FLOWCYCLE: RETHINKING CYCLE CONSISTENCY VIA EXACT BIJECTIVE LATENT FLOWS

Anonymous authors
Paper under double-blind review

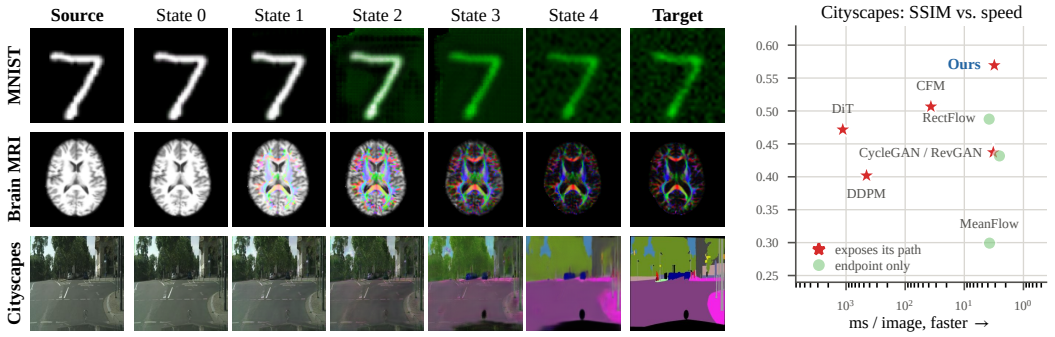


Figure 1: **Exact bijective transport, intrinsic latent trajectories, and single-pass translation.** Left: intermediate states decoded across successive coupling layers on Multimodal MNIST, structural T1-MRI to diffusion imaging, and Cityscapes. Right: translation fidelity versus amortized inference latency evaluated against continuous-time and adversarial baselines.

ABSTRACT

Bidirectional cross-modal translation often entails a compromise between inference efficiency and formal invertibility. Continuous diffusion-based approaches typically require computationally intensive ordinary differential equation (ODE) solvers and are susceptible to integration error accumulation, whereas CycleGAN-type architectures depend on approximate cycle-consistency objectives that provide only indirect and potentially decoupled constraints on reversibility. We introduce FlowCycle, an end-to-end framework that enables single-pass translation while ensuring exact reversibility in latent space. Specifically, we cast cross-domain transport as a discrete normalizing flow in a learned representation space, implemented via spatial affine coupling transformations. This design yields closed-form analytic inversion in one pass, obviates numerical solvers, and exposes an interpretable latent transition trajectory without requiring additional flow evaluations. Empirical results on synthetic digits (Multimodal MNIST), urban scenes (Cityscapes), and clinical multi-sequence brain MRI indicate state-of-the-art translation fidelity. Beyond quantitative performance, FlowCycle offers substantive practical advantages: in medical imaging, its traceable trajectories support transparent, clinician-auditable morphing for reliable synthesis of missing diagnostic sequences; in complex visual environments, its deterministic single-step inference facilitates real-time data simulation and robust cross-domain perception without computational bottlenecks. [1]

1 INTRODUCTION

Cross-modal image translation aims to predict observations in a target modality from an available source domain while preserving invariant semantic identity and structural content (Isola et al., 2017; Zhu et al., 2017). Distinct modalities and representations capture complementary properties of the same underlying subject or scene: in biomedical imaging, structural MRI characterizes macroscopic

[1] Our code is released by `https://anonymous.4open.science/r/FlowCycle-5634/`

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.

anatomical morphology, whereas diffusion MRI probes tissue microstructure and white-matter fiber organization (Dale et al., 1999; Basser et al., 1994; Mori et al., 1999); in urban visual perception, photometric images encode visual appearance while dense semantic maps encode object- and region-level scene structure (Cordts et al., 2016; Isola et al., 2017). Synthesizing missing observations can mitigate unavailable or incomplete acquisitions and enrich downstream analysis (Lee et al., 2019; Liu et al., 2023a). Reliable deployment across physical domains demands both high visual fidelity and consistency, particularly because image-translation models can introduce hallucinated features (Cohen et al., 2018).

Over the past decade, cross-modal generative models have evolved across three successive paradigms, each addressing limitations of its predecessor while introducing new trade-offs. *(1) Heuristic adversarial mapping.* The initial wave relied on conditional and cycle-consistent adversarial networks (Isola et al., 2017; Zhu et al., 2017). While enabling single-pass mapping, their cycle consistency depends on heuristic penalties between independent generators. Lacking an exact mathematical inverse, these models yield inexact round trips and function as opaque systems prone to concealing source information in high-frequency patterns without learning true correspondences. *(2) Direct architectural inversion.* To enforce rigorous consistency, a second paradigm explored architectural reversibility via invertible networks and discrete normalizing flows (van der Ouderaa & Worrall, 2019; Grover et al., 2020; Rombach et al., 2022). By design, the reverse translation is identical to the algebraic inverse of the forward pass. Enforcing invertibility directly in high-dimensional image space requires volume-preserving pixel mappings. Without a compact feature bottleneck to separate invariant semantics from modality-specific appearances, direct invertible architectures often suffer from semantic drift and compromised visual sharpness. *(3) Continuous-time transport.* Most recently, continuous-time formulations, including diffusion models and continuous flow matching (Ho et al., 2020; Lipman et al., 2023; Geng et al., 2025; 2026), reparameterized generation as probability transport along continuous vector fields. While exposing smooth transition trajectories, these models rely on numerical ODE solvers requiring iterative evaluations that accumulate discretization errors, induce inference latency, and sacrifice exact algebraic reversibility. Existing paradigms present an unresolved tripartite tension: heuristic networks offer speed without rigor, pixel-level invertible models offer reversibility without compact representations, and continuous flows offer trajectory transparency at the cost of high latency and numerical error.

In this work, we present **FlowCycle**, an end-to-end framework that unifies single-pass efficiency, exact mathematical reversibility, and transparent transformation trajectories. FlowCycle parameterizes cross-modal transport as a discrete-time, exactly invertible normalizing flow over a shared latent manifold, decoupling feature encoding from domain translation. Because the transport is discrete and algebraically exact, bidirectional translation evaluates in a single forward pass without numerical ODE solvers, while strictly preventing latent information bypass. Intermediate states across successive flow layers expose the true step-wise transformation path at zero additional inference overhead, enabling direct visual inspection through decoding. We validate FlowCycle across synthetic digits (Multimodal MNIST) (LeCun et al., 1998), urban scenes (Cityscapes) (Cordts et al., 2016), and clinical neuroimaging bridging anatomical MRI to diffusion white matter tracts (Alzheimer's Disease Neuroimaging Initiative: ADNI) (Petersen et al., 2010). FlowCycle matches or exceeds baseline structural fidelity while confirming exact latent reversibility and strong downstream label transfer. Beyond cross-modal translation, our formulation demonstrates that traceable trajectories and exact reversibility do not require iterative continuous-time integration, establishing an efficient and verifiable foundation for transparent cross-domain representation learning.

Our main contributions are summarized as follows:

- **Exact Bijective Latent Transport:** We propose a single-pass, closed-form discrete normalizing flow that exactly aligns cross-domain representation manifolds. Without numerical ODE solvers or latent information bypass, it guarantees mathematically exact topological reversibility of the latent transport operator for cross-modal mapping.
- **Intrinsic Transition Dynamics:** We establish an auditable transport mechanism where sequential spatial coupling states natively expose a discrete transition trajectory at zero additional computational overhead. This provides a rigorous analytical foundation that decouples algebraic operator bijectivity from empirical autoencoder projection errors.
- **Transparent Cross-Domain Representation Learning:** We demonstrate the framework's utility across computer vision and clinical domains: synthetic imaging (Multimodal

MNIST), photometric scene generation (Cityscapes), and macro-to-micro neuroimaging mapping (ADNI). Extensive evaluations show FlowCycle is an efficient, clinician-verifiable representation learning paradigm.

2 BACKGROUND

Let modalities A and B have distributions $\mathbb{P}_A, \mathbb{P}_B$ on image space $\mathcal{X}$, supported on disjoint manifolds $\mathcal{X}_A, \mathcal{X}_B$. We map these domains into a compact representation space $\mathcal{Z}$.

A discrete normalizing flow $f = \phi_L \circ \cdots \circ \phi_1$ on $\mathcal{Z}$ provides exact mathematical reversibility. Using spatial affine coupling (Dinh et al., 2017; Kingma & Dhariwal, 2018), an input $\mathbf{z}$ is partitioned into $(\mathbf{z}_{[1]}, \mathbf{z}_{[2]})$ to apply the forward transformation:

$$\mathbf{u}_{[1]} = \mathbf{z}_{[1]}, \qquad \mathbf{u}_{[2]} = e^{\,s(\mathbf{z}_{[1]})} \odot \mathbf{z}_{[2]} + t(\mathbf{z}_{[1]}), \tag{1}$$

where neural networks (s, t) process only the unchanged partition. This ensures the inverse remains in closed-form analytically without numerical solvers. The sequential composition intrinsically generates intermediate states $\mathbf{z}^{(\ell)} = \phi_\ell(\mathbf{z}^{(\ell-1)})$, establishing a discrete transition trajectory at zero additional computational cost.

