid targets, $\mathcal{U} \subseteq \mathcal{P}$ and $\mathcal{G}_s \subseteq \mathcal{G}$, the bidirectional objectives are

$$\begin{aligned} \mathcal{L}_{P\to G} &= |\mathcal{U}|^{-1} \sum\nolimits_{u \in \mathcal{U}} \left[\mathcal{D}(A^s_u, T_u, S_u) + \lambda_x b_x \mathcal{D}(A^x_u, T_u, S_u) \right], \\ \mathcal{L}_{G\to P} &= \Omega_G^{-1} \sum\nolimits_{g \in \mathcal{G}_s} w_g \mathcal{D}_G(A_g^{G\to P}, R_g, S_g^P). \end{aligned} \tag{5}$$

Here λ_x weights the auxiliary cross-attention loss, and $b_x \sim \mathrm{Bernoulli}(p_x)$ gates its activation. The normalizer is $\Omega_G = \sum_{g \in \mathcal{G}_s} w_g > 0$. For videos, $w_g = 1$; for images, it counts valid supervised person locations associated with g.

3.5 Training and Inference

The generation loss $\mathcal{L}_{\text{gen}}$ combines flow matching (Lipman et al., 2023) with an additional mask-weighted velocity-prediction loss (Appendix B.4). On updates with correspondence loss, we optimize

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \omega(\sigma)\big[\lambda_P(n;m)\mathcal{L}_{P\to G} + \lambda_G(n;m)\mathcal{L}_{G\to P}\big], \tag{6}$$

where $\sigma \in [0,1]$ is the noise level and $\omega(\sigma)$ downweights correspondence supervision at high noise. The coefficients λ_P, λ_G follow warmup schedules depending on step n and modality $m \in \{\text{img}, \text{vid}\}$ (Appendix C.1). Other updates use only $\mathcal{L}_{\text{gen}}$. All matching, tracking, target construction, and auxiliary correspondence computations are confined to training, adding no inference overhead.

4 Experiments

4.1 Experimental Setup

Datasets and protocols. We train on paired images from VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022), and paired videos from ViViD (Fang et al., 2024). We evaluate on ViViD-S (Chong et al., 2025b), BiCorr-Test, and the official image test splits of VITON-HD and DressCode. Paired evaluation uses the original garment and source image or video as ground truth; unpaired evaluation uses a different garment without ground truth. Standard benchmarks use their dataset resolutions (Appendix C.3).

BiCorr-Test construction. We curate 150 clips from the official Eevee (Zeng et al., 2026) and ViT-HD (He et al., 2026) test splits to assess fine-grained garment fidelity, with 50 clips each for dresses, upper-body and lower-body garments. Each clip contains 64 consecutive frames with clear garment details and predominantly frontal views, standardized to 624×832 (Appendix C.3).

Implementation details. We initialize BiCorr from Wan2.1-Fun-14B-Control, derived from Wan2.1-I2V-14B (Team Wan et al., 2025). Training progresses from image-only to joint image-video training, followed by joint training at higher resolutions. Inference uses 30 denoising steps and a classifier-free guidance scale of 2.5 (Appendix C.1).

Evaluation metrics. For paired videos, we report frame-averaged structural similarity (SSIM) (Wang et al., 2004) and learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018). Following prior work (Fang et al., 2024; Chong et al., 2025b; Li et al., 2025b), both protocols use Video Fréchet Inception Distance (VFID) with I3D (Carreira & Zisserman, 2017) and 3D-ResNeXt-101 (Hara et al., 2018) features, denoted by VFID_I and VFID_R, respectively.

We further introduce *local high-frequency spectral consistency* (LHFSC) for paired garment detail evaluation. Using a shared binary agnostic mask M, we extract patches at identical locations fully inside M from ground truth X and prediction Y. Let $a(X;M)$ and $a(Y;M)$ concatenate their high-frequency Fourier magnitudes after weighted mean removal and Hann windowing. We define

$$\text{LHFSC}(X,Y;M) = 1 - \frac{\|a(X;M) - a(Y;M)\|_2^2}{\|a(X;M)\|_2^2 + \|a(Y;M)\|_2^2 + \varepsilon}, \tag{7}$$

where $\varepsilon > 0$ ensures numerical stability. Higher scores indicate closer spectral agreement. We average over frames with valid patches, then equally over videos. LHFSC measures magnitude agreement, not phase or temporal consistency (Appendix C.4).

4.2 Comparison with State-of-the-Art Methods

We primarily compare with baseline methods on ViViD-S, excluding variants additionally fine-tuned on method-specific datasets. For this comparison, we use previously reported baseline scores (Li et al., 2025b; He et al., 2026).

On BiCorr-Test, we evaluate the baselines listed in Table 2 on the same clips and garment assignments. Any2AnyTryon (Guo et al., 2025) and CORAL (Kim et al., 2026) outputs are animated with

Table 1: Quantitative comparisons on ViViD-S. FT denotes fine-tuned; I and R denote I3D and 3D-ResNeXt-101 features, respectively. Best scores are **bold**; second-best scores are underlined.

		Paired				Unpaired	
	Method	$\text{VFID}_I\downarrow$	$\text{VFID}_R\downarrow$	SSIM↑	LPIPS↓	$\text{VFID}_I\downarrow$	$\text{VFID}_R\downarrow$
backbone-FT	Wan2.1-I2V (Li et al., 2025b)	18.6245	1.2303	0.7986	0.1786	22.2147	1.0087
Image-based	StableVITON (Kim et al., 2024)	34.2446	0.7735	0.8019	0.1338	36.8985	0.9064
	OOTDiffusion (Xu et al., 2025)	29.5253	3.9372	0.8087	0.1232	35.3170	5.7078
	IDM-VTON (Choi et al., 2025)	20.0812	0.3674	0.8227	0.1163	25.4972	0.7167
Video-based	ViViD (Fang et al., 2024)	17.2924	0.6209	0.8029	0.1221	21.8032	0.8212
	CatV2TON (Chong et al., 2025b)	13.5962	0.2963	0.8727	0.0639	19.5131	0.5283
	KeyTailor (He et al., 2026)	9.5285	0.2761	0.8985	0.0585	14.5381	0.3869
	MagicTryOn (Li et al., 2025b)	8.4030	0.2346	0.9011	0.0602	14.7147	0.3200
	BiCorr	**7.4767**	**0.1178**	**0.9106**	**0.0472**	**14.4684**	**0.1880**

Table 2: Quantitative comparisons on BiCorr-Test. SA denotes StableAnimator. Best scores are **bold**; second-best scores are underlined.

		Paired					Unpaired	
	Method	$\text{VFID}_I\downarrow$	$\text{VFID}_R\downarrow$	SSIM↑	LPIPS↓	LHFSC↑	$\text{VFID}_I\downarrow$	$\text{VFID}_R\downarrow$
Image try-on + animation	Any2AnyTryon + SA	16.8114	0.4200	0.7961	0.1294	0.6337	23.9013	0.6704
	CORAL + SA	14.8401	0.2009	0.8044	0.1178	0.6191	22.2322	0.4307
Video-based	ViViD	15.7350	0.3916	0.8406	0.1181	0.5388	21.4707	0.5409
	CatV2TON	19.0004	0.5999	0.8214	0.1535	0.4997	23.6259	0.7176
	MagicTryOn	11.1093	0.2367	0.8787	0.0735	0.7116	19.1674	0.3778
	BiCorr	**9.7078**	**0.1808**	**0.8801**	**0.0715**	**0.7458**	**18.6819**	**0.3542**

StableAnimator (SA) (Tu et al., 2025). CatV2TON generates at 384×512 using its 512-resolution checkpoint, and resizes outputs to 624×832 for evaluation.

Video try-on. As shown in Tables 1 and 2, BiCorr ranks first across the reported metrics on ViViD-S and BiCorr-Test. Gains are generally larger in paired evaluation, while relative gains in unpaired evaluation are larger for VFID_R than for VFID_I. These results indicate improved reconstruction fidelity and closer agreement with real-video feature distributions, respectively. Higher LHFSC on BiCorr-Test further indicates better preservation of local high-frequency garment appearance.

Qualitative comparison. Figure 3 compares unpaired results at the same time points across methods. In these examples, BiCorr better preserves the floral motifs in ViViD-S, the chest and leaf motifs in BiCorr-Test, and the printed motif in the in-the-wild video. Additional unpaired video comparisons are provided in Appendix A.2.

User study. We conduct a user study with 60 participants rating paired and unpaired BiCorr-Test outputs for temporal consistency, garment-detail fidelity, and overall quality on a five-point scale. BiCorr receives the highest mean ratings under both protocols, with paired/unpaired garment-detail scores of 4.38/4.32 versus MagicTryOn's 3.71/3.65. Across rated paired outputs, LHFSC correlates with mean garment-detail ratings (Spearman's $\rho = 0.719$). Full results and statistical analyses appear in Appendix A.3.

Image try-on. The same model achieves the lowest kernel inception distance (KID) (Bińkowski et al., 2018) under both VITON-HD protocols and leads five of six reported DressCode measures. These results also demonstrate competitive image try-on performance. KeyTailor achieves higher SSIM on both datasets, indicating that our gains do not extend uniformly to structural similarity (Appendix A.1).



