# ADAPTIVELY PIVOTED FLOW MATCHING FOR COMPOSITION-CONSISTENT TRANSLATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Controlled image translation requires jointly achieving target fidelity, source equivariance under appearance interventions, and compositional consistency across sequential paths. Factoring conditional transport through a shared latent pivot ($T_{s\to t} = D_t \circ E_s$) naturally induces a translation groupoid structure: this algebraic formulation inherently ensures compositional consistency under exact integration, guarantees bi-Lipschitz source sensitivity against information collapse, and introduces a structural inductive bias for source equivariance. While this architecture is coupled with a standard Gaussian prior, overlapping interpolation distributions associate the same state, time, and condition with different target velocities, producing conditional velocity ambiguity. To reduce this ambiguity, we introduce Adaptively Pivoted Flow Matching (AP-FM), which replaces the single Gaussian with an adaptive prior comprising trainable multi-Gaussian modes optimized with a prior mode adaptation loss defined using a frozen reference field. At inference, the observed source determines the latent pivot; neither identity labels nor mode lookup is required. On SRN Cars and DeepFashion, AP-FM preserves source appearance under interventions and composed translations, with improved target fidelity over a standard Gaussian prior.

## 1 INTRODUCTION

Controlled image translation renders a source observation under a new condition while preserving its identity and appearance, spanning tasks such as novel-view synthesis under camera changes (Yu et al., 2021; Liu et al., 2023b) and pose-guided person generation under skeletal changes (Ma et al., 2017; Bhunia et al., 2023). Beyond isolated target fidelity, a robust translation system must satisfy two structural properties: *source equivariance*, where appearance interventions on the source induce matching modifications on the target, and *compositional consistency*, where translating through intermediate conditions agrees with direct translation. For instance, modifying an object's intrinsic surface texture or color should faithfully manifest across target configurations, while reaching a target configuration through multiple intermediate steps should yield the same result as a direct translation, without accumulating drift.

The operational role of the source exposes fundamental limitations in existing translation architectures. Standard source-conditioned models generate targets from noise while injecting the source observation as static network context (Liu et al., 2023b; Bhunia et al., 2023); lacking a dynamical sensitivity lower bound in the source input, they frequently attenuate fine-grained appearance changes. Direct data-to-data transport instead integrates continuous paths directly between source and target pairs (Liu et al., 2022; 2023a; Zhou et al., 2024). While invertible for individual condition pairs, independent pairwise fields lack a shared latent pivot, causing composed translations to accumulate structural drift.

Factoring translation through a shared latent space provides a natural way to coordinate transformations across different conditions (Grover et al., 2020; Su et al., 2023). Let $E_v$ encode an observation into the prior space under condition index $v$, and let $D_v = E_v^{-1}$ denote the reverse decoding flow. The set of composite maps $T_{s\to t} = D_t \circ E_s$ naturally forms a **translation groupoid**: intermediate encoding and decoding operations cancel identically, guaranteeing compositional consistency across arbitrary condition paths without path regularization (Proposition 1). Simultaneously, standard flow regularity establishes *bi-Lipschitz source sensitivity*, which guarantees that distinct source features



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.

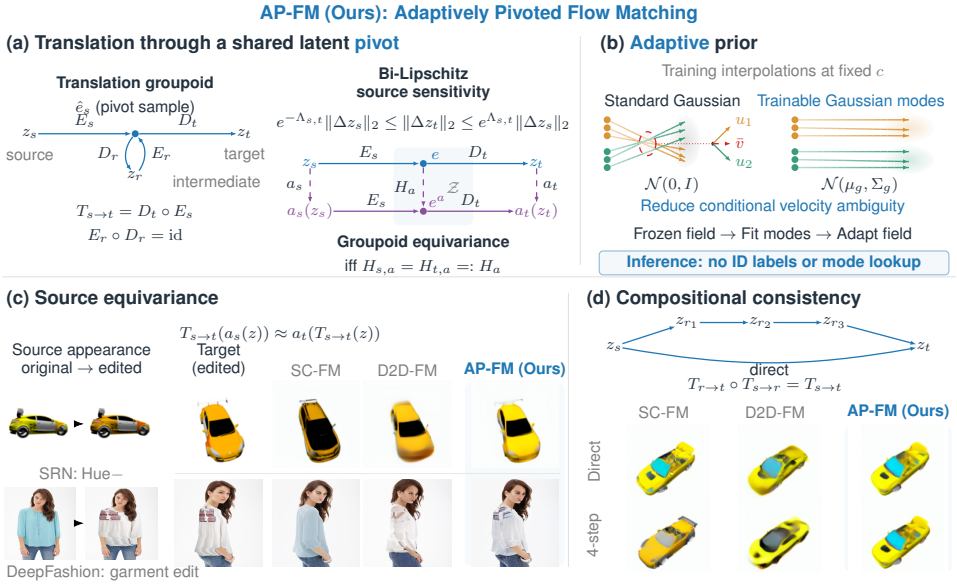


Figure 1: **Overview of Adaptively Pivoted Flow Matching (AP-FM).** (a) Translation via a shared latent pivot induces a translation groupoid, with compositional consistency, bi-Lipschitz source sensitivity and a prior-domain characterization of source equivariance. (b) Fitting adaptive prior modes under a frozen reference field and adapting the field reduce conditional velocity ambiguity; $\bar{v}$ is the posterior mean target velocity. (c,d) On source-appearance interventions and multi-step translation, AP-FM preserves appearance changes while maintaining compositional consistency.

neither collapse nor amplify unboundedly during transport. The encoded intermediate state therefore acts as a *shared latent pivot* across conditions, enabling us to characterize groupoid equivariance as the consistency of *induced prior transformations* (Proposition 2): an appearance intervention is transported consistently across conditions whenever its induced transformation in the prior domain is condition-invariant.

