E-FLOW: ENERGY-GUIDED FLOW MATCHING FOR STRUCTURE-PRESERVING IMAGE STYLE TRANSFER

Anonymous authors
Paper under double-blind review

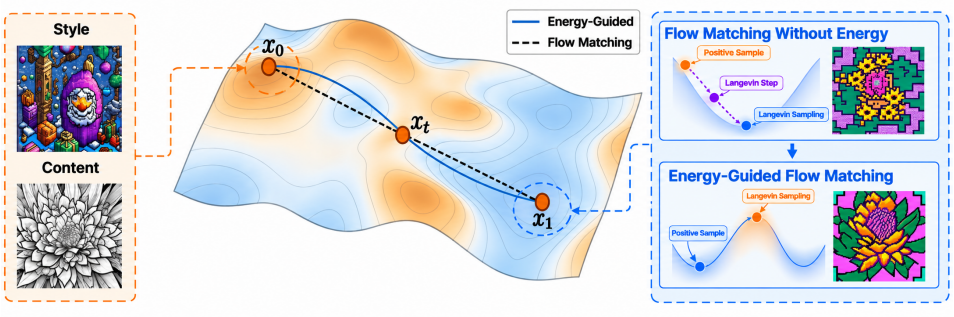


Figure 1: **Comparison of standard and energy-guided Flow Matching**. The energy-guided field alters the transport trajectory by following the compatibility landscape (middle). This proves crucial for Langevin-based local exploration (right), where energy guidance prevents structural degradation and produces high-quality stylizations, unlike the noisy outputs of standard Flow Matching.

ABSTRACT

Reference-based image style transfer aims to reproduce style appearance while preserving content structure. The task remains an underdetermined multi-constraint problem where many intermediate states are plausible but only some are compatible with both content and style. Existing diffusion-based methods lack an explicit scalar criterion to rank these states, and standard Flow Matching, though providing source-to-target geometry, only regresses a coordinate-wise velocity field without scoring state compatibility or coupling all coordinates through a shared objective. In this paper, we propose Energy-Guided Flow Matching (E-Flow) for structure-preserving image style transfer. We introduce a time-dependent energy head that outputs a scalar compatibility energy, and then parameterize the transport field as its negative input gradient, coupling all coordinates through one shared landscape. Thereinto, Flow Matching supplies the geometry, while the energy determines the trajectory. To make the energy discriminative, a local contrastive divergence objective compares the correctly paired target with local alternatives from a short Langevin chain. The stochastic noise helps escape shallow local minima, while the contrastive term forms a low-energy basin around valid stylizations. In addition, we present a discrete Meyer wavelet transform to construct complementary target branches and perform reversible coefficient-space fusion. Experiments show that E-Flow improves the style-content trade-off, yielding higher feature separation with competitive NMI. The code is provided in the supplementary materials.

1 INTRODUCTION

Reference-based image style transfer aims to reproduce the appearance of a style image while preserving the structure and semantics of a content image (Zhang et al., 2023; Deng et al., 2024; Chung et al., 2024). The central challenge is to control the transformation so that object layout and identity remain stable while colour, texture, and local contrast are transported toward the style reference. This task is inherently underdetermined and multi-constraint. Many intermediate states

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.

are visually plausible, but only some remain compatible with both the content layout and the style texture. Therefore, a useful model should represent content and style jointly and be able to assess whether intermediate states remain compatible with both the content layout and the style texture.

Recent diffusion-based methods address this problem through inversion, attention replacement, feature injection, and disentangled conditioning (Zhang et al., 2023; Deng et al., 2024; Chung et al., 2024; Qi et al., 2024; Frenkel et al., 2024). Other directions improve semantic matching or analyze which feature statistics carry style information (He et al., 2026; Yu et al., 2025; Deng et al., 2022). However, they lack an explicit scalar criterion to evaluate whether a nearby stylized state is truly compatible with the given content-style pair, making the style-content trade-off difficult to control and content structure prone to distortion as style strength increases. Flow Matching offers a direct way to study this movement by learning a time-dependent vector field that matches the velocity of a prescribed path, operating on structured feature representations rather than image-space noise trajectories (Lipman et al., 2023). SCFlow (Ma et al., 2025) treats content-style composition as continuous and reversible transport in a joint embedding space. However, SCFlow regresses a coordinate-wise velocity field (Lipman et al., 2023), which provides the source-to-target geometry without scoring state compatibility . A direct velocity head can assign directions independently per coordinate, whereas style transfer requires global coordination because texture changes may affect structure and structural changes may weaken style.

Energy-weighted Flow Matching defines a target distribution and learns the corresponding guided flow without an auxiliary guidance model. Feng et al. (Feng et al., 2025) develop a general theory of guidance for flow matching and show that guidance must account for the chosen flow path rather than simply copy diffusion guidance. Energy matching unifies flow-based and energy-based generation through a time-independent scalar potential (Zhang et al., 2025; Balcerak et al., 2025; Feng et al., 2025).These methods establish general principles for guiding a flow with an energy, with applications that include generic generation, inverse problems, and offline reinforcement learning. However, they do not address reference-based style transfer with an explicitly paired content-style state. Their potentials are often unconditional or task-generic, and they are not conditioned on a specific content-style pair. They also do not learn relative energy among nearby states around a valid stylization. Therefore, simply transplanting an energy-guided flow is insufficient, because the energy cannot rank whether a nearby joint state is compatible with the given style reference.