Standard flows maximize likelihood via the exact change-of-variables:

$$\log q(\mathbf{z}) = \log p_Z\big(f^{-1}(\mathbf{z})\big) + \log\big|\det \nabla f^{-1}(\mathbf{z})\big|. \tag{2}$$

Eq. 2 limits cross-modal translation by working in high-dimensional pixel space (Dinh et al., 2017; Kingma & Dhariwal, 2018) and mapping heterogeneous domains to an uninformative isotropic base p_Z, weakening semantics. Bidirectional translation needs exact invertibility, single-pass inference, and an intrinsic trajectory, but existing methods satisfy these only partly. Decoupled frameworks (Zhu et al., 2017; Zou et al., 2026) (e.g., CycleGAN, CycleDiff, single-step normalizing flows) enable single-pass inference but usually lack exact inverses and explicit paths. Continuous-time methods (Ho et al., 2020; Lipman et al., 2023; Geng et al., 2025) (e.g., diffusion, flow matching) provide trajectories but rely on K-step ODE solvers of order p, incurring $\mathcal{O}(K)$ cost and $\mathcal{O}(K^{-p})$ error per direction. We remove this trade-off by dropping Eq. 2 and applying the discrete flow in Eq. 1 only in a learned representation space, achieving all three properties by construction.

3 METHODOLOGY: FLOWCYCLE

3.1 LATENT TRANSPORT FORMULATION

For each modality $m \in \{A, B\}$, we use an encoder E_m and decoder D_m. The latent distribution is $\pi_m = (E_m)_{\#}\mathbb{P}_m$, supported on a possibly nonlinear manifold $\mathcal{Z}_m \subset \mathcal{Z}$, with no assumed parametric form (Villani et al., 2009; Gabriel & Marco, 2019). Cross-modal translation uses a bijection $f : \mathcal{Z} \to \mathcal{Z}$ satisfying $f_{\#}\pi_A = \pi_B$. For paired data $(\mathbf{x}_A, \mathbf{x}_B) \sim \mathcal{D}$,

$$\begin{aligned} \boldsymbol{\alpha} &= E_A(\mathbf{x}_A), \quad \hat{\boldsymbol{\beta}} = f(\boldsymbol{\alpha}), \qquad T_{A\to B} = D_B \circ f \circ E_A, \\ \boldsymbol{\beta} &= E_B(\mathbf{x}_B), \quad \hat{\boldsymbol{\alpha}} = f^{-1}(\boldsymbol{\beta}), \quad T_{B\to A} = D_A \circ f^{-1} \circ E_B. \end{aligned} \tag{3}$$

There are no skip connections around f, so $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ are the only representations passed between modalities. Let $\Pi_m = E_m \circ D_m$ denote re-encoding through modality m; under perfect reconstruction it is the identity on $\mathcal{Z}_m$ (Alain & Bengio, 2014). The round trip $T_{B\to A} \circ T_{A\to B} = D_A \circ f^{-1} \circ \Pi_B \circ f \circ E_A$ therefore reduces to $D_A \circ E_A$ whenever, on $\mathcal{Z}_A$,

$$\text{(i)}\ \ f^{-1} \circ f = \mathrm{id} \qquad \text{and} \qquad \text{(ii)}\ \ \Pi_B \circ f = f. \tag{4}$$

Condition (i) constitutes an inherent architectural property (Section 3.2), whereas condition (ii) requires f to map strictly into the target manifold and is enforced entirely by the training objective (Section 3.3). Cycle-consistent GANs (Zhu et al., 2017) guarantee neither requirement, instead penalizing only their aggregate discrepancy in pixel space. In FlowCycle, the latent-space cycle is mathematically exact, and the pixel-space round trip departs from $D_A \circ E_A$ exclusively via the residual associated with condition (ii), which is penalized explicitly.[2]

[2]Condition (ii) supersedes the likelihood term in Eq. 2: it requires the transported representation to reside on the empirical target manifold rather than to match a prescribed density, imposing no constraint on the dimensionality of $\mathcal{Z}$.

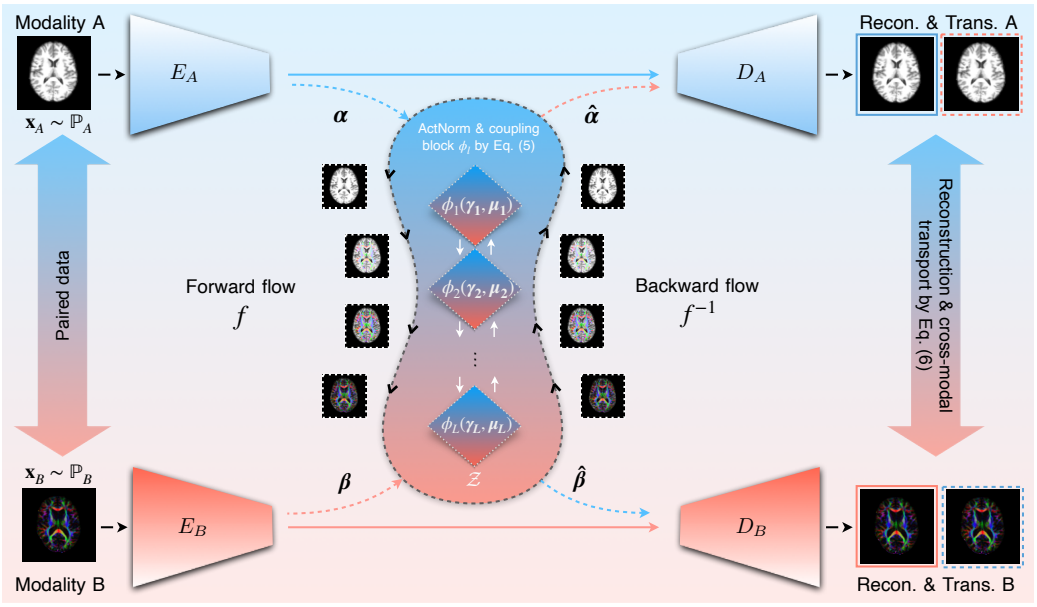


Figure 2: **Overview of FlowCycle.** Encoders E_A and E_B map paired images to latents $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. A discrete normalizing flow f (spatial affine coupling + ActNorm) provides an exact, solver-free bijection in $\mathcal{Z}$ between modalities A and B. Translated codes $\hat{\boldsymbol{\beta}} = f(\boldsymbol{\alpha})$ and $\hat{\boldsymbol{\alpha}} = f^{-1}(\boldsymbol{\beta})$ are decoded by D_B and D_A into the target domains. Coupling-layer intermediate states define an intrinsic transition trajectory (inset). Training jointly optimizes Eq. 9 under a three-stage protocol, with trajectory regularization via Eq. 7 and Eq. 8.

3.2 Exact Spatial Bijection and Intrinsic Trajectory

Each transformation block ϕ_ℓ comprises an ActNorm layer (Kingma & Dhariwal, 2018) followed by an affine coupling layer (Eq. 1), both applied directly to the $C_z \times H_z \times W_z$ feature map; this direct operation on the spatial tensor motivates the term *spatial*. The ActNorm layer is parameterized by per-channel scale and shift parameters, $\boldsymbol{\gamma}_\ell, \boldsymbol{\mu}_\ell \in \mathbb{R}^{C_z}$. The coupling transformation performs a channel-wise split, and a fully convolutional conditioning network yields $(\tilde{\mathbf{s}}_\ell, \mathbf{t}_\ell) = g_\ell(\mathbf{p}_{[1]})$, with the scaling factor bounded via $\mathbf{s}_\ell = \tanh(\tilde{\mathbf{s}}_\ell)$ for numerical stability (Dinh et al., 2017). The forward and inverse mappings for each block are given by:

$$\phi_\ell : \begin{cases} \mathbf{p} = e^{\boldsymbol{\gamma}_\ell} \odot \mathbf{v} + \boldsymbol{\mu}_\ell \\ \mathbf{w}_{[1]} = \mathbf{p}_{[1]} \\ \mathbf{w}_{[2]} = e^{\mathbf{s}_\ell} \odot \mathbf{p}_{[2]} + \mathbf{t}_\ell \end{cases} \iff \phi_\ell^{-1} : \begin{cases} \mathbf{p}_{[1]} = \mathbf{w}_{[1]} \\ \mathbf{p}_{[2]} = e^{-\mathbf{s}_\ell} \odot (\mathbf{w}_{[2]} - \mathbf{t}_\ell) \\ \mathbf{v} = e^{-\boldsymbol{\gamma}_\ell} \odot (\mathbf{p} - \boldsymbol{\mu}_\ell) \end{cases} \tag{5}$$

Exact inversion. The inverse map undoes the forward computation in reverse order. In both directions, g_ℓ is evaluated on the same input $\mathbf{p}_{[1]} = \mathbf{w}_{[1]}$ with strictly positive scaling. Thus $f^{-1} \circ f = f \circ f^{-1} = \mathrm{id}_{\mathcal{Z}}$ up to floating-point round-off, so condition (i) holds by construction without matrix inversion, root finding, or numerical integration.

