MITIGATING EXPOSURE BIAS IN FLOW MATCHING VIA REVERSE ENDPOINT-ANCHORED SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow matching (FM) has emerged as a powerful generative modeling framework with a simple simulation-free training objective. However, its training and inference procedures exhibit an inherent mismatch: training states are analytically constructed along prescribed interpolation paths, whereas inference recursively evolves through states induced by the model's own learned dynamics. This discrepancy leads to exposure bias, limiting generative performance. Although incorporating inference states into training is straightforward, obtaining reliable supervision matched to these states remains challenging, as their exact ground truth endpoints are generally unavailable in advance under nonlinear model dynamics. To address this challenge, we propose *Reverse Endpoint-Anchored Supervision* (RES), a post-training framework that generates the proxy of inference states through the reverse dynamics of an online model and aligns them to real inference states via forward dynamics bridging. Their known trajectory origins then serve as real endpoint targets, thereby allowing the online model to train directly on inference states with their corresponding reliable endpoint supervision. Experiments on representative FM models in both latent and pixel spaces, including SiT and JiT, demonstrate consistent improvements in generation quality, validating the effectiveness and generality of our approach.

1 INTRODUCTION

Flow matching (FM) learns a continuous transport from a simple source distribution to a target data distribution by regressing a time-dependent velocity field along prescribed probability paths (Albergo & Vanden-Eijnden, 2023; Esser et al., 2024; Liu et al., 2024; Albergo et al., 2025). A key advantage of FM is its simulation-free training objective. Under the commonly adopted linear probability path, paired source and target samples analytically specify both the intermediate state and the corresponding target velocity, eliminating the need to integrate the learned dynamics during training (Pooladian et al., 2023; Tong et al., 2024). This formulation reduces generative transport to a straightforward supervised regression problem, while retaining strong generative performance competitive with other generative modeling paradigms (Ma et al., 2024; Ho et al., 2020; Goodfellow et al., 2014; Song et al., 2021).

However, the discrepancy between simulation-free training and iterative inference introduces a fundamental limitation: *exposure bias* (Huang et al., 2026; Ning et al., 2024). During training, FM is optimized on intermediate states analytically constructed along predefined linear interpolation paths. In contrast, during inference, the model recursively evolves through states induced by its own learned velocity field (Lu et al., 2022). Prediction and discretization errors can gradually drive these states away from the interpolation paths observed during training (Karras et al., 2022; Lee et al., 2024). The model must then predict velocities from increasingly out-of-distribution states, causing deviations to compound throughout the trajectory (Ning et al., 2023; Yan et al., 2024). Consequently, exposure bias not merely as errors at individual integration steps, but as a progressively growing mismatch between the input state distributions encountered during training and inference, which can ultimately degrade generation quality. The key difficulty in mitigating this mismatch is not introducing model-induced states during training, which can be readily obtained by model inference, but providing them with reliable supervision. In standard FM, the velocity target is analytically determined by a randomly matched source-target pair (Lipman et al., 2023; Liu et al., 2023). Yet, intermediate states generated during inference generally deviate from these prescribed paths. There-

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

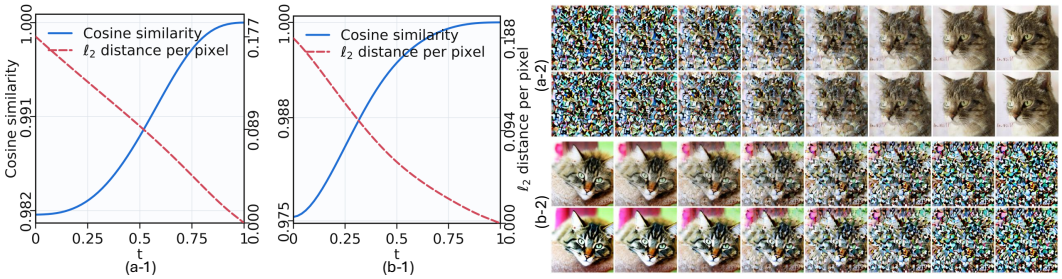

Figure 1: **Quantitative and qualitative evaluation of bidirectional trajectory states.** (a-1) Cosine similarity and pixelwise ℓ_2 distance between intermediate states at complementary time along the *noise-to-image-to-noise* roundtrip trajectory. (a-2) Visualization of the corresponding roundtrip trajectory. The upper row shows N2I inference trajectory (left to right), while the lower row shows I2N inference trajectory (right to left). (b-1, b-2) Analogous quantitative evaluation and trajectory visualization for the reverse *image-to-noise-to-image* roundtrip. Quantitative results are computed over 50K images using SiT-XL/2.

fore, the velocity target defined by the original source-target pair is no longer valid for an inference state, as it describes transport along the prescribed straight-line path rather than from the actual state encountered by the model. In summary, an effective post-training method must satisfy two coupled requirements: *it should expose the model to states induced by its own dynamics while simultaneously assigning those states reliable supervision anchored in real target samples*.

To address this challenge, we introduce Reverse Endpoint-Anchored Supervision (RES), a post-training framework for mitigating exposure bias in FM. RES is motivated by an empirical observation on bidirectional transport. As illustrated in Fig. 1, we consider a noise-to-image (N2I) model that transports Gaussian noise toward the image distribution and an image-to-noise (I2N) model that evolves in the opposite direction. Let $t \in [0, 1]$ denote the normalized integration time and x_t the corresponding intermediate state. Despite their opposite transport directions, we observe a strong correspondence between the two dynamics: states produced by one model at time t closely match those produced by the reverse model at the complementary time $1 - t$. RES builds on this complementary-state correspondence to expose N2I to inference-like states with reliable supervision. Starting from a real image, we partially integrate the frozen I2N dynamics and use the resulting intermediate state as a proxy for the corresponding N2I inference state. We then apply a one-step forward dynamics bridging (FDB) under N2I to further align the proxy with its native inference dynamics. Importantly, the entire construction follows model-induced dynamics without injecting independent noise. Therefore, the originating real image retains a trajectory-consistent correspondence with the bridged state and provides an ideal ground truth (GT) endpoint for supervision.

To exploit this endpoint correspondence without distorting the pretrained transport geometry, we further introduce endpoint-anchored velocity (EAV) supervision. Instead of prescribing a fixed local velocity target, EAV optimizes the online N2I update through the endpoint induced by a short model rollout. This formulation allows the local velocity to adapt to the learned transport geometry while explicitly constraining the resulting trajectory toward the real endpoint. In addition, we retain the original FM objective as a regularizer to preserve the pretrained dynamics. We also confine EAV refinement to the later stage of the N2I trajectory, where exposure bias is more pronounced and the direct influence of inference-time classifier-free guidance (CFG) is empirically weak. This design avoids introducing CFG into the training procedure or modifying the pretrained architecture.

