FLOW MATCHING WITH ARBITRARY AUXILIARY PATHS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce **Flow Matching with Arbitrary Auxiliary Paths (AuxPath-FM)**, a generative modeling framework that makes the auxiliary distribution an explicit choice in probability-path design. AuxPath-FM constructs trajectories $X_t = a(t)X_1 + b(t)X_0 + c(t)\eta$, allowing η to be continuous, discrete, or coupled to the endpoints without requiring Gaussianity or independence. Under explicit moment and coefficient regularity conditions, we establish endpoint preservation, a weak continuity equation, and a tractable regression objective equivalent to marginal flow matching for each fixed path. This formulation accommodates Gaussian, Uniform, Laplace, Rademacher, and learned semantic auxiliaries within a common training framework. Semantic auxiliaries encode labels in the path and enable trajectory-level guidance with one backbone evaluation per sampling step. Experiments reveal how auxiliary choice affects generation across datasets and how semantic guidance improves label alignment. On ImageNet-1k, auxiliary guidance reduces FID from 23.25 to 18.01 at approximately the one-pass CFM cost and roughly half the computation of two-pass classifier-free guidance (CFG) at the same step count. AuxPath-FM connects flexible path design with a practical quality–cost tradeoff in conditional generation.

1 INTRODUCTION

Flow matching Lipman et al. (2023); Liu et al. (2023); Tong et al. (2024) learns the velocity of a probability path from noise to data. Because paths with identical endpoints can present different intermediate states and regression targets, path construction directly shapes the learning problem without changing the target distribution.

Existing work has emphasized endpoint coupling and intermediate geometry. Data-dependent stochastic interpolants develop endpoint coupling Albergo et al. (2024) while retaining an independent Gaussian auxiliary. This leaves the auxiliary law itself as a distinct design axis, spanning discrete perturbations and data-dependent semantic signals.

Discrete auxiliaries lack a Lebesgue density, whereas semantic auxiliaries couple to the data endpoint through their labels. A joint probability measure captures both cases and yields a common transport law and sample-based objective.

We introduce **AuxPath-FM** with additive paths

$$X_t = a(t)X_1 + b(t)X_0 + c(t)\eta, \tag{1}$$

where $X_0 \sim p_0$, $X_1 \sim p_1$, and the joint law $Q(X_0, X_1, \eta)$ may be continuous, singular, or endpoint-coupled, subject to finite-moment conditions. Endpoint constraints on a, b, c preserve p_0 and p_1. Conditioning directly on Q yields weak continuity and fixed-path regression equivalence without an auxiliary density, covering Gaussian, Uniform, Laplace, Rademacher, and semantic auxiliaries within this additive family.

The same construction turns semantic path design into a conditional-control mechanism. Semantic auxiliaries $\eta = F_\phi(Y)$ encode label Y in the path, and their explicit velocity contribution can be rescaled at sampling time. We therefore study both auxiliary-law selection and one-backbone semantic guidance.

Conventional conditional flow matching and CFG place semantic control primarily in conditioned network predictions Tong et al. (2024); Dhariwal & Nichol (2021); Ho & Salimans (2022). AuxPath-

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.

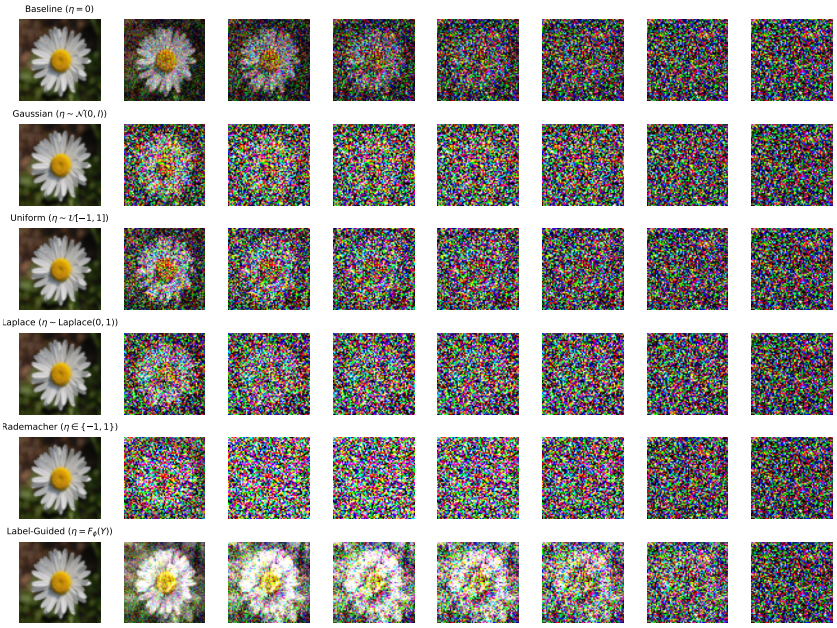


Figure 1: **AuxPath-FM trajectories under diverse auxiliary laws.** The panels compare $X_t = a(t)X_1 + b(t)X_0 + c(t)\eta$ for six choices of η: (a) Baseline ($\eta = 0$), (b) Gaussian $\eta \sim \mathcal{N}(0, I)$, (c) Uniform $\eta \sim \mathcal{U}([-1, 1]^d)$, (d) Laplace $\eta \sim$ Laplace$(0, 1)$, (e) Rademacher $\eta \in \{-1, 1\}^d$, and (f) Label-Guided $\eta = F_\phi(Y)$, illustrating how the auxiliary changes intermediate path geometry.

FM adds a geometric control channel through the path, linking auxiliary-law design and intermediate geometry to sampling-time guidance.

Our contributions are **(i) general auxiliary laws for additive paths**: a joint-measure construction covering continuous, discrete, singular, and endpoint-coupled auxiliaries; **(ii) theory**: weak continuity, fixed-path regression equivalence, and semantic-path translation–residual structure; and **(iii) control**: auxiliary guidance across laws, amplitudes, and temporal profiles.

2 Related Work

Flow-based generative models. Diffusion and flow matching provide the transport setting for our study Song et al. (2021); Ho et al. (2020); Lipman et al. (2023); Liu et al. (2023); Tong et al. (2024). Geometric and discrete variants alter the state space or transitions Kumar et al. (2026); Shaul et al. (2025); AuxPath-FM instead retains continuous states and varies the auxiliary law.

Stochastic paths and auxiliary variables. Auxiliary variables provide a flexible mechanism for shaping probability paths. Stochastic interpolants combine smooth nonlinear interpolation maps with an independent Gaussian latent, yielding density-level transport, Fokker–Planck, and score formulations for regular endpoint densities Albergo et al. (2025). AuxPath-FM develops the complementary law axis by defining the joint law $Q(X_0, X_1, \eta)$ directly and deriving measure-level flow matching for continuous, atomic, and endpoint-coupled auxiliaries. This makes auxiliary structure an explicit design variable for path geometry and semantic control. It complements endpoint-coupling methods, which design the pairing of (X_0, X_1) Albergo et al. (2024); Tong et al. (2024); Pooladian et al. (2023), as well as non-Gaussian transformations and discrete-state dynamics Bansal et al. (2023); Stark et al. (2024); Gat et al. (2024); Campbell et al. (2024).

Conditional modeling and guidance. Network conditioning and classifier-free guidance inject semantics through the velocity model Mirza & Osindero (2014); Dhariwal & Nichol (2021); Ho & Salimans (2022), while auxiliary variables support variance reduction and domain-specific control Bortoli et al. (2023); Huguet et al. (2024); Stoica et al. (2025). Guidance intervals, CADS, and TGATE reduce or schedule guidance computation Kynkäänniemi et al. (2024); Sadat et al. (2024); Liu

et al. (2025). AuxPath-FM separates these roles by scaling a semantic path drift with one conditional backbone evaluation.

3 BACKGROUND

Flow Matching. Flow matching Lipman et al. (2023); Liu et al. (2023) learns a velocity field that transports a base distribution p_0 to a data distribution p_1 along a prescribed probability path p_t.

Let $X_0 \sim p_0$ and $X_1 \sim p_1$. A common choice of interpolation is the linear path $X_t = (1-t)X_0 + tX_1$, whose time derivative is $\dot{X}_t = X_1 - X_0$. A tractable training objective regresses a velocity field $v_\theta(x, t)$ against this sample-wise derivative:

$$\mathcal{L}_{\text{linear}} = \mathbb{E}_{t,X_0,X_1}\left[\|v_\theta(X_t, t) - (X_1 - X_0)\|^2\right]. \tag{2}$$

After training, samples are generated by solving the ordinary differential equation (ODE)

$$\frac{dX_t}{dt} = v_\theta(X_t, t), \tag{3}$$

which realizes the prescribed marginals when the velocity field is exact and the ODE is well posed.

Conditional Flow Matching (CFM). Conditional flow matching uses tractable conditional paths to learn a marginal velocity field Lipman et al. (2023); Tong et al. (2024). For label-conditional generation, the regression additionally observes a class label Y. The model takes the form $v_\theta(x, t, Y)$, with interpolation $X_t = (1-t)X_0 + tX_1$, where $X_0 \sim p_0$ and $X_1 \sim p_1(\cdot|Y)$.

The resulting objective becomes

$$\mathcal{L}_{\text{label}} = \mathbb{E}_{t,X_0,X_1,Y}\left[\|v_\theta(X_t, t, Y) - (X_1 - X_0)\|^2\right]. \tag{4}$$

Here Y affects the velocity parameterization; the interpolation remains the standard endpoint path.

Auxiliary-variable interpolation. Stochastic interpolants Albergo et al. (2025) allow a general smooth endpoint map $I(t, X_0, X_1)$. The spatially linear specialization is

$$X_t = (1-t)X_0 + tX_1 + \sigma(t)Z, \tag{5}$$

where $X_0 \sim p_0$, $X_1 \sim p_1$, and $Z \sim \mathcal{N}(0, I)$ is independent of the endpoints.

Stochastic interpolants broaden map geometry through nonlinear I; their stated density, velocity, and score results use endpoint densities and independent Gaussian Z, while Remark 2 notes other noise laws Albergo et al. (2025). For the additive map, AuxPath-FM proves measure-level transport for general Q, including atomic, singular, and endpoint-coupled auxiliaries (Appendix A.3).

4 METHOD

The method first needs a transport law that remains valid without an auxiliary density. We derive that law and its tractable regression objective for joint auxiliary paths, then specialize the resulting velocity decomposition to semantic auxiliaries and sampling-time guidance.

4.1 ADDITIVE PATHS WITH GENERAL AUXILIARY LAWS

Let $X_0 \sim p_0$ be a base sample and $X_1 \sim p_1$ a data sample. We introduce an auxiliary variable η, which may depend on either endpoint. We denote the joint distribution of (X_0, X_1, η) by Q and define

$$X_t = a(t)X_1 + b(t)X_0 + c(t)\eta, \tag{6}$$

where $a(t), b(t), c(t)$ are time-dependent scalar functions.