Identity initialization. Setting $\boldsymbol{\gamma}_\ell, \boldsymbol{\mu}_\ell$ and the terminal weights of g_ℓ to zero guarantees $f \equiv \mathrm{id}_{\mathcal{Z}}$ at initialization (Zhang et al., 2019). Training thus stably begins with the identity transport on π_A and progressively learns the transformation toward π_B.[3]

Intrinsic trajectory. The sequence of block outputs $\boldsymbol{\alpha}^{(\ell)} = \phi_\ell \circ \cdots \circ \phi_1(\boldsymbol{\alpha})$ defines a trajectory from $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\alpha}$ to $\boldsymbol{\alpha}^{(L)} = \hat{\boldsymbol{\beta}}$, whose induced distributions interpolate between π_A and π_B without any post-hoc interpolation procedure. Intermediate states are synthesized using the shared decoder D_{mix} (Section 4.3), whereas the terminal state is decoded using D_B.

[3]Representations are extracted prior to the encoder's final nonlinearity, because a nonnegative representation cannot support a signed flow output.

Algorithm 1 FlowCycle Optimization

Input: Paired data $\mathcal{D}$; parameters $\theta_E, \theta_D, \theta_f$ with $f \equiv \mathrm{id}_{\mathcal{Z}}$; critics $C_A, C_B, C_{\mathrm{mix}}$
1: Update θ_E, θ_D on $\lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}}$ ▷ Stage 1, $f \equiv \mathrm{id}$
2: Update θ_f on $\lambda_{\mathrm{cross}}\mathcal{L}_{\mathrm{cross}} + \lambda_{\mathrm{traj}}\mathcal{L}_{\mathrm{traj}}$ ▷ Stage 2, θ_E, θ_D frozen
3: **for** $(\mathbf{x}_A, \mathbf{x}_B) \sim \mathcal{D}$ **do** ▷ Stage 3, θ_E frozen
4: $\quad\boldsymbol{\alpha} \leftarrow E_A(\mathbf{x}_A), \boldsymbol{\beta} \leftarrow E_B(\mathbf{x}_B), \hat{\boldsymbol{\beta}} \leftarrow f(\boldsymbol{\alpha}), \hat{\boldsymbol{\alpha}} \leftarrow f^{-1}(\boldsymbol{\beta}), \ell \sim \mathcal{U}\{1, \ldots, L\}$
5: $\quad\{\theta_D, \theta_f\} \leftarrow \{\theta_D, \theta_f\} - \eta \nabla \mathcal{L}$ ▷ Eq. 9
6: $\quad$Update $C_A, C_B, C_{\mathrm{mix}}$ by adversarial ascent (Daskalakis et al., 2017)
7: **end for**
Return: f, f^{-1}; $T_{A\to B}, T_{B\to A}$ (Eq. 3); latent states $\{\boldsymbol{\alpha}^{(\ell)}\}_{\ell=0}^{L}$

3.3 Objective

For unified notation across transport directions, set $\mathbf{z}_A = \boldsymbol{\alpha}$, $\mathbf{z}_B = \boldsymbol{\beta}$, $\hat{\mathbf{z}}_A = \hat{\boldsymbol{\alpha}}$, and $\hat{\mathbf{z}}_B = \hat{\boldsymbol{\beta}}$. Let $\bar{m}$ denote the complementary modality, define $f_{A\to B} = f$ and $f_{B\to A} = f^{-1}$, and assume $\sum_m$ sums over $m \in \{A, B\}$ unless noted. Let $\ell_{\mathrm{adv}}(\cdot\,; C)$ be the least-squares adversarial loss (Mao et al., 2017) for critic C.

Endpoint fidelity. Given a target sample $\mathbf{x}_m$, we define the endpoint fidelity functional

$$\mathcal{E}_m(\mathbf{z}) = \|D_m(\mathbf{z}) - \mathbf{x}_m\|_1 + \lambda_{\mathrm{adv}}\, \ell_{\mathrm{adv}}\big(D_m(\mathbf{z}); C_m\big).$$

Applying $\mathcal{E}_m$ to within-modality reconstructions and cross-modality transports yields

$$\mathcal{L}_{\mathrm{rec}} = \sum_m \mathcal{E}_m(\mathbf{z}_m), \qquad \mathcal{L}_{\mathrm{cross}} = \sum_m \mathcal{E}_m(\hat{\mathbf{z}}_m). \tag{6}$$

Here, $\mathcal{L}_{\mathrm{rec}}$ is used to train the autoencoders, whereas $\mathcal{L}_{\mathrm{cross}}$ supervises the learned bijection and constitutes the only loss term that requires paired observations.

Manifold consistency. Condition (ii) imposes that $\hat{\mathbf{z}}_m$ be a fixed point of the projection operator Π_m (Alain & Bengio, 2014). Defining $G_m = D_{\bar{m}} \circ f_{m\to\bar{m}}$ as the mapping back to the source-image domain, we enforce

$$\mathcal{L}_{\mathrm{lat}} = \sum_m \frac{\|\Pi_m(\hat{\mathbf{z}}_m) - \hat{\mathbf{z}}_m\|_1}{\|\hat{\mathbf{z}}_m\|_1 + \epsilon}, \qquad \mathcal{L}_{\mathrm{cyc}} = \sum_m \big\|G_m\big(\Pi_m(\hat{\mathbf{z}}_m)\big) - G_m(\hat{\mathbf{z}}_m)\big\|_1. \tag{7}$$

The normalization in $\mathcal{L}_{\mathrm{lat}}$ mitigates degenerate solutions in which the loss is artificially reduced by shrinking latent magnitudes. Since $f_{m\to\bar{m}}(\hat{\mathbf{z}}_m) = \mathbf{z}_{\bar{m}}$, it follows that $G_m(\hat{\mathbf{z}}_m) = D_{\bar{m}}(\mathbf{z}_{\bar{m}})$, and thus $\mathcal{L}_{\mathrm{cyc}}$ recovers the standard cycle-consistency penalty; in the present formulation, it specifically targets violations of condition (ii).

Trajectory loss. Intermediate frames $\mathbf{y}^{(k)} = D_{\mathrm{mix}}(\boldsymbol{\alpha}^{(k)})$ are evaluated using a critic C_{mix} trained on the mixture distribution $\frac{1}{2}(\mathbb{P}_A + \mathbb{P}_B)$. We sample $\ell \sim \mathcal{U}\{1, \ldots, L\}$ and define

$$\mathcal{L}_{\mathrm{traj}} = \textstyle\sum_{k\in\{\ell-1,\ell\}} \ell_{\mathrm{adv}}\big(\mathbf{y}^{(k)}; C_{\mathrm{mix}}\big) + \lambda_{\mathrm{s}}\|\mathbf{y}^{(\ell)} - \mathbf{y}^{(\ell-1)}\|_1, \tag{8}$$

where the second term promotes temporal (or pathwise) smoothness between successive intermediate states. The mixed decoder D_{mix} is included in the decoder parameter set θ_D. We refer to this trajectory-smoothness constraint as trajectory regularization (TR), denoting the regularized variant as FlowCycle-TR ($\lambda_{\mathrm{reg}} > 0$) and its unregularized baseline as FlowCycle w/o TR ($\lambda_{\mathrm{reg}} = 0$).

Overall training objective. Let θ_E, θ_D, and θ_f denote the parameters of the encoder, decoder, and flow module, respectively. The complete objective is

$$\mathcal{L} = \mathbb{E}_{\mathcal{D}}\big[\lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{cross}}\mathcal{L}_{\mathrm{cross}} + \lambda_{\mathrm{lat}}\mathcal{L}_{\mathrm{lat}} + \lambda_{\mathrm{cyc}}\mathcal{L}_{\mathrm{cyc}} + \lambda_{\mathrm{traj}}\mathcal{L}_{\mathrm{traj}}\big] \tag{9}$$

where the expectation is taken over samples from the dataset $\mathcal{D}$. Alg. 1 provides a summary of the alternating joint optimization procedure (Bezdek & Hathaway, 2002).

3.4 Complexity Analysis

Computational Complexity & Trajectory Budget. Inference per direction costs $\mathcal{O}(c_E + L\,c_\phi + c_D)$, with c_E, c_D the autoencoder FLOPs and c_ϕ the forward cost of conditioners (s_ℓ, t_ℓ) on $\mathcal{Z}$.