Figure 3: Qualitative comparison of unpaired video try-on on ViViD-S, BiCorr-Test, and an in-the-wild video (left to right). Please zoom in for details.

Table 3: Ablations on ViViD-S with the 1.3B backbone. PCK and tPCK measure Self-P→G and G→P localization on noisy paired ground-truth latents, respectively. Best scores are **bold**; second-best scores are underlined.

	Paired					Unpaired		Correspondence (%)	
Variant	$\text{VFID}_I\downarrow$	$\text{VFID}_R\downarrow$	SSIM↑	LPIPS↓	LHFSC↑	$\text{VFID}_I\downarrow$	$\text{VFID}_R\downarrow$	PCK@2↑	tPCK@2↑
w/o corr	13.2729	0.3687	0.8590	0.0900	0.6390	18.4515	0.4524	3.54	2.75
w/o P→G	11.6829	0.2875	0.8629	0.0862	0.6602	17.8618	0.3558	26.04	71.49
w/o G→P	11.7736	0.3137	0.8631	0.0849	0.6684	17.7161	0.3630	57.78	73.49
w/o Cross-P→G	11.7480	0.2454	0.8617	0.0855	0.6593	18.0535	0.3426	58.82	79.62
Hard single-point targets	12.1122	0.2887	**0.8663**	0.0861	0.6546	17.9681	0.3494	**60.50**	**81.90**
Direct-only pipeline	12.1605	0.3316	0.8641	0.0880	0.6420	17.8823	0.4073	58.49	80.34
Track-only pipeline	12.3035	0.3168	0.8657	0.0875	0.6388	**17.6745**	0.3794	58.55	79.20
Full	**11.4443**	**0.1680**	0.8631	**0.0834**	**0.6733**	17.7479	**0.2846**	58.83	80.52

4.3 Ablation Studies

We conduct ablations on supervision directions, distributional targets, and pseudo-correspondence construction using a 1.3B Wan2.1-Fun-Control backbone. We sample 20% of the full training set, aiming to preserve its data distribution (Appendix C.2). All variants use this subset under otherwise identical experimental settings. Full denotes the complete method, while w/o corr uses only the generation loss.

We measure attention localization using the percentage of correct keypoints (PCK) (Kim et al., 2026). All variants receive the same ground-truth video latents with identical added noise, and adopt the same queries and test pseudo-correspondences. For P→G self-attention, PCK@2 counts a query as correct if its attention peak is within two token spacings of a valid garment match. For G→P, tPCK@2 applies the same criterion to the spatial peak within each pseudo-labeled latent frame.

Bidirectional supervision. Without correspondence supervision, all generation metrics degrade (Table 3). Retaining only one direction also worsens paired LPIPS, LHFSC, and unpaired VFID_R, supporting the complementary roles of P→G and G→P. The exception is a slight improvement

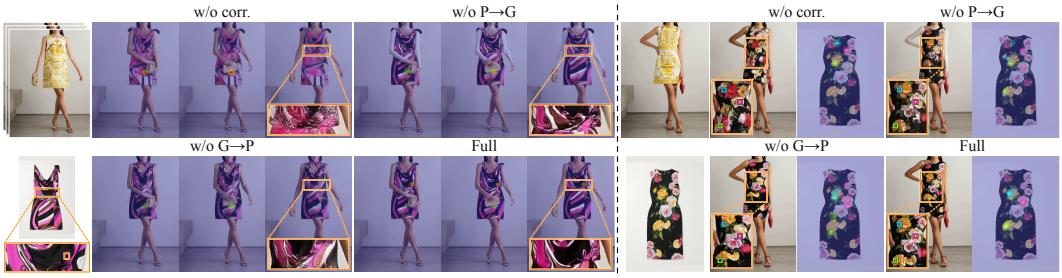

Figure 4: Unpaired try-on results and attention maps for ablation variants. Left: G→P attention on generated frames for a query on the reference garment. Right: P→G self-attention on the reference garment for queries on the person video. Small squares mark query locations; matching colors link each person-video query to its attention map. Please zoom in for details.

in unpaired VFID_I without G→P. Omitting Cross-P→G supervision worsens all seven generation scores despite nearly unchanged Self-P→G PCK@2.

Correspondence targets. Temporal remapping can associate one person-video query with multiple garment locations (Appendix A.4). We evaluate the distributional target formulation by replacing it with confidence-selected hard single-point targets in both directions. This replacement changes center selection, distribution shape, and support regions. As shown in Table 3, hard targets slightly improve SSIM, PCK, and tPCK, but worsen LPIPS, LHFSC, and all VFID scores. These results support the complete target formulation, while showing that attention-peak agreement alone does not capture generation quality.

Pseudo-correspondence construction and selection. Direct-only and track-only retain direct and propagated video matches, respectively, and both omit query selection that balances coverage. Track-only still uses direct matches to initialize tracking. Since retained matches also affect clip sampling, these variants evaluate construction and selection together. Both variants worsen LPIPS, LHFSC, and most VFID scores relative to Full, supporting the complete construction and selection pipeline.

Correspondence analysis. The localization results follow a similar trend: omitting either direction lowers both PCK and tPCK, and omitting all correspondence losses causes a larger decline. These scores measure agreement with test pseudo-correspondences using ground-truth video latents with added noise. To examine whether the differences also appear during generation, Figure 4 compares attention maps from unpaired outputs. In these examples, the ablated variants show less accurate localization alongside poorer preservation of the draped neckline and floral motifs.

5 Conclusion

We presented BiCorr, a framework for bidirectional latent correspondence learning in garment-conditioned video DiTs. Gaussian-based distributional targets accommodate local uncertainty and multiple garment matches per person-video token. Query selection balances supervision across frames and garment regions. Experiments and human evaluations show improved garment-detail fidelity and video quality without additional inference overhead. LHFSC measures local high-frequency spectral agreement in paired evaluation, while BiCorr-Test covers diverse garment appearances.

Limitations and future work. The method relies on pseudo-correspondences that may contain noise, while BiCorr-Test mainly includes short clips with predominantly frontal views. Future work will explore more reliable correspondence estimation and evaluation on longer videos with varied viewpoints.

AI USE STATEMENT

We used generative AI tools for manuscript editing and translation, literature searches, questionnaire preparation, environment setup for baseline methods, and reviewing experimental protocols. The authors reviewed the AI-assisted content and take responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix B describe the method and its implementation. Appendix C provides training configurations, hyperparameters, and ablation settings. Appendix C documents BiCorr-Test construction, evaluation protocols, and LHFSC implementation. The user-study protocol and analysis are provided in Appendix A.3.

REFERENCES

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.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Yushe Cao, Shikun Feng, Fei Shen, Haikuo Peng, Jianqiang Xia, Yiheng Zhu, Dianxi Shi, and Chun Yu. UniVVT: A unified end-to-end framework for high-fidelity video virtual try-on. *arXiv preprint arXiv:2608.05745*, 2026. doi: 10.48550/arXiv.2608.05745. URL https://arxiv.org/abs/2608.05745.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017. doi: 10.1109/CVPR.2017.502.

Tianyu Chang, Xiaohao Chen, Zhichao Wei, Xuanpu Zhang, Qingguo Chen, Weihua Luo, Peipei Song, and Xun Yang. PEMF-VTO: Point-enhanced video virtual try-on via mask-free paradigm. *IEEE Transactions on Consumer Electronics*, 72(1):2267–2279, 2026. doi: 10.1109/TCE.2025.3633891.

Shengyao Chen, Xianbing Sun, Liqing Zhang, and Jianfu Zhang. FlowVVTON: Flow-guided mask-free video virtual try-on. *arXiv preprint arXiv:2608.30450*, 2026. URL https://arxiv.org/abs/2608.30450.

Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. VITON-HD: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14131–14140, 2021. doi: 10.1109/CVPR46437.2021.01391.

Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *Computer Vision – ECCV 2024*, volume 15144 of *Lecture Notes in Computer Science*, pp. 206–235. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-73016-0_13.

Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, Dongmei Jiang, and Xiaodan Liang. CatVTON: Concatenation is all you need for virtual try-on with diffusion models. In *International Conference on Learning Representations*, pp. 66586–66601, 2025a.

Zheng Chong, Wenqing Zhang, Shiyue Zhang, Jun Zheng, Xiao Dong, Haoxiang Li, Yiling Wu, Dongmei Jiang, and Xiaodan Liang. CatV2TON: Taming diffusion transformers for vision-based virtual try-on with temporal concatenation. *arXiv preprint arXiv:2501.11325*, 2025b. doi: 10.48550/arXiv.2501.11325. URL https://arxiv.org/abs/2501.11325.

Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. FW-GAN: Flow-navigated warping GAN for video virtual try-on. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1161–1170, 2019. doi: 10.1109/ICCV.2019.00125. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Dong_FW-GAN_Flow-Navigated_Warping_GAN_for_Video_Virtual_Try-On_ICCV_2019_paper.html.