However, coupling this architecture with a shared standard Gaussian prior introduces a different problem: interpolation state distributions from distinct observations can overlap, assigning different target velocities to the exact same state and flow time. We formalize this phenomenon as *conditional velocity ambiguity*, which induces irreducible velocity-regression error, ultimately degrading target fidelity and attenuating transferred appearance. To resolve this ambiguity, we introduce **Adaptively Pivoted Flow Matching (AP-FM)** (Figure 1), which replaces the single Gaussian with adaptive priors comprising trainable Gaussian modes—shared across views of an identity while accommodating shared modes across identity groups. This design reshapes the *training coupling* without altering the shared-pivot translation structure: compositional consistency remains algebraically inherited from the translation groupoid, while the mitigated velocity ambiguity eases flow regression, yielding superior target fidelity and more faithful, coherent appearance transport. At inference, the observed source is routed through the latent pivot, requiring neither identity labels nor mode retrieval.

Experiments on SRN Cars and DeepFashion evaluate target fidelity, source equivariance, and compositional consistency, with coupling-geometry and field-fitting diagnostics assessing how adaptive priors alter interpolation geometry and velocity-regression residuals.

Our contributions are:

- **Translation groupoid formulation**: We formalize controlled translation as a translation groupoid that guarantees compositional consistency and bi-Lipschitz source sensitivity, characterizing how a shared latent pivot provides the structural inductive bias for groupoid equivariance.

- **Adaptively Pivoted Flow Matching (AP-FM)**: We identify conditional velocity ambiguity under independent Gaussian couplings as the primary bottleneck, and AP-FM employs a two-stage procedure that first optimizes adaptive prior modes under a frozen reference field to establish an ambiguity-reducing training coupling, and then adapts the translation field.
- **Empirical and diagnostic validation**: Across novel-view synthesis and articulated pose transfer, we demonstrate marked advantages in target fidelity, source equivariance under appearance interventions, and multi-step composition over conditioning and unpivoted baselines, substantiating the mitigation of conditional velocity ambiguity via coupling geometry and field-fitting diagnostics.

## 2 Related Work

**Consistent image translation.** Consistency across image domains is classically enforced through cycle-consistency losses (Zhu et al., 2017) or learned pairwise diffusion bridges (Liu et al., 2023a; Zhou et al., 2024). Such pairwise formulations do not by themselves coordinate translations across multiple conditions. AlignFlow (Grover et al., 2020) instead uses invertible domain-specific normalizing flows through a shared latent space, guaranteeing exact cycle consistency, while DDIB (Su et al., 2023) composes independently trained diffusion probability-flow ODEs through a common latent distribution, with practical consistency limited by ODE discretization (Su et al., 2023). Our formulation adopts this shared-pivot structure. We make the resulting multi-condition composition law explicit through a translation-groupoid formulation and characterize its source sensitivity and appearance equivariance in Section 3.1. Our main departure is then to study the *training coupling* within this shared-pivot structure: a common Gaussian endpoint can induce conditional velocity ambiguity, which motivates the adaptive prior.

**Flow matching and prior design.** Flow matching trains velocity fields along prescribed paths (Lipman et al., 2023; Liu et al., 2022). Beyond improving sample pairings via optimal transport (Pooladian et al., 2023; Cheng & Schwing, 2025), prior design alters the endpoint geometry: MM-FM fits Gaussian mixtures to target representations (Luo et al., 2026), while CSFM (Kim et al., 2026) learns condition-dependent source distributions. Unlike these methods, AP-FM optimizes identity-shared prior modes via coupling optimization under a frozen reference field, cleanly decoupling prior fitting from field adaptation. This structured coupling separates intermediate interpolation densities across flow time, reducing conditional velocity ambiguity ($C_{\text{vel}}$) without the instability of joint training.

**View- and pose-conditioned generation.** Novel-view synthesis coordinates source appearances with viewing rays (Yu et al., 2021; Liu et al., 2023b; Zheng & Vedaldi, 2024), whereas pose-guided generation maps identity references onto articulated body skeletons (Bhunia et al., 2023; Fan et al., 2025). Both domains demand that the source identity and fine-grained appearance remain strictly preserved across dramatic geometric changes. However, existing architectures largely treat these as one-way conditional generation tasks, overlooking whether the learned transport satisfies composition across sequential conditions or preserves controlled appearance interventions. This work provides a unified lens to assess how the transport geometry and training prior govern fidelity, source equivariance, and path consistency beyond isolated one-step generation. Appendix H details related multiview and virtual try-on models.

## 3 Methodology

AP-FM separates the *translation structure* from the *training coupling*. We first formulate translation through a shared latent pivot, whose invertible conditional flows define a translation groupoid with compositional consistency, bi-Lipschitz source sensitivity, and a structural bias for groupoid equivariance (Section 3.1). We then show how a standard Gaussian prior induces conditional velocity ambiguity and introduce adaptive Gaussian modes that reshape this coupling before field adaptation, while leaving the shared-pivot inference structure unchanged (Section 3.2).

### 3.1 Flow Translation through a Shared Latent Pivot

Let $v \in \mathcal{V}$ index conditions, with $c_v$ the corresponding network conditioning representation. Source and target condition indices are denoted by $s, t \in \mathcal{V}$. For identity $k \in \{1, \ldots, K\}$, let $z_{k,v} \in \mathbb{R}^D$ denote its latent observation under condition index $v$. We reserve $\tau \in [0, 1]$ for flow time and $v_\theta$ for the velocity field. Coupling this architecture with a standard Gaussian prior yields the Gaussian-Prior Flow Matching (GP-FM) (Appendix B), which trains the conditional field $v_\theta$ via:

$$\mathcal{L}_{\mathrm{GP}}(\theta) = \mathbb{E}_{k,v,\tau,\epsilon}\left[\frac{1}{D}\left\|v_\theta(x_\tau, \tau, c_v) - u\right\|_2^2 \,\middle|\, x_\tau = (1-\tau)z_{k,v} + \tau\epsilon,\ u = \epsilon - z_{k,v}\right], \quad (1)$$

where $\epsilon \sim \mathcal{N}(0, I_D)$ is sampled independently of the observation and $\tau \sim \mathcal{U}[0, 1]$. Data lie at $\tau = 0$ and the prior at $\tau = 1$.