For structure-preserving image style transfer, we propose Energy-Guided Flow Matching, shown in Fig. 1. Standard flow Matching supplies a source-to-target geometric path but only regresses a coordinate-wise velocity field. It neither scores whether intermediate states are compatible with both content and style nor couples all embedding coordinates through a shared objective. To address this, we devise a time-dependent energy head that outputs a scalar compatibility energy over the joint content-style state, and we parameterize the transport field as its negative input gradient. This retains Flow Matching as the geometric backbone while making the trajectory subject to one shared landscape, so style-related and content-related dimensions are no longer updated independently.

However, a joint-state energy alone is insufficient for reference-based transfer, because compatibility is pair-specific. Existing energy-guided models often rely on unconditional or task-generic potentials and cannot rank states with respect to a particular content-style pair. We therefore condition the energy on the pair representation and on time. The preferred direction can then change along the path. Early transport resolves coarse source-to-target displacement, while later transport refines local compatibility between style and structure. The energy thus scores states relative to a specific pair rather than a generic data distribution.

Gradient matching alone cannot determine the relative energy of nearby states, since local perturbations may appear plausible while breaking content or style. To make the landscape discriminative, we introduce a local contrastive divergence objective. The paired target serves as the positive sample, and a negative sample is generated by Gaussian perturbation followed by a short Langevin refinement under the same energy. Comparing their energies encourages lower energy for valid targets and higher energy for nearby alternatives. The Langevin noise helps escape shallow local minima, while the contrastive term forms a low-energy basin around valid stylizations, preventing similar scores for visually close but semantically incompatible states.

Frequency-aware diffusion and wavelet-based stylization provide related evidence that separating coarse structure from detail can help preserve content while transferring appearance (Gao et al.,

2025; Wang et al., 2025). Therefore, a discrete Meyer wavelet transform is used only to construct complementary target branches and perform reversible coefficient-space fusion before frozen unCLIP decoding. In summary, our contributions are as follows:

- We formulate reference-based image style transfer as energy-guided conditional Flow Matching (E-Flow), where a scalar compatibility landscape parameterizes the transport field through its negative input gradient.
- We introduce a pair- and time-conditioned energy together with a local contrastive objective, which uses Langevin-refined negative alternatives to shape a discriminative landscape and form a low-energy basin around valid stylizations.
- We employ a discrete Meyer wavelet transform to construct complementary target branches and perform coefficient-space fusion. The experiments show the effectiveness of E-Flow.

2 Related Work

Reference-based style transfer. The classical formulation matches deep feature statistics between content and style (Gatys et al., 2016), and instance normalization enables fast arbitrary style transfer (Huang & Belongie, 2017). Later work studies feature alignment, attention, and explicit style control. Diffusion-based methods explore inversion, attention replacement, and reference feature injection. Representative examples include inversion-based trajectory manipulation (Zhang et al., 2023), Plug-and-Play feature transfer (Tumanyan et al., 2023), attention-based style injection (Chung et al., 2024), DEADiff for semantic-style separation (Qi et al., 2024), and B-LoRA for independent style-content control (Frenkel et al., 2024). Recent work further addresses selective style control (Lei et al., 2025), attention distillation (Zhou et al., 2025), start-point selection (Xu et al., 2025), and content-style composition (Xing et al., 2025).

Flow Matching learns a vector field by regressing the velocity of a chosen probability path without simulating the path during training (Lipman et al., 2023). Rectified Flow studies related straightening objectives for efficient transport and transfer (Liu et al., 2023). SCFlow applies this principle to style–content disentanglement with reversible representation mappings (Ma et al., 2025).

Energy-guided generative models. Energy-based learning represents a distribution through a scalar score and has long used contrastive divergence to compare observed and locally explored states (Hinton, 2002). Modern formulations study implicit generation (Du & Mordatch, 2020), improved contrastive-divergence optimization (Du et al., 2021), and scalable energy models (Yin et al., 2026). Energy-guided Flow Matching introduces a scalar preference over states and uses its gradient to alter a transport field. Recent work develops path-aware guidance theory (Feng et al., 2025), energy-weighted Flow Matching objectives (Zhang et al., 2025), and unified energy–flow formulations (Balcerak et al., 2025).

3 Method

Given a content image $I_c \in \mathbb{R}^{H\times W\times 3}$ and a style image $I_s \in \mathbb{R}^{H'\times W'\times 3}$, reference-based image style transfer aims to generate a stylized image I_{cs} that preserves the semantic structure of I_c while adopting the visual appearance of I_s. The content image defines the object layout, identity, and spatial arrangement that should remain stable. The style image defines the colour distribution, texture, brushwork, and local contrast that should be transferred. The desired output should satisfy both requirements simultaneously, which makes the task inherently multi-constraint Zheng et al. (2024).

Formally, the goal is to learn a mapping

$$F : (I_c, I_s) \mapsto I_{cs}, \tag{1}$$

such that I_{cs} remains structurally faithful to I_c and visually similar to I_s. Unlike image-to-image translation with a unique ground truth, this mapping is underdetermined. A single content–style pair admits many visually plausible stylizations, and there is no single correct output. Different trade-offs between style strength and content preservation can all be reasonable. Therefore, a successful method must not only produce a stylized image but also control the balance between the two factors.