We evaluate RES by post-training two representative FM baselines: SiT (Ma et al., 2024) in latent space and JiT (Li & He, 2026) in pixel space, spanning multiple model scales. Across both architectures and scales, RES consistently improves multi-step generation quality. These results demonstrate that exposing the model to inference-time states while providing reliable endpoint supervision effectively mitigates exposure bias in FM.

2 Related Work

Training Time Simulation of Inference States. A line of work seeks to mitigate exposure bias by exposing the model to perturbed or generated intermediate states during training. Input Pertur-

bation (Ning et al., 2023) injects additional noise into interpolated states during training to mimic accumulated prediction errors. Imagine Flash (Kohler et al., 2024) calibrates the student on states generated by its own inference, while its Shifted Reconstruction Loss adapts teacher supervision across different time steps. DMD2 (Yin et al., 2024) similarly unrolls a few-step student generator during training and uses student generated intermediate states as inputs. DDIL (Garrepalli et al., 2024) combines data-noise interpolants with teacher and student inference trajectories. The interpolants maintain diversity and coverage, teacher trajectories provide stable supervision, and student trajectories allow the student to correct accumulated trajectory errors using teacher feedback. ARD (Kim et al., 2025) further conditions a diffusion transformer on its predicted ODE history to reduce error accumulation in few step generation. These methods typically use teacher predictions as surrogate targets, which can propagate teacher bias and limit the reliability of supervision on model-induced states. Moreover, they are largely formulated for few-step distillation and do not directly improve conventional multi-step sampling. As multi-step generation quality often sets a practical performance ceiling for subsequent distillation and other downstream tasks, mitigating exposure bias along multi-step trajectories is particularly consequential.

Inference Time State Calibration. A complementary line of work retains the trained model and corrects the generation trajectory during sampling. Time Shift Sampler (Li et al., 2024) selects a nearby effective time index whose noise level better matches the current generated state. Epsilon Scaling (Ning et al., 2024) rescales the predicted noise so that the inference trajectory remains closer to the vector field represented during training. Frequency Regulation (Yu & Zhan, 2025) corrects spectral discrepancies between analytically noised training states and generated states by independently regulating low and high frequency components. Reflected Diffusion Models (Lou & Ermon, 2023) replace heuristic projection with a reflected stochastic process that enforces consistent support constraints during training and inference. These methods can be applied to pretrained models without additional optimization, but they correct deviations only after they arise and do not improve the learned velocity field on model induced states.

3 Reverse Endpoint-Anchored Supervision

Our goal is to develop a post-training framework that mitigates exposure bias between training and inference, thereby improving generation quality. Unlike approaches that rely on additional teacher (or reward) models whose supervision may inherit systematic bias Clark et al. (2024); Black et al. (2024); Kirstain et al. (2023); Xu et al. (2023), our framework derives supervision directly from real data. This is enabled by the complementary-time correspondence between bidirectional dynamics, which associates model-induced inference states with exact endpoints from the data distribution.

In this section, we first review FM and characterize the exposure bias arising from the discrepancy between training and inference trajectories (Sec. 3.1). We then introduce a reverse I2N model and establish the correspondence between N2I and I2N intermediate states at complementary timesteps (Sec. 3.2). Building on this correspondence, we introduce RES, which combines FDB, EAV supervision, and FM regularization for post-training (Sec. 3.3). Finally, we describe the training configuration and temporal strategy used in practice (Sec. 3.4). An overview of RES is shown in Fig. 2.

3.1 Preliminary

Flow Matching. FM formulates generative modeling as an Ordinary Differential Equations (ODE)-based transport problem, learning a continuous mapping from the source distribution π_{sou} to the target distribution π_{tar}. For empirical observations $x_0 \sim \pi_{sou}$ and $x_1 \sim \pi_{tar}$ on time $t \in [0, 1]$, FM is defined as:

$$dx_t = v_\theta(x_t, t)dt, \tag{1}$$

during training, the intermediate state x_t is the linear interpolation between x_0 and x_1, given by:

$$x_t = tx_1 + (1 - t)x_0, \tag{2}$$

$v_\theta : \mathbb{R}^d \times [0, 1] \to \mathbb{R}^d$ denotes the neural velocity field defined on the data-time domain. Taking the derivative of x_t with respect to t yields the supervised velocity target $v_{tar} = x_1 - x_0$. Therefore, during training, v_θ is optimized to fit v_{tar}:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\left[d(v_\theta(x_t, t), v_{tar})\right], \tag{3}$$

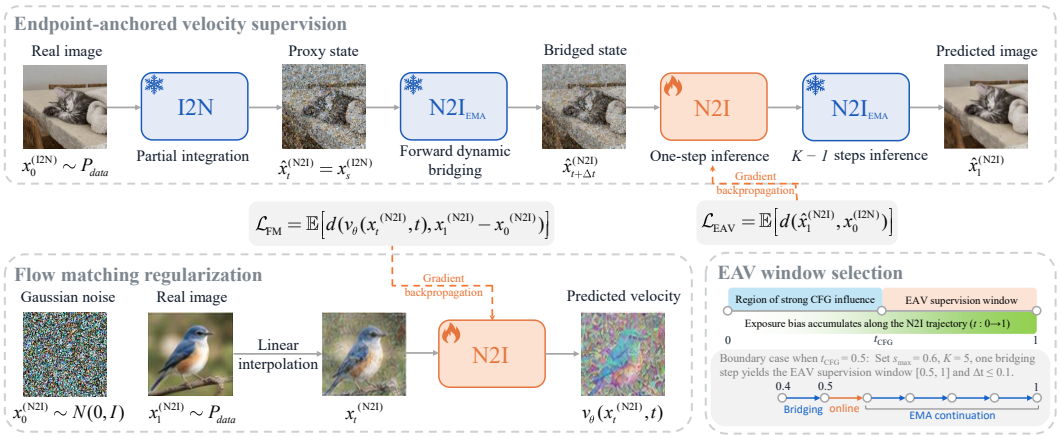

Figure 2: **Overview of RES.** RES post-trains N2I using EAV supervision together with FM regularization. A frozen I2N model first generates a proxy state from a real image, which is further aligned with the N2I inference dynamics through a one-step FDB using the EMA of N2I. The online N2I model then performs one trainable update, followed by $K-1$ gradient-free inference steps with the frozen N2I EMA model to obtain the predicted endpoint, which is supervised by the originating real image. FM regularization is retained to preserve the pretrained transport dynamics, while EAV supervision is restricted to the later trajectory region where exposure bias is more pronounced and CFG influence is reduced.

where $d(\cdot,\cdot)$ is a metric function that satisfies $\forall a, b : d(a,b) \geq 0$ and $d(a,b) = 0$ if and only if $a = b$. During the inference process, the ODE usually needs to be discretized and solved via the Euler method (Hairer et al., 1993), expressed as:

$$x_{t+\Delta t} = x_t + \Delta t \cdot v_\theta(x_t, t), \tag{4}$$

where $\Delta t = \frac{1}{N}$, N represents the total number of sampling steps, and $t = k\Delta t,\ k = 0, 1, \ldots, N$.