Zixun Fang, Wei Zhai, Aimin Su, Hongliang Song, Kai Zhu, Mao Wang, Yu Chen, Zhiheng Liu, Yang Cao, and Zheng-Jun Zha. ViViD: Video virtual try-on using diffusion models. *arXiv preprint arXiv:2405.11794*, 2024. doi: 10.48550/arXiv.2405.11794. URL https://arxiv.org/abs/2405.11794.

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.

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.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d CNNs retrace the history of 2d CNNs and ImageNet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6546–6555, 2018.

Qingdong He, Xueqin Chen, Yanjie Pan, Peng Tang, Pengcheng Xu, Zhenye Gan, Chengjie Wang, Xiaobin Hu, Jiangning Zhang, and Yabiao Wang. The devil is in the details: Enhancing video virtual try-on via keyframe-driven details injection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9182–9191, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/He_The_devil_is_in_the_details_Enhancing_Video_Virtual_Try-On_CVPR_2026_paper.html.

Zijian He, Peixin Chen, Guangrun Wang, Guanbin Li, Philip H. S. Torr, and Liang Lin. WildVidFit: Video virtual try-on in the wild via image-based controlled diffusion models. In *Computer Vision – ECCV 2024*, volume 15075 of *Lecture Notes in Computer Science*, pp. 123–139. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-72643-9_8. URL https://link.springer.com/chapter/10.1007/978-3-031-72643-9_8.

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. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Hyeonho Jeong, Chun-Hao P. Huang, Jong Chul Ye, Niloy J. Mitra, and Duygu Ceylan. Track4Gen: Teaching video diffusion models to track points improves video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7276–7287, 2025. doi: 10.1109/CVPR52734.2025.00682. URL https://openaccess.thecvf.com/content/CVPR2025/html/Jeong_Track4Gen_Teaching_Video_Diffusion_Models_to_Track_Points_Improves_Video_CVPR_2025_paper.html.

Jianbin Jiang, Tan Wang, He Yan, and Junhui Liu. ClothFormer: Taming video virtual try-on in all module. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10799–10808, 2022. doi: 10.1109/cvpr52688.2022.01053. URL https://openaccess.thecvf.com/content/CVPR2022/html/Jiang_ClothFormer_Taming_Video_Virtual_Try-On_in_All_Module_CVPR_2022_paper.html.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6013–6022, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Karaev_CoTracker3_Simpler_and_Better_Point_Tracking_by_Pseudo-Labelling_Real_Videos_ICCV_2025_paper.html.

Johanna Karras, Yingwei Li, Nan Liu, Luyang Zhu, Innfarn Yoo, Andreas Lugmayr, Chris Lee, and Ira Kemelmacher-Shlizerman. Fashion-VDM: Video diffusion model for virtual try-on. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11. Association for Computing Machinery, 2024. doi: 10.1145/3680528.3687623. URL https://doi.org/10.1145/3680528.3687623.

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.

Jiyoung Kim, Youngjin Shin, Siyoon Jin, Dahyun Chung, Jisu Nam, Tongmin Kim, Jongjae Park, Hyeonwoo Kang, and Seungryong Kim. CORAL: Correspondence alignment for improved virtual try-on. *arXiv preprint arXiv:2602.17636*, 2026.

Minkyung Kwon, Jinhyeok Choi, Jiho Park, Seonghu Jeon, Jinhyuk Jang, Junyoung Seo, Minseop Kwak, Jin-Hwa Kim, and Seungryong Kim. Correspondence-attention alignment for multi-view diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2316–2326, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Kwon_Correspondence-Attention_Alignment_for_Multi-View_Diffusion_Models_CVPR_2026_paper.html.

JoungBin Lee, Jaewoo Jung, Jongmin Lee, Tongmin Kim, Hyunsung Kim, Takuya Narihira, Kazumi Fukuda, Jahyeok Koo, Jisang Han, Yuki Mitsufuji, and Seungryong Kim. MVTrack4Gen: Multi-view point tracking as geometric supervision for 4D video generation. *arXiv preprint arXiv:2606.26087*, 2026. doi: 10.48550/arXiv.2606.26087. URL https://arxiv.org/abs/2606.26087.

Dong Li, Wenqi Zhong, Wei Yu, Yingwei Pan, Dingwen Zhang, Ting Yao, Junwei Han, and Tao Mei. Pursuing temporal-consistent video virtual try-on via dynamic pose interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22648–22657, 2025a.

Guangyuan Li, Siming Zheng, Hao Zhang, Jinwei Chen, Junsheng Luan, Binkai Ou, Lei Zhao, Bo Li, and Peng-Tao Jiang. MagicTryOn: Harnessing diffusion transformer for garment-preserving video virtual try-on. *arXiv preprint arXiv:2505.21325*, 2025b. doi: 10.48550/arXiv.2505.21325. URL https://arxiv.org/abs/2505.21325.

Siqi Li, Zhengkai Jiang, Jiawei Zhou, Zhihong Liu, Xiaowei Chi, and Haoqian Wang. RealVVT: Towards photorealistic video virtual try-on via spatio-temporal consistency. *arXiv preprint arXiv:2501.08682*, 2025c. URL https://arxiv.org/abs/2501.08682.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *Computer Vision – ECCV 2022*, volume 13668 of *Lecture Notes in Computer Science*, pp. 345–362. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-20074-8_20. URL https://link.springer.com/chapter/10.1007/978-3-031-20074-8_20.

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. doi: 10.1145/3581783.3612137.

Jisu Nam, Soowon Son, Dahyun Chung, Jiyoung Kim, Siyoon Jin, Junhwa Hur, and Seungryong Kim. Emergent temporal correspondences from video diffusion transformers. In *Advances in Neural Information Processing Systems*, volume 38, pp. 123683–123727, 2025. doi: 10.52202/085713-4122. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b2fea79b1137d917e8b7cce9434ab5fa-Abstract-Conference.html.

Dingbao Shao, Song Wu, Xinyu Chen, Qian Wang, Jiahang Li, Kuai Jiang, Jiang Lin, Yuhang Liu, Ziyu Chen, Duo Li, Jiaxin Hu, Shengrong Gu, Ziheng Tang, Rongrong Liu, Yanlun Peng, Liang Li, Junlan Feng, Lujia Jin, Ting Zhang, Jian Yang, and Zili Yi. InstructVVT: Instruction-driven video virtual try-on without auxiliary spatial priors. *arXiv preprint arXiv:2608.14070*, 2026. doi: 10.48550/arXiv.2608.14070. URL https://arxiv.org/abs/2608.14070.

Qingyu Shi, Jianzong Wu, Jinbin Bai, Jiangning Zhang, Lu Qi, Yunhai Tong, and Xiangtai Li. Decouple and track: Benchmarking and improving video diffusion transformers for motion transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10995–11005, 2025. doi: 10.1109/ICCV51701.2025.01023. URL https://openaccess.thecvf.com/content/ICCV2025/html/Shi_Decouple_and_Track_Benchmarking_and_Improving_Video_Diffusion_Transformers_For_ICCV_2025_paper.html.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=2NlGyqNjns.

Kosuke Takemoto and Takafumi Koshinaka. SIFT-VTON: Geometric correspondence supervision on cross-attention for virtual try-on. In *Pattern Recognition*, volume 16817 of *Lecture Notes in Computer Science*, pp. 464–477. Springer Nature Switzerland, 2027. doi: 10.1007/978-3-032-31673-8_31.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. doi: 10.48550/arXiv.2503.20314. URL https://arxiv.org/abs/2503.20314.

Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. StableAnimator: High-quality identity-preserving human image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21096–21106, 2025.

Siqi Wan, Jingwen Chen, Yingwei Pan, Ting Yao, and Tao Mei. Incorporating visual correspondence into diffusion model for virtual try-on. In *International Conference on Learning Representations*, pp. 42557–42578, 2025.

Jinjuan Wang, Wenzhang Sun, Ming Li, Yun Zheng, Fanyao Li, Zhulin Tao, Donglin Di, Hao Li, Wei Chen, and Xianglin Huang. ChronoTailor: Harnessing attention guidance for fine-grained video virtual try-on. *arXiv preprint arXiv:2506.05858*, 2025. URL https://arxiv.org/abs/2506.05858.

Yuanbin Wang, Weilun Dai, Long Chan, Huanyu Zhou, Aixi Zhang, and Si Liu. GPD-VVTO: Preserving garment details in video virtual try-on. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7133–7142, 2024. doi: 10.1145/3664647.3680701. URL https://doi.org/10.1145/3664647.3680701.

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. doi: 10.1109/TIP.2003.819861.

Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. GP-VTON: Towards general purpose virtual try-on via collaborative local-flow

global-parsing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23550–23559, 2023.

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*, volume 39, pp. 8996–9004, 2025. doi: 10.1609/aaai.v39i9.32973.