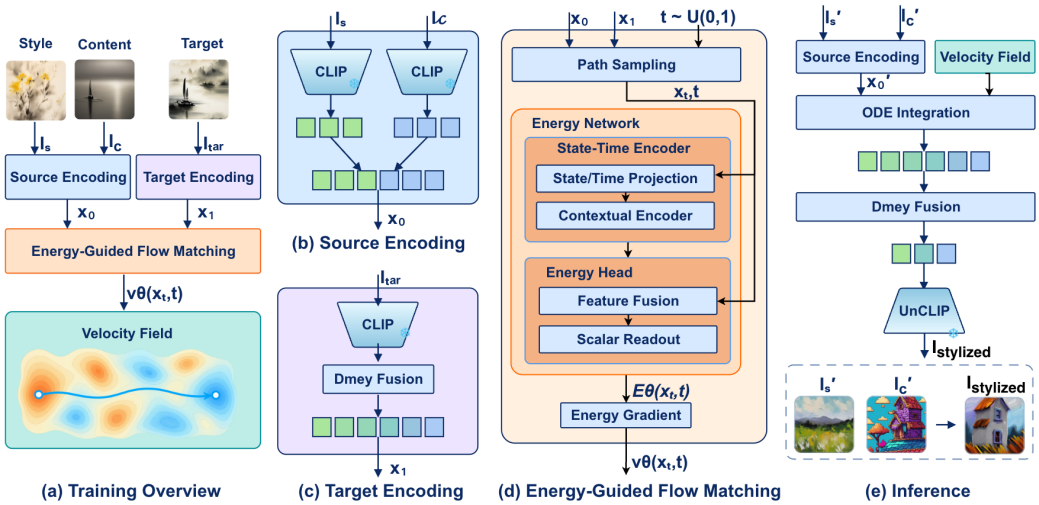


Figure 2: Overview of the proposed energy-guided conditional Flow Matching model. Content and style images enter CLIP separately; the joint representation is transported by an energy-guided field, and the transported branches are fused before frozen unCLIP decoding.

3.1 Overview

We propose Energy-Guided Flow Matching (E-Flow) for structure-preserving image style transfer. As shown in Fig. 2, the overall pipeline consists of a training stage and an inference stage. During training, a content image I_c, a style image I_s, and a paired stylized target I_t are encoded by a frozen CLIP encoder Radford et al. (2021). The content and style embeddings are concatenated into a source state x_0, while the target embedding is transformed by a discrete Meyer wavelet transform into a joint target state x_1. Flow Matching constructs a conditional path between x_0 and x_1, and an energy network outputs a scalar compatibility energy along this path, whose negative gradient defines the transport field. At inference, only I_c and I_s are provided. The model integrates the energy-gradient field from x_0 to the endpoint, splits the transported state into two branches, fuses them in wavelet coefficient space, and decodes with a frozen unCLIP decoder Ramesh et al. (2022) to produce the stylized image I_{cs}.

3.2 Source Encoding

The content image I_c and the style image I_s are encoded independently by the same frozen CLIP encoder, yielding a content embedding $e_c \in \mathbb{R}^d$ and a style embedding $e_s \in \mathbb{R}^d$. The source state is defined as their channel-wise concatenation:

$$x_0 = [e_c \parallel e_s] \in \mathbb{R}^{2d}. \tag{2}$$

The concatenation is explicit, with the first d dimensions encoding content and the remaining d dimensions encoding style, so that content and style coordinates remain identifiable at the input of the transport model. The encoder is kept frozen throughout training and inference.

3.3 Target Encoding

Target encoding is used only during training. The paired stylized target I_t is encoded by the frozen CLIP encoder into a target embedding $e_t \in \mathbb{R}^d$. To preserve the multi-scale separation between structure and texture before decoding, we apply a one-level discrete Meyer wavelet transform to e_t:

$$(a_t, d_t) = \mathrm{DWT}_{\mathrm{dmey}}(e_t), \tag{3}$$

where a_t and d_t denote approximation and detail coefficients. As shown in Fig. 3, the two training branches are reconstructed with complementary coefficient weights:

$$\begin{aligned} e_t^{(0)} &= \mathrm{IDWT}_{\mathrm{dmey}}(\alpha_a a_t,\ \alpha_d d_t), \\ e_t^{(1)} &= \mathrm{IDWT}_{\mathrm{dmey}}((1-\alpha_a) a_t,\ (1-\alpha_d) d_t), \\ x_1 &= [e_t^{(a)} \parallel e_t^{(d)}] \in \mathbb{R}^{2d}. \end{aligned} \tag{4}$$

Here $\alpha_a = 1$ and $\alpha_d = 0$, so the two branches carry complementary approximation and detail information. $e_t^{(0)}$ retains the low-frequency semantic structure, while $e_t^{(1)}$ retains the high-frequency edge and texture cues. Concatenating the two branches gives the joint target state $x_1 \in \mathbb{R}^{2d}$. This construction exposes complementary channels to the transport model. Meyer wavelets are orthogonal, compactly supported, and well localized in frequency, making them effective at separating approximation and detail information without redundancy.

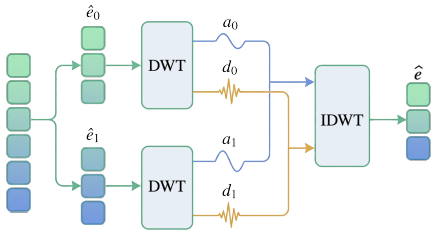


Figure 3: Reversible coefficient-space fusion via discrete Meyer wavelet. Two branches are decomposed into approximation and detail coefficients.