To ensure correct boundary behavior, the coefficients satisfy $a(0) = 0, b(0) = 1, c(0) = 0$ and $a(1) = 1, b(1) = 0, c(1) = 0$, so the path has endpoint laws p_0 and p_1. These identities hold for each

sampled endpoint pair and auxiliary. The coefficient $c(t)$ controls the auxiliary displacement between the endpoints.

Conditional probability paths. Write $p_t = \mathrm{Law}(X_t)$ as a probability measure, and q for the law of (X_0, η). Conditional laws exist even for discrete auxiliaries or deterministic couplings.

Marginal probability path. For every Borel set A, disintegration gives

$$p_t(A) = \int p_t(A \mid x_0, \eta)\, q(dx_0, d\eta). \tag{7}$$

Marginal vector field. With $\dot{X}_t = \dot{a}X_1 + \dot{b}X_0 + \dot{c}\eta$, define, for almost every t,

$$u_t(x) = \mathbb{E}_Q[\dot{X}_t \mid X_t = x]. \tag{8}$$

This conditional-expectation definition specifies the field p_t-almost everywhere and applies to discrete as well as continuous laws.

Algorithm 1 AuxPath-FM: Training

```
# v_theta: velocity model
# a(t), b(t), c(t): path coefficients
# Q: joint law with finite moments
x0, x1, eta = sample_joint_Q()
t = sample_uniform()

X_t = a(t)* x1 +  b(t)* x0 +  c(t)* eta
v_t = dot_a(t)* x1 +  dot_b(t)* x0 +
     dot_c(t)* eta

loss = metric(v_theta(X_t,t) -  v_t)
```

Algorithm 2 AuxPath-FM: Sampling

```
x = sample_p0(x_shape)

for t in schedule:
    x = x +  v_theta(x,t) *  dt
```

Theorem 1 (Weak Continuity of the Marginal Path)**.** *Let Q have finite first moments and let a, b, c be absolutely continuous with the stated endpoint values. Then $p_t = \mathrm{Law}(X_t)$ is narrowly continuous and u_t above satisfies*

$$\partial_t p_t + \nabla \cdot (u_t p_t) = 0 \tag{9}$$

in the distributional sense, with endpoints p_0, p_1.

Differentiating test functions along the interpolation gives the weak identity for the full joint-law family. Spatial Lipschitz and growth bounds on u_t then upgrade the weak solution to a unique ODE flow. Appendix A proves both statements and compares the additive law classes.

Implication. The theorem places continuous, discrete, and coupled auxiliaries in the same transport formulation. At each state, the marginal field averages the sample-wise interpolation velocities; under the stated ODE conditions, integrating this field reproduces the marginal path. The sampler follows this conditional-mean field, which generally differs from any single sample-wise training curve.

Choice of auxiliary distribution. A discrete auxiliary selects a displacement, not a discrete state space: differentiable a, b, c still yield differentiable curves. Thus, auxiliary discreteness is compatible with continuous-time transport. Section 4.6 compares representative continuous, discrete, and semantic choices.

4.2 Velocity Field and Training Objective

We parameterize the velocity by $v_\theta(x, t)$ and use the sample derivative

$$\dot{X}_t = \dot{a}(t)X_1 + \dot{b}(t)X_0 + \dot{c}(t)\eta. \tag{10}$$

From marginal to conditional objectives. The marginal flow matching objective regresses against u_t along the probability path p_t:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{t,x\sim p_t}\left[\|v_\theta(x,t) - u_t(x)\|^2\right], \tag{11}$$

Samples from p_t are available, but u_t is generally intractable.

AuxPath Conditional Flow Matching. We instead regress against the available sample-wise derivative:

$$\mathcal{L}_{\mathrm{AuxPath}} = \mathbb{E}_{t,X_0,X_1,\eta}\left[\|v_\theta(X_t,t) - \dot{X}_t\|^2\right], \tag{12}$$

where $t \sim \mathcal{U}[0,1]$ independently of $(X_0, X_1, \eta) \sim Q$. Sampling (t, X_0, X_1, η) gives an unbiased Monte Carlo estimate of this loss. Conditioning the derivative on (t, X_t) yields the marginal velocity.

Equivalence to marginal flow matching. For a fixed path, conditional regression and marginal flow matching differ by an irreducible variance term.

Theorem 2 (Fixed-Path Regression Equivalence). *Fix Q and the coefficients independently of θ. For independent $T \sim \mathcal{U}[0,1]$, assume $\mathbb{E}\|\dot{X}_T\|^2 < \infty$ and $\mathbb{E}\|v_\theta(X_T, T)\|^2 < \infty$. Then*

$$\mathcal{L}_{\text{AuxPath}}(\theta) = \mathcal{L}_{\text{FM}}(\theta) + \underbrace{\mathbb{E}\|\dot{X}_T - u_T(X_T)\|^2}_{C \geq 0}, \tag{13}$$

where C is independent of θ. If differentiation under expectation is valid,

$$\nabla_\theta \mathcal{L}_{\text{AuxPath}} = \nabla_\theta \mathcal{L}_{\text{FM}}. \tag{14}$$

Implication. For fixed Q and coefficients, sample-wise MSE recovers the marginal-velocity objective directly from joint-law samples, so the two objectives share their population minimizers over θ. The same projection applies with observed labels; for learned paths, Appendix B tracks the path-dependent variance explicitly.

4.3 Semantic Auxiliaries: A Conditional-Generation Application

For conditional generation, we set $\eta = F_\phi(Y)$, where a lightweight network maps label Y to an auxiliary vector. Sharing Y couples η with the data endpoint X_1, allowing class information to shape the training path directly.

We learn F_ϕ in the backbone's input representation, either pixels or latents. Sample-wise MSE targets the class mean in expectation:

$$F_\phi(Y) \approx \mathbb{E}[X_1 \mid Y]. \tag{15}$$

We fit this map with the mean-squared-error (MSE) objective

$$\mathcal{L}_F = \mathbb{E}_{X_1, Y}\left[\|F_\phi(Y) - X_1\|^2\right]. \tag{16}$$

Minimizing this MSE alone gives the class mean in the unrestricted population limit. Two-stage training first fits F_ϕ and then freezes it while training the velocity model. With X_0 independent of Y and $X_t^y = a(t)X_1 + b(t)X_0 + c(t)F_\phi(y)$, the conditional velocity is

$$u_t^y(x) = \mathbb{E}[\dot{a}(t)X_1 + \dot{b}(t)X_0 \mid X_t^y = x, y] + \dot{c}(t)F_\phi(y). \tag{17}$$

This decomposition assigns the residual to $v_\theta(x, t, y)$ and the explicit drift to the auxiliary. Given y, the translated path preserves optimal regression variance. The residual captures class structure beyond the prototype: identical class means can coexist with distinct velocity fields, motivating label-conditioned residuals (Appendix B). Algorithms 3–5 use $s = y$ for network conditioning and omit s for the path-only variant.

Thus the auxiliary represents class-level path translation, while the residual models within-class structure Mirza & Osindero (2014); Tong et al. (2024). This division explains why path-only and network-plus-path conditioning can differ in accuracy and sample quality.

4.4 Auxiliary Guidance

The semantic path decomposes the sampling velocity into a learned residual and an explicit auxiliary drift that can be controlled independently. We write w_{CFG} for backbone guidance and w_{aux} for auxiliary guidance; $\emptyset$ denotes the null condition, with $F_\phi(\emptyset) = \mu$, the global data mean in the model input space. The backbone argument is $s = y$ for the conditional branch and $s = \emptyset$ for the null branch; a label-blind residual omits s. For auxiliary label y, the composite velocity is

$$v(x, t, y) = v_\theta(x, t, s) + \dot{c}(t)F_\phi(y). \tag{18}$$

Setting $w_{\text{CFG}} = 1$ selects $s = y$ for a conditioned backbone and isolates auxiliary-drift guidance:

$$v_{\text{aux}}(x, t, y) = v_\theta(x, t, s) + \dot{c}(t)\big[F_\phi(\emptyset) + w_{\text{aux}}(F_\phi(y) - F_\phi(\emptyset))\big]. \tag{19}$$

This controls the drift with one backbone evaluation per step. Appendix D derives the two-scale dynamics and shows when auxiliary guidance admits a CFG interpretation over surrogate fields.

Algorithm 3 AuxPath-FM (Cond.): Training

```
# Stage 1: train semantic model
x1, y = sample_data()
pred = F_phi(y)
loss_F = metric(pred - x1)

# Stage 2: train flow model
x0, x1, y = sample_batch()
eta = stopgrad(F_phi(y))
t = sample_uniform()
X_t = (a(t)* x1
       + b(t)* x0
       + c(t)* eta)
v_t = (dot_a(t)* x1
       + dot_b(t)* x0)
loss = metric(
    v_theta(X_t,t,s) - v_t)
```

Algorithm 4 AuxPath-FM (Cond.): Sampling

```
x = randn(x_shape)
eta = F_phi(y)
for t in schedule:
    v = v_theta(x,t,s) + dot_c(t)* eta
    x = x + v * dt
```

Algorithm 5 AuxPath-FM: Auxiliary Guidance

```
x = randn(x_shape)
eta_c = F_phi(y)
eta_u = F_phi(empty)
eta_aux = eta_u + w_aux * (eta_c - eta_u)
for t in schedule:
    v = v_theta(x,t,s) + dot_c(t) * eta_aux
    x = x + v * dt
```

4.5 Toy Experiment: Visualizing Auxiliary Guidance

We test auxiliary guidance on a bimodal Gaussian ring using Algorithm 5. Stronger guidance separates trajectories toward the two modes (Figure 2). Its scale changes the explicit drift, intermediate states, and subsequent transport with one backbone evaluation, distinct from standard two-pass CFG Ho & Salimans (2022).

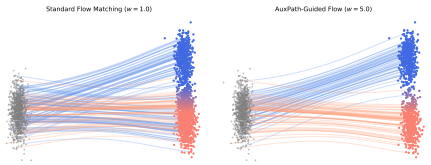


Figure 2: **Auxiliary guidance.** Comparison between $w_{\text{aux}} = 1.0$ and $w_{\text{aux}} = 5.0$. Higher guidance leads to clear trajectory separation.

4.6 Toy Experiment: Ring-64 Dataset

We next use ring-64, a 64-mode unit circle, to test whether different valid auxiliary paths preserve conditional alignment. All variants use the same label-conditioned backbone $v_\theta(X_t, t, Y)$ and training budget. CFM uses the linear path $X_t = (1-t)X_0 + tX_1$, whereas AuxPath-FM adds $t(1-t)\eta$ with Gaussian, Uniform, Laplace, Rademacher, or label-guided $\eta = F(Y)$.