**Translation through a shared latent pivot.** Write $\mathcal{Z}$ for the latent state space. Let $\Phi^v_{a,b}$ solve the velocity field ODE $\dot{x}_\tau = v_\theta(x_\tau, \tau, c_v)$ from time $a$ to $b$. The forward encoding flow $E_v = \Phi^v_{0,1} : \mathcal{Z} \to \mathcal{Z}$ and backward decoding flow $D_v = \Phi^v_{1,0} : \mathcal{Z} \to \mathcal{Z}$ implement latent-pivoted translation (Grover et al., 2020; Su et al., 2023):

$$z_s \xrightarrow{E_s} \hat{e}_s \xrightarrow{D_t} \hat{z}_t, \qquad T_{s\to t} = D_t \circ E_s. \quad (2)$$

Crucially, each flow operates strictly under its own condition, translating the observed source deterministically through the shared pivot $\hat{e}_s \in \mathcal{Z}$ without drawing fresh noise during inference.

**Flow regularity.** Throughout our analysis, each conditional velocity field is continuous in time and locally Lipschitz in state, ensuring unique forward and backward solutions across $[0, 1]$ on the domains considered. Consequently, $E_v$ and $D_v$ are homeomorphisms onto $\mathcal{Z}$, and ODE reversibility guarantees $D_v = E_v^{-1}$.

#### 3.1.1 Compositional Consistency via Translation Groupoids

Controlled translation across varying conditions naturally realizes a translation groupoid (Brown, 1987), for which objects are index-sample pairs $(v, z_v)$ and morphisms are pairwise translation operators $T_{s\to t} = D_t \circ E_s : \mathcal{Z} \to \mathcal{Z}$. By routing transport through the shared latent pivot, intermediate encoding and decoding flows cancel identically under exact integration, algebraically enforcing compositional consistency across arbitrary condition paths without auxiliary path constraints or round-trip regularizers.

**Proposition 1** (Compositional consistency)**.** *The shared translation operators $\{T_{s\to t}\}_{s,t}$ define a translation groupoid satisfying:*

$$T_{s\to s} = \mathrm{id}_{\mathcal{Z}}, \qquad T_{s\to t}^{-1} = T_{t\to s}, \qquad T_{r\to t} \circ T_{s\to r} = T_{s\to t}. \quad (3)$$

*Proof.* Identity and invertibility follow directly from ODE reversibility $D_v \circ E_v = \mathrm{id}_{\mathcal{Z}}$. For any sequence of conditions $s \to r \to t$, composition cancels at the intermediate pivot: $T_{r\to t} \circ T_{s\to r} = (D_t \circ E_r) \circ (D_r \circ E_s) = D_t \circ (E_r \circ D_r) \circ E_s = D_t \circ E_s = T_{s\to t}$. Associativity is inherited from standard function composition. □

By induction, exact cancellation holds along any finite path of arbitrary length. Consequently, $V$ conditional flows fully parameterize all $V(V-1)$ directed translations.

#### 3.1.2 Source Sensitivity, Groupoid Equivariance, and Latent Pivoting

**Bi-Lipschitz source sensitivity.** Unlike static source conditioning that injects context into noise evolution, latent-pivoted translation dynamically transports the source state itself. For velocity fields with integrable state-Lipschitz bounds $L_v(\tau)$, Gronwall's inequality (Teschl, 2012, Section 2.4) establishes a bi-Lipschitz guarantee:

$$e^{-\Lambda_{s,t}}\|z - z'\|_2 \le \|T_{s\to t}(z) - T_{s\to t}(z')\|_2 \le e^{\Lambda_{s,t}}\|z - z'\|_2, \quad (4)$$

where $\Lambda_{s,t} = \int_0^1 (L_s(\tau) + L_t(\tau))\, d\tau$. The positive lower bound guarantees that distinct source observations cannot collapse to identical target representations, preserving fine-grained appearance information across translation.

**Groupoid equivariance.** Let $a_v : \mathcal{Z} \to \mathcal{Z}$ denote a prescribed appearance transformation conditioned on view or pose $c_v$. The condition-dependent appearance transformation $a_v$ is *translation groupoid equivariant* if transforming the source before translation agrees with transforming the translated target:

$$T_{s\to t} \circ a_s = a_t \circ T_{s\to t} \qquad \text{for all } s, t. \tag{5}$$

Let $H_{v,a} := E_v \circ a_v \circ D_v : \mathcal{Z} \to \mathcal{Z}$ denote the transformation induced in the common latent pivot space. Proposition 2 characterizes this equivariance directly via pivot alignment.

**Proposition 2** (Prior-domain characterization)**.** *The condition-dependent appearance transformation $a_v$ is translation groupoid equivariant, satisfying $T_{s\to t} \circ a_s = a_t \circ T_{s\to t}$ for all $s, t$, if and only if its induced prior transformations, defined by*

$$H_{v,a} := E_v \circ a_v \circ D_v : \mathcal{Z} \to \mathcal{Z}, \tag{6}$$

*coincide across conditions:*

$$H_{s,a} = H_{t,a} \qquad \textit{for all } s, t. \tag{7}$$

*Proof.* Since $T_{s\to t} = D_t \circ E_s$ and flows are bijective with $D_v = E_v^{-1}$, we expand $T_{s\to t} \circ a_s = D_t \circ H_{s,a} \circ E_s$ and $a_t \circ T_{s\to t} = D_t \circ H_{t,a} \circ E_s$. Composing with $E_t$ on the left and $D_s$ on the right immediately yields $H_{s,a} = H_{t,a}$. □

Proposition 2 formalizes the structural inductive bias of latent pivoting: appearance transformations are preserved across condition shifts whenever they induce consistent representations in the shared pivot space $\mathcal{Z}$. Appendix A.1 provides the formal ODE flow regularity and bi-Lipschitz sensitivity proofs, while Appendix C.1 details the protocol connecting latent groupoid equivariance to empirical image-space interventions.