3.4 Energy-Guided Flow Matching

Conditional Path and Energy Parameterization Flow Matching Lipman et al. (2023) provides the geometric backbone of the transport. For $t \sim \mathcal{U}[0,1]$ and a source state $\epsilon = x_0$, the intermediate state is defined by the linear interpolation

$$x_t = tx_1 + [1 - (1 - \sigma_{\min})t]\epsilon, \tag{5}$$

where $\sigma_{\min}$ controls the residual source contribution at the terminal time. The corresponding conditional velocity is

$$u_t(x_0, x_1) = x_1 - (1 - \sigma_{\min})x_0. \tag{6}$$

A standard Flow Matching model regresses this velocity with a coordinate-wise vector head, but such a head does not explicitly score whether nearby states are compatible, nor does it couple all embedding coordinates through a shared objective. We retain the same path and velocity target but parameterize the learned field through a scalar energy.

Given (x_t, t), a state-time encoder computes a time-dependent context:

$$h_t = f_\phi(x_t, t) \in \mathbb{R}^m. \tag{7}$$

An energy head then combines h_t, the current state x_t, and a time embedding, and reduces them to a single scalar:

$$E_\theta(x_t, t) = g_\theta(h_t, x_t, t) \in \mathbb{R}. \tag{8}$$

The transport field is defined as the negative state gradient of this scalar:

$$v_\theta(x_t, t) = -\nabla_{x_t} E_\theta(x_t, t). \tag{9}$$

Coordinates share one compatibility landscape, changing one coordinate influences the preferred direction of the others. The landscape is conditioned on state-time contexts and time t, so the preferred direction can change along the path. Early transport resolves coarse source-to-target displacement, while later transport refines local compatibility between style and structure. Flow Matching supplies the geometry Lipman et al. (2023), and the energy determines how the trajectory moves through that geometry Feng et al. (2025).

Langevin Chain and Local Contrastive Divergence The velocity-matching loss only constrains the gradient direction of the energy along the path. It does not specify how the energy should behave around a valid stylization. Many local perturbations remain visually close to the target yet break either content or style. If the energy assigns similar scores to these alternatives, the transport may drift toward incompatible states. To shape the relative ranking of nearby states, we introduce a local contrastive divergence objective.

Starting from a positive target state x^+, we initialize a local alternative with Gaussian perturbation and refine it using a short Langevin chain under the same energy:

$$\begin{aligned} x_0^- &= x^+ + \sigma_{\text{neg}}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \\ x_{k+1}^- &= x_k^- - \eta_k \nabla_x E_\theta(x_k^-, 1) + \sqrt{2\eta_k}\xi_k, \quad \xi_k \sim \mathcal{N}(0, I), \end{aligned} \tag{10}$$

where σ_{neg} controls the perturbation scale, η_k is the step size, and ξ_k is the Langevin noise. The deterministic gradient term pushes the state along the energy descent direction, while the stochastic

term explores the target neighborhood. The stochastic noise in the Langevin chain helps escape shallow local minima. The negative sample x^- is used only during training to compare energies and does not constitute an additional generation path at inference.

By comparing the energies of positive and negative samples, the landscape forms a low-energy basin around valid stylizations, while local perturbations that break either content or style are pushed to higher energy. This mechanism directly addresses the bottleneck that gradient matching alone cannot distinguish between compatible and incompatible nearby states. A few unstable Langevin chains may produce extreme negative energies that dominate the objective, so we trim extreme values and clamp the contrastive statistic to stabilize training. The resulting energy provides a reliable compatibility score for intermediate states along the transport path.

3.5 Training Objective

Training is performed on a dataset of paired content–style–stylized triplets $\{(I_c, I_s, I_t)\}$. The CLIP embeddings of a content–style pair provide x_0, and the paired target provides x_1 through the Meyer transform. A random time $t \sim \mathcal{U}[0,1]$ is sampled, and x_t is formed by Eq. 5. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \mathcal{L}_{\mathrm{CD}}, \tag{11}$$

where the velocity-matching loss is

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{t,x_0,x_1}\left[\|v_\theta(x_t,t) - u_t(x_0,x_1)\|_2^2\right]. \tag{12}$$

The contrastive divergence loss compares the energies of positive and negative samples:

$$\mathcal{L}_{\mathrm{CD}} = \lambda_{\mathrm{CD}}\left(\mathbb{E}[E_\theta(x^+,1)] - \mathbb{E}[E_\theta(x^-,1)]\right). \tag{13}$$

Minimizing this term encourages lower energy for correctly paired targets and higher energy for nearby alternatives, forming a low-energy basin around valid stylizations. Negative sampling and contrastive divergence are used only during training. Embedding MSE is recorded only as a validation measure and is not part of the energy objective.

3.6 Inference

At inference, only the content image I_c and the style image I_s are provided. Source encoding yields $x_0 = [e_c \parallel e_s]$, consistent with training. The model integrates the learned energy-gradient field $v_\theta(x_t,t)$ from $t=0$ to $t=1$, obtaining a transported joint state. This state is split into two predicted branches $\hat{e}^{(a)}$ and $\hat{e}^{(d)}$. The two branches are decomposed in the same wavelet domain, and their coefficients are added before reconstruction:

$$\begin{aligned}(a_0, d_0) &= \mathrm{DWT}_{\mathrm{dmey}}(\hat{e}^{(a)}),\\ (a_1, d_1) &= \mathrm{DWT}_{\mathrm{dmey}}(\hat{e}^{(d)}),\\ \hat{e} &= \mathrm{IDWT}_{\mathrm{dmey}}(a_0 + a_1, d_0 + d_1).\end{aligned} \tag{14}$$

The fused embedding $\hat{e}$ is passed to the frozen unCLIP decoder to produce the stylized image I_{cs}. Inference does not construct negative samples or use the contrastive divergence term. The image encoder and decoder remain frozen, isolating the effect of energy-guided representation transport.