Exposure Bias. Due to discretized numerical integration at inference time, sampling trajectories progressively accumulate discretization errors, which compound with the model's intrinsic prediction errors. As t increases, the resulting trajectories increasingly deviate from the straight interpolation paths observed during training, giving rise to exposure bias. Consequently, the model is required to predict velocities at out-of-distribution input states during inference. With CFG, this distribution shift becomes increasingly pronounced at larger guidance scales, ultimately degrading generation quality.

3.2 Bidirectional Dynamics Construction

To mitigate exposure bias, we seek to expose the model to realistic inference-time states during training while providing reliable GT supervision for such states. In generative tasks with an explicit target distribution, real samples from the target distribution naturally serve as GT targets. As a result, the key challenge is to obtain inference-like intermediate states during training while preserving a valid correspondence with real image targets. We identify a useful trajectory duality between a standard N2I FM model v_θ^{N2I} and an additionally pretrained reverse I2N model v_ϕ^{I2N}. Despite evolving in opposite directions, the two dynamics produce highly consistent intermediate states at complementary timesteps. As shown in Fig. 1, states generated through both "*noise to image to noise*" and "*image to noise to image*" roundtrip trajectories exhibit consistently high cosine similarity and low pixelwise ℓ_2 distance to their counterparts. The relatively larger discrepancy in ℓ_2 distance reflects variations in unavoidable low-level characteristics (for example, the generated image will have higher contrast and color saturation than the real image (Jia et al., 2025; Saharia et al., 2022; Shum et al., 2025; Chu et al., 2025)), whereas the high cosine similarity indicates that the underlying state structures of the two are largely consistent. The trajectory visualizations further show close correspondence between intermediate states generated by the two opposite models.

This observation provides a direct mechanism for constructing inference-like training states: I2N can generate intermediate states at complementary timesteps that closely match those encountered by N2I during inference. Crucially, since I2N trajectories originate from real samples drawn from

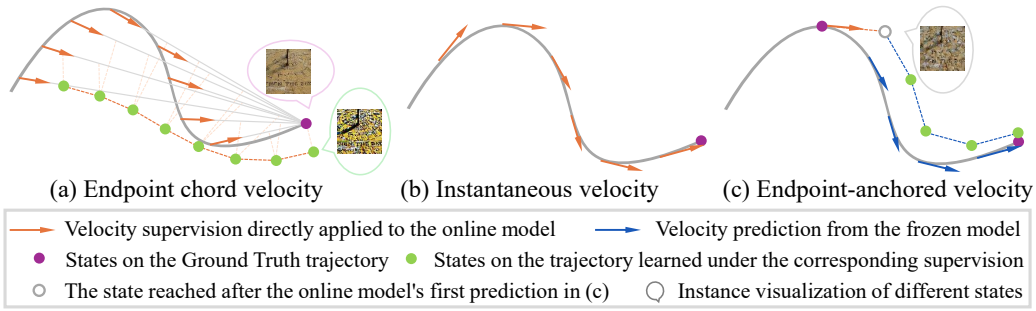

Figure 3: **Different forms of velocity supervision.** (a) *Endpoint chord velocity* directly points from an intermediate state x_t to the endpoint x_1. (b) *Instantaneous velocity* denotes the local velocity of the trajectory at x_t, which is generally unavailable in closed form and is estimated by the model. (c) *Endpoint-anchored velocity* is optimized through a K-step continuation to the endpoint. The online model predicts the first step, while the frozen EMA model completes the remaining steps. Supervision is defined by the discrepancy between the resulting endpoint and the GT endpoint.

the target distribution, these samples provide GT targets for the corresponding trajectories. This allows us to expose N2I to inference-like states during training while retaining supervision anchored in real data, without relying on model-generated pseudo targets.

N2I and I2N share the same architecture, training objective, optimization scheme, and hyperparameters, differing only in the exchange of source and target distributions. For optimal reverse dynamics fidelity, we recommend training a dedicated I2N model. When computational resources are limited, I2N inference may instead be approximated by N2I inversion (Mokady et al., 2023), albeit with reduced performance. Specifically, standard inversion reuses the forward velocity field for backward integration, whereas the forward and reverse velocities at corresponding states need not coincide under nonlinear learned dynamics and discrete integration. The resulting velocity mismatch accumulates across solver steps, progressively altering the course of the reverse trajectory. As shown in Fig. 4(a), states near the end of the inversion trajectory retain substantial image information instead of approaching the noise distribution, and this effect is more pronounced in pixel space.

3.3 Endpoint-Anchored Bidirectional Training

RES refines N2I using inference-aligned states with GT endpoint supervision. First, a frozen I2N model generates complementary states from real images as proxies for N2I inference states. Next, a one-step N2I FDB further aligns these proxies with the native N2I sampling dynamics and preserves their correspondence with the originating real images. We then introduce EAV supervision, which optimizes the local N2I update through the endpoint induced by a short rollout rather than a predefined velocity target. Finally, the original FM objective is retained as a regularizer to preserve the pretrained transport dynamics during post-training. The detailed procedure is described below.

Proxy State Construction. Since our framework involves two opposing dynamics, We use $x_t^{(\cdot)}$ to denote the state at time t along the trajectory specified by $(\cdot) \in \{\mathrm{N2I}, \mathrm{I2N}\}$, i.e., $x_0^{(\mathrm{N2I})}$, $x_1^{(\mathrm{I2N})} \sim \pi_{\mathrm{sou}}$ and $x_1^{(\mathrm{N2I})}$, $x_0^{(\mathrm{I2N})} \sim \pi_{\mathrm{tar}}$. As shown in Fig. 2, starting from a real image $x_0^{(\mathrm{I2N})}$, we integrate the frozen I2N model from time 0 to s:

$$x_s^{(\mathrm{I2N})} = \Phi_{v_\phi^{\mathrm{I2N}}}^{0\to s}\left(x_0^{(\mathrm{I2N})}\right). \tag{5}$$

Based on the observed complementary trajectory correspondence, $x_s^{(\mathrm{I2N})}$ serves as a proxy for the N2I state at $t = 1 - s$.

Forward Dynamics Bridging. Although $x_s^{(\mathrm{I2N})}$ provides a strong proxy for the corresponding N2I state, a residual mismatch remains between the two dynamics. We further align the proxy with N2I inference through a one-step FDB. Specifically, we define

$$\hat{x}_t^{(\mathrm{N2I})} = x_s^{(\mathrm{I2N})}, \qquad t = 1 - s, \tag{6}$$

and advance it by one solver step Δt using the frozen exponential moving average (EMA) copy of N2I:

$$\hat{x}_{t+\Delta t}^{(\mathrm{N2I})} = \Phi_{v_{\theta_{\mathrm{ema}}}^{\mathrm{N2I}}}^{t \to t+\Delta t}\left(\hat{x}_t^{(\mathrm{N2I})}\right). \tag{7}$$

The FDB evolves the proxy state under the velocity field used during N2I inference, producing a state more representative of those encountered along actual sampling trajectories.