### 3.2 Adaptively Pivoted Flow Matching

Within the translation groupoid, an independent Gaussian prior $\mathcal{N}(0, I_D)$ induces overlapping interpolation distributions that pair identical states and conditions with conflicting velocities. We formalize their conditional variance as *conditional velocity ambiguity* (Appendix A.2), which sets an irreducible Bayes risk and forces squared-error regression to average divergent targets. To mitigate this ambiguity while strictly preserving the groupoid structure, Adaptively Pivoted Flow Matching (AP-FM) replaces the single Gaussian with adaptive priors optimized under a frozen reference field.

#### 3.2.1 Adaptive Prior Coupling

A fixed assignment $g : \{1, \dots, K\} \to \{1, \dots, M\}$ maps each identity to a diagonal Gaussian mode:

$$Q_{g(k)} = \mathcal{N}\Big(\mu_{g(k)}, \operatorname{diag}(\sigma^2_{g(k)})\Big), \qquad \epsilon = \mu_{g(k)} + \sigma_{g(k)} \odot \xi, \quad \xi \sim \mathcal{N}(0, I_D), \tag{8}$$

with trainable centroids and coordinate-wise scales $\mathcal{M} = \{(\mu_j, \sigma_j)\}_{j=1}^M$. Views of an identity share a mode, with independent noise draws across training pairs; multiple identities may share a mode. Neither identity nor mode labels enter the network condition. Fixing $M = 1$, $\mu_1 = 0$, and $\sigma_1 = \mathbf{1}$ recovers GP-FM.

Trainable mode centroids and scales reshape component likelihoods and posterior weights across flow time, providing a mechanism for reducing conditional velocity ambiguity by separating interpolation components with conflicting target velocities (Appendix A.2).

#### 3.2.2 Prior Mode Adaptation Loss

We first train a GP-FM reference field $v_{\theta_0}$. With its parameters frozen, we optimize the prior parameters $\mathcal{M}$ by minimizing the following prior mode adaptation loss:

$$\mathcal{L}_{\mathrm{B}}(\mathcal{M}; \theta_0) = \mathbb{E}_{k,v,\tau,\xi} \frac{1}{D} \left\| v_{\theta_0}\big((1-\tau) z_{k,v} + \tau\epsilon, \tau, c_v\big) - (\epsilon - z_{k,v}) \right\|_2^2. \tag{9}$$

Fresh draws of $\xi$ and $\tau \sim \mathcal{U}[0, 1]$ evaluate candidate priors across flow time. Gradients pass through both the field's state input and the velocity target to update mode centroids and scales. The prior mode adaptation loss favors couplings whose target velocities are well matched by the frozen reference field. Appendix E gives the Bridge fitting variants, initialization, and regularization.

#### 3.2.3 Field Adaptation and Inference

With $\mathcal{M}$ fixed, initialize $\theta$ from $\theta_0$ and fine-tune the field using the same squared-error formulation with fresh prior draws and interpolation times:

$$\mathcal{L}_{\mathrm{FM}}(\theta; \mathcal{M}) = \mathcal{L}_{\mathrm{B}}(\mathcal{M}; \theta). \tag{10}$$

With the coupling fixed, this stage optimizes the field's approximation to the conditional mean target velocity; intrinsic ambiguity remains fixed. Inference evaluates $\hat{z}_t = D_t(E_s(z_s))$ using the adapted conditional field.

## 4 Experiments

We compare AP-FM with source-conditioned FM (SC-FM), direct data-to-data FM (D2D-FM), and Gaussian-Prior Flow Matching (GP-FM), which implements the groupoid factorization $T_{s\to t}$ under a standard Gaussian prior. Appendix B provides their formulations and experiment specifications. We choose novel-view synthesis and pose-guided appearance transfer as complementary controlled-translation settings. Both provide multiple observations of the same underlying identity under changing conditions, enabling direct evaluation of source equivariance and multi-step composition. At the same time, they stress different transformation regimes: SRN changes rigid 3D viewpoint under camera conditioning, whereas DeepFashion changes non-rigid articulated structure under dense pose conditioning. This allows us to test whether the proposed transport structure generalizes across substantially different condition spaces rather than being specific to one task. Appendices F and G include further ablations and more qualitative comparisons.

### 4.1 Novel-View Synthesis on SRN Cars

We evaluate novel-view synthesis on SRN Cars (Sitzmann et al., 2019), parameterizing continuous viewpoint conditioning with Plücker ray embeddings.

#### 4.1.1 Evaluation Protocol

**Fidelity.** Target reconstruction quality is evaluated on unperturbed source-target view pairs using PSNR and SSIM (Wang et al., 2004).

**Source equivariance.** We probe groupoid equivariance (Proposition 2) using source appearance interventions at fixed camera conditions: hue shifts and RGB-to-GBR channel permutations. Let $\hat{y}, \hat{y}^a$ denote model predictions from the clean and intervened sources, and $y_t, y_t^a$ the corresponding ground-truth targets. Following the transformation-response paradigm of Lenc & Vedaldi (2015), we define a foreground RGB descriptor $h_{\mathrm{rgb}} \in \mathbb{R}^6$ from channel means and standard deviations and compare the induced appearance responses $\Delta_{\mathrm{pred}} = h_{\mathrm{rgb}}(\hat{y}^a) - h_{\mathrm{rgb}}(\hat{y})$ and $\Delta_{\mathrm{gt}} = h_{\mathrm{rgb}}(y_t^a) - h_{\mathrm{rgb}}(y_t)$. For each intervention, response cosine $C_{\mathrm{rgb}}$ measures directional agreement, while relative response discrepancy $D_{\mathrm{rgb}}$ measures the normalized response error. We additionally report SSIM and LPIPS (Zhang et al., 2018) against the transformed target $y_t^a$.