Zhengze Xu, Mengting Chen, Zhao Wang, Linyu Xing, Zhonghua Zhai, Nong Sang, Jinsong Lan, Shuai Xiao, and Changxin Gao. Tunnel Try-on: Excavating spatial-temporal tunnels for high-quality virtual try-on in videos. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3199–3208, 2024. doi: 10.1145/3664647.3680836. URL https://doi.org/10.1145/3664647.3680836.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 4210–4220, 2023. URL https://openaccess.thecvf.com/content/ICCV2023W/CV4Metaverse/html/Yang_Effective_Whole-Body_Pose_Estimation_with_Two-Stages_Distillation_ICCVW_2023_paper.html.

Jianhao Zeng, Yancheng Bai, Ruidong Chen, Xuanpu Zhang, Lei Sun, Dongyang Jin, Ryan Xu, Nannan Zhang, Dan Song, and Xiangxiang Chu. Eevee: Towards close-up high-resolution video-based virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 4614–4624, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Zeng_Eevee_Towards_Close-up_High-resolution_Video-based_Virtual_Try-on_CVPRF_2026_paper.html.

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. doi: 10.1109/CVPR.2018.00068.

Jun Zheng, Fuwei Zhao, Youjiang Xu, Xin Dong, and Xiaodan Liang. VITON-DiT: Learning in-the-wild video try-on from human dance videos via diffusion transformers. *arXiv preprint arXiv:2405.18326*, 2024. doi: 10.48550/arXiv.2405.18326. URL https://arxiv.org/abs/2405.18326.

Xiaojing Zhong, Zhonghua Wu, Taizhe Tan, Guosheng Lin, and Qingyao Wu. MV-TON: Memory-based video virtual try-on network. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 908–916, 2021. doi: 10.1145/3474085.3475269. URL https://doi.org/10.1145/3474085.3475269.

Zijian Zhou, Shikun Liu, Xiao Han, Haozhe Liu, Kam Woh Ng, Tian Xie, Yuren Cong, Hang Li, Mengmeng Xu, Juan-Manuel Perez-Rua, Aditya Patel, Tao Xiang, Miaojing Shi, and Sen He. Learning flow fields in attention for controllable person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2491–2501, 2025.

Cheng Zou, Senlin Cheng, Bolei Xu, Dandan Zheng, Xiaobo Li, Jingdong Chen, and Ming Yang. Video virtual try-on with conditional diffusion transformer inpainter. *arXiv preprint arXiv:2506.21270*, 2025. doi: 10.48550/arXiv.2506.21270. URL https://arxiv.org/abs/2506.21270.

Tongchun Zuo, Zaiyu Huang, Shuliang Ning, Ente Lin, Chao Liang, Zerong Zheng, Jianwen Jiang, Yuan Zhang, Mingyuan Gao, and Xin Dong. DreamVVT: Mastering realistic video virtual try-on in the wild via a stage-wise diffusion transformer framework. *arXiv preprint arXiv:2508.02807*, 2025. doi: 10.48550/arXiv.2508.02807. URL https://arxiv.org/abs/2508.02807.

APPENDIX CONTENTS

A Additional Results and Analyses 15
A.1 Image Results 15
A.2 Additional Video Results 15
A.3 User Study and Metric Agreement 20
A.4 Ablation and Correspondence Analysis 20
B Additional Method Details 23
B.1 Pseudo-Correspondences 23
B.2 Latent Remapping 24
B.3 Distributional Targets 24
B.4 Attention Supervision and Generation Loss 25
C Training and Evaluation 26
C.1 Main-Model Training 26
C.2 Ablation Setup 27
C.3 BiCorr-Test and Evaluation Protocols 27
C.4 LHFSC Implementation 28

A ADDITIONAL RESULTS AND ANALYSES

A.1 IMAGE RESULTS

Table 4 reports our model's results on VITON-HD and DressCode at 768×1024, alongside published baseline scores (Xie et al., 2023; Morelli et al., 2023; Choi et al., 2025; Xu et al., 2025; Chong et al., 2025a;b; Li et al., 2025b; He et al., 2026). Paired evaluation uses SSIM, LPIPS, Fréchet Inception Distance (FID) (Heusel et al., 2017), and KID (Bińkowski et al., 2018); unpaired evaluation uses FID and KID.

Table 4: Image try-on results on VITON-HD and DressCode. p and u denote paired and unpaired evaluation, and KID is multiplied by 10^3. Best and second-best results are **bold** and underlined.

	Metric	Methods								
		GP-VTON	LaDI-VTON	IDM-VTON	OOTDiff.	CatVTON	CatV2TON	MagicTryOn	KeyTailor	**BiCorr**
VITON-HD	$FID^p\downarrow$	8.726	11.386	6.338	9.305	6.139	8.095	**4.959**	5.293	5.315
	$KID^p\downarrow$	3.944	7.248	1.322	4.086	0.964	2.245	0.572	0.720	**0.502**
	SSIM↑	0.8701	0.8603	0.8806	0.8187	0.8691	0.8902	0.9104	**0.9201**	0.8985
	LPIPS↓	0.0585	0.0733	0.0789	0.0876	0.0973	0.0572	**0.0429**	0.0566	0.0540
	$FID^u\downarrow$	11.844	14.648	9.611	12.408	9.143	11.222	9.079	**8.528**	8.819
	$KID^u\downarrow$	4.310	8.754	1.639	4.689	1.267	2.986	1.032	0.788	**0.756**
DressCode	$FID^p\downarrow$	9.927	9.555	6.821	4.610	3.992	5.722	6.550	2.746	**2.395**
	$KID^p\downarrow$	4.610	4.683	2.924	0.955	0.818	2.338	0.725	0.517	**0.403**
	SSIM↑	0.7711	0.7656	0.8797	0.8854	0.8922	0.9222	0.9295	**0.9621**	0.9314
	LPIPS↓	0.1801	0.2366	0.0563	0.0533	0.0455	0.0367	0.0301	0.0347	**0.0272**
	$FID^u\downarrow$	12.791	10.676	9.546	12.567	6.137	8.627	11.727	5.147	**4.322**
	$KID^u\downarrow$	6.627	5.787	4.320	6.627	1.549	3.838	1.544	1.012	**0.735**

Figure 5 compares unpaired image results, including CatVTON (Chong et al., 2025a) and Any2AnyTryon (Guo et al., 2025), with insets highlighting lettering and fine patterns. ViViD uses a single input frame, and CatV2TON uses its public 384×512 inference configuration.

A.2 ADDITIONAL VIDEO RESULTS

Figures 6–8 compare unpaired results at shared time points and crop locations. ViViD-S source videos use external garments. CatV2TON uses native 384×512 outputs. Video comparisons and additional results are provided in the supplementary demo.



Figure 5: Unpaired image try-on comparison on VITON-HD and DressCode. Please zoom in for details.