A key advantage of this construction is the preservation of sample-specific endpoint correspondence. Under standard regularity conditions, an ODE with a fixed velocity field admits a unique solution for each state-time pair (Teschl, 2012). Consequently, each intermediate state lies on a unique integral curve and inherits a well-defined endpoint correspondence under the induced flow. I2N starts from a real image and evolves exclusively along its native image-to-noise dynamics. Therefore, its intermediate states remain deterministically coupled to the originating image through the I2N flow (Liu et al., 2023). The subsequent FDB introduces no independent noise and advances the image-coupled proxy solely under the native N2I velocity field, bringing its evolution into closer alignment with N2I inference dynamics.

This property fundamentally distinguishes our construction from the random interpolation. Independently sampled noise and a real image define only a geometric pairing that is generally not induced by the learned N2I flow (Pooladian et al., 2023). Therefore, the interpolated state is not guaranteed to lie on the unique N2I trajectory associated with the designated real image. Its flow-induced endpoint may instead differ from the assigned GT target. This discrepancy introduces an endpoint-coupling mismatch between the constructed state and its supervision. In contrast, RES constructs intermediate states exclusively through the native dynamics of I2N and N2I. This improves alignment with N2I inference while largely preserving a flow-consistent correspondence with the real image endpoint.

Endpoint-Anchored Velocity. The remaining challenge is to supervise N2I on model-induced intermediate states. Since trajectories induced by learned neural velocity fields are generally curved, the chord direction from an intermediate state to the GT endpoint does not necessarily match the local tangent of the underlying trajectory. Therefore, directly regressing this endpoint chord biases the local velocity toward the target at every training step. Repeated redirection can induce systematic over-correction and pixel-value overshooting. As a result, generated samples may exhibit excessive saturation, contrast, and sharpening, as shown in Fig. 3(a). Alternatively, the negated instantaneous velocity predicted by the frozen I2N model can serve as the supervision target, as illustrated in Fig. 3(b). This target better preserves the local transport geometry. However, it does not explicitly constrain the endpoint obtained after subsequent N2I integration. Consequently, local velocity errors can accumulate along the trajectory and progressively shift the terminal state away from the GT endpoint. In severe cases, this endpoint drift destabilizes training and leads to collapse.

Instead, we propose endpoint-anchored velocity (EAV) supervision. EAV does not prescribe an explicit local velocity target, it supervises the current online update through the endpoint predicted by a few inference steps of the N2I itself. Given the bridged state $\hat{x}_t^{(\mathrm{N2I})}$, we roll out the N2I dynamics to $t = 1$ using K uniform Euler steps following Eq. (4), with $\Delta t = \frac{1-t}{K}$. The first step is taken by the online N2I model $v_\theta^{(\mathrm{N2I})}$, while the remaining $K-1$ steps are completed by the frozen EMA model $v_{\theta_{\mathrm{ema}}}^{(\mathrm{N2I})}$. We stop gradients through the EMA velocity predictions in the continuation steps, treating them as constant increments when forming the final image prediction. The EAV loss thus backpropagates only through the online N2I update. The GT endpoint target corresponding to $\hat{x}_1^{(\mathrm{N2I})}$ is the real image $x_1^{(\mathrm{N2I})\star} = x_0^{(\mathrm{I2N})}$ which initializes the corresponding I2N trajectory, so we then optimize

$$\mathcal{L}_{\mathrm{EAV}} = \mathbb{E}\left[d\left(\hat{x}_1^{(\mathrm{N2I})}, x_1^{(\mathrm{N2I})\star}\right)\right]. \tag{8}$$

EAV supervises the terminal prediction induced by the online update without prescribing a local velocity target. Therefore, the online model retains the flexibility to adapt its local velocity to the learned transport geometry. Meanwhile, the GT endpoint provides an explicit global constraint on the resulting trajectory. As a result, EAV mitigates endpoint drift without repeatedly forcing the local velocity toward the chord direction. This design provides a more stable and effective supervision signal, as illustrated in Fig. 3(c).

Flow Matching Regularization. EAV supervision exposes the model to inference-induced states

with GT endpoint targets. However, optimizing exclusively on these states may gradually distort the transport dynamics inherited from the pretrained FM model. To preserve the original transport structure during post-training, we additionally retain the standard FM objective Eq. (3) as a regularizer. The total loss is

$$\mathcal{L}_{\text{RES}} = \mathcal{L}_{\text{EAV}} + \mathcal{L}_{\text{FM}}. \tag{9}$$

3.4 Implementation Details

We jointly select the I2N integration horizon and the EAV rollout length based on the model type and its sensitivity to CFG. EAV supervision focuses on later N2I states, where exposure bias has accumulated and the direct influence of CFG is weaker. Since the interval over which CFG substantially affects the final image varies across backbones (Fig. 4(b)), we empirically identify a backbone-specific cutoff t_{CFG} beyond which this effect becomes negligible, and restrict EAV online updates to $[t_{\text{CFG}}, 1]$. We partially integrate the I2N dynamics up to time s to produce an intermediate state, which is assigned as a proxy at the complementary N2I time $t = 1 - s$. Starting from this state, the rollout of FDB and EAV forms $K + 1$ consecutive Euler steps with uniform step size $\Delta t = \frac{s}{K+1}$. The frozen EMA model performs the first step, the online N2I model performs the second, and the frozen EMA model completes the remaining $K - 1$ steps. For velocity-prediction models, we require $\Delta t \leq 0.1$, as larger steps can distort the local transport geometry and produce visible artifacts. As shown in Fig. 2, we constrains the joint selection of K and $s_{\max}$ together with the CFG cutoff. Specifically, sampling $s \sim \mathcal{U}(0, s_{\max})$ requires

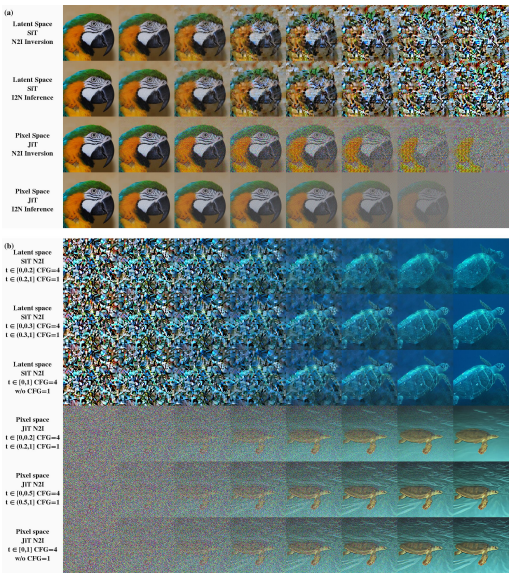

Figure 4: **Visualization of different types of trajectories.** (a) N2I inversion and I2N inference trajectories for the same image in the latent and pixel space, using 50 Euler steps with CFG = 1. (b) N2I generation trajectories using 50 Euler steps. All rows share the same initial noise, while CFG = 4 is applied during early trajectory intervals of different lengths.