**Compositional consistency.** To test path independence (Proposition 1), we report target fidelity (SSIM/LPIPS against ground-truth $y_t$) for direct ($L = 1$) and composed translations ($L \in \{2, 4\}$) across path lengths.

Metric definitions and task-specific evaluation details, including interventions, path construction, and dataset splits, are provided in Appendix C.

#### 4.1.2 Comparison with Baselines

Injecting the source as static context, SC-FM severely attenuates appearance interventions with low response cosine $C_{\mathrm{rgb}}$ and high discrepancy $D_{\mathrm{rgb}}$ (Table 1), reflecting the lack of a dynamical sensitivity lower bound. D2D-FM restores dynamical sensitivity by evolving the source state directly, achieving higher response cosine $C_{\mathrm{rgb}}$. However, lacking a shared latent pivot, its compositional translations accumulate severe drift (Figure 3), leading to significant degradation in SSIM and LPIPS. GP-FM's shared latent pivot cancels intermediate steps, providing path stability and preserving global structure (stable SSIM) across $L = 4$ compositional translation. Yet, coupling to a single Gaussian

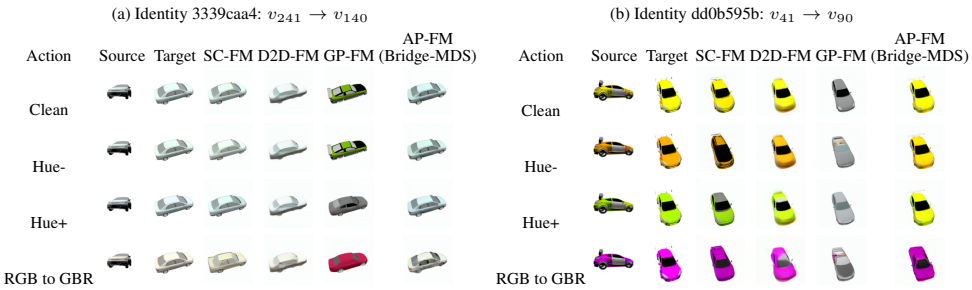


Figure 2: SRN source equivariance examples under hue shifts and channel permutation. Matched interventions are applied to the source and ground-truth target.

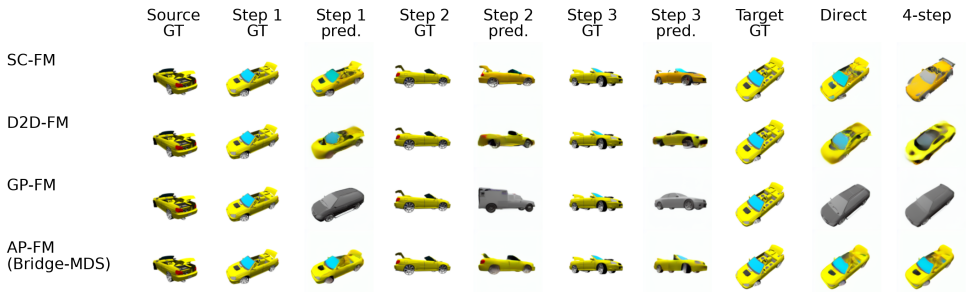


Figure 3: Four-step SRN composition along $82 \rightarrow 121 \rightarrow 41 \rightarrow 191 \rightarrow 241$.

Table 1: SRN Cars target fidelity, source equivariance, and compositional consistency.

| Method | Fidelity | Source equivariance | | | Compositional consistency |
|---|---|---|---|---|---|
| | Clean PSNR↑/SSIM↑ | Hue− | Hue+ | RGB→GBR | GT: direct/$L$=2/$L$=4 Rows: SSIM↑/LPIPS↓ |
| | | SSIM ↑ /LPIPS ↓ /$C_{\text{rgb}}$ ↑ /$D_{\text{rgb}}$ ↓ | | | |
| SC-FM | 23.17/.9005 | .8472/.1400 | .8558/.1255 | .8203/.1767 | .9224/.9131/.8964 |
| | | .481/2.31 | .497/2.04 | .672/1.69 | **.0717**/.0853/.1128 |
| D2D-FM | 19.40/.8518 | .8444/.1651 | .8461/.1594 | **.8288/.1756** | .8875/.8700/.8486 |
| | | .533/2.40 | .636/1.44 | **.742/1.23** | .1431/.1752/.2153 |
| GP-FM | 15.39/.7970 | .8016/.2300 | .7994/.2366 | .7922/.2428 | .8365/.8348/.8336 |
| | | .000/9.30 | .054/9.47 | .227/6.80 | .2362/.2581/.2606 |
| AP-FM | **24.23/.9136** | **.8813/.1056** | **.8809/.1048** | .8244/.1778 | **.9314/.9304/.9265** |
| | | **.554/1.38** | **.668/1.11** | .721/1.59 | .0722/**.0757/.0826** |

prior inflates conditional velocity ambiguity $C_{\text{vel}}$. This regression averaging alters target appearance and leaves the field unresponsive to source color interventions (driving $C_{\text{rgb}}$ near zero), as exemplified in Figure 2 where some predictions collapse into generic gray bodies.

