Learning Straight Flows: Variational Flow Matching for Efficient Generation

Supplementary Material

1. Proofs for Straight, Non-Intersecting Interpolations

This appendix provides a self-contained development of the analytical concepts and proofs underlying our method. We introduce a *straight, non-intersecting interpolation* Z that is *compatible* with the linear interpolation induced by a given coupling X, and we establish the following properties: (i) preservation of time marginals, (ii) reduction of convex transport costs, (iii) equivalent characterizations of straight (non-intersecting) couplings, and (iv) equivalence with vanishing time derivative.

Notation. For a random process $X = \{X_t\}_{t\in[0,1]}$, write $\mathrm{Law}(X_t)$ for its marginal law at time t, and $\mathbb{E}[\cdot]$ for expectation. For a coupling (X_0, X_1) we denote $\Delta^X := X_1 - X_0$ and the linear interpolant $X_t = (1-t)X_0 + tX_1$. Conditional expectations such as $\mathbb{E}[\Delta^X \mid X_t = x]$ are understood whenever they exist.

1.1. Preliminaries

Definition 1 (Coupling, linear interpolation, and conditional velocity)**.** Let (X_0, X_1) be any coupling on $\mathbb{R}^d$ with joint density $\rho(x_0, x_1) = \rho_0(x_0)\,\rho_1(x_1)$. Define the linear interpolation

$$X_t = (1-t)X_0 + tX_1, \qquad t \in [0,1]$$
$$\Delta^X = v^X(X_t, t \,|\, (X_0, X_1)) = X_1 - X_0.$$

The *marginal velocity* associated with X is

$$v^X(x,t) \;=\; \mathbb{E}\big[\Delta^X \,\big|\, X_t = x\big], \qquad (x,t) \in \mathbb{R}^d \times [0,1]$$
$$= \int \Delta^X \cdot p(X_0, X_1 \,|\, X_t = x)\, d(X_0, X_1).$$

When $x \notin \mathrm{supp}(X_t)$, we set $v^X(x,t) = 0$. All statements below compare v^X only on sets where it is evaluated against a marginal law: $\mathrm{Law}(X_t)$.

Definition 2 (Rectifiability of X)**.** We say that X is *rectifiable* [2] if $v^X(\cdot, t)$ is locally bounded for each t and the continuity equation

$$\partial_t \pi_t + \nabla\cdot\big(v^X(\cdot,t)\,\pi_t\big) \;=\; 0, \qquad \pi_{t=0} = \mathrm{Law}(X_0),$$

admits a unique solution $\{\pi_t\}_{t\in[0,1]}$. Equivalently, the ordinary differential equation $\dot{X}_t = v^X(X_t, t)$ admits a unique flow of characteristics.

Definition 3 (Non-intersection functional [2])**.** For any coupling (X_0, X_1) with linear interpolant X_t, define

$$V\big((X_0, X_1)\big) \;:=\; \int_0^1 \mathbb{E}\Big[\big\|\, \Delta^X - \mathbb{E}[\Delta^X \mid X_t]\,\big\|^2\Big]\; dt.$$

Lemma 1 (Non-intersection if and only if zero conditional variance)**.** *For a coupling (X_0, X_1) with linear interpolation X_t, the following are equivalent:*

1. *For two independent identically distributed couplings (X_0, X_1) and (X_0', X_1'),*

$$\exists t \in (0,1) : (1-t)X_0 + tX_1 = (1-t)X_0' + tX_1'$$
$$\mathbb{P}[(X_0, X_1) \neq (X_0', X_1')] = 0.$$

2. $V\big((X_0, X_1)\big) = 0$*; equivalently* $\Delta^X = \mathbb{E}[\Delta^X \mid X_t]$ *for* $t \in [0,1]$.