$$s_{\max} \leq \min\left\{1,\ 0.1(K+1),\ \frac{K+1}{K}(1 - t_{\text{CFG}})\right\}. \tag{10}$$

The second bound ensures $\Delta t \leq 0.1$, while the third places every online update at $t + \Delta t \geq t_{\text{CFG}}$. For example, when $t_{\text{CFG}} = 0.5$, choosing $K = 5$ and $s_{\max} = 0.6$ yields $\Delta t \leq 0.1$ and $t + \Delta t \in [0.5, 1]$. For image-prediction models, we set $K = 1$ and fix the FDB step to 0.1. The online model then predicts the terminal image from the bridged state. Accordingly, we sample $s \geq 0.1$ and choose $s_{\max}$ such that the bridged time $t + 0.1$ remains within $[t_{\text{CFG}}, 1]$.

4 Experiments

Experimental Setup. We evaluate RES on class-conditional ImageNet at 256×256 resolution using SiT-B/2 and SiT-XL/2 in latent space, and JiT-B/16 and JiT-H/16 in pixel space. The N2I models are initialized from pretrained checkpoints, and the corresponding I2N models are trained for 300K steps. For RES post-training, we set $(K, s_{\max}) = (7, 0.8)$ for SiT and $(1, 0.6)$ for JiT, without modifying the model architectures. Further experimental details are provided in the Appendix A.1.

Evaluation Metrics. We evaluate generation quality using Fréchet Inception Distance (FID) (Heusel et al., 2017), Inception Score (IS) (Salimans et al., 2016), Precision, and Recall on 50K generated samples. FID measures the distributional discrepancy between generated and real images in the Inception feature space, with lower values indicating better distributional alignment. IS evaluates both the quality and diversity of generated samples, where higher values are preferred. Precision and Recall provide complementary measures of fidelity and diversity: Precision quantifies the proportion of generated samples that lie within the support of the real data distribution, while Recall measures how well the generated distribution covers the real data distribution. Unless otherwise specified, we follow the sampling and evaluation protocols of the corresponding baselines.



Figure 5: **Qualitative comparison of baselines and RES on ImageNet** 256×256**.** Left: SiT-XL/2 with Dopri5 (adaptive NFE, CFG $= 1.5$). Right: JiT-H/16 with Heun (NFE $= 50$, CFG $= 2.2$).

Table 1: Comparison of Baseline and RES under different configurations on ImageNet 256×256.

Model	Method	Sampler	CFG	NFE	IS ↑	FID ↓	Prec. ↑	Rec. ↑
SiT-B/2	Baseline	Dopri5	1.5	Adaptive	75.01	19.28	0.64	0.55
		Euler	1.0	50	36.18	42.18	0.49	0.62
		Euler	1.5	10	64.70	30.19	0.59	0.48
		Euler	1.5	50	76.43	20.07	0.65	0.55
	RES	Dopri5	1.5	Adaptive	93.10	14.63	0.67	0.57
		Euler	1.0	50	43.24	35.99	0.52	0.63
		Euler	1.5	10	72.66	27.20	0.60	0.50
		Euler	1.5	50	92.72	15.45	0.67	0.55
SiT-XL/2	Baseline	Dopri5	1.5	Adaptive	258.07	2.15	0.80	0.60
		Euler	1.0	50	118.65	11.18	0.66	0.67
		Euler	1.5	10	181.23	8.62	0.73	0.51
		Euler	1.5	50	247.88	2.40	0.80	0.59
	RES	Dopri5	1.5	Adaptive	268.51	1.91	0.80	0.62
		Euler	1.0	50	137.65	8.24	0.68	0.67
		Euler	1.5	10	200.24	6.26	0.74	0.55
		Euler	1.5	50	262.04	1.98	0.80	0.60
JiT-B/16	Baseline	Heun	3.0	50	267.97	3.64	0.74	0.50
		Euler	3.0	10	194.46	10.75	0.62	0.38
		Euler	3.0	50	278.35	4.13	0.76	0.47
	RES	Heun	3.0	50	260.91	3.57	0.73	0.54
		Euler	3.0	10	184.40	10.55	0.62	0.42
		Euler	3.0	50	268.66	4.07	0.76	0.50
JiT-H/16	Baseline	Heun	2.2	50	294.06	1.91	0.71	0.63
		Euler	2.2	10	195.38	9.41	0.56	0.50
		Euler	2.2	50	296.56	2.13	0.72	0.61
	RES	Heun	2.2	50	291.68	1.85	0.78	0.63
		Euler	2.2	10	192.75	9.21	0.66	0.52
		Euler	2.2	50	292.88	2.02	0.72	0.62

4.1 Main Results

As shown in Tab. 1, RES consistently improves generation quality across the evaluated models and sampling configurations, achieving lower FID in every matched comparison. For SiT, these improvements are accompanied by higher IS and generally maintained or improved Precision and Recall. The gains extend across adaptive and fixed-step solvers, including low-NFE sampling. Within the SiT family, SiT-B/2 exhibits larger gains than SiT-XL/2. Under Dopri5 (Dormand & Prince, 1980), FID decreases from 19.28 to 14.63 for SiT-B/2 and from 2.15 to 1.91 for SiT-XL/2. This difference may reflect larger velocity estimation errors in the smaller model, which accumulate during sampling and leave greater scope for RES to mitigate inference state mismatch. The stronger pretrained SiT-XL/2 baseline offers less room for improvement but still benefits from post-training.

The improvements on JiT are smaller and more metric-dependent. FID decreases and Recall increases across the evaluated settings, whereas IS does not improve uniformly and Precision varies. Thus, the JiT results point primarily to improved distributional fit and coverage rather than uniform gains across all metrics. A possible explanation is the difference in prediction target: JiT constructs

its sampling velocity from the current state and a predicted clean image, keeping each update referenced to an endpoint estimate. SiT predicts velocity directly, leaving the endpoint to emerge from successive integration steps. This distinction may affect how local errors accumulate and then vary the benefit of endpoint supervision. A detailed analysis of the differences between these model families is beyond the scope of this work. The qualitative comparisons in Fig. 5 complement the quantitative results. Relative to paired baseline samples, RES generally produces fewer visible artifacts and more coherent local textures in both SiT-XL/2 and JiT-H/16, yielding a more natural photographic appearance.