Table 1 reports nearest-mode accuracy and distance error. Relative to CFM, label-guided η raises accuracy from 57.20% to 60.60% and lowers error from 0.3398 to 0.2955, whereas all four random auxiliaries reduce accuracy and increase error. Improving both metrics indicates better conditional placement. The shared backbone and training budget isolate the training path, showing that path validity alone does not determine conditional alignment. This contrast separates validity from task alignment: random endpoint-preserving displacements need not organize states by label, whereas the semantic auxiliary couples the path geometry to the intended mode. Figure 3 shows the trajectories.

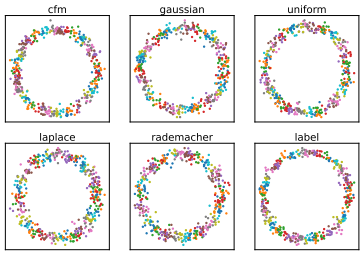


Figure 3: Ring-64 trajectories under auxiliary laws η with the same conditioned backbone and budget; label-guided η is semantically coupled, unlike independent Gaussian stochastic-interpolant noise Albergo et al. (2025).

Table 1: Ring-64 evaluation. Higher accuracy and lower error are better.

Auxiliary Type	Acc (%) ↑	Error ↓
CFM (baseline)	57.20	0.3398
Gaussian	45.60	0.3956
Uniform	49.80	0.3586
Laplace	47.20	0.4228
Rademacher	51.00	0.3911
Label-guided (Ours)	**60.60**	**0.2955**

5 EXPERIMENTS

Our experiments ask how the auxiliary law affects generation, whether semantic auxiliaries improve conditional alignment, and how guidance trades quality against computation. We study MNIST, CIFAR-10, and ImageNet-1k on one NVIDIA RTX 4090 GPU; the appendices isolate guidance variants, coefficient shapes, unconditional-to-conditional adaptation, trajectories, and matched fine-tuning.

5.1 EXPERIMENTAL SETUP

Datasets. We use MNIST Lecun et al. (1998), CIFAR-10 Krizhevsky & Hinton (2009), and ImageNet-1k Deng et al. (2009) for conditional generation and scalability tests.

Evaluation Metrics. We measure label agreement using classifier accuracy (Acc), with classifiers achieving 99.1% on MNIST (CNN) An et al. (2020) and 92% on CIFAR-10 (ResNet-56) He et al. (2016). Generation metrics are Fréchet Inception Distance (FID), Kernel Inception Distance (KID) Binkowski et al. (2018), and Inception Score (IS) Heusel et al. (2017); Salimans et al. (2016), each computed from 50k samples. Because results are single runs, small differences indicate comparable performance rather than stable rankings across seeds.

5.2 SEMANTIC AUXILIARY DISTRIBUTIONS ON MNIST AND CIFAR-10

Experimental setup. We use a standard **U-Net** Ronneberger et al. (2015); Ho et al. (2020) for CIFAR-10 and **LightningDiT** Yao et al. (2025) for MNIST. All methods train for 400K steps under identical settings. AuxPath-FM follows Algorithm 3 and samples with Algorithm 4; v_θ regresses the residual target, while $\eta = F_\phi(Y)$ enters through the training path and the sampling drift. Figure 4 compares CFM, auxiliary guidance, and their combination.

In Table 2, CFM uses network conditioning; CFM+η additionally uses the auxiliary path and drift; and η alone uses a label-blind residual backbone. Numeric auxiliaries broadcast the numerical label value to every entry of a tensor with the same shape as the model input, $\eta = y\mathbf{1}$, regardless of the input dimensions. Learned auxiliaries use $F_\phi(Y)$.

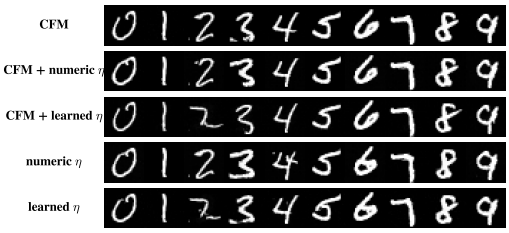


Figure 4: **Qualitative comparison of conditional generation on MNIST.** We compare standard CFM, fixed numeric auxiliary paths (η = numeric), and learned label prototypes ($\eta = F_\phi(Y)$).

Table 2: Results on MNIST and CIFAR-10. Acc is reported using pretrained classifiers.

	MNIST				CIFAR-10			
Method	Acc(%) ↑	FID ↓	KID ↓	IS ↑	Acc(%) ↑	FID ↓	KID ↓	IS ↑
CFM	95.5	**5.90**	**0.006**	2.099	82.24	4.71	0.0032	9.18
CFM + numeric η	**98.1**	7.33	0.007	2.096	83.61	4.54	0.0030	9.19
CFM + learned η	97.1	7.57	0.008	2.112	**87.28**	**3.68**	**0.0017**	**9.20**
numeric η	94.3	10.04	0.009	2.097	79.18	5.73	0.0038	8.89
learned η	90.6	13.82	0.015	**2.128**	77.03	5.90	0.0043	8.93

Results. Table 2 shows that auxiliary paths improve label agreement when combined with network conditioning. On CIFAR-10, learned η raises accuracy from 82.24% to 87.28%, lowers FID from 4.71 to 3.68, and improves KID and IS. On MNIST, numeric η gives the highest accuracy (98.1%), while CFM retains the best FID (5.90), so stronger label alignment need not imply lower FID. Joint path and network conditioning outperforms path-only conditioning in accuracy and FID; the separate guidance sweep reaches 85.82% accuracy and its best FID at a moderate scale (Appendix D.3).

5.3 Continuous, Discrete, and Learned Auxiliary Distributions

Setup. We compare auxiliary laws using the full-velocity objective and sampler in Algorithms 1–2; inference evaluates the learned velocity without an auxiliary input. For ImageNet-1k at 256×256, DiT-B/2 Peebles & Xie (2023) starts from a pre-trained CFM checkpoint Dao et al. (2023) and is fine-tuned for 300K steps, about 3% of its original 10M-step budget. CFM follows the matched protocol in Appendix H.1.

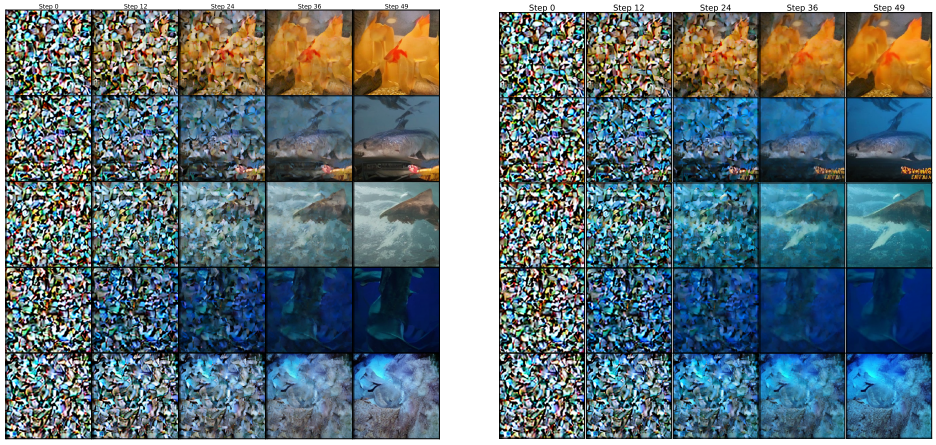
Results. Table 3 gives CIFAR-10 FIDs of 4.38–4.59 across all auxiliaries versus 4.71 for CFM, including 4.46 for discrete Rademacher. Gaussian has the lowest measured FID, while label-guided η has the best KID and IS among auxiliary variants. On ImageNet-1k, however, all auxiliary FIDs (23.45–24.71) exceed the matched CFM value of 23.25. Continuous and discrete laws therefore both support generation, but their quality effects depend on the dataset and training regime; the small CIFAR-10 gaps do not establish a stable ordering.

Table 3: Auxiliary-law comparison. Bold marks the best auxiliary in each column; CFM is the reference.

	CIFAR-10			ImageNet-1k
Method	FID ↓	KID ↓	IS ↑	FID ↓
CFM	4.71	0.0032	9.18	23.25
Gaussian	**4.38**	0.0029	9.21	24.19
Uniform	4.40	0.0028	9.17	23.48
Laplace	4.49	0.0031	9.19	24.71
Rademacher	4.46	0.0031	9.24	24.23
Label-guided	4.59	**0.0024**	**9.32**	**23.45**

5.4 Joint Training and Auxiliary Guidance on ImageNet-1k

Joint-training setup. We use joint training to test whether adapting the auxiliary map with the backbone creates a useful one-pass guidance channel. The objective combines residual flow loss with $\lambda_{\text{prior}}\mathcal{L}_F$, using $c_\alpha(t) = \alpha t(1-t)$: α controls training-path amplitude, while w_{aux} controls sampling guidance. Appendix H.1 gives details. Figure 5 changes only w_{aux} from 1.0 to 1.5 at $\alpha = 1$, $\lambda_{\text{prior}} = 0.5$, reducing FID from 22.01 to 19.43. Thus, one trained model supports distinct guided trajectories and evaluation states for the fixed residual network.



(a) $w_{\text{aux}} = 1.0$ (b) $w_{\text{aux}} = 1.5$

Figure 5: **Joint-training trajectories on ImageNet-1k.** Both panels use $\alpha = 1$, $\lambda_{\text{prior}} = 0.5$, and one backbone evaluation per step ($w_{\text{CFG}} = 1$). Rows trace $t = 0$ to $t = 1$. With the backbone and solver fixed, changing w_{aux} changes intermediate states and hence subsequent inputs to the residual network. It rescales a label-coupled deterministic drift, rather than a diffusion coefficient or score term as in stochastic interpolants Albergo et al. (2025).

Quantitative results. Every $(\alpha, \lambda_{\text{prior}})$ configuration in Table 4 improves for some $w_{\text{aux}} > 1$, but the preferred scale depends on path amplitude. FID decreases throughout the sweep at $\alpha = 0.2$, reaches its row minimum at $w_{\text{aux}} = 3.0$ or 3.5 for $\alpha = 0.5$, and reaches its minimum at 2.0 for $\alpha = 1$. Thus, larger training-path amplitudes favor lower sampling-time guidance scales. The lowest

measured FID is 18.01 at $(\alpha, \lambda_{\text{prior}}, w_{\text{aux}}) = (1, 1, 2)$; Appendix D.5 further isolates the temporal profile.