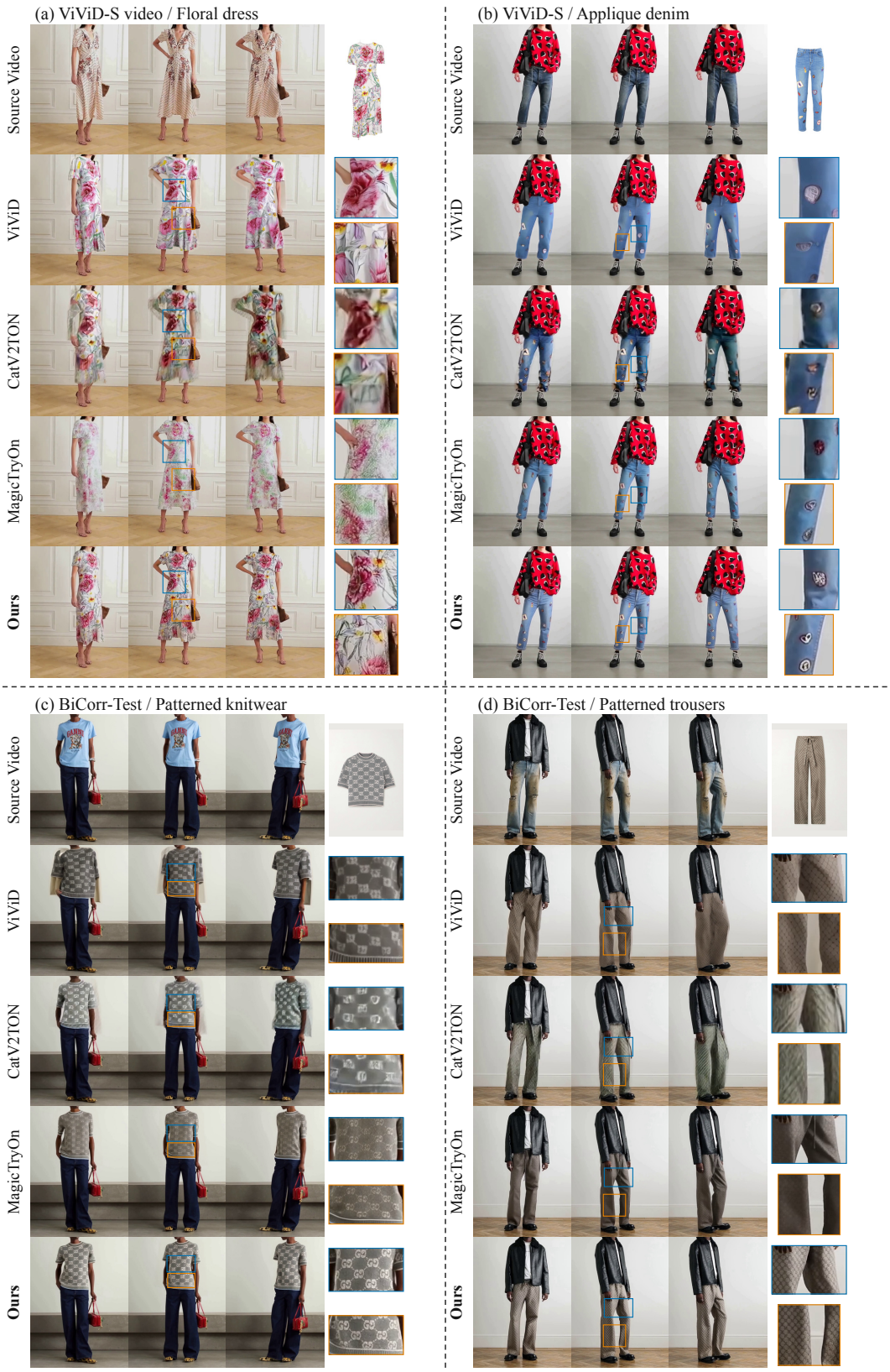

Figure 6: Unpaired video try-on comparisons using ViViD-S source videos (a,b) and BiCorr-Test (c,d). Please zoom in for details.



Figure 7: Unpaired video try-on in the wild (e,f) and person-to-video garment transfer (g,h). Please zoom in for details.



Figure 8: Unpaired video try-on with a ViViD-S garment and an external person video (i), and in-the-wild inputs (j–l). Please zoom in for details.

A.3 User Study and Metric Agreement

Each of the 60 participants rates four randomly assigned paired and three unpaired BiCorr-Test cases, selected independently of model performance. Case and baseline exposure is balanced independently of scores. Participants consented to the research use of their anonymous ratings and could discontinue participation at any time. Each trial shows anonymized outputs from BiCorr and 1–3 baselines in random order. References are the target garment and the ground-truth (paired) or source (unpaired) video. Ratings cover temporal consistency, garment-detail fidelity, and overall quality from 1 (severe problems) to 5 (almost no visible problems). Garment detail covers patterns, text, stripes, and material appearance. Ties are allowed; pilot responses are excluded. We average ratings within each method-case combination, then equally across cases rated for all five methods (Table 5).

Table 5: Mean user ratings on BiCorr-Test (1–5, higher is better). TC, GDF, and OQ denote temporal consistency, garment-detail fidelity, and overall try-on quality. Best and second-best means are **bold** and underlined.

	Paired			Unpaired		
Method	TC	GDF	OQ	TC	GDF	OQ
ViViD	2.27	2.31	2.19	2.62	2.47	2.24
CatV2TON	2.04	1.60	1.61	1.88	1.69	1.62
MagicTryOn	<u>4.14</u>	<u>3.71</u>	<u>3.92</u>	<u>4.26</u>	<u>3.65</u>	<u>3.82</u>
CORAL+SA	2.73	2.90	2.65	2.93	3.27	3.06
BiCorr	**4.59**	**4.38**	**4.48**	**4.65**	**4.32**	**4.44**

We analyze within-trial BiCorr-baseline rating differences using linear mixed-effects models with crossed participant and case random intercepts. Across four baselines, two protocols, and three criteria, all 24 comparisons favor BiCorr (two-sided Wald tests, Holm-adjusted $p < 0.01$). Against MagicTryOn, estimated garment-detail rating differences are 0.64 (pointwise 95% confidence interval $[0.31, 0.97]$) for paired outputs and 0.55 ($[0.18, 0.93]$) for unpaired outputs.

Agreement with human judgments. We analyze 175 paired video-method outputs from the five methods in Table 5, with 720 ratings. Each output is one observation, represented by its mean garment-detail rating. Spearman correlations with LHFSC, SSIM, and negated LPIPS are pooled across outputs or computed within cases and averaged equally. Agreement compares metric and human pairwise rankings within each participant's trial, excluding human ties. We average agreement within cases, then equally across cases. Confidence intervals use 5,000 case-bootstrap resamples with ratings held fixed.

Table 6: Association with human garment-detail ratings on paired BiCorr-Test outputs. Agreement measures consistency between metric and human pairwise rankings. Brackets give 95% case-bootstrap intervals for pooled Spearman correlation. Best and second-best point estimates are **bold** and underlined.

Metric	Pooled ρ	Within-case ρ	Agreement (%)
LHFSC	**0.719** [0.658, 0.781]	**0.847**	**89.3**
SSIM	0.247 [0.167, 0.344]	0.589	78.2
$-$LPIPS	<u>0.634</u> [0.569, 0.705]	<u>0.839</u>	<u>88.4</u>

LHFSC has the highest pooled correlation, which partly reflects between-method differences. Its within-case correlation and agreement are close to LPIPS, with confidence intervals for the differences spanning zero. Ratings assess garment detail rather than high-frequency perception alone.

A.4 Ablation and Correspondence Analysis

In Figure 9, Full better preserves the central motif, white separation bands, and internal contours.

Multiple locations after temporal remapping. Before query selection and top-K truncation, 20.51% of valid mapped person-video queries in ViViD-S and 15.95% in BiCorr-Test have mul-

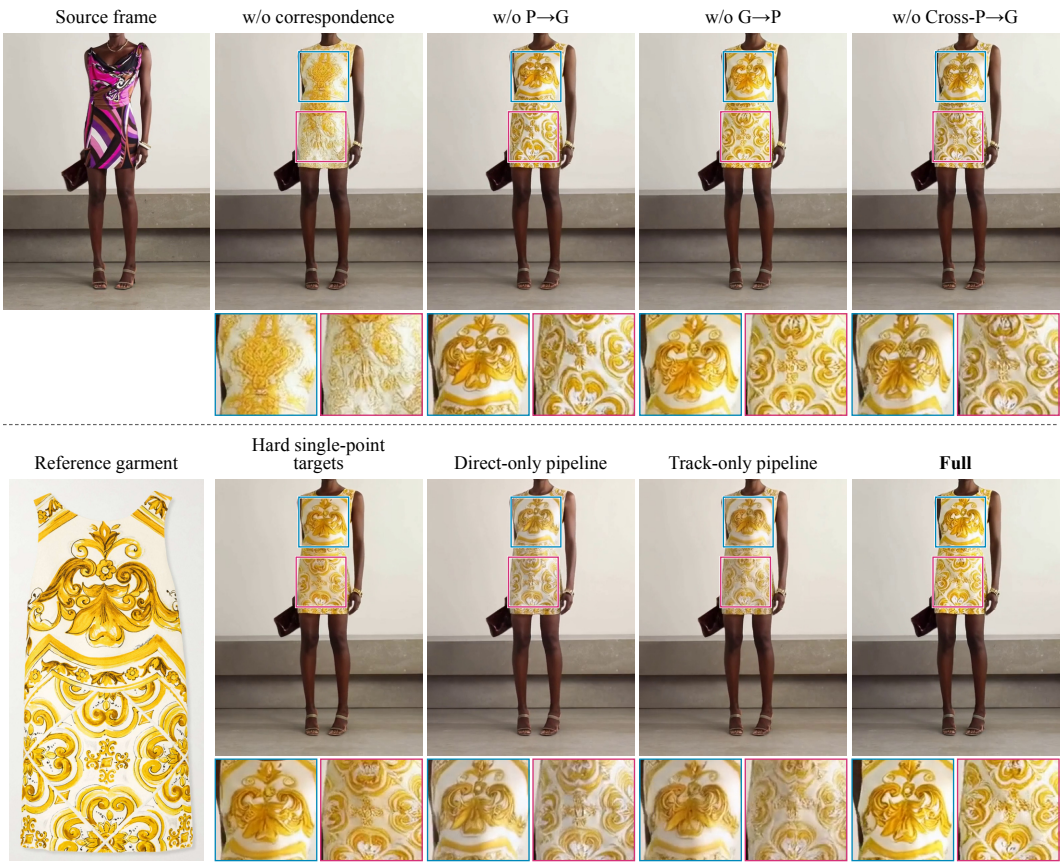

Figure 9: Unpaired BiCorr-Test comparison of the 1.3B ablation variants, using the same frame and seed. Please zoom in for details.

tiple garment locations (Figure 10). These queries combine observations from different source frames. Most locations are nearby. After transitively merging centers within four token spacings, 3.32% and 1.86% of queries retain multiple separated groups. These statistics describe test pseudo-correspondences, which may contain errors, rather than selected training targets.