Because ϕ_ℓ and ϕ_ℓ^{-1} run the same network on the invariant partition $\mathbf{z}_{[1]}$, runtime is direction-invariant: $\mathcal{T}_{A,B} \equiv \mathcal{T}_{B,A}$. Unlike continuous-time methods requiring K solver steps and truncation error $\mathcal{O}(K^{-p})$ (Hairer et al., 1993), FlowCycle performs deterministic transport with $\mathrm{NFE} = 1$. Logging the latent trajectory $\{\mathbf{z}^{(\ell)}\}_{\ell=1}^{L}$ adds no flow cost; generating m intermediate states costs only $m\, c_D$ for decoding.

Memory Footprint. Tab. 1 summarizes structural characteristics across paradigms. Because spatial affine coupling layers are bijective, backpropagation gradients can be computed by reconstructing intermediate activations during the backward pass instead of storing them (Gomez et al., 2017). Thus, training-time memory complexity is $\mathcal{O}(1)$ in network depth L, avoiding the usual $\mathcal{O}(L)$ activation-caching cost. Throughput is reported in Section 4.5.

Table 1: Inference complexity per direction. NFE: sequential network evaluations; δ: learned cycle residual; K, p: ODE steps and order. ‡Core latent operator only.

Method	NFE	$\|\mathbf{z} - f^{-1}(f(\mathbf{z}))\|$	Path
CycleGAN	1	$\mathcal{O}(\delta)$	✗
RevGAN	1	$0^{\ddagger}$	✗
MeanFlow	1	$\mathcal{O}(\delta)$	✗
Diff. / FM	K	$\mathcal{O}(K^{-p})$	✓
FlowCycle	**1**	**0**	✓

Table 2: **Quantitative evaluation of bidirectional translation fidelity.** Metrics report mean ± standard deviation on the held-out test sets (5,000 MNIST pairs, 500 Cityscapes pairs, and 430 ADNI slices). Ours with MS ($\lambda_s > 0$) denotes the full framework regularized with the trajectory smoothness penalty, while ours w/o MS ($\lambda_s = 0$) ablates this sequential constraint. Higher values indicate superior generation quality (PSNR is reported in dB). Best and second-best results are **bolded** and underlined, respectively.

	Synthetic MNIST MRI/PET				Cityscapes				ADNI			
	Syn-MRI → Syn-PET		Syn-PET → Syn-MRI		Photo → Labels		Labels → Photo		T1 → FA		FA → T1	
Method	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
CycleGAN	0.8280±0.020	28.47±0.60	0.9463±0.018	25.67±1.79	0.5566±0.067	14.08±1.48	0.3068±0.074	13.26±2.12	0.6798±0.027	17.87±0.82	0.8035±0.034	19.39±1.37
RevGAN	0.8266±0.021	28.37±0.56	0.9404±0.019	25.29±1.87	0.5693±0.067	14.33±1.51	0.3049±0.068	13.02±1.92	0.6614±0.022	17.55±0.74	0.7983±0.028	19.67±1.27
DDPM	0.1953±0.024	25.46±0.57	0.4416±0.081	22.82±1.35	0.5050±0.109	14.28±2.31	0.2987±0.086	11.25±2.41	0.2095±0.039	17.18±1.41	0.3919±0.037	19.99±1.24
MeanFlow	0.8158±0.018	30.33±0.64	0.4055±0.054	18.66±1.44	0.2525±0.048	10.54±1.00	0.3459±0.068	13.53±1.65	0.6912±0.024	18.61±0.64	0.4644±0.086	18.56±1.25
CFM	0.8412±0.023	31.12±0.44	0.9028±0.031	24.66±1.82	0.5948±0.089	16.66±2.02	0.4184±0.082	16.55±2.51	0.7288±0.026	18.87±0.78	0.7897±0.039	19.92±1.27
RectFlow	0.8226±0.025	**31.25**±0.42	0.8960±0.030	25.06±1.82	0.5652±0.085	17.09±1.87	0.4100±0.080	16.50±2.47	0.7263±0.029	19.43±0.79	0.7716±0.062	20.30±1.29
DiT	0.8314±0.022	28.76±0.57	0.9196±0.023	23.71±1.57	0.5563±0.075	14.51±1.47	0.3870±0.080	14.94±2.22	0.7133±0.027	18.33±0.79	0.7991±0.036	19.34±1.63
Ours w/o TR	0.8657±0.018	30.55±0.63	**0.9525**±0.016	**26.54**±1.79	0.6965±0.069	18.59±2.22	**0.4422**±0.099	**16.64**±2.91	**0.7627**±0.030	**19.88**±0.94	**0.8421**±0.028	**21.04**±1.36
Ours	**0.8677**±0.017	30.78±0.53	0.9484±0.017	26.27±1.72	**0.7074**±0.068	**18.65**±2.27	0.4261±0.084	16.29±2.69	0.7520±0.033	19.53±0.98	0.8217±0.031	20.56±1.29

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate FlowCycle across three diverse visual domains. Synthetic Multimodal MNIST pairs (LeCun et al., 1998) grayscale structural digits with functional PET (Positron emission tomography)-like variants featuring nonlinear intensity shifts and additive noise (12,000 training and 5,000 test pairs at 64×64). Cityscapes (Cordts et al., 2016) translates between dense urban semantic label maps and complex photographic street scenes (2,677 training and 500 held-out test pairs at 256×256). For clinical neuroimaging, we utilize the ADNI dataset (Weiner et al., 2010) to bridge macro-structural morphology and microstructural tissue architecture, pairing structural T1-weighted MRI with diffusion-derived fractional anisotropy (FA) maps (1,700 training and 430 test samples; ten 112×112 axial slices per subject).

Baselines and Evaluation. We benchmark against cycle-consistent adversarial models (CycleGAN (Zhu et al., 2017), RevGAN (van der Ouderaa & Worrall, 2019)), continuous score-based diffusion models (DDPM (Ho et al., 2020)), continuous flow matching architectures (CFM (Lipman et al., 2023), RectFlow (Liu et al., 2023b)), and single-step frameworks (MeanFlow (Geng et al., 2025), DiT (Peebles & Xie, 2023)). Performance is quantified via structural similarity (SSIM) and peak signal-to-noise ratio (PSNR) on identical test splits.

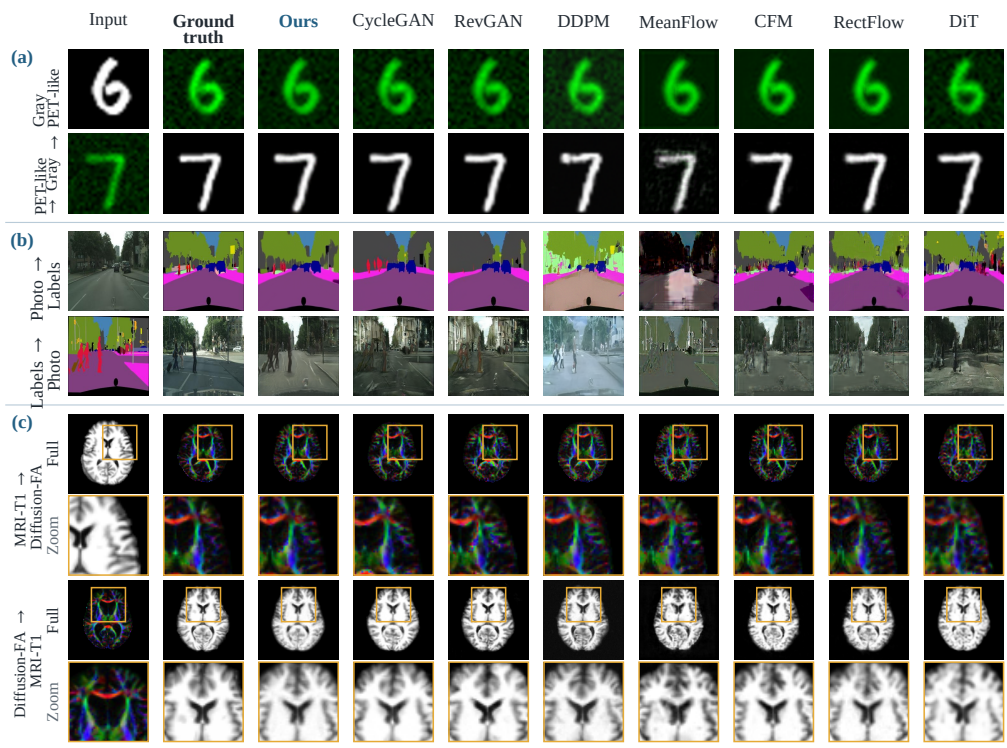


Figure 3: **Qualitative comparison of bidirectional image translation.** (a) Synthetic MNIST MRI/PET cross-modal synthesis. (b) Cityscapes photometric scene generation and semantic mapping. (c) Brain scan translation between T1-weighted MRI and diffusion FA maps, with matched zoomed insets highlighting white matter boundary preservation.