Table 4: Joint-training FID versus guidance scale w. For CFM ($\alpha = 0$), $w = w_{\text{CFG}}$; otherwise $w = w_{\text{aux}}$ with $w_{\text{CFG}} = 1$. Bold marks the lowest FID for each $(\alpha, \lambda_{\text{prior}})$ configuration; gray denotes the CFM baseline.

		w								
α	λ_{prior}	1.0	1.5	2.0	2.5	3.0	3.5	4.0	Pass	GFLOPs
0	0	23.25	5.31	5.18	8.61	12.07	14.77	16.80	2(1)	4600.24(2300.12)
0.2	0.5	22.77	21.96	21.22	20.52	19.88	19.37	**18.98**	1(1)	2303.08(2301.60)
	1.0	22.61	21.75	20.93	20.22	19.60	19.14	**18.67**	1(1)	2303.08(2301.60)
	2.0	22.93	22.01	21.18	20.51	19.90	19.33	**18.90**	1(1)	2303.08(2301.60)
0.5	0.5	22.84	21.37	19.33	19.33	18.92	**18.67**	18.72	1(1)	2303.08(2301.60)
	1.0	22.61	20.75	19.46	18.55	**18.10**	18.14	18.49	1(1)	2303.08(2301.60)
	2.0	22.52	20.70	19.36	18.49	**18.06**	18.13	18.59	1(1)	2303.08(2301.60)
1.0	0.5	22.01	19.43	**18.30**	18.60	20.29	23.17	27.21	1(1)	2303.08(2301.60)
	1.0	21.84	19.13	**18.01**	18.33	19.95	22.94	27.06	1(1)	2303.08(2301.60)
	2.0	22.18	19.52	**18.60**	19.16	21.25	24.67	29.35	1(1)	2303.08(2301.60)

Comparison with standard CFG. Table 4 reports backbone evaluations per step and GFLOPs per image over 50 steps; parentheses give the cost at $w = 1$, and the outer values give the cost at $w > 1$. At the one-pass operating point, auxiliary guidance lowers FID from 23.25 to 18.01 while increasing GFLOPs by about 0.13%. This improves quality near the one-pass budget and uses roughly half the computation of two-pass standard CFG at the same step count.

6 Discussion

Auxiliary-path design couples three choices: the displacement law, its temporal profile, and its sampling strength. Random auxiliaries improve measured CIFAR-10 FID but degrade ImageNet-1k FID and ring-64 alignment, whereas semantic auxiliaries improve conditional alignment and the one-pass ImageNet operating point. Path amplitude and shape also shift the preferred guidance scale.

The translation analysis explains why these effects arise beyond population consistency. Deterministic label auxiliaries preserve Bayes regression variance but change the finite network's parameterization and add an explicit sampling drift (Appendix B.4). This one-backbone control complements standard guidance: combining both scales lowers FID in the tested grid (Appendix D.4).

Training amplitude and sampling strength therefore define a joint operating point. The former determines the auxiliary displacement seen during training, whereas the latter rescales it at inference; the nonmonotonic FID at large scales is consistent with guided trajectories entering states less represented by the training path.

Limitations. Our evaluation establishes a stronger one-pass ImageNet operating point for auxiliary guidance; standard CFG still attains a lower minimum FID with roughly twice the backbone computation. Random auxiliaries do not improve consistently across datasets. The text-to-image study in Appendix F covers one SANA/LoRA configuration, one training and sampling seed, 30 held-out prompts, and a local BLIP-VQA evaluator. It establishes compatibility in this setting, while large-scale text guidance remains unevaluated.

7 Conclusion

AuxPath-FM establishes flow matching for general auxiliary laws in an additive path family, including continuous, discrete, singular, and endpoint-coupled laws under finite-moment conditions. Semantic auxiliaries provide one-pass control of the explicit path drift, complementing endpoint coupling and network guidance Albergo et al. (2024); Ho & Salimans (2022).

AI USE STATEMENT

In this work, we used large language models (LLMs) solely to edit author-written text for clarity, grammar, and academic style. We did not use generative AI tools for research conception, methodological or theoretical development, experimental design, analysis or interpretation of results, or the formulation of scientific conclusions; other tasks requiring disclosure under the ICLR AI policy were not applicable to this work. All AI-assisted edits were reviewed by the authors against the original text to ensure accuracy and preserve the intended scientific meaning. We take responsibility for the final content of this work, including any text produced with the aid of generative AI.

ETHICS STATEMENT

This work uses standard public image benchmarks and does not involve human subjects, private data, or deployment of a decision-making system. The method may inherit biases from its training data, and improved controllable generation may be misused; our experiments are limited to research benchmarks and we do not claim readiness for high-stakes applications. We adhere to the ICLR Code of Ethics and report the method, limitations, and results transparently.

REPRODUCIBILITY STATEMENT

The main paper specifies the AuxPath formulation, training objectives, algorithms, datasets, metrics, and experimental settings. Full proofs are provided in Appendices A–B, additional experiments in Appendices D–E, and an anonymous implementation is described in Appendix H.

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. URL https://jmlr.org/papers/v26/23-1605.html.

Michael Samuel Albergo, Mark Goldstein, Nicholas Matthew Boffi, Rajesh Ranganath, and Eric Vanden-Eijnden. Stochastic interpolants with data-dependent couplings. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 921–937. PMLR, 2024. URL https://proceedings.mlr.press/v235/albergo24a.html.

Sanghyeon An, Minjun Lee, Sanglee Park, Heerin Yang, and Jungmin So. An ensemble of simple convolutional neural network models for mnist digit recognition, 2020. URL https://arxiv.org/abs/2008.10400.

Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. In *Neural Information Processing Systems (NeurIPS)*, 2023. URL https://openreview.net/forum?id=XH3ArccntI.

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

Valentin De Bortoli, Guan-Horng Liu, Tianrong Chen, Evangelos A. Theodorou, and Weilie Nie. Augmented bridge matching. *arXiv preprint arXiv:2311.06978*, 2023.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In *International Conference on Machine Learning (ICML)*, 2024. URL https://openreview.net/forum?id=kQwSbv0BR4.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *arXiv preprint arXiv:2310.18235*, 2023.

Quan Dao, Hao Phung, Binh Nguyen, and Anh Tran. Flow matching in latent space. *arXiv*, 2023. URL https://arxiv.org/abs/2307.08698.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Neural Information Processing Systems (NeurIPS)*, 34, 2021.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In *Neural Information Processing Systems (NeurIPS)*, 2024. URL https://openreview.net/forum?id=GTDKo3Sv9p.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems (NeurIPS)*, NIPS'17, pp. 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Neural Information Processing Systems (NeurIPS)*, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023.

Guillaume Huguet, James Vuckovic, Kilian FATRAS, Eric Thibodeau-Laufer, Pablo Lemos, Riashat Islam, Cheng-Hao Liu, Jarrid Rector-Brooks, Tara Akhound-Sadegh, Michael M. Bronstein, Alexander Tong, and Joey Bose. Sequence-augmented SE(3)-flow matching for conditional protein generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=paYwtPBpyZ.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Shivam Kumar, Yixin Wang, and Lizhen Lin. Flow matching is adaptive to manifold structures. *arXiv preprint arXiv:2602.22486*, 2026.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2404.07724.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

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

Haozhe Liu, Wentian Zhang, Jinheng Xie, Francesco Faccio, Mengmeng Xu, Tao Xiang, Mike Zheng Shou, Juan-Manuel Perez-Rua, and Jürgen Schmidhuber. Faster diffusion via temporal attention decomposition. *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2404.02747.

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 (ICLR)*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. In *arXiv preprint arXiv:1411.1784*, 2014.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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, pp. 28100–28127. PMLR, 2023. URL https://proceedings.mlr.press/v202/pooladian23a.html.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zMoNrajk2X.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Neural Information Processing Systems (NeurIPS)*, NIPS'16, pp. 2234–2242, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Neta Shaul, Uriel Singer, Itai Gat, and Yaron Lipman. Transition matching: Scalable and flexible generative modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=An0ePypuOJ.

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 (ICLR)*, 2021.

Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet flow matching with applications to DNA sequence design. In *International Conference on Machine Learning (ICML)*, 2024. URL https://openreview.net/forum?id=syXFAVqx85.

George Stoica, Vivek Ramanujan, Xiang Fan, Ali Farhadi, Ranjay Krishna, and Judy Hoffman. Contrastive flow matching. In *IEEE International Conference on Computer Vision (ICCV)*, 2025.

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. URL https://openreview.net/forum?id=CD9Snc73AW.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712, 2025.

A PROOF OF AUXPATH-FM WEAK-TRANSPORT CONSISTENCY

We prove consistency directly for continuous, discrete, and coupled auxiliary laws. Let $(X_0, X_1, \eta) \sim Q$ in $(\mathbb{R}^d)^3$ have finite first moments. Let a, b, c be absolutely continuous on $[0,1]$ with the endpoint values in Sec. 4.1, and set $X_t = a(t)X_1 + b(t)X_0 + c(t)\eta$. Every expectation below uses this joint law, which permits dependence among all three variables.

A.1 MEASURE-VALUED CONTINUITY EQUATION

Proof of Theorem 1. Absolute continuity gives almost surely absolutely continuous sample paths and

$$\mathbb{E}\int_0^1 \|\dot{X}_t\|\,dt \le \|\dot{a}\|_{L^1}\mathbb{E}\|X_1\| + \|\dot{b}\|_{L^1}\mathbb{E}\|X_0\| + \|\dot{c}\|_{L^1}\mathbb{E}\|\eta\| < \infty. \tag{20}$$

To obtain one jointly measurable velocity rather than unrelated fixed-time versions, work on the product space $([0,1]\times\Omega, dt\otimes Q)$ and define $\Gamma(t,\omega) = (t, X_t(\omega))$. Its pushforward is the space–time measure $\rho = \Gamma_{\#}(dt \otimes Q)$, equivalently $\rho(dt, dx) = dt\, p_t(dx)$. Define the finite vector measure

$$m(B) = \int_0^1 \mathbb{E}[\mathbf{1}_B(t, X_t)\dot{X}_t]\,dt \tag{21}$$

for Borel $B \subseteq [0,1]\times\mathbb{R}^d$. The preceding bound makes every component of m finite. Moreover, $m \ll \rho$: if $\rho(B) = 0$, then $\mathbf{1}_B(t, X_t) = 0$ for $dt \otimes Q$-almost every (t,ω). Applying the Radon–Nikodym theorem componentwise gives a jointly Borel field $u = dm/d\rho$. For every bounded Borel vector field g,

$$\int g(t,x)\cdot u(t,x)\,\rho(dt,dx) = \int_0^1 \mathbb{E}[g(t,X_t)\cdot\dot{X}_t]\,dt. \tag{22}$$

This is exactly the defining identity for the space–time conditional expectation of $\dot{X}_t$ given (t, X_t). Hence $u_t(x) = \mathbb{E}[\dot{X}_t \mid X_t = x]$ for $dt\,p_t(dx)$-almost every (t,x). Conditional Jensen, followed by Fubini, yields $\int_0^1\int \|u_t(x)\|p_t(dx)\,dt \le \mathbb{E}\int_0^1 \|\dot{X}_t\|dt < \infty$. For $\psi \in C_c^1([0,1]\times\mathbb{R}^d)$, the chain rule and Fubini's theorem imply

$$\begin{aligned}
&\int \psi(1,x)p_1(dx) - \int \psi(0,x)p_0(dx) \\
&\quad= \int_0^1 \mathbb{E}[\partial_t\psi(t,X_t) + \nabla\psi(t,X_t)\cdot\dot{X}_t]\,dt \\
&\quad= \int_0^1\int [\partial_t\psi(t,x) + \nabla\psi(t,x)\cdot u_t(x)]\,p_t(dx)\,dt.
\end{aligned} \tag{23}$$

Equation (22) with $g = \nabla\psi$ gives the last equality. This is the weak continuity equation. The endpoint identities hold sample-wise. Finally, if $t_n \to t$, absolute continuity gives $X_{t_n} \to X_t$ almost surely; thus $\mathbb{E}[f(X_{t_n})] \to \mathbb{E}[f(X_t)]$ by bounded convergence for every $f \in C_b(\mathbb{R}^d)$. This proves narrow continuity. □

The measure formulation includes atomic auxiliaries and singular couplings such as $\eta = F(X_0)$. When p_t admits a density, the weak identity recovers the usual density form of the continuity equation. The argument applies the conditional-expectation principle used in flow matching and stochastic interpolants Lipman et al. (2023); Albergo et al. (2025) directly to the joint auxiliary law.