Table 2: RES ablations on SiT-B/2.

Variant	IS ↑	FID ↓
Baseline	75.01	19.28
w/ FDB → w/o FDB	92.88	14.77
I2N inference → random interpolation	85.43	17.10
I2N inference → N2I inversion	91.16	15.31
RES (w/ FDB and I2N inference)	93.10	14.63

4.2 Ablation Analysis

Tab. 2 examines FDB and proxy state construction. With the remaining components fixed, N2I inversion outperforms random interpolation, while I2N inference achieves the best results, highlighting the importance of proxy state construction. I2N inference without FDB also surpasses N2I inversion with FDB. Adding FDB to I2N states provides a modest further improvement, reducing FID from 14.77 to 14.63. This limited marginal gain is consistent with I2N proxy states already being closely aligned with N2I dynamics, leaving relatively little residual mismatch for FDB to correct.

Table 3: Effect of K and $s_{\max}$ on SiT-B/2.

$s_{\max}$	K	FID ↓	Time / 100 steps (min)
0.8	10	14.57	3.58
0.8	7	14.63	3.28
0.8	4	15.53	3.05
0.8	1	16.85	2.90
0.6	5	16.53	3.15
0.3	2	18.27	3.04
1.0	9	19.07	4.37

Table 3 shows a clear tradeoff between rollout length and training cost. With $s_{\max}$ fixed, increasing K substantially improves FID until the gain saturates, while training time continues to rise. When K is small, the online update spans a large time interval. Although EAV constrains the terminal image, this constraint does not uniquely determine the velocity at an intermediate state: the model may reduce endpoint error through a coarse displacement without learning a local velocity suitable for multi-step inference. A longer rollout makes the online update more local and better resolves the subsequent transport. Beyond a moderate K, additional steps primarily increase frozen EMA computation, since each rollout still contains only one trainable update.

At a matched maximum step size, increasing $s_{\max}$ from a short to an intermediate I2N horizon broadens proxy state coverage and improves performance. Further extending the horizon reverses this gain. This decline may partly arise from extending online updates to early N2I states, whose distributions are particularly sensitive to the inference-time CFG scale. Proxies constructed without CFG may not adequately cover this scale-dependent variation. I2N proxy states also approach the noise endpoint. Since the learned N2I and I2N dynamics inevitably produce terminal distributions that differ to some extent from the real-image and Gaussian distributions, respectively, proxies near the noise end may not faithfully represent states encountered during Gaussian-initialized N2I sampling. The decline at longer horizons is consistent with the combined influence of these effects.

5 Conclusion

This paper introduces RES, a post-training framework for mitigating exposure bias in FM. RES trains the model directly on inference-aligned states with corresponding GT endpoint supervision, avoiding potentially biased surrogate targets. These states are constructed using complementary I2N dynamics followed by a one-step FDB, while EAV target provides a reliable local supervision. An FM regularization objective further preserves the pretrained transport dynamics during post-training. Experiments across models of latent and pixel space and multiple parameter scales consistently demonstrate the effectiveness of RES without requiring architectural modifications.

AI USE STATEMENT

We used generative AI tools to improve the clarity, grammar, and readability of author-written text. The authors reviewed all AI-assisted edits for technical accuracy and take full responsibility for the final manuscript.

REFERENCES

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024.

Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. Fire: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12830–12839. IEEE, 2025.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, 2024.

John R Dormand and Peter J Prince. A family of embedded runge-kutta formulae. *Journal of computational and applied mathematics*, 6(1):19–26, 1980.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024.

Risheek Garrepalli, Shweta Mahajan, Munawar Hayat, and Fatih Porikli. Ddil: Diversity enhancing diffusion distillation with imitation learning. *arXiv preprint arXiv:2410.11971*, 2024.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2672–2680, 2014.

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

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Guanbo Huang, Jingjia Mao, Fanding Huang, Fengkai Liu, Xiangyang Luo, Yaoyuan Liang, Jiasheng Lu, Xiaoe Wang, Pei Liu, Ruiliu Fu, Ruqi Huang, and Shao-Lun Huang. Exposure bias can alleviate itself via directional and frequency rectification in flow matching. *arXiv preprint arXiv:2606.28226*, 2026. doi: 10.48550/arXiv.2606.28226.

Zexi Jia, Chuanwei Huang, Yeshuang Zhu, Hongyan Fei, Xiaoyue Duan, Zhiqiang Yuan, Ying Deng, Jiapei Zhang, Jinchao Zhang, and Jie Zhou. Secret lies in color: Enhancing ai-generated images detection with color distribution analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13445–13454. IEEE, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577, 2022. doi: 10.52202/068431-1926.

Yeongmin Kim, Sotiris Anagnostidis, Yuming Du, Edgar Schönfeld, Jonas Kohler, Markos Georgopoulos, Albert Pumarola, Ali Thabet, and Artsiom Sanakoyeu. Autoregressive distillation of diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15745–15756, 2025.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 36652–36663, 2023. doi: 10.52202/075280-1594.

Jonas Kohler, Albert Pumarola, Edgar Schönfeld, Artsiom Sanakoyeu, Roshan Sumbaly, Peter Vajda, and Ali Thabet. Imagine flash: Accelerating emu diffusion models with backward distillation. *arXiv preprint arXiv:2405.05224*, 2024.

Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. In *Advances in Neural Information Processing Systems*, volume 37, pp. 63082–63109, 2024. doi: 10.52202/079017-2014.

Mingxiao Li, Tingyu Qu, Ruicong Yao, Wei Sun, and Marie-Francine Moens. Alleviating exposure bias in diffusion models through sampling with shifted time steps. In *International Conference on Learning Representations*, volume 2024, pp. 16816–16838, 2024.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

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.

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.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. InstaFlow: One step is enough for high-quality diffusion-based text-to-image generation. In *International Conference on Learning Representations*, 2024.

Aaron Lou and Stefano Ermon. Reflected diffusion models. In *International Conference on Machine Learning*, pp. 22675–22701. PMLR, 2023.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5775–5787, 2022. doi: 10.52202/068431-0418.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6038–6047. IEEE, 2023.

Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input perturbation reduces exposure bias in diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26245–26265. PMLR, 2023.

Mang Ning, Mingxiao Li, Jianlin Su, Albert Ali Salah, and Itir Onal Ertugrul. Elucidating the exposure bias in diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 15167–15189, 2024.

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Ka Chun Shum, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Color alignment in diffusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28446–28455. IEEE, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Gerald Teschl. *Ordinary differential equations and dynamical systems*, volume 140. American Mathematical Soc., 2012.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15903–15935, 2023. doi: 10.52202/075280-0700.