4.2 Bidirectional Cross-Modal Translation

Tab. 2 reports quantitative metrics of translation fidelity. Over all translation directions and datasets, method achieves state-of-the-art SSIM. In terms of PSNR, the proposed framework attains the best performance in five of the six evaluated mappings, whereas RectFlow yields a slight improvement for grayscale-to-PET translation. We additionally evaluate a regularized variant incorporating a trajectory-smoothness penalty ($\lambda_s > 0$). While promoting sequential coherence leads to a small decrease in pixel-level fidelity in exchange for increased topological stability, the regularized formulation remains competitive and outperforms baseline methods on specific tasks, including semantic-to-photographic synthesis. To support interpretable latent transitions, all qualitative results and subsequent analyzes are therefore reported using the regularized model.

Fig. 3 provides qualitative visual comparisons across datasets. For Synthetic MNIST, our framework synthesizes emission-intensity patterns with high fidelity while maintaining the intrinsic topological structure of the digits. For Cityscapes, it generates globally consistent scene layouts and well-defined object boundaries from sparse semantic conditioning. For clinical ADNI imaging, magnified anatomical insets indicate that translating structural MRI observations into FA maps preserves microstructural tissue interfaces of axonal organization. Conversely, the inverse translation reconstructs ventricular morphology accurately without introducing spurious structural artifacts.

4.3 Trajectory Analysis and Latent Transport

Image trajectories. Fig. 4(a–c) illustrates the continuous transformation dynamics across representative translation tasks by decoding intermediate latent states. While the diffusion baseline (DiT) exhibits substantial trajectory fragmentation and chaotic intermediate estimates in the early stages, FlowCycle formulation produces a smooth and monotonic transport path. Across diverse structural paradigms, our method preserves global geometric layout and fine anatomical topology throughout the entire trajectory, while target-domain appearance emerges progressively and coherently. These

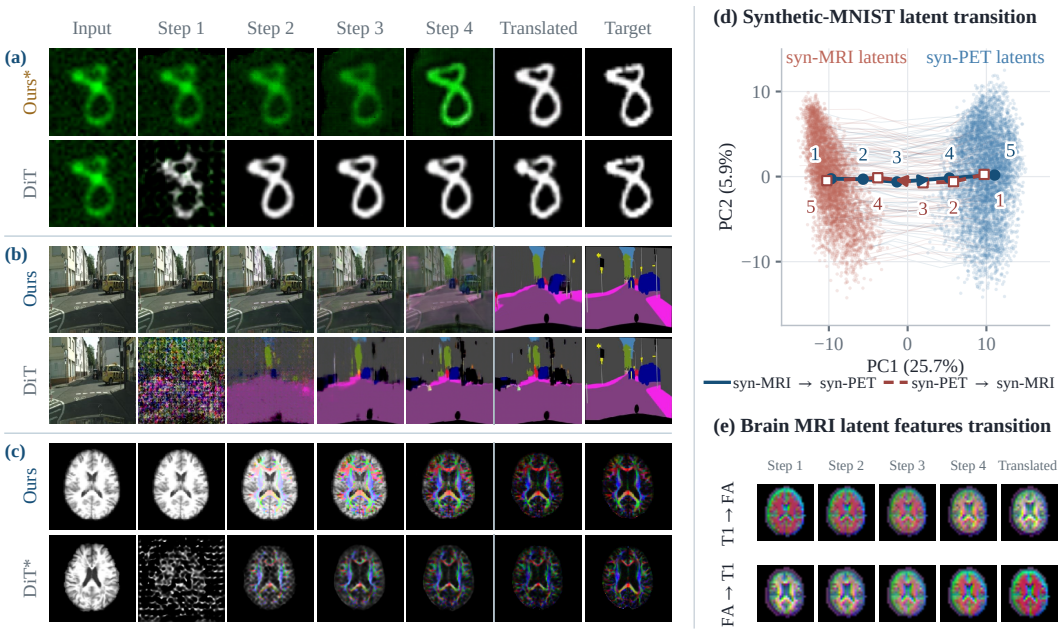


Figure 4: **Comparison of decoded trajectories generated by the proposed model, the strongest baseline, and latent transport. (a–c)** Decoded image trajectories for Synthetic MNIST MRI/PET, Cityscapes, and ADNI. The step index denotes the snapshot shown. *Input/Ground truth* correspond to paired real images, and *Translated* refers to the final predicted output. **(d)** Trajectory characterization using principal component analysis (PCA) of the proposed synthesis, showing real latent codes and the corresponding individual (thin) and mean (bold) trajectories. **(e)** Visualization of the latent representations across latent states in both directions.

visualizations demonstrate that our approach establishes a topologically stable vector field that offers mechanistic interpretability during translation, with its downstream utility evaluated in Section 4.5.

Latent trajectories. Fig. 4(d,e) examines transport directly in latent space. Mean MNIST paths move toward target-domain latent code clouds in both directions (d), while shared PCA colors reveal gradual spatial changes in ADNI feature maps (e). These views support progressive changes in the learned representations independently of the visualization decoder. Reversibility is evaluated in Section 4.4. Additional baseline trajectories, decoder controls, and latent-transport diagnostics are provided in Appendices B and C.

4.4 Latent Reversibility and Image-Cycle Consistency

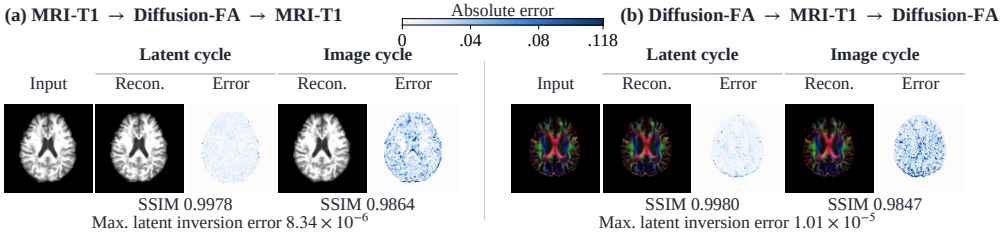


Figure 5: **Latent-space reversibility and image-cycle consistency in brain imaging.** (a) Forward cycle: MRI → diffusion FA → reconstructed MRI. (b) Backward cycle: diffusion FA → MRI → reconstructed diffusion. Latent cycles measure reversibility of transported latent representations before decoding; full image cycles also include autoencoding and reconstruction. Residual maps show absolute voxel-wise intensity differences on a common scale across conditions.

Fig. 5 disaggregates exact algebraic reversibility and image-space cycle consistency on the ADNI dataset. For the latent-space round trip, the maximum absolute residual error remains at the level of floating-point precision ($\varepsilon_{\text{inv}} \approx 0$), indicating that the discrete spatial coupling layers imple-

ment closed-form inversion without measurable numerical dissipation. Assessment of the complete image-space cycle through the autoencoding projections shows negligible anatomical drift and spatially confined, low-amplitude structural residuals. These results support that the learned latent manifolds remain well constrained and stably anchored under bidirectional mapping.

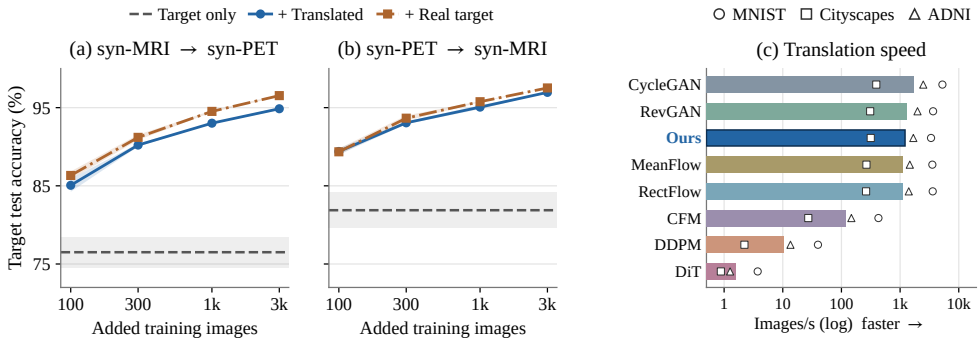


Figure 6: **Downstream utility and throughput.** (a, b) Target-domain classification accuracy in low-label settings (10 labels/class), comparing augmentation with synthesized cross-modal observations vs. real target-domain data. (c) Translation throughput (images/s, log scale; higher is faster): bars show the geometric mean across three datasets, markers show per-dataset results, all on a single NVIDIA A100 GPU.

4.5 Downstream Transfer and Computational Complexity

Fig. 6(a,b) quantifies the downstream utility of synthesized observations in a label-scarce regime. When the target-domain training set is limited (10 labeled instances per class), augmenting it with cross-modally synthesized samples yields a consistent increase in classification accuracy. Moreover, performance saturates toward that of models trained solely on fully paired, authentic target data, indicating that the proposed framework transfers task-relevant discriminative structure across domains rather than overfitting to superficial appearance statistics.

Fig. 6(c) shows amortized per-sample inference latency. Using a closed-form discrete transport formulation, FlowCycle performs bidirectional inference in one forward pass, yielding orders-of-magnitude lower latency than iterative continuous-time diffusion and flow-matching baselines. While heuristic decoupled-network methods run similarly fast, only FlowCycle offers exact analytic reversibility and interpretable intermediate representations without continuous-time numerical integration overhead.