AP-FM resolves velocity ambiguity via trainable multi-Gaussian modes. Relative to GP-FM, prior adaptation substantially recovers surface textures and geometric crispness (marked gains in PSNR and SSIM), while aligning color shifts ($C_{\text{rgb}}$) and halving intervention discrepancy ($D_{\text{rgb}}$). Simultaneously, it maintains high path fidelity under multi-step composition ($L = 4$), preventing shape distortion and visual drift (stable SSIM and LPIPS). The shared latent pivot thus secures multi-step path invariance, while adaptive prior coupling ensures faithful velocity regression.

### 4.2 Pose-Guided Appearance Transfer on DeepFashion

We evaluate articulated pose transfer on DeepFashion-MultiModal-Parts2Whole (Liu et al., 2016; Jiang et al., 2022; Huang et al., 2024), with OpenPose skeleton maps (Cao et al., 2017) specifying body configuration as the condition.

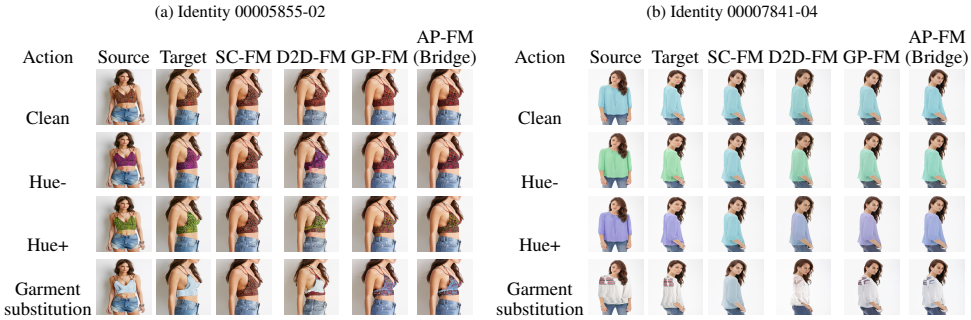


Figure 4: DeepFashion source equivariance examples under garment hue changes and substitution.

Table 2: DeepFashion target fidelity, source equivariance, and compositional consistency.

| Method | Fidelity | Source equivariance | | | | Compositional consistency |
|---|---|---|---|---|---|---|
| | Clean PSNR↑/SSIM↑ | Hue− | Hue+ | RGB→GBR | Garment substitution | GT: direct/$L$=2/$L$=4 Rows: SSIM↑/LPIPS↓ |
| | | SSIM ↑ /LPIPS ↓ /$C_{\text{app}}$ ↑ /$D_{\text{app}}$ ↓ | | | | |
| SC-FM | 21.60/.7841 | .7782/.1286<br>.102/**.993** | .7768/.1532<br>.101/**.996** | .7593/.1913<br>.211/.984 | **.7226**/.1995<br>.324/**.957** | .9369/.9349/.9350<br>.0276/.0288/.0287 |
| D2D-FM | 20.94/.7661 | .7569/.1402<br>.274/1.219 | .7560/.1634<br>.343/1.158 | .7268/.1994<br>.588/.919 | .6855/.2081<br>.525/1.098 | .8743/.8000/.7304<br>.0556/.1004/.1608 |
| GP-FM | 21.80/.7878 | .7802/.1264<br>.281/1.135 | .7790/.1489<br>**.356**/1.076 | .7633/.1796<br>**.608/.847** | .7145/**.1902**<br>**.529**/.972 | **.9671/.9695/.9708**<br>**.0151/.0141/.0137** |
| AP-FM | **21.92/.7901** | **.7818/.1257**<br>**.282**/1.142 | **.7815/.1474**<br>**.356**/1.081 | **.7639/.1780**<br>.607/.852 | .7140/.1916<br>.528/.984 | .9637/.9658/.9676<br>.0164/.0155/.0149 |

### 4.2.1 Evaluation Protocol

**Fidelity.** As on SRN, target fidelity is measured by PSNR/SSIM on pose-transfer pairs.

**Source equivariance.** We evaluate color shifts and garment substitution at fixed poses. We use spatially pooled features $h_{\text{app}}$ from the frozen DINOv2-based RAE (Oquab et al., 2024; Zheng et al., 2025) as an appearance descriptor. Using the same response formulation as above, we calculate response scores $C_{\text{app}}$ and $D_{\text{app}}$. For garment substitution, matched clothing edits are applied to both source and target; SSIM and LPIPS assess the predicted appearance against the modified target.

**Compositional consistency.** We evaluate direct and composed translations across path lengths $L \in \{1, 2, 4\}$, propagating predicted latents along matched view sequences. All outputs are evaluated against the decoded ground-truth target latents.

### 4.2.2 Comparison with Baselines

As observed on SRN, transporting the source state enhances response to appearance interventions: under garment substitutions, D2D-FM aligns latent transfer directions more faithfully than SC-FM (higher response cosine $C_{\text{app}}$, Table 2). However, D2D-FM fails to preserve anatomical structure under severe deformations (lower target SSIM), where duplicated arm and sleeve contours can be observed in Figure 4. D2D-FM also accumulates drift under compositional translation ($L = 4$), verified by a steep drop in SSIM and a more than two-fold increase in perceptual LPIPS. In contrast, GP-FM preserves both anatomical structure and fine appearance details across sequential poses while maintaining competitive SSIM and stable $L = 4$ path metrics, confirming the advantage of latent groupoid factorization.

Adaptive priors further refine fine-grained appearance details. AP-FM achieves the highest clean fidelity, consistently improves transformed-target SSIM and LPIPS across hue shifts and garment edits, and matches GP-FM's strong compositional stability ($L = 4$). The more moderate gain relative to SRN reflects the conditioning structure: dense OpenPose skeletons already constrain global body topology, leaving primarily texture and identity variation for the adaptive modes to disambiguate. Section 4.3 substantiates this interplay via coupling geometry and empirical field residuals.