A.2 DETERMINISTIC ODE REALIZATION

Proposition 1 (ODE realization under additional regularity)**.** *Let (p_t, u_t) be the narrowly continuous solution constructed in Theorem 1. Suppose u has a Borel representative such that, for almost every t and every $x, z \in \mathbb{R}^d$,*

$$\|u_t(x) - u_t(z)\| \le L(t)\|x - z\|, \qquad \|u_t(0)\| \le B(t), \qquad L, B \in L^1([0,1]). \tag{24}$$

Then the Carathéodory ODE $\dot{x}_t = u_t(x_t)$ generates a unique global flow $\Phi_{s,t}$, and

$$p_t = (\Phi_{0,t})_{\#}p_0, \qquad t \in [0,1]. \tag{25}$$

Proof. The Lipschitz bound and $\|u_t(x)\| \le B(t) + L(t)\|x\|$ give global existence and uniqueness by the Carathéodory theorem and Grönwall's inequality. The pushforward $(\Phi_{0,t})_{\#}p_0$ is a narrowly continuous finite-first-moment solution of the continuity equation. The same integrable Lipschitz bound gives uniqueness in this class, so it equals the weak solution p_t constructed above. □

Theorem 1 establishes weak marginal transport from first moments and absolutely continuous coefficients. Proposition 1 upgrades this result to an exact deterministic sampler when the marginal field satisfies the stated Lipschitz and growth bounds, including at the endpoints. The resulting trajectories follow the conditional mean velocity, which averages the velocities of training interpolations passing through each state.

A.3 Auxiliary-law generality for additive paths

The two constructions can be written side by side:

$$\text{independent additive path:} \quad X_t^{\mathrm{ind}} = I(t, X_0, X_1) + \gamma(t)Z, \qquad Z \perp (X_0, X_1),$$
$$\text{AuxPath:} \quad X_t^{Q} = a(t)X_1 + b(t)X_0 + c(t)\eta, \quad (X_0, X_1, \eta) \sim Q.$$

Stochastic interpolants generalize the map I and develop density-level machinery, commonly with an additive variable independent of the endpoints; their framework also discusses non-Gaussian noise laws Albergo et al. (2025). At the level of weak transport, the law-axis contribution is the dependence structure between the auxiliary and the endpoints. Accordingly, set $I(t, x_0, x_1) = a(t)x_1 + b(t)x_0$, $\gamma(t) = c(t)$, and define the independent class to include arbitrary auxiliary noise laws:

$$\mathcal{Q}_{\mathrm{ind}} := \left\{ \nu(dx_0, dx_1)\, \xi(d\eta) : \nu \in \mathcal{P}_1((\mathbb{R}^d)^2),\ \xi \in \mathcal{P}_1(\mathbb{R}^d) \right\},$$
$$\mathcal{Q}_{\mathrm{Aux}} := \mathcal{P}_1((\mathbb{R}^d)^3).$$

Proposition 2 (Law classes for the additive path)**.** *For absolutely continuous coefficients,*

$$\mathcal{Q}_{\mathrm{ind}} \subsetneq \mathcal{Q}_{\mathrm{Aux}}.$$

On $\mathcal{Q}_{\mathrm{ind}}$, the two additive constructions have identical paths, sample velocities, conditional-mean fields, and quadratic-objective minimizers whenever $\mathbb{E}\|\dot{X}_T\|^2 < \infty$. Every $Q \in \mathcal{Q}_{\mathrm{Aux}}$ yields the narrow path and weak continuity equation of Theorem 1; under the space–time second-moment condition of Appendix B.1, it also yields the L^2 velocity projection. The proposition isolates the law-axis extension at the level of weak transport. Density smoothness, score identities, and Fokker–Planck conclusions follow under their corresponding regularity assumptions.

Proof. Substitution gives $X_t^{\mathrm{ind}} = X_t^Q$ and $\dot{X}_t^{\mathrm{ind}} = \dot{X}_t^Q$ for every product law in $\mathcal{Q}_{\mathrm{ind}}$. If the stochastic-interpolant objective is written as $\mathcal{L}_{\mathrm{SI}}(V) = \mathbb{E}[\frac{1}{2}\|V\|^2 - D \cdot V]$ and $\mathcal{L}_{\mathrm{Aux}}(V) = \mathbb{E}\|V - D\|^2$, then the exact relation is

$$\mathcal{L}_{\mathrm{SI}}(V) = \tfrac{1}{2}\mathcal{L}_{\mathrm{Aux}}(V) - \tfrac{1}{2}\mathbb{E}\|D\|^2,$$

so their minimizers coincide. The general weak statement follows from Theorem 1. Inclusion is strict: choose any $\nu \in \mathcal{P}_1((\mathbb{R}^d)^2)$ whose X_0 marginal is non-degenerate and let $\eta = X_0$. The resulting joint law belongs to $\mathcal{Q}_{\mathrm{Aux}}$ but cannot factor as $\nu \otimes \xi$, because η is a non-constant measurable function of X_0. Semantic auxiliaries give the same kind of strict witness whenever they are statistically dependent on the endpoint pair. □

B AuxPath Flow Matching Loss Functions

B.1 Projection identity for a fixed path

Let $T \sim \mathcal{U}[0, 1]$ be independent of the joint sample, $D = \dot{X}_T$, and $U = \mathbb{E}[D \mid T, X_T]$. Assume D and $V = v_\theta(X_T, T)$ are square-integrable. With $\mathcal{L}_{\mathrm{AuxPath}} = \mathbb{E}\|V - D\|^2$ and $\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\|V - U\|^2$, expand

$$\mathbb{E}\|V - D\|^2 = \mathbb{E}\|V - U\|^2 + \mathbb{E}\|D - U\|^2 - 2\mathbb{E}\langle V - U, D - U\rangle. \tag{26}$$

The last term vanishes because $\mathbb{E}[D - U \mid T, X_T] = 0$. This proves Theorem 2, with $C = \mathbb{E}\|D - U\|^2 \geq 0$. The same projection argument holds after additionally conditioning on an observed label Y.

The required integrability is over both time and the joint law. Sufficient conditions are finite second moments of Q and $\dot{a}, \dot{b}, \dot{c} \in L^2([0,1])$. Finite $\mathbb{E}\|\dot{X}_t\|^2$ separately for almost every t is insufficient: with $X_0 = X_1 = 0$, a Rademacher η, and $c(t) = \sqrt{t}(1-t)$, each interior-time moment is finite but its time integral diverges. The space-time integrability condition ensures that the MSE and its variance decomposition are finite.

B.2 Residual regression and label information

Fix ϕ and let $R_T = \dot{a}(T)X_1 + \dot{b}(T)X_0$. Assume $X_0 \sim p_0$ independently of Y, and use $X_T = a(T)X_1 + b(T)X_0 + c(T)F_\phi(Y)$. Define

$$r_T^Y(X_T) = \mathbb{E}[R_T \mid T, X_T, Y], \qquad r_T(X_T) = \mathbb{E}[R_T \mid T, X_T]. \tag{27}$$

The exact label-conditional velocity is $u_T^Y(X_T) = r_T^Y(X_T) + \dot{c}(T)F_\phi(Y)$. A label-conditioned residual network trained against R_T therefore targets the correct residual. For any square-integrable label-blind residual v_θ, orthogonality yields

$$\begin{aligned}&\mathbb{E}\|v_\theta(X_T, T) + \dot{c}(T)F_\phi(Y) - u_T^Y(X_T)\|^2 \\ &\quad = \mathbb{E}\|v_\theta(X_T, T) - r_T(X_T)\|^2 + \underbrace{\mathbb{E}\|r_T(X_T) - r_T^Y(X_T)\|^2}_{\text{residual variation across labels}}.\end{aligned} \tag{28}$$

The second term quantifies the residual information supplied by the label beyond the observed state and time. It vanishes when the label-conditioned and label-blind conditional means agree almost surely; otherwise it is the irreducible error of label-blind residual regression.

Classes with identical means. Let equiprobable classes have $X_1 \mid Y = 0 \sim \mathcal{N}(0,1)$ and $X_1 \mid Y = 1 \sim \mathcal{N}(0,4)$, independently of $X_0 \sim \mathcal{N}(0,1)$. Their optimal MSE prototypes both equal zero. For $a(t) = t$, $b(t) = 1 - t$, and $\sigma_y^2 \in \{1, 4\}$,

$$r_t^y(x) = \frac{t\sigma_y^2 - (1-t)}{t^2\sigma_y^2 + (1-t)^2}\, x. \tag{29}$$

Although the prototypes coincide, the class variances yield different residual fields at interior times. With identical prototypes and a shared base law, a label-blind sampler produces the same law for both labels. A label-conditioned residual captures this additional class dependence with one backbone evaluation at $w_{\mathrm{CFG}} = 1$.

B.3 Joint optimization of the path and velocity

For every fixed ϕ, the preceding identities remain valid. During joint optimization, however, the path X_T^ϕ, its conditional residual, and the variance term can all depend on ϕ:

$$\mathcal{L}_{\mathrm{AuxPath}}(\theta, \phi) = \mathcal{L}_{\mathrm{FM}}(\theta, \phi) + C(\phi). \tag{30}$$

Where differentiable, the path-parameter gradient therefore includes $\nabla_\phi C(\phi)$, which can be nonzero for general learned paths, including label-blind regression. Deterministic label auxiliaries with label-conditioned residuals form a special case: the translation result below makes $C(\phi)$ constant.

Minimizing $\mathbb{E}\|F(Y) - X_1\|^2$ alone gives the class mean by conditional least squares. For a finite residual network, adding the flow loss can change the optimum. At any fixed learned ϕ, endpoint preservation and the fixed-path projection result hold under their moment assumptions.

B.4 Deterministic semantic auxiliaries as translated paths

Fix the joint law of (X_0, X_1, Y) and coefficient functions independently of ϕ. Let T be independent and uniform, set $Z_t = a(t)X_1 + b(t)X_0$, $R_t = \dot{a}(t)X_1 + \dot{b}(t)X_0$, and assume the required square integrability. Define $h_\phi(t, y) = c(t)F_\phi(y)$ and $X_t^\phi = Z_t + h_\phi(t, Y)$.