3.7 Why a scalar energy guides a vector field

The energy parameterization provides a useful separation between geometry and preference. Let $\gamma(t)$ denote the learned state trajectory. Along this trajectory, the energy changes according to

$$\frac{\mathrm{d}}{\mathrm{d}t}E_\theta(\gamma(t),t) = \nabla_x E_\theta^\top \dot{\gamma}(t) + \partial_t E_\theta. \tag{15}$$

The state-gradient term is the quantity used by the transport field, while the explicit time derivative allows the preferred direction to vary over the path (Feng et al., 2025). In early integration steps, the state is close to the source representation and the field must resolve the large content–style displacement. Near the target, the same energy can emphasize local compatibility and suppress directions that alter the content structure. This interpretation does not require a separate image-space reward: the scalar is evaluated directly on the representation being transported.

Table 1: Quantitative comparison of style and content separation across different methods and class counts. NMI and FDR are reported for style classes and content classes. E-Flow is compared with CLIP, DEADiff, CSD, SCFlow, and ablations without energy or without dmey fusion.

Method	Style Classes						Content Classes			
	10		25		51		25		200	
	NMI↑	FDR↑	NMI↑	FDR↑	NMI↑	FDR↑	NMI↑	FDR↑	NMI↑	FDR↑
CLIP Radford et al. (2021)	0.49	0.24	0.58	0.31	0.40	0.30	0.37	0.33	0.54	0.43
DEADiff Qi et al. (2024)	0.39	0.26	0.51	0.34	0.41	0.34	0.65	0.44	0.51	0.56
CSD Somepalli et al. (2024)	0.69	0.52	0.82	0.67	0.72	0.63	0.19	0.27	0.33	0.31
SCFlow Ma et al. (2025)	0.81	2.60	**0.92**	3.73	**0.87**	3.52	0.84	1.80	0.84	2.17
E-Flow (w/o energy)	0.83	0.93	0.81	1.11	0.76	1.11	0.87	1.18	0.81	1.25
E-Flow (w/o dmey fusion)	0.86	2.75	0.84	3.12	0.81	3.12	0.88	2.38	0.83	2.51
E-Flow (Ours)	**0.88**	**3.47**	0.86	**4.06**	0.83	**4.06**	**0.88**	**2.47**	**0.84**	**2.61**

The Flow Matching target remains important because an energy alone does not specify a unique transport schedule. The prescribed path determines which states are visited and supplies the velocity target in Eq. 6 (Lipman et al., 2023); the energy objective selects among locally plausible directions around that path. In this sense, the two terms are complementary Balcerak et al. (2025). Removing the Flow Matching term would leave the model without the intended endpoint geometry, while removing the energy term would recover a purely velocity-matching baseline.

4 Experiments

4.1 Experimental setup

We conduct experiments on the SCFlow dataset (Ma et al., 2025), adopting the same train/validation split, frozen CLIP/unCLIP pair, preprocessing, solver, and evaluation scripts. The data are organized by style category and content instance. Using c^* and s^* for alternate content and style identities, each sampled triplet has the form $(I_{c^*,s}, I_{c,s}, I_{c,s^*})$: it contains shared-style and shared-content pairs together with their corresponding style–content target. These cases are mixed in training to expose the factor substitutions needed for disentanglement while keeping the encoder and decoder fixed. NMI measures agreement between representation clusters and factor labels; FDR is the between-class to within-class scatter ratio Nguyen et al. (2010).

4.2 Main Results

Tab. 1 shows the quantitative comparison across style and content class counts. Relative to SCFlow, E-Flow improves FDR in all five settings, indicating that the learned representation achieves better class separation for both style and content. We compared with the methods, including CLIP Radford et al. (2021), DEADiff Qi et al. (2024), CSD Somepalli et al. (2024) and SCFlow Ma et al. (2025). For style NMI, E-Flow is higher on 10 classes (0.88 vs. 0.81) but slightly lower on 25 and 51 classes (0.86 vs. 0.92 and 0.84 vs. 0.87). For content NMI, E-Flow improves on 25 classes (0.88 vs. 0.84) and matches SCFlow at 200 classes at the displayed precision. The energy term acts as a compatibility preference that sharpens feature separation, while the Flow Matching geometry remains the primary driver of factor recoverability. The consistent FDR gains suggest that energy guidance helps form more compact and better-separated clusters, which is desirable for structure-preserving style transfer. The decrease in style NMI on 25 and 51 classes can be attributed to the energy term's emphasis on local compatibility between content and style, which may slightly reduce the compactness of style clusters in favor of better content preservation. Unlike SCFlow, our energy-guided field prioritizes ranking states by compatibility, a criterion that does not always align with maximizing style cluster separation. Consequently, the model trades a small amount of style NMI for consistent FDR gains and improved structural preservation.

4.3 Ablation Study

Tab. 1 shows the results of E-Flow without energy and E-Flow without dmey fusion, which demonstrates the effectiveness of the complete E-Flow. Removing either the energy term or the dmey fusion

Table 2: Results on different fusion methods for the target encoder. The wavelet domain is more suitable for the integration and decomposition of styles and contents.