5 Discussion & Conclusion

This work introduces FlowCycle, an end-to-end generative framework for bidirectional cross-modal translation that combines single-pass deterministic inference with an analytically exact reversible operator. It formulates cross-domain transport as an exact discrete normalizing flow on a shared latent manifold, avoiding pixel-space cycle heuristics and iterative continuous-time diffusion/flow-matching. Spatial affine coupling layers create an explicit bottleneck that limits shortcuts, enables closed-form inversion independent of solver tolerances, and exposes intermediate geometric trajectory states at no extra cost. Across synthetic benchmarks, urban street scenes, and clinical neuroimaging (macrostructural contrasts and diffusion-based microstructure), FlowCycle achieves state-of-the-art fidelity while remaining exactly reversible and preserving downstream-relevant features. Currently, FlowCycle is trained mainly with paired data. Although the latent discrete flow is exactly invertible, the outer autoencoder is a continuous projection, so pixel-space cycle fidelity is limited by reconstruction error rather than a global bijection. Future work will target fully unpaired training via distribution matching and/or optimal-transport regularization, and add domain-specific inductive biases (e.g., diffusion biophysics or task-conditioned semantic priors) to improve fidelity and clinical/perceptual utility. FlowCycle's deterministic latent trajectories also offer a principled basis for modeling state evolution and continuous transformations on geometric manifolds (Bronstein et al., 2017).

AI USE STATEMENT

In this work, we used generative AI tools (an LLM coding assistant) to implement and debug parts of the experimental code. Additionally, we used generative AI tools to polish the manuscript for clarity and formatting. We have reviewed all AI-assisted work: all code was checked and its outputs verified by the authors, and every reported number was regenerated from the saved evaluation outputs. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only existing, publicly available datasets and involves no new data collection or interaction with human participants. The ADNI neuroimaging data were obtained under the ADNI data use agreement; ADNI collected them with institutional review board approval at each participating site and informed consent from all participants, and distributes them de-identified. We report only aggregate statistics and individual slices, and release no subject identifiers. Cityscapes is used under its research license, and the synthetic MNIST-based pairs contain no personal data. Translated medical images are research outputs and are not intended for clinical diagnosis without further validation.

REPRODUCIBILITY STATEMENT

Sections 3.1–3.3 and Algorithm 1 specify the architecture, objectives, and three-stage training procedure. Appendix A describes dataset construction, the train/test splits, and the evaluation protocol; and Appendices B detail the trajectory and cycle analyses. All reported numbers are computed from saved model outputs on the held-out test sets. We will release the code, configuration files, and trained checkpoints upon publication.

REFERENCES

Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(1):3563–3593, 2014.

Peter J Basser, James Mattiello, and Denis LeBihan. Mr diffusion tensor spectroscopy and imaging. *Biophysical journal*, 66(1):259–267, 1994.

James C Bezdek and Richard J Hathaway. Some notes on alternating optimization. In *AFSS international conference on fuzzy systems*, pp. 288–300. Springer, 2002.

Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

Joseph Paul Cohen, Margaux Luck, and Sina Honari. Distribution matching losses can hallucinate features in medical image translation. In *International conference on medical image computing and computer-assisted intervention*, pp. 529–536. Springer, 2018.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Anders M Dale, Bruce Fischl, and Martin I Sereno. Cortical surface-based analysis: I. segmentation and surface reconstruction. *Neuroimage*, 9(2):179–194, 1999.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.

Peyré Gabriel and Cuturi Marco. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025. URL https://arxiv.org/abs/2505.13447.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30467–30476, 2026.

Aidan N Gomez, Mengye Ren, Raquel Urtasun, and Roger B Grosse. The reversible residual network: Backpropagation without storing activations. *Advances in neural information processing systems*, 30, 2017.

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*, 2020.

Ernst Hairer, Gerhard Wanner, and Syvert P Nørsett. *Solving ordinary differential equations I: Nonstiff problems*. Springer, 1993.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5967–5976. Ieee, 2017.

Durk P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, 2018.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Dongwook Lee, Junyoung Kim, Won-Jin Moon, and Jong Chul Ye. Collagan: Collaborative gan for missing image data imputation. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2482–2491. IEEE, 2019.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Jiang Liu, Srivathsa Pasumarthi, Ben Duffy, Enhao Gong, Keshav Datta, and Greg Zaharchuk. One model to synthesize them all: Multi-contrast multi-scale transformer for missing data imputation. *IEEE transactions on medical imaging*, 42(9):2577–2591, 2023a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023b. URL https://arxiv.org/abs/2209.03003.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, 2017.

Susumu Mori, Barbara J Crain, Vadappuram P Chacko, and Peter CM Van Zijl. Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 45 (2):265–269, 1999.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, Clifford R Jack Jr, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. Alzheimer's disease neuroimaging initiative (adni) clinical characterization. *Neurology*, 74(3): 201–209, 2010.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Tycho F. A. van der Ouderaa and Daniel E. Worrall. Reversible GANs for memory-efficient image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Michael W Weiner, Paul S Aisen, Clifford R Jack Jr, William J Jagust, John Q Trojanowski, Leslie Shaw, Andrew J Saykin, John C Morris, Nigel Cairns, Laurel A Beckett, et al. The alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimer's & Dementia*, 6(3): 202–211, 2010.

Hongyi Zhang, Yann N Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization. *arXiv preprint arXiv:1901.09321*, 2019.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 2242–2251. Ieee, 2017.

Shilong Zou, Yuhang Huang, Renjiao Yi, Chenyang Zhu, and Kai Xu. Cyclediff: Cycle diffusion models for unpaired image-to-image translation. *IEEE Transactions on Image Processing*, 2026.

A Datasets and evaluation protocols

Synthetic MNIST MRI/PET is a controlled benchmark where digit geometry and class cues derive from MNIST (LeCun et al., 1998). The "MRI" domain replicates grayscale digits across three channels. The synthetic PET domain applies class-conditional hotspot enhancement, point-spread blurring ($\sigma = 1.2$), an SUVR-like uptake nonlinearity simulating localized tracer accumulation, additive Gaussian noise ($\sigma = 0.05$), and a pseudo-color tint, producing $3 \times 64 \times 64$ images. Modality labels denote these synthetic transformations, not clinical scans.

Cityscapes (Cordts et al., 2016) evaluates bidirectional translation between street photos and color-coded semantic label maps. We use paired pix2pix (Isola et al., 2017) at 256×256. Training applies the same resize, random crop, and horizontal flip to both images in each pair to preserve alignment; evaluation resizes inputs to 256×256 without augmentation.

ADNI evaluates cross-modality translation between T1-weighted MRI (structure) and fractional anisotropy (FA) from diffusion MRI (diffusion anisotropy). Paired T1/FA volumes are coregistered to MNI space and resampled to 112×112 in-plane. We split subjects 80/20 (seed 42): 170 train and 43 test. We extract 10 axial slices per subject (1,700 train; 430 test). In direction-encoded color FA (DEC-FA), hue encodes principal diffusion direction and brightness encodes FA magnitude; diffusion orientation comes from acquisition and is not predicted.

Parameters & Experimental Setup. All models are trained on identical splits and evaluated via test PSNR/SSIM without EMA. CycleGAN (Zhu et al., 2017) and RevGAN (van der Ouderaa & Worrall, 2019) use 64-channel ResNets (6 blocks; 9 for Cityscapes) with LSGAN, Adam (lr 2×10^{-4}), $\lambda_{\text{cyc}} = 10$, and $\lambda_{\text{id}} = 5$. DDPM (Ho et al., 2020) (50 DDIM steps), CFM (Lipman et al., 2023) (10 Euler steps), RectFlow (Liu et al., 2023b) (1-step reflow), and MeanFlow (Geng et al., 2025) (1-step) share a 64-channel U-Net trained with AdamW (lr 2×10^{-4}). DiT (**?**) (width 384, depth 12; patch 2, 4 on Cityscapes) uses $1{,}000$ ancestral steps over a $4\times$-downsampled latent space. FlowCycle uses Adam ($\beta = (0.5, 0.999)$, clip 5.0) trained across three stages: Stage 1 pretrains domain autoencoders (lr 2×10^{-4}); Stage 2 trains four coupling blocks (width 128, lr 2×10^{-4}); Stage 3 jointly refines flow+decoders (lr 10^{-4}) and critics (lr 2×10^{-4}) with $\lambda_{\text{rec}} = \lambda_{\text{cross}} = \lambda_{\text{cyc}} = 10$ and adversarial weight 1.0. FlowCycle-TR adds $\lambda_{\text{lat}} = 2$ and $\lambda_s = 1$ in Stage 3 ($\lambda_s = 0.3$, adversarial weight 0.1 for MNIST-MRI/PET), whereas FlowCycle w/o TR sets both to 0.