Proposition 3 (Conditional translation and invariant regression variance). *Let $r_t^0(z, y) = \mathbb{E}[R_t \mid Z_t = z, Y = y]$. When the residual network observes Y, its optimal field on the auxiliary path is*

$$r_t^\phi(x, y) = r_t^0(x - h_\phi(t, y), y), \qquad u_t^\phi(x, y) = r_t^\phi(x, y) + \partial_t h_\phi(t, y). \tag{31}$$

The Bayes residual risk is independent of ϕ:

$$C_R(\phi) = \mathbb{E}\|R_T - r_T^\phi(X_T^\phi, Y)\|^2 = \mathbb{E}\|R_T - r_T^0(Z_T, Y)\|^2. \tag{32}$$

The same invariance holds for the conditional full-velocity regression variance.

Proof. Given (T, Y), the translation $Z_T \mapsto X_T^\phi$ is invertible. Hence $\sigma(T, X_T^\phi, Y) = \sigma(T, Z_T, Y)$ and their conditional expectations of R_T coincide. This gives the residual identity and its invariant variance. The added derivative $\partial_t h_\phi(T, Y)$ is measurable with respect to either conditioning sigma-algebra, so it cancels from the full-velocity regression error. □

If the exact ODEs are well posed, $x_t = z_t + h_\phi(t, y)$ transforms $\dot{z}_t = r_t^0(z_t, y)$ into $\dot{x}_t = u_t^\phi(x_t, y)$. Since $h_\phi(0, y) = h_\phi(1, y) = 0$, they have identical endpoint maps. Hence deterministic semantic paths are translated parameterizations of the same ideal label-conditioned endpoint transport.

Consequence for joint MSE training. For unrestricted label-conditioned residual functions v, minimizing Eq.(38) over v yields

$$\inf_v \mathcal{L}_{\text{joint}}(v, \phi) = C_R + \lambda_{\text{prior}} \mathbb{E}\|F_\phi(Y) - X_1\|^2. \tag{33}$$

For $\lambda_{\text{prior}} > 0$ and an unrestricted auxiliary function, the population optimum is the class mean; at $\lambda_{\text{prior}} = 0$, all admissible auxiliaries share the same optimal residual risk. In finite networks, approximation error and optimization dynamics can depend on the translation, providing the setting studied in our experiments. The invariance relies on observing Y, a deterministic auxiliary given Y, and an endpoint law fixed independently of ϕ. Label-blind, additionally randomized, or jointly changing endpoint constructions instead follow the general identity in Appendix B.3.

C Compositional and Degenerate Auxiliary Variables

C.1 Compositional Auxiliary Distributions

Distributional flexibility includes mixtures $\sum_k \pi_k p_\eta^{(k)}$, with $\pi_k \geq 0$ and $\sum_k \pi_k = 1$, provided the resulting joint law has the requisite first moments (and space-time second moments for MSE). For an infinite mixture, finite moments of each component separately do not ensure a finite mixture moment.

C.2 Deterministic Auxiliary Variables

Deterministic auxiliaries are also covered: for $\eta = F(X_0)$ the path is $X_t = a(t)X_1 + b(t)X_0 + c(t)F(X_0)$. The joint law may be singular, but the measure proof applies whenever the moments hold. Its marginal field is

$$u_t(x) = \mathbb{E}[\dot{a}(t)X_1 + \dot{b}(t)X_0 + \dot{c}(t)F(X_0) \mid X_t = x]. \tag{34}$$

A deterministic auxiliary changes the intermediate path while retaining the endpoint laws. For sufficiently regular F, its contribution can also be absorbed into a deterministic interpolant. The joint-law formulation additionally covers auxiliaries that lack this regularity or retain randomness conditional on the endpoints.

D Auxiliary Guidance

D.1 Branch velocities and two guidance scales

For a label-conditioned residual, define $v_y = v_\theta(x, t, y) + \dot{c}F_\phi(y)$ and $v_\emptyset = v_\theta(x, t, \emptyset) + \dot{c}F_\phi(\emptyset)$.

Proposition 4 (Matched-branch interpretation). *For $s \in \{y, \emptyset\}$, let $X_t^s = Z_t + c(t)F_\phi(s)$ denote the path associated with that branch, where $Z_t = a(t)X_1 + b(t)X_0$ and $R_t = \dot{a}(t)X_1 + \dot{b}(t)X_0$. If*

$$v_\theta(x, t, s) = \mathbb{E}[R_t \mid X_t^s = x, s]$$

under the distribution assigned to branch s, then v_s is its exact conditional velocity. Consequently, equal guidance scales combine two exact branch velocities by the standard CFG rule.

Proof. Because $\dot{X}_t^s = R_t + \dot{c}(t)F_\phi(s)$ and the second term is branch-measurable, conditioning on (X_t^s, s) gives $\mathbb{E}[\dot{X}_t^s \mid X_t^s = x, s] = v_s$. □

Separating guidance scales gives

$$\begin{aligned} v^{\text{dual}}(x, t, y) = v_\theta(x, t, \emptyset) + w_{\text{CFG}}[v_\theta(x, t, y) - v_\theta(x, t, \emptyset)] \\ + \dot{c}(t)[F_\phi(\emptyset) + w_{\text{aux}}(F_\phi(y) - F_\phi(\emptyset))]. \end{aligned} \tag{35}$$

If $w_{\text{CFG}} = w_{\text{aux}} = w$, the expression equals $v_\emptyset + w(v_y - v_\emptyset)$; Proposition 4 gives its exact two-branch velocity interpretation. At $w_{\text{CFG}} = 1$, the residual reduces to $v_\theta(x, t, y)$, requiring one backbone evaluation even when $w_{\text{aux}} > 1$. At $w_{\text{aux}} = 1$ one auxiliary evaluation suffices; otherwise both auxiliary branches are used in the reported accounting. Conventional non-unit backbone guidance uses two backbone evaluations.

For a label-blind residual, the backbone terms coincide and w_{CFG} has no effect. In this case, auxiliary guidance is algebraically equivalent to CFG applied to the surrogate fields v_y and $v_\emptyset$; Eq.(28) states the approximation limit of those fields. Algorithms 3–5 use $s = y$ for a conditioned backbone and omit s for the path-only variant. Algorithm 3 shows two-stage training; joint training also differentiates through the auxiliary-dependent path and includes the prior loss.

Null branch and marginalization. We set $F_\phi(\emptyset) = \mu = \mathbb{E}[X_1]$, the global data mean. Auxiliary guidance uses $\mu + w_{\text{aux}}(F_\phi(y) - \mu)$ to scale the class displacement from this reference. The marginal velocity of the semantic training path is

$$u_t(x) = \mathbb{E}[R_t \mid X_t = x] + \dot{c}(t)\mathbb{E}[F_\phi(Y) \mid X_t = x], \tag{36}$$

where the posterior auxiliary mean depends on the state, unlike the fixed reference μ. Pairing $F_\phi(\emptyset) = \mu$ with a residual matched to the null-condition path yields the exact unconditional branch characterized by Proposition 4.

Guided sampling dynamics. The two scales independently control the residual and auxiliary components of the sampling velocity. Changing the auxiliary drift also changes the states at which the trained residual is evaluated. Consequently, the guided dynamics determine the output distribution; endpoint preservation of the training interpolation alone does not identify it. We evaluate these dynamics through generation quality and computational cost.

D.2 Auxiliary Guidance after Two-Stage Training

Table 5 reports auxiliary guidance after two-stage training with DiT-B/2 Peebles & Xie (2023). With $w_{\text{CFG}} = 1$, increasing w_{aux} from 1.0 to 2.0 lowers FID from 22.49 to 19.81 while retaining one backbone evaluation per step. Joint-training and standard-CFG results are reported in Table 4.

Table 5: Auxiliary guidance after two-stage training on ImageNet-1k, with $w_{\text{CFG}} = 1$. Pass counts backbone evaluations per step; GFLOPs are per image over 50 steps.

w_{aux}	Pass	GFLOPs	FID ↓
1.0	1	2301.60	22.49
1.5	1	2303.08	20.43
2.0	1	2303.08	19.81

D.3 AUXILIARY GUIDANCE ON CIFAR-10

We evaluate auxiliary guidance on CIFAR-10 using two-stage training (Algorithm 3) and auxiliary-guided sampling (Algorithm 5). The backbone guidance scale is fixed at $w_{\mathrm{CFG}} = 1$, so all settings require one backbone evaluation per step. Table 6 reports results at four auxiliary guidance scales.

Table 6: Effect of auxiliary guidance on CIFAR-10 with $w_{\mathrm{CFG}} = 1$. Bold marks the best value in each metric.

w_{aux}	FID ↓	KID ↓	IS ↑	Acc (%) ↑
1.0	4.14	0.0026	9.24	83.68
1.2	**3.85**	0.0024	9.27	84.47
1.5	4.10	**0.0022**	9.27	85.39
2.0	6.36	0.0029	**9.34**	**85.82**

Results. Increasing w_{aux} from 1.0 to 2.0 raises label accuracy from 83.68% to 85.82%, a gain of 2.14 percentage points. Moderate guidance improves both alignment and distributional quality: at 1.2, accuracy rises to 84.47% and FID falls from 4.14 to 3.85; at 1.5, KID reaches its minimum of 0.0022. At 2.0, accuracy and IS are highest, while FID rises to 6.36. Thus, the auxiliary scale adjusts label alignment and sample quality without adding a second backbone evaluation, with moderate values preferred when both FID and label agreement matter.

D.4 JOINT EVALUATION OF THE TWO GUIDANCE SCALES

Table 7 separates the two guidance scales. Raising w_{aux} from 1.0 to 1.2 lowers FID from 22.49 to 21.47 at $w_{\mathrm{CFG}} = 1$, and from 12.75 to 11.87 at $w_{\mathrm{CFG}} = 1.2$. Auxiliary guidance therefore improves FID at both backbone settings in this grid. Enabling standard CFG produces the larger reduction but adds a second backbone evaluation. Combining both scales gives the grid minimum, 11.87 FID at 4603.20 GFLOPs, whereas auxiliary-only guidance uses 2303.08 GFLOPs.

Table 7: Joint 2×2 evaluation of auxiliary guidance and standard CFG on ImageNet-1k using DiT-B/2. GFLOPs are per image over 50 sampling steps.

w_{aux}	w_{CFG}	Pass	GFLOPs	FID ↓
1.0	1.0	1	2301.60	22.49
1.0	1.2	2	4601.72	12.75
1.2	1.0	1	2303.08	21.47
1.2	1.2	2	4603.20	11.87

D.5 EQUAL-PEAK AUXILIARY COEFFICIENT SHAPES

We compare coefficient shapes on ImageNet-1k at 256×256 with $\alpha = 1$ and $\lambda_{\mathrm{prior}} = 1.0$. Write $c_\alpha(t) = \alpha h(t)$, where all profiles in Table 8 vanish at the endpoints and reach the same maximum value, $1/4$. Their peak times differ: Quadratic, Sine, Sine squared, and Quartic peak at $t = 1/2$, while Early and Late peak at $t = 1/3$ and $2/3$, respectively.