Proof. $(1) \Rightarrow (2)$: Non-intersection implies that the slope Δ^X is a measurable function of (X_t, t), hence $\mathrm{Var}(\Delta^X \mid X_t) = 0$; integrating over t gives $V = 0$. $(2) \Rightarrow (1)$: If $\Delta^X = \mathbb{E}[\Delta^X \mid X_t]$, then the slope through any X_t is unique, two distinct lines cannot share a point at the same time unless they coincide. □

1.2. The straight interpolation Z compatible with X

Definition 4 (Straight interpolation compatible with X)**.** A process $Z = \{Z_t\}_{t\in[0,1]}$ on the same probability space as X is called a *straight interpolants* compatible with Flow Matching trajectories X if the following hold:

(Z1) **Linear paths.** There exist random endpoints (Z_0, Z_1) with $Z_t = (1-t)Z_0 + tZ_1$ and $\Delta^Z := Z_1 - Z_0$.

(Z2) **Non-intersection.** $V\big((Z_0, Z_1)\big) = 0$. Equivalently, $\Delta^Z = \mathbb{E}[\Delta^Z \mid Z_t]$ for $t \in [0,1]$.

(Z3) **Velocity-field matching.** With $\mu_t = \mathrm{Law}(Z_t)$, for $x \sim \mu_t$ and $t \in [0,1]$.

$$v^Z(x,t) := \mathbb{E}[\Delta^Z \mid Z_t = x] = v^X(x,t).$$

(Z4) **Initialization.** $Z_0 = X_0$.

Remark 2 (Immediate consequences). From (Z2), $v^Z(Z_t, t) = \Delta^Z$, hence

$$Z_1 - Z_0 = \int_0^1 v^Z(Z_t, t)\, dt.$$

By (Z3), $v^Z = v^X$ at $x \sim \mu_t$, so we also have

$$\Delta^Z = \int_0^1 v^X(Z_t, t)\, dt.$$

1.3. Main results

Theorem 3 (Marginal preservation)**.** *Assume X is rectifiable and Z satisfies Definition 4. Then* $\mathrm{Law}(Z_t) = \mathrm{Law}(X_t)$ *for all* $t \in [0,1]$.

Proof. Let $h : \mathbb{R}^d \to \mathbb{R}$ be any compactly supported continuously differentiable test function. Using (Z1) and (Z2),

$$\frac{d}{dt}\mathbb{E}[h(Z_t)] = \mathbb{E}[\nabla h(Z_t)^\top \dot{Z}_t] = \mathbb{E}[\nabla h(Z_t)^\top \Delta^Z] = \mathbb{E}[\nabla h(Z_t)^\top v^Z(Z_t, t)].$$

By (Z3), $v^Z(\cdot, t) = v^X(\cdot, t)$ for μ_t, therefore

$$\frac{d}{dt}\mathbb{E}[h(Z_t)] = \mathbb{E}[\nabla h(Z_t)^\top v^X(Z_t, t)].$$

This shows that $\mu_t = \mathrm{Law}(Z_t)$ solves the continuity equation $\partial_t \mu_t + \nabla \cdot \big(v^X(\cdot, t)\mu_t\big) = 0$ with $\mu_0 = \mathrm{Law}(Z_0) = \mathrm{Law}(X_0)$ by (Z4). By rectifiability of X, the solution is unique and equals $\pi_t = \mathrm{Law}(X_t)$. Hence $\mu_t = \pi_t$ for all $t \in [0, 1]$. □

Theorem 4 (Convex transport-cost reduction)**.** *Under the assumptions of Theorem 3, for any convex $c : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathbb{E}[c(Z_1 - Z_0)] \;\le\; \mathbb{E}[c(X_1 - X_0)]\,.$$

If c is strictly convex, equality holds if and only if $V\big((X_0, X_1)\big) = 0$.