Hanshu Yan, Xingchao Liu, Jiachun Pan, Jun Hao Liew, Qiang Liu, and Jiashi Feng. PeRFlow: Piecewise rectified flow as universal plug-and-play accelerator. In *Advances in Neural Information Processing Systems*, volume 37, pp. 78630–78652, 2024. doi: 10.52202/079017-2497.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024.

Meng Yu and Kun Zhan. Frequency regulation for exposure bias mitigation in diffusion models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10370–10378, 2025.

A Appendix

A.1 Experimental Details

Training Setup. We instantiate our method using SiT and JiT at an image resolution of 256×256. The SiT variant operates in the latent space of a pretrained SD-VAE, whereas the JiT variant is trained directly in RGB pixel space without a VAE. Both variants are trained on ImageNet-1K using AdamW with zero weight decay. For SiT, we use a global batch size of 256, a learning rate of 5×10^{-5}, and Adam coefficients $(\beta_1, \beta_2) = (0.9, 0.999)$. For JiT, we use a global batch size of 512, a learning rate of 1×10^{-5}, and $(\beta_1, \beta_2) = (0.9, 0.95)$. The EMA decay is set to 0.9999 for both variants. SiT is trained with FP16 autocasting and gradient scaling, while JiT uses BF16 autocasting without gradient scaling. We disable attention and projection dropout for JiT-H/16. We recommend 10 epochs of RES post-training for SiT-B/2 and JiT-B/16, and 2 epochs for SiT-XL/2 and JiT-H/16. We note that the pretrained baselines have already converged, and further training with their original objectives yields no additional improvement in generation quality under the evaluated settings. Detailed hyperparameter settings are provided in Tab. A1, and the complete training procedure is presented in Alg. A1.

Continuous-Time Sampling. At each iteration, the global batch is divided equally between the standard FM branch and the EAV branch. All time variables are sampled independently for each example and remain continuous throughout training. For the standard FM branch, SiT samples

$$t_{\text{int}} \sim \mathcal{U}(0, 1),$$

whereas JiT follows the Logit-Normal time parameterization

$$t_{\text{int}} = \text{sigmoid}(z), \qquad z \sim \mathcal{N}(-0.8, 0.8^2).$$

Given Gaussian noise x_0 and a real-data endpoint x_1, we construct

$$x_t = (1 - t_{\text{int}})x_0 + t_{\text{int}}x_1$$

and regress the velocity target $x_1 - x_0$ using mean-squared error.

For the EAV branch, we independently sample a continuous inverse duration

$$s \sim \mathcal{U}(0, s_{max}).$$

A frozen I2N transports each real-data endpoint to its sample-specific time s. The I2N trajectory is evaluated using a base Euler step size of $\Delta h = 0.04$. Importantly, the sampled time is not quantized to the Euler grid. For each example, we apply $\lfloor s/\Delta h \rfloor$ full Euler updates followed, when necessary, by one residual update of magnitude

$$\Delta t_{\text{end}} = s - \left\lfloor \frac{s}{\Delta h} \right\rfloor \Delta h.$$

Thus, the I2N trajectory reaches the exact continuously sampled target time.

Dropout Probability of Label. Both branches use a class-label dropout probability of 0.1. For each example, a single classifier-free dropout mask is shared across the I2N rollout and all subsequent endpoint-prediction operations. The training-time CFG scale is fixed to 1.0. The final training objective is

$$\mathcal{L}_{\text{RES}} = \mathcal{L}_{\text{EAV}} + \mathcal{L}_{\text{FM}},$$

where $\mathcal{L}_{\text{FM}}$ and $\mathcal{L}_{\text{EAV}}$ are the branch-wise mean interpolation and backward endpoint losses, respectively, and are assigned equal weights.

A.2 Interpretation of Ground-Truth Endpoint Supervision

RES is motivated by the empirically observed similarity between states of the N2I and I2N dynamics at complementary times. Starting from a real image, reverse integration and forward dynamics bridging construct an intermediate state associated with that image, which is then used as the endpoint supervision target. Throughout this paper, the term *ground-truth (GT) endpoint target* refers to this originating real image drawn from the data distribution. It identifies the source of supervision, rather than asserting that the image is the exact endpoint obtained by continuing the bridged state under the current N2I dynamics.

In practice, the N2I and I2N dynamics are represented by separately learned neural networks and evaluated using numerical solvers. Approximation and discretization errors therefore prevent exact time-reversal consistency from being guaranteed. In particular, the deterministic construction of a bridged state establishes an association with its originating image, but does not by itself imply that subsequent N2I integration will recover that image exactly. Nor do we claim that the originating image is an optimal target under every possible criterion. Nevertheless, it provides a directly available, sample-specific supervision target grounded in real data, without requiring an unknown ideal endpoint or a model-generated endpoint target.

Accordingly, RES should be understood as an empirically motivated post-training framework, rather than a method whose validity requires exact equivalence between the two learned flows. Its working hypothesis is that the observed complementary-state correspondence makes the originating real image a useful endpoint anchor for refining N2I dynamics. The effectiveness of this supervision is evaluated experimentally; exact endpoint correspondence and mathematical optimality are not assumed or established.

A.3 Discussion about CFG and Future Directions

Exposure bias is shaped by both accumulated sampling errors and inference-time modifications to the learned dynamics. Prediction and discretization errors progressively alter the states encountered during generation, while CFG introduces an additional shift through velocity extrapolation. This shift can have competing effects. For SiT, FID initially decreases with increasing guidance strength and rises again under stronger guidance. Moderate CFG can therefore improve agreement with the data distribution despite departing from the unguided dynamics. Consequently, guidance-induced deviations cannot uniformly be treated as harmful errors. Mitigating exposure bias requires preserving beneficial transport while reducing deviations that degrade generation quality.

This distinction motivates separating EAV supervision from the most CFG-sensitive portion of the trajectory. Early N2I states depend strongly on the inference-time guidance scale, whereas RES constructs proxy states through unguided I2N integration. Supervising this region would require accounting for guidance-dependent changes in both state coverage and the correspondence with real-image targets. Concentrating EAV updates in later N2I states reduces this dependence while targeting accumulated exposure bias. This choice also limits interference with the early extrapolation through which CFG improves generation. The separation concerns the direct influence of guidance: earlier CFG still shapes the states entering the refinement window. RES thus refines their subsequent evolution in a region where the opposing dynamics exhibit approximate agreement and local guidance sensitivity is weaker.

We hypothesize that moderate CFG and RES can be complementary. I2N proxy construction starts directly from real images, so its initial distribution is exactly the target data distribution. Under the weak guidance sensitivity near the image end, early I2N trajectories are expected to remain relatively stable with respect to guidance. Early N2I trajectories are substantially more sensitive, and moderate CFG yields the closest terminal agreement with real data among the evaluated scales, as measured by FID. Given the approximate agreement between the opposing dynamics near the image end, this improved terminal alignment may be accompanied by greater overlap between late N2I states and early I2N proxy states at complementary times. Moderate CFG would then steer N2I sampling toward regions represented by the proxies, while RES refines the subsequent transport using the associated real-image targets. This provides a possible mechanism for their complementary benefits.