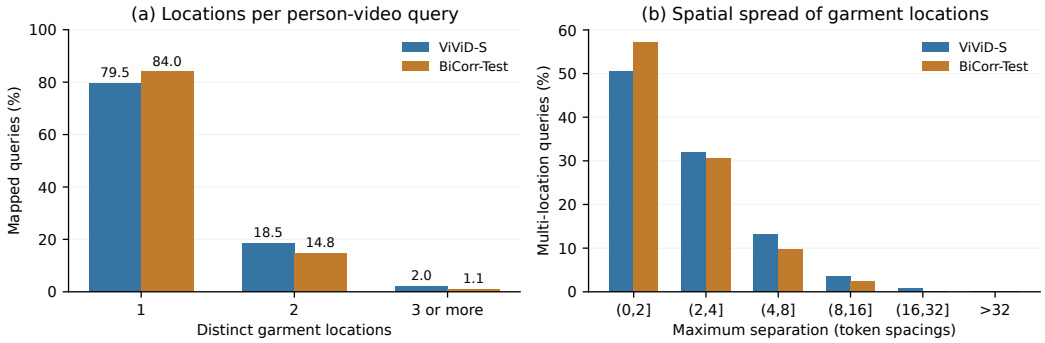

Figure 10: Statistics of mapped test pseudo-correspondences before query selection and top-K truncation. Proportions are averaged equally across videos; (b) uses only multi-location queries, excluding videos without them.

Probe protocol. We probe paired ground-truth latents at $\sigma \in \{0.25, 0.5, 0.75\}$ using Eq. 17, with shared noise, queries, and full-pipeline test pseudo-correspondences. Evaluation supports use these matches and a fixed radius, independently of training targets. Readouts aggregate 30 blocks: generation-only uses uniform weights, while other variants learn noise-dependent weights.

Localization. Self- and Cross-P→G PCK@2 measure the fraction of queries whose attention peaks lie within two token spacings (32 image pixels) of any valid garment match. G→P tPCK@2 applies this spatial criterion separately within each labeled latent frame.

Coverage. For mode coverage (MC), centers within $2r$ are grouped transitively, where $r = 2$ tokens. The radius-r neighborhoods of each group form disjoint supports $B_1, \ldots, B_C$, where C is the number of groups. For $C \geq 2$, MC is $\sum_{k=1}^{C} \min\{A_u^s(B_k), 1/C\}$. Let F_g be the set of labeled latent frames for garment query g, with C_f components in frame f. For G→P temporal coverage, we normalize support masses by total attention mass over labeled frames, then cap each component's mass at $1/(|F_g|C_f)$. This measures balanced coverage without penalizing mass outside labeled frames. Neither diagnostic directly measures output-video temporal consistency. We average queries within latent frames for P→G, labeled frames within garment queries for G→P, then noise levels and videos.

Table 7: Correspondence diagnostics on shared noisy paired ground-truth latents from ViViD-S. MC denotes mode coverage. All values are percentages; higher is better.

Variant	Self-P→G PCK@2	Cross-P→G PCK@2	G→P tPCK@2	Self-P→G MC	G→P temporal coverage
A0: Full	58.83	4.14	80.52	24.48	35.02
A1: w/o correspondence	3.54	0.09	2.75	2.90	1.48
A2: w/o P→G	26.04	0.19	71.49	12.56	26.49
A3: w/o G→P	57.78	4.04	73.49	23.81	24.20
A4: w/o Cross-P→G	58.82	0.31	79.62	24.20	34.26
A5: Hard single-point targets	60.50	4.10	81.90	26.64	40.06
A7: Direct-only pipeline	58.49	3.72	80.34	23.96	33.79
A8: Track-only pipeline	58.55	4.05	79.20	24.22	34.12

Allocation within valid supports. For $m_k = A_u^s(B_k)$, the dominant share is $\max_k m_k / \sum_k m_k$. Queries with zero denominators are excluded. On the shared multi-mode queries, A0 and A5 have dominant shares of 71.73% and 73.63%, respectively, using the same averaging. Their total support masses are 24.48% and 26.64%. A5 has higher PCK and coverage, whereas A0 distributes

valid-support mass more evenly. This does not establish a causal explanation for their generation differences.

B Additional Method Details

B.1 Pseudo-Correspondences

Direct matching and refinement. Frozen DINOv3 ViT-L/16 provides normalized final-layer descriptors. Matching is restricted to resized, eroded person-side and reference-garment masks. Let $C_t(i,j)$ be the cosine similarity between person descriptor i in frame t and garment descriptor j. Valid descriptors are matched by $j^*(i) = \arg\max_j C_t(i,j)$ and $i^*(j) = \arg\max_i C_t(i,j)$. Their 2D descriptor-grid coordinates are $c_D^P(i)$ and $c_D^G(j)$. The round-trip errors are

$$\begin{aligned} e_P(i) &= \left\| c_D^P(i) - c_D^P(i^*(j^*(i))) \right\|_2, \\ e_G(i) &= \left\| c_D^G(j^*(i)) - c_D^G(j^*(i^*(j^*(i)))) \right\|_2. \end{aligned} \tag{8}$$

A strict match satisfies $\max(e_P, e_G) \le \gamma$, where γ is the round-trip error threshold. Ambiguity filtering requires a similarity margin $m(i) \ge 0.02$ over candidates outside a neighborhood of $j^*(i)$. The softmax-weighted concentration $\ell(i)$ of top-ranked candidates inside that neighborhood must be at least 0.55. Anchor matching may relax γ when both tests pass; refinement uses the strict threshold. Candidate confidence is

$$\begin{aligned} q(i) = \operatorname{sigm}\left(\frac{C_t(i,j^*(i)) - 0.35}{0.08}\right) \operatorname{sigm}\left(\frac{m(i) - 0.03}{0.02}\right) \\ \times \exp\left(-\frac{e_P(i) + e_G(i)}{4\kappa}\right) \ell(i)^{1/4}, \end{aligned} \tag{9}$$

where $\operatorname{sigm}(x) = (1 + e^{-x})^{-1}$ and κ is 1 at 448-pixel resolution and 1.8 at 832. Confidence guides filtering and query selection, not target-component weights.

Anchor selection prioritizes DWPose (Yang et al., 2023) frontal visibility and frame quality, then DINO matching quality and temporal separation. At 832-pixel resolution, we select up to four anchors at least 12 frames apart and up to 96 anchor queries. Refinement applies the thresholds in Table 8, followed by confidence-ordered non-maximum suppression on both grids.

Table 8: Direct-match filtering settings at 448/832-pixel resolutions. Radii use descriptor-token units; sharpness is a robust within-video score.

Setting	Value
Quality pass: frontal score / sharpness / mask ratio	$0.26/-1.5/0.03$
Refinement: frontal score / sharpness / mask ratio	$0.60/0/0.05$
Refinement cycle radius (448/832)	$1.8/3.24$
Refinement minimum confidence (448/832)	$0.70/0.75$
Refinement suppression radius, both grids (448/832)	$1/2$

Trajectory propagation and fusion. At 832-pixel resolution, each anchor patch supplies 12 points for CoTracker3. Tracking uses a 60-frame offline window with backward tracking. Aggregation requires at least three points with visibility above 0.5. Tracking is reinitialized at intermediate frames to check agreement at reliable anchors. The validity score is the product of scores for visibility, spatial compactness, agreement with reinitialized tracks, appearance consistency, frontal visibility, and person-side garment-mask validity. Thresholds are 0.25 for frontal score, 0.5 for appearance similarity, and 0.05 for validity.

Fusion prioritizes anchors. Direct refinement fills uncovered queries. It replaces conflicting trajectory targets beyond the spatial tolerance only when appearance, tracking agreement, or visibility is weak. Reliable conflicting tracks keep their targets with lower sampling priority; matches within tolerance keep the existing target.

Training-query selection. Person-video queries are partitioned by latent frame and a 4×4 spatial grid, retaining each cell's highest-confidence query in the candidate pool. Selection allocates 60% of the budget to frame-balanced coverage, 15% to uniform random sampling, and the rest by confidence with a density penalty. Garment-query selection balances spatial coverage and preserves the highest-confidence represented group per latent frame. Stages 2/3 cap P→G queries at 640/1,024 and G→P groups at 256/512.

B.2 Latent Remapping

Temporal indices. For a clip of T_{src} sampled frames starting at s with stride d, source indices are $b_j = s + dj$, $j = 0, \ldots, T_{\mathrm{src}} - 1$. We reindex retained observations by j in Eq. 1 and exclude observations outside these frames. Let T_{enc} be the VAE input length after padding. The Wan VAE (Team Wan et al., 2025) handles the first frame separately, then groups four frames:

$$\tau_{\mathrm{VAE}}(j) = \begin{cases} 0, & j = 0, \\ 1 + \lfloor (j-1)/4 \rfloor, & j > 0, \end{cases} \qquad F_P = 1 + \lfloor (T_{\mathrm{enc}} - 1)/4 \rfloor. \tag{10}$$

Valid labels satisfy $0 \le j < T_{\mathrm{src}}$ and $0 \le \tau_{\mathrm{VAE}}(j) < F_P$; unencoded trailing groups are excluded. Probes repeat the last of 64 source frames to encode 65 frames, giving $F_P = 17$. All real frames remain eligible; padding creates no additional labels. Images use $T_{\mathrm{src}} = T_{\mathrm{enc}} = F_P = 1$.