B Additional trajectory analysis and decoder evaluation

Latent-transport analysis details. Fig. 4(d) shows PCA on latent codes from real images in both MNIST domains after 4×4 spatial pooling; the first two components explain 31.6% of the variance. Mean trajectories use all 5,000 held-out paired samples, and thin trajectories show 60 random pairs per direction. Discrete states are linked by piecewise-linear segments, but distances in the projection do not imply correspondence in the full latent space. Panel (e) visualizes three shared principal components of $256 \times 28 \times 28$ ADNI feature maps via RGB. Forward and reverse trajectories start from independently encoded real T1- and FA-derived inputs. Fig. 7(c) shows another held-out subject with the same encoding.

Classifier transfer. Fig. 7(a) shows digit-classification on 5,000 MNIST pairs withheld from translator training. In five stratified folds, a standardized linear classifier is trained on real Syn-PET latent codes from four folds and evaluated (fixed) on the remaining fold under three PET-probe conditions. Accuracies are 91.3% on real Syn-PET codes, 10.6% on Syn-MRI codes, and 88.6% after latent transport. A probe trained and tested on Syn-MRI codes reaches 92.9%. Thus, transported codes are mostly compatible with a classifier tuned to the target (PET) representation, but this does not assess semantic fidelity at intermediate transport states.

Distance to the paired target. Fig. 7(b) plots the saved per-state ADNI test-set means of $\|z^{(\ell)} - z_{\text{target}}\|_2 / \|z_{\text{target}}\|_2$ over 430 slices from 43 subjects. The target is the paired real image's encoded representation. The mean relative distance decreases from 1.29 to 0.83 for T1$\rightarrow$FA and from 1.29 to 0.78 in reverse. The curves summarize discrete-state alignment. Forward per-block RMS changes are 0.47, 0.55, 0.63, and 0.66.

Shared-decoder protocol. The four-block flow maps $z^{(0)} = E_A(x)$ to $z^{(4)} = f(E_A(x))$. Reverse traversal starts from $E_B(y)$ and applies inverse blocks in reverse order. The encoders, flow, and

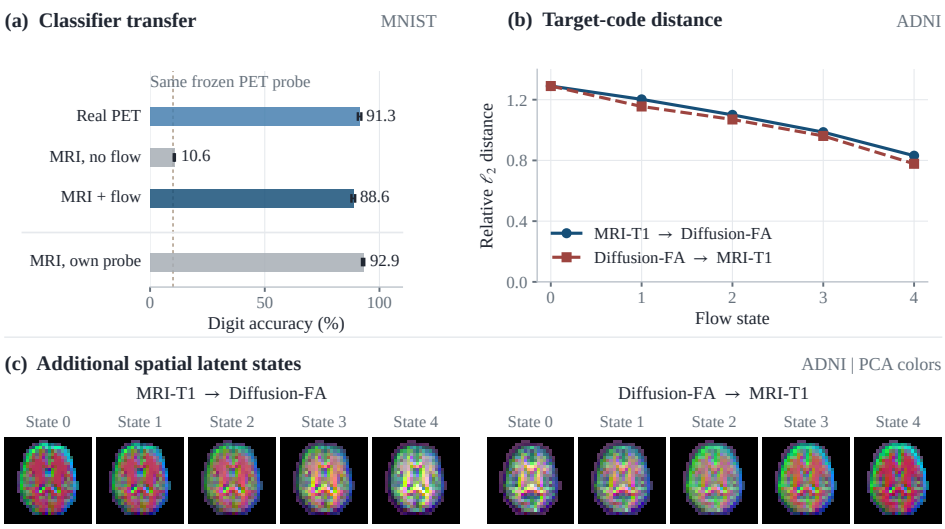


Figure 7: **Supporting latent-transport analyses.** **(a)** MNIST classifier transfer; error bars show standard deviations across five folds, and the dashed line marks 10% chance accuracy. **(b)** Mean relative distance to the paired ADNI target code at each native flow state. **(c)** An additional held-out ADNI subject, shown in both traversal directions with shared PCA colors.

original decoders are frozen. The shared decoder G is supervised on four mappings: $E_A(x) \mapsto x$, $f(E_A(x)) \mapsto y$, $E_B(y) \mapsto y$, and $f^{-1}(E_B(y)) \mapsto x$. Its loss uses only source/target images (not intermediate states) and it is used only for visualization.

Fig. 8 shows the native intermediate states ($\ell = 0, \ldots, 4$) and the original decoder's final output (separate column). For visualization, the scalar prediction sets intensity, principal diffusion directions set RGB, and FA provides the reveal mask. The foreground is the union of voxels with valid T1 and FA > 0.025. The mask is Gaussian-smoothed ($\sigma = 0.8$ pixels); core regions are thresholded at the 70th percentile and filtered to connected components $\geq$ six pixels. Reveal uses a distance-ranked smoothstep (width 0.08) plus Gaussian feathering ($\sigma = 0.75$ pixels). Luminance blending interpolates between grayscale and direction-colored views with $m = \ell/4$ (forward) and $m = 1 - \ell/4$ (reverse). These settings are for visualization only; the model does not predict diffusion orientation.

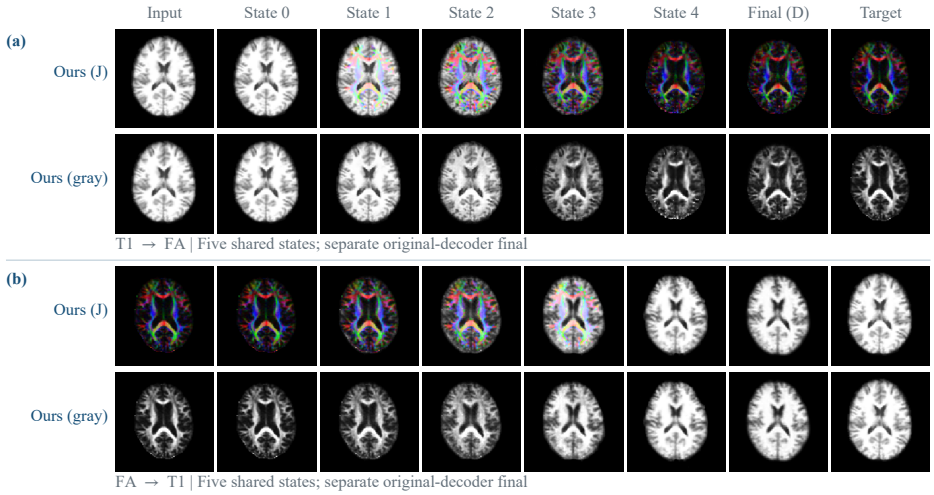


Figure 8: **Visualization of additional brain trajectories** derived from the shared decoder and the native decoder, presented for both directions in the paired sample.

Latent trajectory compared with baselines. Fig. 9 to Fig.10 illustrate additional examples in each direction, against baselines with the native transition path.