Interface	Style Classes						Content Classes			
	10		25		51		25		200	
	NMI↑	FDR↑	NMI↑	FDR↑	NMI↑	FDR↑	NMI↑	FDR↑	NMI↑	FDR↑
Mean	0.86	2.75	0.84	3.12	0.81	3.12	0.88	2.38	0.83	2.51
sym4	0.87	**3.53**	0.86	4.11	0.83	**4.10**	0.88	**2.52**	0.83	**2.66**
coif2	0.88	3.31	0.86	3.87	0.83	3.83	0.88	2.49	0.83	2.64
bior2.2	0.88	3.51	0.86	**4.16**	**0.84**	4.09	0.87	2.48	0.83	2.61
dmey	**0.88**	3.47	**0.86**	4.06	0.83	4.06	**0.88**	2.47	**0.83**	2.61

Table 3: Results on different meyer coefficient weights. Among them, when the coefficient weight is 0.5, FDR achieved satisfactory results (4.57 for style, 3.25 for content).

Weight	Style Classes						Content Classes			
	10		25		51		25		200	
	NMI↑	FDR↑	NMI↑	FDR↑	NMI↑	FDR↑	NMI↑	FDR↑	NMI↑	FDR↑
Original	0.88	3.47	0.86	4.06	0.83	4.06	0.88	2.47	0.83	2.61
$\alpha = 0.20$	0.87	3.35	0.86	3.89	0.83	3.89	0.88	2.83	0.84	2.99
$\alpha = 0.35$	0.87	3.43	0.86	3.98	0.83	3.97	0.88	2.91	0.84	3.08
$\alpha = 0.50$	**0.88**	**3.97**	**0.86**	**4.60**	**0.84**	**4.57**	**0.89**	**3.06**	**0.85**	**3.25**

Table 4: Energy parameterization ablation under mean fusion. Comparison of branch-context and state-context-time parameterizations across style and content class counts.

Parameterization	Style classes						Content classes			
	10		25		51		25		200	
	NMI↑	FDR↑	NMI↑	FDR↑	NMI↑	FDR↑	NMI↑	FDR↑	NMI↑	FDR↑
Branch-context	0.86	**2.74**	**0.85**	**3.14**	**0.81**	**3.11**	**0.88**	2.14	**0.83**	2.28
State–context–time	**0.87**	2.67	**0.85**	3.06	**0.81**	3.03	**0.88**	**2.16**	**0.83**	**2.30**

consistently reduces FDR, confirming that both components contribute to better style-content separation. Tab. 2 compares different fusion interfaces for the target encoder. The dmey interface achieves the best overall balance, reaching an FDR of 4.06 for style and 2.61 for content, outperforming mean fusion (3.12 and 2.51) and other wavelet bases. This indicates that the Meyer basis is particularly suitable for decomposing and recombining style and content information in the coefficient domain. Tab. 3 studies different Meyer coefficient weights. The default setting corresponds to $\alpha = 1$ for approximation and $\alpha = 0$ for detail, while varying α adjusts the relative contribution of the two channels. The results show that $\alpha = 0.50$ yields the best FDR for both style (4.57) and content (3.25).

4.4 Qualitative comparison

We use fixed input pairs and compare decoded outputs at the same image resolution. It examines whether the content arrangement remains stable, whether the transferred appearance is visible, and whether local texture is coherent across the image. It is not used to replace the representation metrics. Fig. 4 compares fixed content–style pairs processed by SCFlow and the proposed model. The same inputs are used for both methods; only the transport and fusion interface changes. The forward rows show the final stylizations, while the reverse rows show the corresponding factor reconstructions.

4.5 Transport trajectories

To visualize the role of the energy gradient, we decode intermediate fused states at matched integration times and project the complete trajectories into a shared two-dimensional visualization basis. Fig. 5 shows the resulting paths for SCFlow and the proposed model. The proposed trajectory reflects the state-dependent energy correction while the dmey fusion preserves complementary coefficient channels before decoding. The plotted coordinates are a joint visualization projection, not the coordinates used by the model. We therefore use the trajectory figure to explain qualitative motion and branch fusion, while the tables remain the evidence for quantitative separation. Intermediate

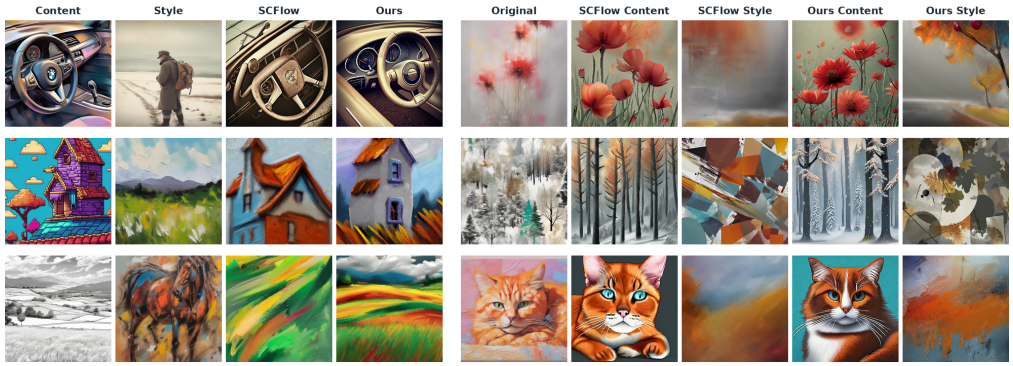


Figure 4: Qualitative comparison on matched content–style pairs. Forward stylization and reverse factor reconstruction use the same inputs for SCFlow and the proposed method.