Table 8: Equal-peak coefficient profiles and their derivatives. The actual coefficient and derivative are $c_\alpha = \alpha h$ and $\dot{c}_\alpha = \alpha h'$.

Profile	$h(t)$	$h'(t)$	Peak time
Quadratic	$t(1-t)$	$1-2t$	$1/2$
Sine	$\frac{1}{4}\sin(\pi t)$	$\frac{\pi}{4}\cos(\pi t)$	$1/2$
Sine squared	$\frac{1}{4}\sin^2(\pi t)$	$\frac{\pi}{4}\sin(2\pi t)$	$1/2$
Quartic	$4[t(1-t)]^2$	$8t(1-t)(1-2t)$	$1/2$
Early	$\frac{27}{16}t(1-t)^2$	$\frac{27}{16}(1-t)(1-3t)$	$1/3$
Late	$\frac{27}{16}t^2(1-t)$	$\frac{27}{16}t(2-3t)$	$2/3$

Equal peak amplitude controls the maximum displacement coefficient. As Figure 6 shows, this leaves both the peak time and derivative profile free to vary, changing when and how strongly the auxiliary affects the velocity.

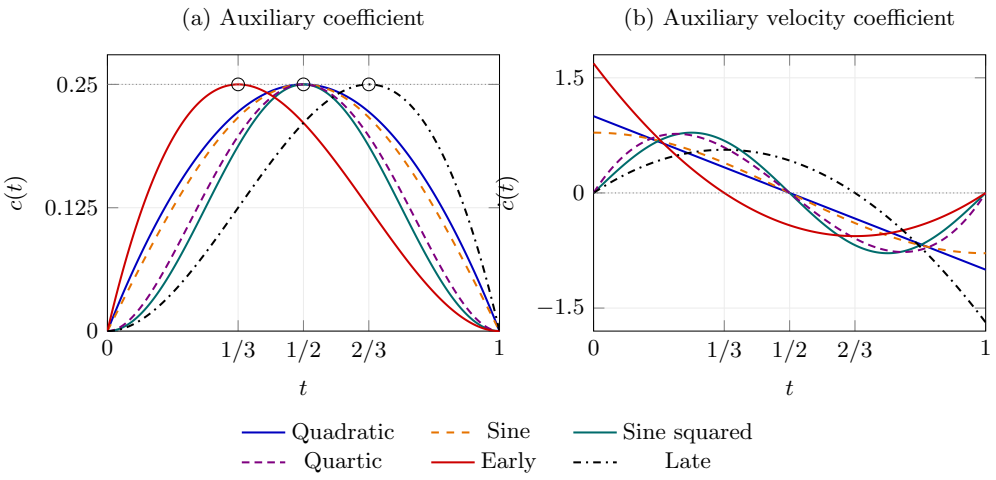


Figure 6: **Auxiliary coefficients and derivatives at $\alpha = 1$.** All six $c(t)$ profiles have peak value 0.25. Circles mark the three peak times $1/3$, $1/2$, and $2/3$; four symmetric profiles share the middle point. The derivative curves show how equal-peak profiles distribute auxiliary velocity differently over time.

Table 9: ImageNet-1k FID for equal-peak auxiliary coefficients at $\alpha = 1$ and $\lambda_{\text{prior}} = 1.0$. Columns sweep w_{aux} with $w_{\text{CFG}} = 1$. The Quadratic row reproduces the corresponding configuration in Table 4. Bold marks the lowest FID in each row.

	w_{aux}						
Profile	1.0	1.5	2.0	2.5	3.0	3.5	4.0
Quadratic	21.84	19.13	**18.01**	18.33	19.95	22.94	27.06
Sine	22.16	19.60	**18.40**	18.75	20.24	23.05	26.93
Sine squared	22.54	20.66	19.64	**19.31**	19.66	20.71	22.48
Quartic	22.47	20.50	19.42	**19.17**	19.70	21.05	23.12
Early	22.30	20.88	19.92	19.54	**19.50**	20.03	21.06
Late	21.06	18.96	**18.19**	19.01	21.52	25.55	30.53

Results. The Quadratic profile attains the lowest FID in this comparison, 18.01 at $w_{\text{aux}} = 2.0$. Sine and Late also reach their minima at 2.0, with FIDs of 18.40 and 18.19. Sine squared and Quartic favor 2.5, attaining 19.31 and 19.17, while Early reaches 19.50 at 3.0. At $w_{\text{aux}} = 4.0$, Early gives the lowest FID (21.06), followed by Sine squared (22.48) and Quartic (23.12); Late rises to 30.53. Late therefore performs well at moderate guidance but deteriorates more sharply at larger scales, whereas Early retains better quality at the upper end. Equal peak amplitude does not fix either the preferred guidance scale or the response to stronger guidance.

E From Unconditional to Conditional Generation via Auxiliary Paths

We next test whether a semantic auxiliary can add label control to a pre-trained unconditional flow model without changing its backbone architecture.

Method. Starting from an unconditional velocity model $v_\theta(x, t)$, we fine-tune on auxiliary-shifted states using a label-dependent map $F(y)$. This map broadcasts the label for numeric auxiliaries, $F(y) = y\mathbf{1}$, or uses a learned semantic module, $F(y) = F_\phi(y)$. The backbone regresses the residual target, while the auxiliary supplies the label-dependent path displacement. At inference time, conditional generation is achieved by augmenting the dynamics as

$$v(x, t, y) = v_\theta(x, t) + \dot{c}(t)F(y), \tag{37}$$

without modifying the original model architecture.

Results. Starting from the same pre-trained unconditional model, we fine-tune with the auxiliary module. Table 10 reports conditional accuracies of 79.18% and 77.03% for numeric and learned auxiliaries. Their FIDs are 5.73 and 5.90, compared with 3.72 for the unconditional checkpoint. The auxiliary path therefore introduces label control while retaining the backbone architecture, at the cost of higher FID than the unconditional checkpoint.

Table 10: Unconditional-to-conditional fine-tuning results.

Method	Acc(%) ↑	FID ↓	KID ↓	IS ↑
Unconditional	N/A	3.72	0.0017	9.06
+ numeric η	79.18	5.73	0.0038	8.89
+ learned η	77.03	5.90	0.0043	8.93

Discussion. These path-only settings, also reported in Table 2, add label control without a label input to the backbone. Numeric auxiliaries give higher accuracy and lower FID than learned auxiliaries in this experiment. Adaptation therefore depends on the interaction between the auxiliary representation and the residual model.

F Text-to-Image Validation with AuxPath-LoRA

We also test auxiliary paths in latent text-to-image generation. We adapt the 600M-parameter SANA backbone Xie et al. (2024) at 512×512 on 20,000 cached COCO pairs Lin et al. (2014) using LoRA Hu et al. (2021). Evaluation covers 30 held-out compositional prompts with 129 atomic checks following TIFA and Davidsonian Scene Graph protocols Hu et al. (2023); Cho et al. (2023). All samples share prompts, seeds, CFG scale 4.5, Euler sampling, and 20 steps. A local BLIP-VQA-base model Li et al. (2022) supplies yes/no judgments, so the VQA and semantic scores are local proxies rather than official benchmark results.

Relative to Standard LoRA, $w_{\text{aux}} = 1$ raises Atom and All-Correct by 1.55 and 3.33 points, while $w_{\text{aux}} = 2$ raises them by 4.65 and 13.33 points. The stronger setting also gives the highest Object, Attr., VQA, and VIE scores and matches Standard LoRA on Relation.

Because both AuxPath rows use the same trained checkpoint, their difference isolates sampling-time auxiliary strength. Raising w_{aux} from 1 to 2 increases All-Correct from 43.33 to 53.33 and also raises Atom and Object, showing that the auxiliary drift remains controllable after LoRA adaptation.

Table 11: Text-to-image prompt-following results on 30 held-out prompts. "$w_{\text{aux}} = 1$" and "$w_{\text{aux}} = 2$" evaluate the same AuxPath-trained LoRA at two auxiliary strengths. The two Δ rows report absolute percentage-point changes relative to Standard LoRA for every metric. The atomic checks use a TIFA/DSG-style decomposition Hu et al. (2023); Cho et al. (2023).

Method	Atom	All	Object	Attr.	Relation	VQA	VIE
Standard LoRA	80.62	40.00	90.20	77.27	67.74	90.00	85.31
AuxPath-LoRA ($w_{\text{aux}} = 1$)	82.17	43.33	92.16	81.82	**70.97**	86.67	84.42
AuxPath-LoRA ($w_{\text{aux}} = 2$)	**85.27**	**53.33**	**96.08**	**86.36**	67.74	**93.33**	**89.30**
Δ AuxPath $w = 1$ vs. Standard	+1.55	+3.33	+1.96	+4.55	+3.23	−3.33	−0.89
Δ AuxPath $w = 2$ vs. Standard	+4.65	+13.33	+5.88	+9.09	+0.00	+3.33	+3.99