Spatial indices. For domain $D \in \{P, G\}$, let the source descriptor grid be $\widehat{H}_D \times \widehat{W}_D$ and the target token grid be $H_D \times W_D$. For source index $a = y\widehat{W}_D + x$, we remap each axis from length n to length m using ρ:

$$\rho(i; n, m) = \begin{cases} 0, & n = 1 \text{ or } m = 1, \\ \mathrm{clip}_{[0, m-1]}\Big(\mathrm{round}\, \frac{i(m-1)}{n-1}\Big), & \text{otherwise}, \end{cases}$$
$$\mathcal{R}_D(a) = W_D\, \rho(y; \widehat{H}_D, H_D) + \rho(x; \widehat{W}_D, W_D). \tag{11}$$

Rounding uses the nearest integer with ties to even; invalid indices are excluded. Equal-sized grids map identically. The VAE's spatial stride of eight and $1 \times 2 \times 2$ DiT patches give a 16-pixel stride, matching ViT-L/16. Images use stored dense correspondences directly. Flattened indices are $fH_PW_P + yW_P + x$ for person-video tokens and $yW_G + x$ for garment tokens. Here f is the latent-frame index and (x, y) are spatial token coordinates.

B.3 Distributional Targets

Let $\mathcal{I}_{\mathrm{clip}}$ index valid mapped observations and $e_{u,g} = \sum_{i \in \mathcal{I}_{\mathrm{clip}}} \mathbf{1}[u_i = u, \bar{g}_i = g]$. P→G retains up to $K = 4$ centers with the largest nonzero counts: $g_{u,1}, \ldots, g_{u,K_u}$, where $K_u = |\mathcal{M}_u|$. Counts select centers but do not weight components. Reverse groups $\mathcal{Q}_g = \{u \mid e_{u,g} > 0\}$ use all mapped observations before P→G query selection or top-K truncation. Query selection is independent in each direction; empty groups are omitted.

For $D \in \{P, G\}$, let $c_D(a)$ return spatial token coordinates, omitting time. The unnormalized Gaussian $\phi_D(a; b, s) = \exp(-\|c_D(a) - c_D(b)\|_2^2/(2s^2))$ has center b and bandwidth $s > 0$. We write $G_s(g; g_0) = \phi_G(g; g_0, s)$.

Video P→G targets. Let s_s and s_m be the single- and multi-center bandwidths, and let $\mu_G \in [0, 1]$ be the single-center mixing coefficient. The unnormalized target is

$$\widetilde{T}_u(g) = \begin{cases} \mu_G\, \mathbf{1}[g = g_{u,1}] + (1 - \mu_G) G_{s_s}(g; g_{u,1}), & K_u = 1, \\ \dfrac{1}{K_u} \displaystyle\sum_{k=1}^{K_u} G_{s_m}(g; g_{u,k}), & K_u > 1. \end{cases} \tag{12}$$

We normalize after mixing: $T_u(g) = \widetilde{T}_u(g) / \sum_{g' \in \mathcal{G}} \widetilde{T}_u(g')$. In Eq. 2, κ_G is the single-center mixture for $K_u = 1$ and G_{s_m} otherwise; equal coefficients apply before normalization. The support S_u is the union of Euclidean neighborhoods around these centers, using the single- or multi-center radius in Table 9.

Video G→P targets. For a nonempty $\mathcal{Q}_g$, bandwidth $s_P > 0$, and mixing coefficient $\mu_P \in [0, 1]$, the person-side component is

$$\kappa_P(v; q) = \begin{cases} \mu_P \mathbf{1}[v = q] + (1 - \mu_P)\phi_P(v; q, s_P), & |\mathcal{Q}_g| = 1, \\ \phi_P(v; q, s_P), & |\mathcal{Q}_g| > 1. \end{cases} \tag{13}$$

The single-center condition applies to the entire garment group. Eq. 2 restricts each component to its center's latent frame. It combines components with equal coefficients and normalizes over the full $\mathcal{P}$. The support S_g^P is the union of radius-ρ_P spatial neighborhoods within the centers' latent frames. Neither S_u nor S_g^P truncates the target distributions.

Image targets. For images, $F_P = 1$. Let g_u be the garment target for query u and $\mu_G^{\text{img}} \in [0, 1]$ its mixing coefficient. We normalize the Gaussian first, $\widehat{G}_s(g; g_u) = G_s(g; g_u) / \sum_{g' \in \mathcal{G}} G_s(g'; g_u)$, then use

$$T_u^{\text{img}}(g) = \mu_G^{\text{img}} \mathbf{1}[g = g_u] + (1 - \mu_G^{\text{img}})\widehat{G}_s(g; g_u). \tag{14}$$

Image G→P uses the same component construction with $F_P = 1$ and may contain multiple person locations.

Table 9: Target settings by training stage. ρ denotes the support radius; bandwidths and radii are in token spacings. Stage 3 continuation uses Stage 3 settings.

Target / parameters	Stage 1	Stage 2	Stage 3
Image P→G: $(s, \mu_G^{\text{img}}, \rho)$	$(0.80, 0.65, 1.75)$	$(0.60, 0.68, 1.75)$	$(1.10, 0.68, 3.50)$
Video P→G, single: (s_s, μ_G, ρ)	–	$(0.55, 0.70, 1.75)$	$(0.95, 0.70, 3.0)$
Video P→G, multiple: (s_m, ρ)	–	$(0.75, 2.25)$	$(1.25, 4.0)$
G→P: (s_P, μ_P, ρ_P)	$(0.75, 0.60, 2.0)$	$(0.75, 0.60, 2.0)$	$(1.2, 0.60, 3.2)$

B.4 Attention Supervision and Generation Loss

Self-P→G and G→P reuse generation self-attention queries and keys after root mean square (RMS) normalization and rotary positional embeddings (RoPE). Cross-P→G uses shared RMS-normalized cross-attention queries and garment-specific RMS-normalized keys, without a positional transform.

For $B = 8$ noise bins and L supervised blocks, we jointly learn logits $\eta \in \mathbb{R}^{B \times L}$ with the DiT. Noise level σ selects $b(\sigma) = \text{clip}(\lfloor B(1 - \sigma) \rfloor, 0, B - 1)$. Shared block weights use a unit-temperature softmax:

$$\pi_{b,:} = \text{softmax}(\eta_{b,:}), \qquad \bar{A}_\sigma^\chi = \sum_{\ell=1}^{L} \pi_{b(\sigma),\ell} A_\ell^\chi, \tag{15}$$

where $\chi \in \{s{:}P \to G, x{:}P \to G, G \to P\}$ identifies the readout.

Alignment losses. Eq. 4 evaluates $\log p$ as $\log \max\{p, \varepsilon\}$, with $\varepsilon = 10^{-8}$. The entropy gate is

$$\gamma_{\text{ent}}(T, S) = \text{clip}\left(1 - \frac{\text{Ent}(T)}{\log(\max(|S|, 2))}, 0, 1\right). \tag{16}$$

For a distribution P, let $P(S) = \sum_{j \in S} P(j)$. Exact conditioning gives $P^S(i) = P(i)\mathbf{1}[i \in S]/P(S)$ when $P(S) > 0$. The implementation uses the denominator $\max\{P(S), \varepsilon\}$ and floored log arguments. The restricted weights sum to one when $P(S) \geq \varepsilon$ and to less than one otherwise. The implemented KL is thus a stabilized approximation. Before evaluating $\mathcal{D}_G$, we floor the reverse row at ε and renormalize it over its full key domain.

Eq. 5 uses single-sample notation. Minibatch reduction sums numerator terms, then divides by total valid-query count for P→G or total garment-group weight for G→P. Sample-specific noise weights scale numerators only.

Generation loss. Let z_0 be the video VAE latent of the ground-truth sample and let $\epsilon \sim \mathcal{N}(0, I)$. Following flow matching, we form

$$z_\sigma = (1-\sigma)z_0 + \sigma\epsilon, \qquad v^* = \epsilon - z_0, \qquad \sigma \in [0,1]. \tag{17}$$

Let $\hat{v}_\theta$ be the predicted velocity and $e_\theta = \hat{v}_\theta - v^*$ its error. Let D count latent elements per sample and C count channels. We spatially downsample binary agnostic masks by nearest-neighbor interpolation. Their union within each latent-frame group forms $\widetilde{M}$. Broadcasting this mask over channels, with mask-loss weight $\lambda_{\text{mask}} \geq 0$, gives

$$\mathcal{L}_{\text{gen}} = \mathbb{E}\left[\frac{\|e_\theta\|_2^2}{D} + \lambda_{\text{mask}}\frac{\|\widetilde{M} \odot e_\theta\|_2^2}{\max\{C\|\widetilde{M}\|_1, \varepsilon\}}\right]. \tag{18}$$

C Training and Evaluation

C.1 Main-Model Training

Training uses 11,647 VITON-HD images, 48,392 DressCode images, and 7,759 ViViD videos, with AdamW and bfloat16 on eight NVIDIA A800 80GB GPUs (Table 10). Stages 1–2 and the first 3,500 Stage 3 updates use learning rate 10^{-5} and 100 warmup steps. Before the final 2,000 updates, we reset the optimizer and scheduler and increase the video repeat factor from 3 to 4. We use 50 warmup steps followed by cosine decay from 5×10^{-6}. We evaluate the final checkpoint.