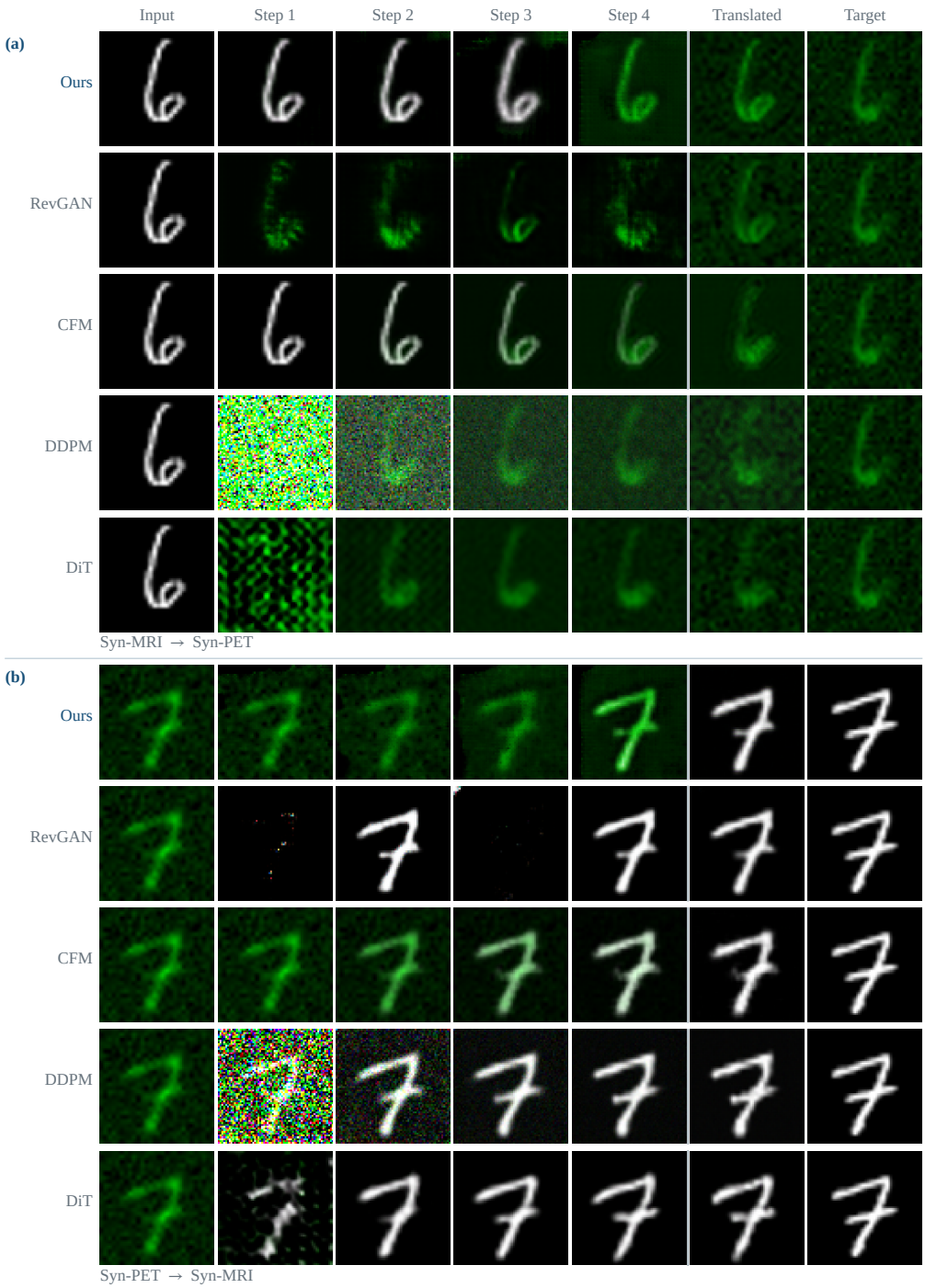


Figure 9: **Comprehensive comparison of trajectories across all models on the synthetic MNIST-MRI/PET dataset**. Two representative samples are shown in two transformation directions; each group includes results from Ours, RevGAN, CFM, DDPM, and DiT.

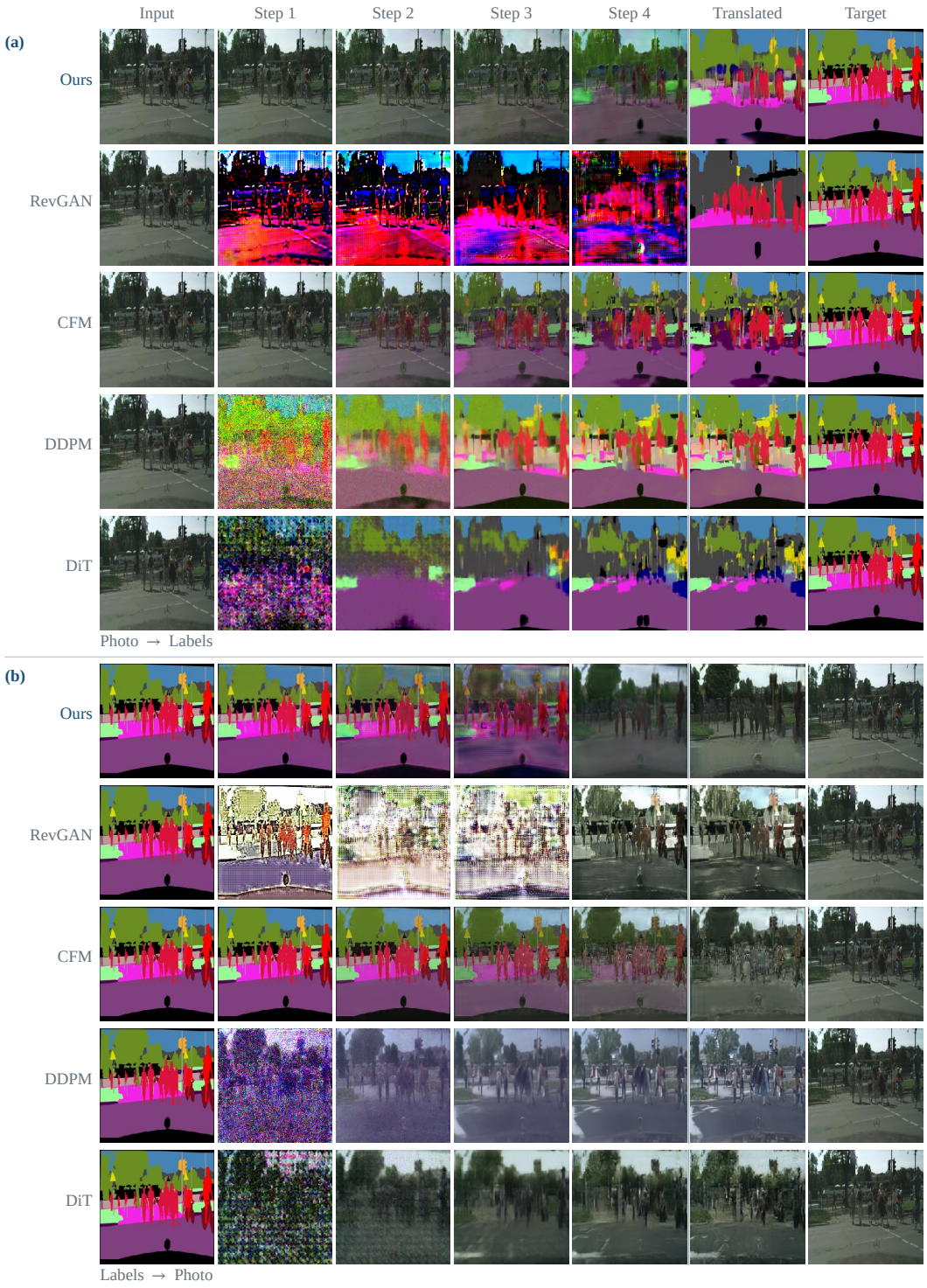


Figure 10: **Additional Cityscapes trajectories.** One sample; two directions; each group contains Ours, RevGAN, CFM, DDPM, and DiT.

C SPATIAL AFFINE-COUPLING DIAGNOSTICS

Fig. 11 presents the effective coupling parameters of FlowCycle-TR for each transformation direction, evaluated on two held-out ADNI image slices. For the subset of channels updated via an affine coupling transformation, we express $w^{(2)} = a \odot p^{(2)} + b$, with $a = \exp(s)$ and $b = t$. The visualizations aggregate statistics over the 128 updated channels at each spatial location: the signed channel-wise mean of $\log a$ characterizes the average local expansion or contraction, whereas the channel-wise root-mean-square (RMS) of b quantifies the magnitude of the additive shift.

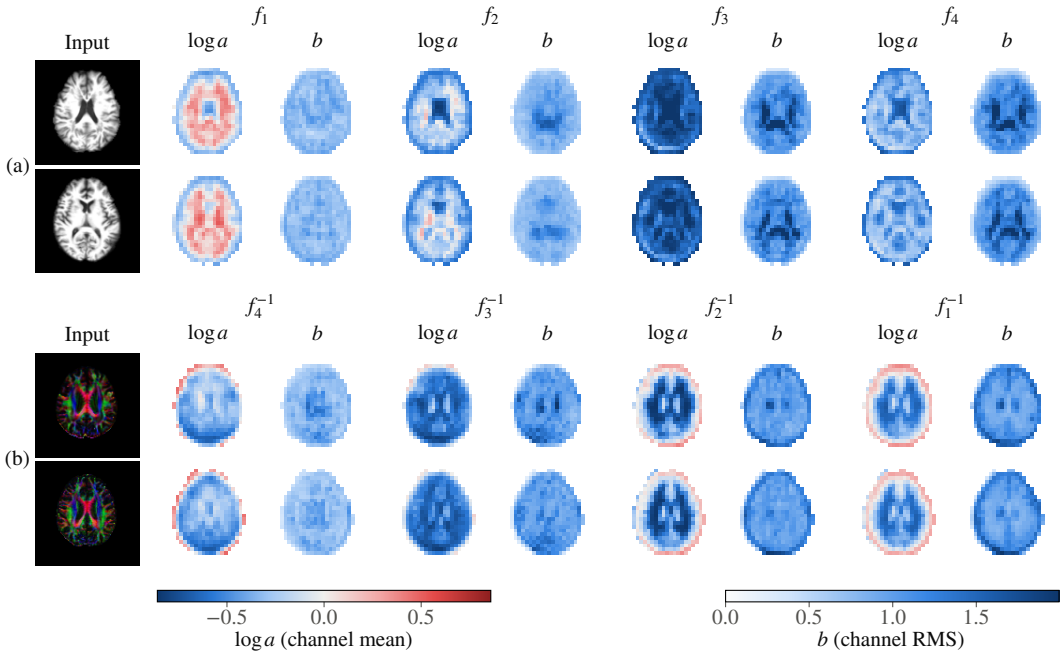


Figure 11: **Spatial affine-coupling terms on a neuroimaging dataset. (a)** MRI-T1 to Diffusion-FA and **(b)** Diffusion-FA to MRI-T1, with two illustrative subjects per direction. These summarize latent coupling terms rather than image intensities.