This complementarity hypothesis raises a further question: how should guidance change as RES improves the learned dynamics? Part of the benefit of moderate CFG may arise from compensating for model inaccuracies, while another part may reflect its influence on conditional fidelity and sample diversity. If RES reduces inaccuracies that guidance previously compensated for, the preferred guidance strength and temporal support may also change. The interaction between CFG and RES may therefore depend on the model's remaining errors and their distribution along the trajectory. Studying how the incremental benefit of CFG evolves during post-training could clarify which effects remain complementary and which become less necessary as the dynamics improve.

This perspective also informs the extension of RES to earlier N2I states. Supervision in the guidance-sensitive region should preserve beneficial steering while mitigating the additional mismatch introduced by extrapolation. Achieving this requires understanding how guidance strength, temporal placement, and endpoint supervision jointly affect generation quality and distribution coverage. The current temporal separation provides a starting point by limiting their direct interaction. Future work will investigate coordinated guidance and refinement schedules that account for the evolving model dynamics, enabling earlier trajectory correction without suppressing the useful effects of CFG.

Table A1: Training hyperparameters for the SiT- and JiT-based variants.

Hyperparameter	**SiT**	**JiT**
Basic training configuration		
Backbone	SiT	JiT
Training space	VAE latent	RGB pixels
Image resolution	256^2	256^2
Global batch size	256	512
Optimizer	AdamW	AdamW
Learning rate	5×10^{-5}	1×10^{-5}
Adam (β_1, β_2)	$(0.9, 0.999)$	$(0.9, 0.95)$
Weight decay	0	0
Training precision	FP16	BF16
Gradient scaling	Yes	No
EMA decay	0.9999	0.9999
Maximum epochs	10 for B/2; 2 for XL/2	10 for B/16; 2 for H/16
Attention/projection dropout	Default	0.0/0.0
Method-specific configuration		
Branch batch split	128/128	256/256
Interpolation time for FM regularization	$t \sim \mathcal{U}(0, 1)$	$\mathrm{sigmoid}\big(\mathcal{N}(-0.8, 0.8^2)\big)$
Interpolation loss for FM regularization	Velocity MSE	Velocity MSE
Inverse duration	$\tau \sim \mathcal{U}(0, 0.8)$	$\tau \sim \mathcal{U}(0.1, 0.6)$
N2I time range	$t = 1 - \tau \in (0.2, 1)$	$t = 1 - \tau \in (0.4, 1)$
Inverse teacher	Frozen I2N EMA	Frozen I2N EMA
Inverse solver	Euler, 25-step grid	Euler, 25-step grid
I2N Euler step size	$\Delta h = 0.04$	$\Delta h = 0.04$
EMA bridge steps	1	1
Endpoint prediction	1 online + 6 EMA steps	Single online prediction
Backward loss	EAV loss	EAV loss
Label dropout	0.1	0.1
Training CFG scale	1.0	1.0
Loss weighting	$1:1$	$1:1$

Algorithm A1 Training procedure of RES

Require: Online N2I model v_θ, N2I EMA model $v_{\theta_{\text{ema}}}$, and frozen I2N model u_ϕ
Require: I2N inference stepsize $\Delta h = 0.04$, maximum sampling time for I2N $s_{\max}$, EAV inference steps K, and EMA decay μ

while not converged **do**
  Sample a image batch $\{x\}$ and split it equally into $\mathcal{B}_{\text{FM}}$ and $\mathcal{B}_{\text{EAV}}$
  FM branch
  Sample $t \sim \mathcal{U}(0,1)$ and $\epsilon \sim \mathcal{N}(0, I)$
  $x_t \leftarrow (1-t)\epsilon + tx, \quad v^\star \leftarrow x - \epsilon.$
  $\hat{v}_{\text{FM}} \leftarrow v_\theta(x_t, t, y).$
  EAV branch
  if $K > 1$ **then**
    Sample $s \sim \mathcal{U}(0, s_{\max})$ independently per example
  else
    Sample $s \sim \mathcal{U}(0.1, s_{\max})$ independently per example
  end if
  Compute the number of full pre-inference steps and the residual:

$$N = \left\lfloor \frac{s}{\Delta h} \right\rfloor, \qquad \Delta h_{\text{end}} = s - N\Delta h.$$

  Initialize the I2N trajectory with $\hat{x} \leftarrow x$
  for $n = 0, \ldots, N-1$ **do**
    $\hat{x} \leftarrow \hat{x} + \Delta h\, u_\phi(\hat{x}, n\Delta h, y).$
  end for
  if $\Delta h_{\text{end}} > 0$ **then**
    $\hat{x} \leftarrow \hat{x} + \Delta h_{\text{end}}\, u_\phi(\hat{x}, N\Delta h, y).$
  end if
  Set the N2I starting time and EAV inference stepsize:

$$r \leftarrow 1 - s, \qquad \Delta t \leftarrow \frac{s}{1+K}.$$

  if $K > 1$ **then**
    Apply one no-gradient FDB:

$$\hat{x} \leftarrow \hat{x} + \Delta t\ \text{sg}\left[v_{\theta_{\text{ema}}}(\hat{x}, r, y)\right], \qquad r \leftarrow r + \Delta t.$$

    Apply one gradient-carrying online step:

$$\hat{x} \leftarrow \hat{x} + \Delta t\, v_\theta(\hat{x}, r, y), \qquad r \leftarrow r + \Delta t.$$

    for $k = 2, \ldots, K$ **do**
      $\hat{x} \leftarrow \hat{x} + \Delta t\ \text{sg}\left[v_{\theta_{\text{ema}}}(\hat{x}, r, y)\right]$
      $r \leftarrow r + \Delta t.$
    end for
  else
    Apply one no-gradient FDB:

$$\hat{x} \leftarrow \hat{x} + 0.1\ \text{sg}\left[v_{\theta_{\text{ema}}}(\hat{x}, r, y)\right], \qquad r \leftarrow r + 0.1.$$

    Apply one gradient-carrying online step:

$$\hat{x} \leftarrow \hat{x} + (1-r)\, v_\theta(\hat{x}, r, y).$$

  end if
  Compute

$$\mathcal{L}_{\text{RES}} = \underbrace{\|\hat{v}_{\text{FM}} - v^\star\|_2^2}_{\mathcal{L}_{\text{FM}}} + \underbrace{\|\hat{x} - x\|_2^2}_{\mathcal{L}_{\text{EAV}}}.$$

  Update θ using $\nabla_\theta \mathcal{L}_{\text{RES}}$, then update $\theta_{ema} \leftarrow \mu\theta_{ema} + (1-\mu)\theta$
end while