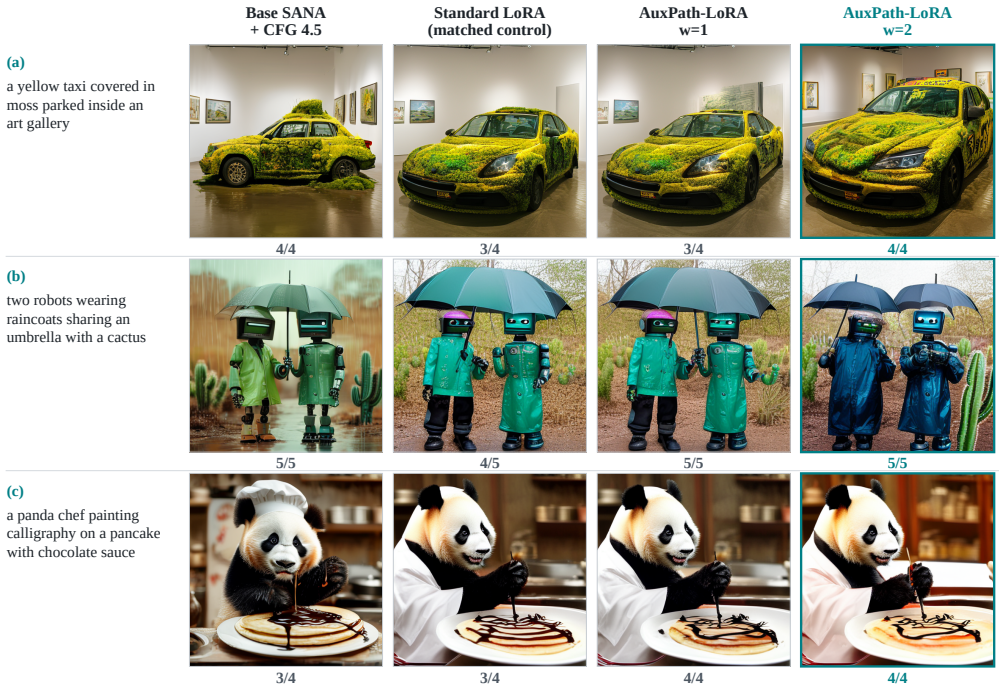
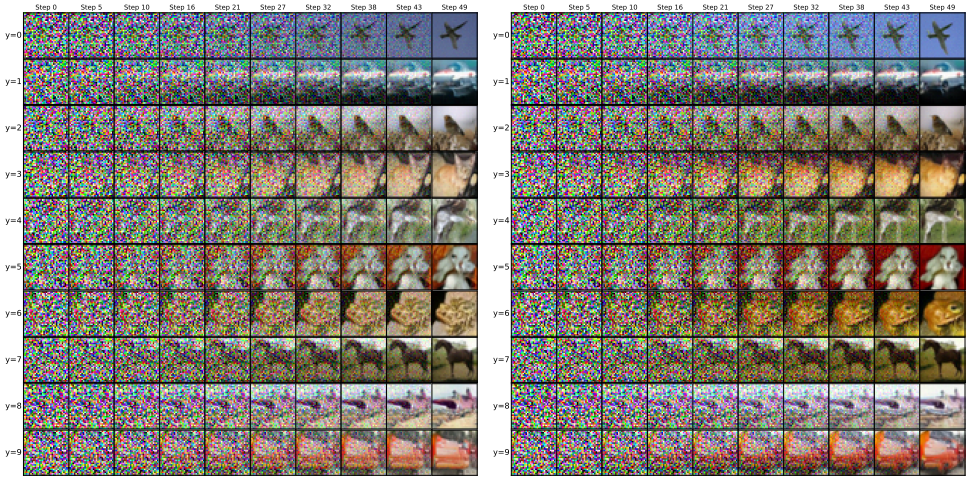


Figure 7: Qualitative text-to-image comparison under matched prompts, seeds, CFG scale, sampler, and denoising steps. The four columns are Base SANA, Standard LoRA, AuxPath-LoRA with $w_{\text{aux}} = 1$, and the same AuxPath-LoRA with $w_{\text{aux}} = 2$. The numbers below each sample are the correctly satisfied atomic checks out of the prompt-specific total according to the local BLIP-VQA evaluator Li et al. (2022).

Reproducibility. The accompanying experiment repository records the checkpoint configurations, prompt set, sampling protocol, alignment evaluator, and aggregation procedure used to produce Table 11. We report the resulting aggregate values directly rather than embedding machine-specific invocation details in the paper.

The reported scope is one training seed, one sampling seed, 30 prompts, and a local evaluator. We omit counting accuracy because only three atomic counting questions make each answer worth 33.33 points.



(a) Trajectory with $w = 1.0$ (b) Trajectory with $w = 2.0$

Figure 8: **Generation trajectories of AuxPath-FM with auxiliary guidance on CIFAR-10.** Each row shows the evolution of a sample along the probability path from the initial state ($t = 0$) to the final generated image ($t = 1$). The panels compare auxiliary guidance scales $w_{\text{aux}} = 1.0$ and 2.0, with one backbone evaluation per sampling step.

G Additional Qualitative Results

Figure 8 shows how auxiliary scale changes CIFAR-10 trajectories, while Figure 9 compares ImageNet-1k trajectories across auxiliary distributions. Together, they connect the main quantitative results to the intermediate states induced by path design.

H Anonymous Code Release and Reproducibility

An anonymous implementation of AuxPath-FM is available at:

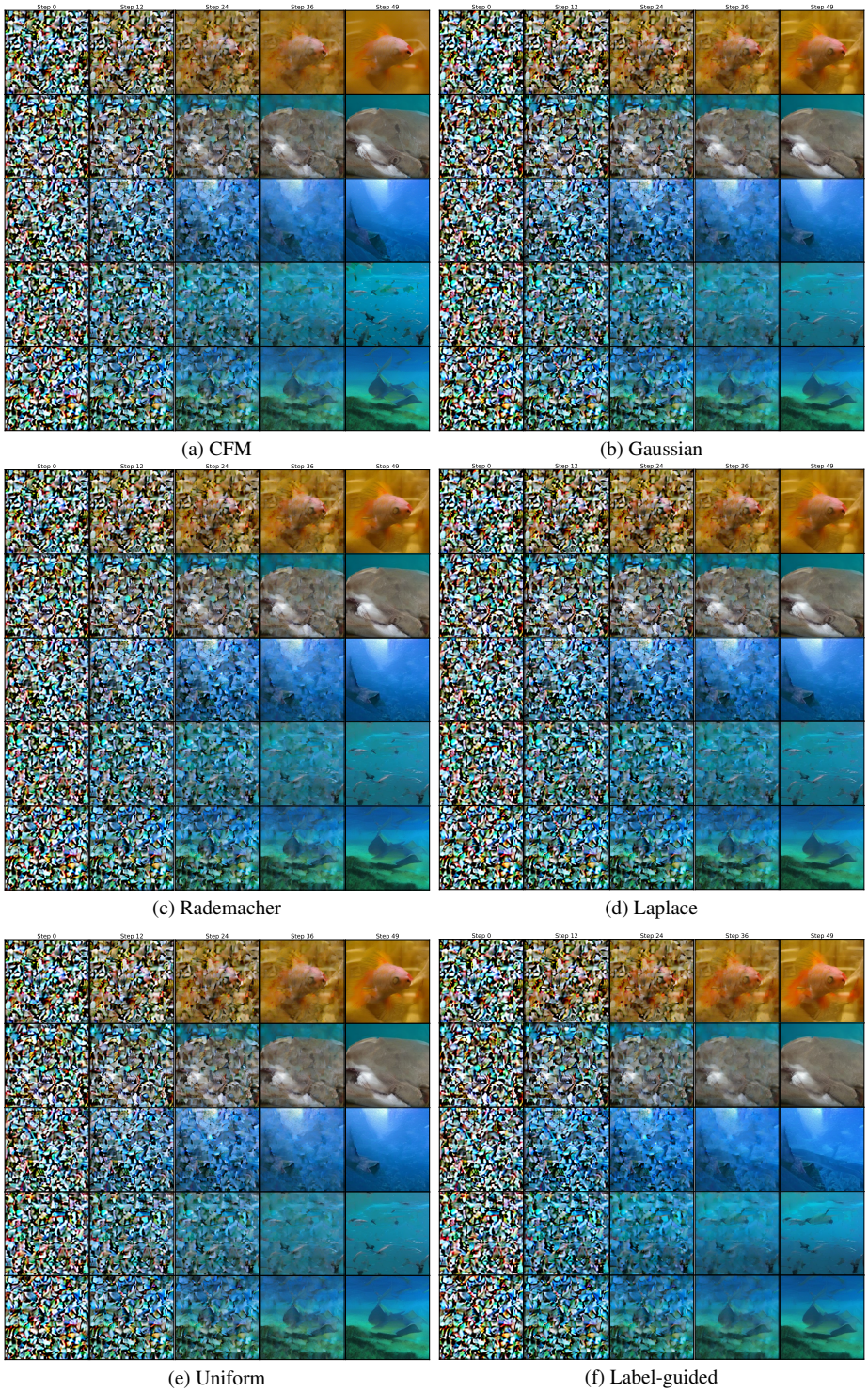
`https://anonymous.4open.science/r/AuxPath-FM-9DFE`

Evaluation protocol. We report single runs with fixed backbones, solvers, and coefficient schedules. The measured differences characterize these runs; training-seed variability remains unestimated. The ImageNet-1k joint-training study uses 256×256 images.

Implementation. The repository contains the core auxiliary-path implementation, including auxiliary construction and path coefficients. The null auxiliary is the global mean in the model input space (latent space for latent models). The ImageNet-1k training entry point is `train_imagenet256_DIT_AFM.py`; the experimental objective and matched baseline settings are specified below.

H.1 ImageNet-1k Fine-Tuning and Matched CFM Baseline

The ImageNet-1k experiments use DiT-B/2 at 256×256, initialized from the same pre-trained CFM checkpoint. Fine-tuning uses a learning rate of 2×10^{-5}, bfloat16 precision, and the 8-bit



(a) CFM (b) Gaussian

(c) Rademacher (d) Laplace

(e) Uniform (f) Label-guided

Figure 9: **Effect of auxiliary perturbations in CFM.** We visualize trajectories under standard CFM and different auxiliary signals using $t(1-t)\eta$. Gaussian, Rademacher, Laplace, and Uniform perturbations correspond to stochastic noise injections, while label-guided signals use $\eta = F_\phi(Y)$ for semantic conditioning. All variants use the same probability flow construction.

AdamW (`AdamW8bit`) optimizer. Each matched baseline shares the training budget (300K updates for the auxiliary-distribution comparison), data preprocessing, batch size, learning-rate schedule, and checkpoint-selection rule of its corresponding auxiliary-path run. Joint-training comparisons use matching update budgets and training settings.

For joint training in Table 4, the auxiliary prior uses MSE in the latent representation. With $X_t^\phi = tX_1 + (1-t)X_0 + c_\alpha(t)F_\phi(Y)$ and $R = X_1 - X_0$, the objective is

$$\mathcal{L}_{\text{joint}}(\theta, \phi) = \mathbb{E}\Big[\|v_\theta(X_t^\phi, t, Y) - R\|^2\Big] + \lambda_{\text{prior}}\mathbb{E}\big[\|F_\phi(Y) - X_1\|^2\big]\,. \tag{38}$$

Gradients from the residual flow loss also reach F_ϕ through X_t^ϕ. The explicit drift $\dot{c}_\alpha(t)F_\phi(Y)$ is added at sampling. Joint optimization adapts the auxiliary to both the prior MSE and the finite backbone's residual regression loss (Appendix B).

For the CFM control, we set $\alpha = 0$ in $c_\alpha(t) = \alpha\, t(1-t)$ and $\lambda_{\text{prior}} = 0$. The auxiliary module is inactive; fine-tuning retains the original CFM conditioning interface and flow objective. In Table 4, the $\alpha = 0$ row sweeps standard w_{CFG}, whereas the $\alpha > 0$ rows sweep w_{aux} with $w_{\text{CFG}} = 1$. Standard CFG combines conditional and unconditional backbone predictions, requiring two backbone evaluations per step for $w_{\text{CFG}} > 1$; at $w_{\text{CFG}} = 1$, only the conditional prediction is needed. Auxiliary guidance uses one backbone evaluation plus the auxiliary computation. The standard-CFG sweep is evaluated from the fine-tuned CFM checkpoint using the same solver, 50 sampling steps, 50k generated samples, and evaluation pipeline as the compared method.

Tables 3 and 4 report the fine-tuned CFM results; inference costs follow Sec. 5.4.