Table 10: Main-model training schedule. Resolutions are width × height; batch sizes are effective global values. Update counts include auxiliary-resolution generation-only updates.

Stage	Primary image resolution	Primary video resolution / nominal frames	Updates	Batch
1	384×512	–	2,400	64
2	384×512	336×448 / 81	6,000	16
3	768×1024	624×832 / 65	5,500	16

Table 11 lists auxiliary updates. Resolution bases scale aspect-ratio buckets; dimensions are rounded down to multiples of 16. Base b gives $3b/4 \times b$ pixels for the standard portrait bucket. In Stages 2–3, each auxiliary stream uses a deterministic $2:1$ video:image update ratio.

Table 11: Auxiliary generation-only update schedule. Each cycle runs primary, low-, then high-resolution updates in the listed ratios. Bases are in pixels; parentheses give integer repetition weights. Video frame counts are nominal. Stage 3 includes continuation.

Stage / stream	Update cycle primary:low:high	Image bases	Video bases / frames
1 / low	9 : 1 : 0	256, 320, 384 (3:1:1)	–
2 / low	10 : 1 : 0	256, 320, 384, 448 (5:1:1:3)	256, 320, 384 / 81 (3:1:1)
3 / low	16 : 2 : 1	256, 320, 384, 448, 832 (5:1:1:1:3)	256, 320, 384 / 81 (3:1:1)
3 / high		1088	896, 960, 1024, 1088 / 41 (1:1:2:2)

Correspondence schedules. The correspondence weight decreases from 1 to $\omega_{\min}$ as noise rises from σ_{low} to σ_{high}:

$$\omega(\sigma) = \omega_{\min} + (1-\omega_{\min})\,\text{clip}\left(\frac{\sigma_{\text{high}} - \sigma}{\sigma_{\text{high}} - \sigma_{\text{low}}}, 0, 1\right). \tag{19}$$

Let $s(n)$ count completed primary-resolution updates at optimization step n within the current stage. Let a_m be the scale for modality $m \in \{\text{img}, \text{vid}\}$. Let c_P, c_G be base coefficients, S_P, S_G warmup

Table 12: Main-model loss and schedule settings. Warmup and start values count primary-resolution updates; c_P, c_G precede modality scaling.

Parameter	Stage 1	Stage 2	Stage 3	Stage 3 cont.
c_P	0.024	0.024	0.016	0.016
c_G	0.008	0.008	0.0072	0.0045
$(a_{\text{img}}, a_{\text{vid}})$	$(1, -)$	$(0.5, 1.5)$	$(0.4, 0.75)$	$(0.4, 0.75)$
S_P	500	500	400	400
S_0	400	200	200	200
S_G	800	500	400	400
λ_x	0.6	0.6	0.6	0.4
$(\lambda_{\text{ent}}, \lambda_{\text{out}})$	$(0.015, 0.16)$	$(0.015, 0.16)$	$(0.012, 0.12)$	$(0.012, 0.12)$
$(\omega_{\min}, \sigma_{\text{low}}, \sigma_{\text{high}})$	$(0.1, 0.6, 0.9)$			
$(\lambda_{\text{shape}}, \lambda_{\text{region}})$	$(0.25, 0.25)$			
p_x	0.6			
λ_{mask}	1.0			

lengths, and S_0 the G→P start step. The scheduled outer coefficients are

$$\lambda_P(n; m) = a_m c_P \min\left(\frac{s(n)}{S_P}, 1\right),$$
$$\lambda_G(n; m) = a_m c_G \,\text{clip}\left(\frac{s(n) - S_0}{S_G}, 0, 1\right). \tag{20}$$

Table 12 lists the settings. Auxiliary-resolution updates use only $\mathcal{L}_{\text{gen}}$ and do not advance s. The counter resets between stages, but not at Stage 3 continuation.

C.2 Ablation Setup

The 1.3B ablations use a fixed 20% subset of 13,560 samples, stratified by dataset, garment category, and pseudo-label availability. Proportional cluster quotas and density-regularized farthest-point selection encourage diversity in garment appearance and person/video structure.

Stages 1–3 use 500/1,200/1,600 updates, learning rates $(1.2, 1.2, 1.0) \times 10^{-5}$, and effective batch sizes 96/24/16 on four GPUs. Each stage uses 20 warmup steps, then a constant learning rate. Variants continue from their own preceding-stage checkpoints. Evaluation uses Stage 3 update 1,600 and 25 denoising steps.

All conditioning pathways remain active. Enabled Self-P→G and G→P losses have unit internal multipliers; Cross-P→G uses $\lambda_x = p_x = 0.6$. Outer coefficients and alignment weights follow Stages 1–3 of Table 12, with (S_P, S_0, S_G) shortened to $(100, 80, 160)$, $(100, 40, 100)$, and $(80, 40, 80)$.

After remapping, A5 selects the garment token with the highest summed observation confidence per person-video query. Each garment query then selects one person-video token from these retained matches by the same rule. Both directions use one-hot targets and singleton supports, with unchanged loss coefficients. Exact conditional KL is zero; the stabilized implementation is unchanged.

A7/A8 retain Full's training records and image supervision. They use independently obtained direct or propagated non-seed video matches, respectively. All valid positive-confidence queries are retained without coverage-balanced selection. Retained correspondence frames guide clip sampling within the same videos. A8 still initializes tracking from direct matches.

C.3 BiCorr-Test and Evaluation Protocols

BiCorr-Test selection. BiCorr-Test contains 123 Eevee (Zeng et al., 2026) and 27 ViT-HD (He et al., 2026) test clips with clear garment details and predominantly frontal views. Back and strict profile views are excluded; mild three-quarter views are allowed. Each clip contains 64 consecutive

source frames, resized with aspect ratio preserved and centrally padded to 624×832. Clip identities and frame intervals were fixed before inspecting model outputs or scores.

Both sources use MediaPipe Pose (Bazarevsky et al., 2020) shoulder/hip geometry and landmark visibility to screen orientation. Eevee adds a DensePose (Güler et al., 2018) front-torso check; ViT-HD adds shoulder/hip ordering and facial visibility checks. We interpolate brief detection gaps and smooth pose scores. Frontal windows are ranked by orientation scores and manually checked for detection errors.

Evaluation. Baseline scores in Tables 1 and 4 come from MagicTryOn (Li et al., 2025b) (arXiv v3) and KeyTailor (He et al., 2026). The fine-tuned Wan2.1-I2V scores come from MagicTryOn's Table 2.

For the BiCorr-Test baselines, we use publicly released weights and inference configurations based on the authors' recommendations (Table 13). CORAL and Any2AnyTryon generate a try-on reference from the first source-video frame and the assigned garment. StableAnimator animates this reference using the source video's DWPose (Yang et al., 2023) sequence.

Table 13: Released checkpoints and generation sizes for the BiCorr-Test baselines. Dimensions are width $\times$ height.

Method	Checkpoint	Size
ViViD	Official release	624×832
CatV2TON	`512-64K`	384×512
MagicTryOn	Official 14B release	624×832
Any2AnyTryon (image)	FLUX.1-dev + released LoRA	624×832
CORAL (image)	FLUX.1-Fill-dev + `coral_dc`	624×832
StableAnimator (animation)	Released SVD + Animation weights	624×832

Our model outputs 624×832 videos and 768×1024 images (width $\times$ height). Video inputs have 64 frames at 25 fps and are padded to 65 for VAE alignment. Saved outputs retain 64 frames. BiCorr-Test methods share clips and fixed garment assignments; unpaired garments come from the same category. Video SSIM and SqueezeNet-based LPIPS use full-resolution frames.

C.4 LHFSC Implementation

We use aligned RGB frames X, Y with pixel values in $[0, 1]$. We sample 32×32 patches fully inside the shared input agnostic mask M, using stride 16 from the image origin. Let $X_{p,c}$ denote patch p in channel c, w the separable periodic Hann window, and $\mu^w_{p,c}(X)$ the weighted patch mean. The high-frequency magnitudes are

$$A_{p,c}(X) = \left|\mathcal{F}_2\left[w \odot \left(X_{p,c} - \mu^w_{p,c}(X)\right)\right]\right|_{\mathcal{H}}. \tag{21}$$

Here $\mathcal{F}_2$ is the orthonormal 2D Fourier transform, and $\mathcal{H} = \{f : \|f\|_2 \geq 1/8\}$ selects frequencies in cycles per pixel. Concatenating magnitudes across patches, RGB channels, and frequency bins gives $a(X; M)$ in Eq. 7, with $\varepsilon = 10^{-12}$. Errors and energies are pooled before division, without radial averaging. We average scores over frames with at least one complete in-mask patch, then equally over videos. Videos without valid frames are excluded.

Pairs with zero high-frequency magnitudes score 1. The agnostic mask may include skin or background, and high-frequency responses can reflect folds, boundaries, or noise as well as garment patterns.