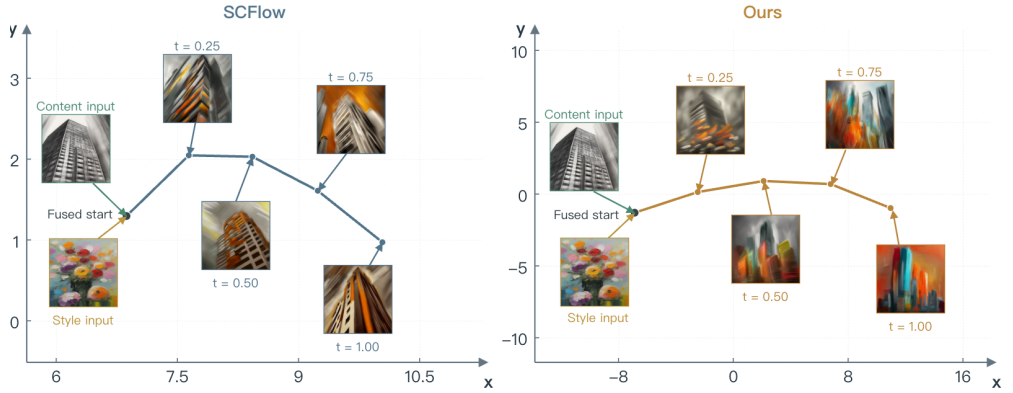


Figure 5: Decoded intermediate states along fused transport paths for the same content–style pair. The energy-guided model differs in its state-dependent field and target-feature fusion. x denotes progress along the shared direction, and y denotes the orthogonal deviation.

thumbnails are included at matched integration times to show whether a visually coherent stylization appears gradually along the route.

For clarity, let $\bar{z}_0$ and $\bar{z}_1$ be the averages of the two methods' initial and final fused states. The horizontal coordinate is the normalized common transport direction $x = (z - \bar{z}_0)^\top u$, where $u \propto \bar{z}_1 - \bar{z}_0$. The vertical coordinate is $y = (z - \bar{z}_0)^\top v$, where v is the dominant transverse direction of the combined trajectory curvature after removing its component along u. Thus, x reports progress toward the shared terminal representation, whereas y records the main deviation from that direction. These coordinates are computed after inference solely to make the high-dimensional motion readable. The shared projection is fitted after collecting both methods' states, so the axes have the same orientation for the two trajectories.

5 Conclusion and Limitation

We studied reference-based style transfer as conditional representation transport and introduced an energy-guided Flow Matching field. The scalar energy couples the joint content–style state through its input gradient, allowing the transport to score intermediate states and suppress directions that break either content or style. A discrete Meyer interface provides complementary target channels and coefficient-space fusion before decoding. Controlled ablations and matched visualizations separate the effect of energy guidance from the effect of the feature interface. The energy term consistently improves FDR while maintaining competitive NMI, and the dmey interface achieves the balance among the evaluated fusion bases. Several directions remain open. The current energy field is learned in a frozen CLIP embedding space, extending it to other visual encoders or to video and 3D content could broaden its applicability. We hope this work encourages further study of explicit energy landscapes for controllable and structure-preserving generation.

REPRODUCIBILITY STATEMENT

To ensure the full reproducibility of our work, we have submitted the core code of our experiment to the supplementary materials. After the manuscript is accepted, the complete code and weights will be uploaded to the open-source community.

AI USE STATEMENT

We used generative AI tools (large language models) to polish and edit the manuscript text.

REFERENCES

Michal Balcerak, Tamaz Amiranashvili, Antonio Terpin, Suprosanna Shit, Lea Bogensperger, Sebastian Kaltenbach, Petros Koumoutsakos, and Bjoern Menze. Energy matching: Unifying flow matching and energy-based models for generative modeling. In *Advances in Neural Information Processing Systems 38*, 2025. doi: 10.52202/085713-0291.

Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8795–8805, 2024. doi: 10.1109/CVPR52733.2024.00840.

Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11326–11336, 2022.

Yingying Deng, Xiangyu He, Fan Tang, and Weiming Dong. Z*: Zero-shot style transfer via attention reweighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6934–6944, 2024.

Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models, 2020. URL `https://arxiv.org/abs/1903.08689`.

Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy-based models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2837–2848. PMLR, 2021.

Ruiqi Feng, Chenglei Yu, Wenhao Deng, Peiyan Hu, and Tailin Wu. On the guidance of flow matching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16993–17029. PMLR, 2025.

Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora. In *Computer Vision – ECCV 2024*, pp. 181–198, 2024. doi: 10.1007/978-3-031-72684-2_11.

Zheng Gao, Jifei Song, Zhensong Zhang, Jiankang Deng, and Ioannis Patras. Frequency-guided diffusion for training-free text-driven image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19195–19205, 2025.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016. doi: 10.1109/CVPR.2016.265.

Boyu He, Yunfan Ye, Chang Liu, Weishang Wu, Fang Liu, and Zhiping Cai. Stylegallery: Training-free and semantic-aware personalized style transfer from arbitrary image references. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29092–29102, 2026.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. doi: 10.1162/089976602760128018.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1501–1510, 2017. doi: 10.1109/ICCV.2017.167.