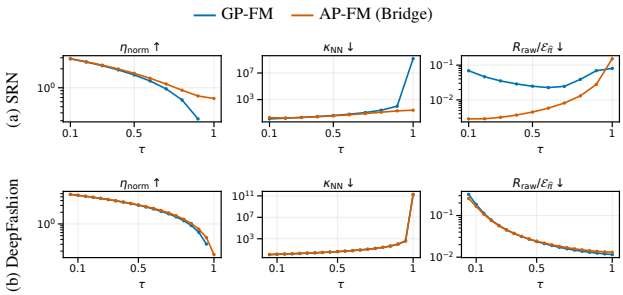


Figure 5: Coupling geometry and field fitting. Fixed-scale center separation $\eta_{\text{norm}}$, local stiffness proxy $\kappa_{\text{NN}}$, and energy-normalized empirical field residual $R_{\text{raw}}/\mathcal{E}_{\hat{\pi}}$ across flow time from data ($\tau = 0$) to prior ($\tau = 1$). Comparison uses identical camera conditions on SRN and local pose neighborhoods on DeepFashion.

### 4.3 Coupling Geometry and Field-Fitting Diagnostics

We analyze how adaptive priors alter the training coupling to resolve conditional velocity ambiguity, and the resulting regression fit after adaptation (Figure 5). Appendix D provides the derivations, diagnostic protocols, and detailed results.

**Adaptive priors separate competing interpolation paths.** Whereas GP-FM collapses all interpolation paths to a single origin as $\tau \to 1$ despite divergent velocity targets, AP-FM maintains finite separation between identity groups via mode centroids $\mu_{g(k)}$. This geometric separation directly suppresses the local velocity conflict quantified by the stiffness proxy $\kappa_{\text{NN}}$. This advantage is most pronounced in the near-prior regime: at representative time $\tau = 0.9$, AP-FM expands center separation $\eta_{\text{norm}}$ while reducing stiffness $\kappa_{\text{NN}}$. This geometric effect is more pronounced on SRN, where $\kappa_{\text{NN}}$ drops by several orders of magnitude under identity-level modes ($M = 300$), compared to the more moderate reduction on DeepFashion under shared within-category modes ($M = 8$).

**Adaptive priors reduce empirical field regression residual.** Figure 5 reports the energy-normalized field residual $R_{\text{raw}}/\mathcal{E}_{\hat{\pi}}$. Normalization by target velocity energy makes regression errors comparable across couplings with different velocity scales. On SRN, prior adaptation reduces this normalized residual by roughly half, directly underpinning the substantial fidelity and structural leaps over GP-FM in PSNR and SSIM (Table 1). On DeepFashion, the residual decreases more moderately, mirroring the moderate yet consistent fidelity gains (Table 2). Across both tasks, AP-FM preserves the translation groupoid structure of latent pivoting while optimizing the underlying velocity regression landscape.

## 5 Conclusion

We showed that controlled translation benefits from separating two roles: a shared latent pivot enforces the composition law and empirically observed source equivariance, while adaptive endpoint coupling improves the associated velocity regression problem. AP-FM demonstrates this separation on novel-view and pose-guided translation, with the largest gains when the standard Gaussian coupling induces strong ambiguity.

**Limitations and Future Work.** While this work establishes the translation groupoid formulation and its adaptive prior coupling, several promising frontiers emerge for future investigation. First, extending AP-FM to long-horizon settings, such as camera-controlled video generation and pose-guided animation, requires controlling the accumulation of ODE discretization errors and image codec drift over repeated translations. Furthermore, Proposition 2 characterizes exact groupoid equivariance in the prior domain through $H_{s,a} = H_{t,a}$. A natural next step is to explicitly design representation spaces that satisfy this conjugacy condition for specified appearance transformations $a_v$. Such representations could provide stronger guarantees for applications such as multi-view material editing and pose-consistent garment manipulation. Finally, while our current mode assignments structure identity and category groupings effectively, automatically inferring mode cardinality, assignments, and anisotropic covariances from the conditional data manifold would extend AP-FM to more heterogeneous and open-world settings, including category-level novel-view synthesis and cross-identity avatar rendering.

## REFERENCES


Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5968–5976, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Bhunia_Person_Image_Synthesis_via_Denoising_Diffusion_Model_CVPR_2023_paper.html.

Ronald Brown. From groups to groupoids: a brief survey. *Bulletin of the London Mathematical Society*, 19(2):113–134, March 1987. ISSN 0024-6093. doi: 10.1112/blms/19.2.113. URL http://dx.doi.org/10.1112/blms/19.2.113.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Ho Kei Cheng and Alexander Schwing. The curse of conditions: Analyzing and improving optimal transport for conditional flow-based generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15875–15884. IEEE, 2025.

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*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/11626_ECCV_2024_paper.php.

Hongxing Fan, Zehuan Huang, Lipeng Wang, Haohua Chen, Li Yin, and Lu Sheng. Parts2Whole: Generalizable multi-part portrait customization. *IEEE Transactions on Image Processing*, 34: 5241–5256, 2025. doi: 10.1109/TIP.2025.3597037. URL https://huanngzh.github.io/Parts2Whole/.

Aditya Grover, Christopher Chute, Rui Shu, Zhangjie Cao, and Stefano Ermon. AlignFlow: Cycle consistent learning from multiple domains via normalizing flows. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4028–4035, 2020. doi: 10.1609/aaai.v34i04.5820. URL https://ojs.aaai.org/index.php/AAAI/article/view/5820.

Zehuan Huang, Hongxing Fan, Lipeng Wang, and Lu Sheng. From parts to whole: A unified reference framework for controllable human image generation. *arXiv preprint arXiv:2404.15267*, 2024.

Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2Human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4): 1–11, 2022. doi: 10.1145/3528223.3530104.

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. URL https://openaccess.thecvf.com/content/CVPR2024/html/Kim_StableVITON_Learning_Semantic_Correspondence_with_Latent_Diffusion_Model_for_Virtual_CVPR_2024_paper.html.