Mingkun Lei, Xue Song, Beier Zhu, Hao Wang, and Chi Zhang. Stylestudio: Text-driven style transfer with selective control of style elements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23443–23452, 2025. doi: 10.1109/CVPR52734.2025.02183.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Matthew Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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*, 2023. URL https://arxiv.org/abs/2209.03003.

Pingchuan Ma, Xiaopei Yang, Yusong Li, Ming Gui, Felix Krause, Johannes Schusterbauer, and Bj ”orn Ommer. Scflow: Implicitly learning style and content disentanglement with flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14919–14929, 2025.

Xuan Vinh Nguyen, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.*, 11:2837–2854, 2010. doi: 10.5555/1756006.1953024. URL https://dl.acm.org/doi/10.5555/1756006.1953024.

Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8693–8702, 2024. doi: 10.1109/CVPR52733.2024.00830.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Marie Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. doi: 10.48550/ARXIV.2204.06125. URL https://doi.org/10.48550/arXiv.2204.06125.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *CoRR*, abs/2404.01292, 2024. doi: 10.48550/ARXIV.2404.01292. URL https://doi.org/10.48550/arXiv.2404.01292.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1921–1930, 2023.

Xinbo Wang, Wenju Xu, Qing Zhang, and Wei-Shi Zheng. Domain generalizable portrait style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15802–15811, 2025.

Peng Xing, Haofan Wang, Yanpeng Sun, wangqixun, Baixu, Hao Ai, Jen-Yuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. In *Advances in Neural Information Processing Systems 38*, 2025. doi: 10.52202/085713-3363.

Ruojun Xu, Weijie Xi, XiaoDi Wang, Yongbo Mao, and Zach Cheng. Stylessp: Sampling startpoint enhancement for training-free diffusion-based method for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18260–18269, 2025.

Xuwang Yin, Claire Zhang, Julie Steele, Nir Shavit, and Tony T. Wang. Scalable energy-based models via adversarial training: Unifying discrimination and generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6fcb1afcc1e9c2c82c8ddddf03bcf0f6-Abstract-Conference.html.

Xi Yu, Xiang Gu, Zhihao Shi, and Jian Sun. Wasserstein style distribution analysis and transform for stylized image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17496–17505, 2025. doi: 10.1109/ICCV51701.2025.01625.

Shiyuan Zhang, Weitong Zhang, and Quanquan Gu. Energy-weighted flow matching for offline reinforcement learning. In *International Conference on Learning Representations*, 2025.

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10146–10156, 2023. doi: 10.1109/CVPR52729.2023.00978.

Sizhe Zheng, Pan Gao, Peng Zhou, and Jie Qin. Puff-net: Efficient style transfer with pure content and style feature fusion network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 8059–8068. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00770. URL https://doi.org/10.1109/CVPR52733.2024.00770.

Yang Zhou, Xu Gao, Zichong Chen, and Hui Huang. Attention distillation: A unified approach to visual characteristics transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18270–18280, 2025.

A Interpreting the ablations

The fusion-interface results show that the choice of target representation affects FDR more strongly than NMI. This is expected: NMI primarily measures whether factor assignments remain recoverable, whereas FDR also reflects the compactness and margin of the resulting clusters. The dmey interface retains the same content and style labels while separating approximation and detail coefficients, so its gain is interpreted as a change in how target information is presented to the transport model rather than as an additional supervision signal.

The coefficient-weighting study gives a more specific view of this effect. Increasing the detail contribution changes the balance between semantic structure and localized texture. The best row is therefore not treated as a universal wavelet rule; it is the setting selected for the final model under the fixed data and decoder protocol. The important observation is that the interface can be tuned without changing the energy architecture, which makes the two sources of variation experimentally separable.

The energy parameterization table complements these observations. The branch-context and state–context–time variants have similar NMI values, but differ in FDR. This is consistent with the role of x_t and t: the state supplies the current location along the path, while time tells the energy head whether the model is resolving a global displacement or a local refinement. The final variant is chosen for its path-aware formulation, not because it dominates every individual column.

B Reproducibility and scope

The comparison is designed to isolate the transport mechanism. We retain the same CLIP feature dimensionality, the same unCLIP decoder, and the same input ordering for all rows. The baseline and proposed outputs are generated from identical content–style pairs and the same solver schedule. The only changes in the ablation tables are the wavelet interface, the coefficient weighting, or the information supplied to the energy head. This separation matters because changing the image decoder or the sampling schedule could improve visual appearance independently of the energy field.

The representation metrics should also be read with their scope in mind. NMI and FDR measure factor organization in the learned feature space; they do not measure pixel-level fidelity or aesthetic

preference. We therefore avoid treating a single metric as a complete quality criterion. The qualitative panel tests whether the factor organization is reflected in decoded images, and the trajectory panel tests whether the effect of the energy gradient is visible along integration rather than only at the endpoint. Together, these evaluations distinguish a representation-level improvement from a change that is visible only after decoding.

There are three limitations. First, the endpoint construction relies on paired content/style supervision, so the method is not a fully unpaired style-transfer system. Second, differentiating the scalar energy adds training cost and requires the energy scale to be calibrated with the Flow Matching term. Third, the study fixes the CLIP/unCLIP pair and the wavelet interface during the main comparison; a different encoder or decoder could change the relative behavior of the transport and energy terms. These limitations define the intended scope of the experiments and motivate future work on adaptive energy schedules, unpaired endpoint construction, and decoder-independent evaluation.