Junwan Kim, Jiho Park, Seonghu Jeon, and Seungryong Kim. Better source, better flow: Learning condition-dependent source distribution for flow matching, 2026. URL https://arxiv.org/abs/2602.05951.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Lenc_Understanding_Image_Representations_2015_CVPR_paper.html.

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. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos Theodorou, Weili Nie, and Anima Anandkumar. $I^2$SB: Image-to-image schrödinger bridge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22042–22062. PMLR, 2023a. URL https://proceedings.mlr.press/v202/liu23ai.html.

Jiaqi Liu, Jichao Zhang, Paolo Rota, and Nicu Sebe. Multi-focal conditioned latent diffusion for person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16019–16028, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_Multi-focal_Conditioned_Latent_Diffusion_for_Person_Image_Synthesis_CVPR_2025_paper.html.

Qihao Liu, Xi Yin, Alan Yuille, Andrew Brown, and Mannat Singh. Flowing from words to pixels: A noise-free framework for cross-modality evolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2755–2765, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_Flowing_from_Words_to_Pixels_A_Noise-Free_Framework_for_Cross-Modality_CVPR_2025_paper.html.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9298–9309, 2023b. URL https://openaccess.thecvf.com/content/ICCV2023/html/Liu_Zero-1-to-3_Zero-shot_One_Image_to_3D_Object_ICCV_2023_paper.html.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Generating multiview-consistent images from a single-view image. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/753d9584b57ba01a10482f1ea7734a89-Abstract-Conference.html.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Yanzuo Lu, Manlin Zhang, Andy J Ma, Xiaohua Xie, and Jianhuang Lai. Coarse-to-fine latent diffusion for pose-guided person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6420–6429, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lu_Coarse-to-Fine_Latent_Diffusion_for_Pose-Guided_Person_Image_Synthesis_CVPR_2024_paper.html.

Gaoxiang Luo, Frank Cole, Sihang Zhang, Yuxiang Wan, Yulong Lu, and Ju Sun. Flow matching for multimodal distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23260–23271, June 2026.

Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, volume 30, pp. 406–416, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/34ed066df378efacc9b924ec161e7639-Abstract.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28100–28127. PMLR, 2023. URL https://proceedings.mlr.press/v202/pooladian23a.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, 2019. URL https://www.vincentsitzmann.com/srns/.

George Stoica, Vivek Ramanujan, Xiang Fan, Ali Farhadi, Ranjay Krishna, and Judy Hoffman. Contrastive flow matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1185–1194, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Stoica_Contrastive_Flow_Matching_ICCV_2025_paper.html.

Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=5HLoTvVGDe.

Xavier Sumba, Carles Balsells-Rodas, and Yingzhen Li. Structured coupling for flow matching, 2026. URL https://arxiv.org/abs/2605.07676.

Gerald Teschl. *Ordinary Differential Equations and Dynamical Systems*. American Mathematical Society, August 2012. ISBN 9781470425869. doi: 10.1090/gsm/140. URL http://dx.doi.org/10.1090/gsm/140.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. ISSN 1941-0042. doi: 10.1109/tip.2003.819861. URL http://dx.doi.org/10.1109/TIP.2003.819861.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4578–4587, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yu_pixelNeRF_Neural_Radiance_Fields_From_One_or_Few_Images_CVPR_2021_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 (CVPR)*, June 2018.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders, 2025. URL https://arxiv.org/abs/2510.11690.

Chuanxia Zheng and Andrea Vedaldi. Free3D: Consistent novel view synthesis without 3D representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9720–9731, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zheng_Free3D_Consistent_Novel_View_Synthesis_without_3D_Representation_CVPR_2024_paper.html.

Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/20e45668fefa793bd9f2edf19be12c4b-Abstract-Conference.html.

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*, 2017. URL https://junyanz.github.io/CycleGAN/.

## AI USE STATEMENT

In this work, we used generative AI tools for code implementation assistance and preliminary drafting of manuscript sections and proof sketches. We have not used generative AI tools for experimental execution, data collection, or generating evaluation outcomes. Additionally, we used generative AI tools for exploratory literature retrieval, language polishing, and scripting the vector layout code for Figure 1. We have thoroughly reviewed and verified all AI-assisted work: the lead author specified the software architecture and verified all script implementations via unit tests; the lead author and mentor independently derived, cross-checked, and finalized all mathematical proofs and theoretical claims; and all reported empirical metrics and visual outputs are derived directly from the unmodified outputs of the trained models. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

## ETHICS STATEMENT

The experiments in this work involve pose-guided appearance transfer on human images from the DeepFashion and Parts2Whole benchmarks. These datasets were collected from public domains and are utilized strictly in accordance with their respective non-commercial academic research licenses. Controlled image translation technologies carry potential dual-use risks, including unauthorized manipulation of personal identity or clothing without consent, and the generation of misleading visual content. This research evaluates the mathematical properties of continuous flow transport in controlled academic settings and is not intended for unconstrained deployment on identifiable individuals. To mitigate misuse, any practical application of such systems should enforce explicit informed consent, adhere to privacy and copyright regulations, and incorporate provenance-tracking mechanisms such as digital watermarking and clear modification disclosure.

## REPRODUCIBILITY STATEMENT

To ensure complete reproducibility of our theoretical and empirical contributions:

1. **Theory and Assumptions**: Appendices A details the underlying assumptions, regularities, and complete proofs for the groupoid composition laws, bi-Lipschitz sensitivity, and groupoid equivariance.
2. **Architectures and Training**: Appendices B and E specify complete network backbones, condition projections, optimization schedules, loss formulations, and hyperparameter configurations across both tasks.
3. **Evaluation and Diagnostics**: Appendix C formalizes the evaluation protocals and metrics, while Appendix D.4 outlines the numerical sampling protocols for coupling geometry and field residuals.