Proof. From (Z2),

$$Z_1 - Z_0 = \int_0^1 v^Z(Z_t, t)\, dt.$$

By Jensen's inequality in time,

$$\mathbb{E}[c(Z_1 - Z_0)] \le \int_0^1 \mathbb{E}\big[c\big(v^Z(Z_t, t)\big)\big]\; dt.$$

Using (Z3) and Theorem 3,

$$\int_0^1 \mathbb{E}\big[c\big(v^Z(Z_t, t)\big)\big]\, dt = \int_0^1 \mathbb{E}\big[c\big(v^X(X_t, t)\big)\big]\, dt = \int_0^1 \mathbb{E}\big[c\big(\mathbb{E}[\Delta^X \mid X_t]\big)\big]\, dt.$$

A conditional Jensen inequality (conditioning on X_t) yields

$$\mathbb{E}\big[c\big(\mathbb{E}[\Delta^X \mid X_t]\big)\big] \le \mathbb{E}\big[\mathbb{E}\big[c(\Delta^X) \mid X_t\big]\big] = \mathbb{E}\big[c(\Delta^X)\big]\,.$$

Combining gives the inequality. If c is strictly convex, equality forces tightness of the conditional Jensen step, hence $\Delta^X = \mathbb{E}[\Delta^X \mid X_t]$, which by Lemma 1 is $V((X_0, X_1)) = 0$. □

Theorem 5 (Equivalent characterizations of straight interpolants)**.** *Assume X is rectifiable and Z satisfies Definition 4. The following are equivalent: when they hold, we say X yielded from couplings (X_0, X_1) are* straight interpolants*, which coincide with Z.*

(i) ***Tight convex cost.*** *There exists a strictly convex c with $\mathbb{E}[c(Z_1 - Z_0)] = \mathbb{E}[c(X_1 - X_0)]$.*
(ii) ***Endpoint coincidence.*** *$(Z_0, Z_1) = (X_0, X_1)$.*
(iii) ***Pathwise equality.*** *$Z_t = X_t$ for all $t \in [0, 1]$.*
(iv) ***Non-intersection for X.*** *$V\big((X_0, X_1)\big) = 0$.*

Proof. 1. (iii) $\implies$ (ii) is immediate.
2. (ii) $\implies$ (i) turns the inequality of Theorem 4 into equality.
3. (i) $\implies$ (iv): In Theorem 4 with strictly convex c, equality forces tightness of the conditional Jensen step; hence $\Delta^X = \mathbb{E}[\Delta^X \mid X_t]$ for $t \in [0, 1]$, which is $V((X_0, X_1)) = 0$ by Lemma 1.
4. (iv) $\implies$ (iii): If $V((X_0, X_1)) = 0$, then for $t \in [0, 1]$, $\dot{X}_t = \Delta^X = v^X(X_t, t)$. For Z, (Z2)–(Z3) give $\dot{Z}_t = \Delta^Z = v^Z(Z_t, t) = v^X(Z_t, t)$. Hence both X and Z solve $\dot{Y}_t = v^X(Y_t, t)$ with the same initial value $Y_0 = X_0 = Z_0$ by (Z4). Uniqueness of rectifiability in Definition 2 yields $Z \equiv X$.

□

Corollary 6 (One-step generation)**.** *If any, and therefore all, of the items in Theorem 5 hold, then along each path $v^X(X_t, t) \equiv \Delta^X$ is constant in t and the ordinary differential equation $\dot{y}_t = v^X(y_t, t)$ integrates in one step:*

$$X_1 \;=\; X_0 + \int_0^1 v^X(X_t, t)\, dt \;=\; X_0 + \Delta^X.$$

1.4. Equivalence with vanishing time derivative

Definition 5 (Time derivative)**.** For a differentiable vector field $v : \mathbb{R}^d \times [0, 1] \to \mathbb{R}^d$, the time derivative along its characteristics is

$$\frac{d}{dt}v(x, t) = D_t v(x, t) \;:=\; \frac{\partial v}{\partial t}\frac{dt}{dt} \;+\; \frac{\partial v}{\partial x}\frac{dx}{dt} = \frac{\partial}{\partial t}v(x, t) \;+\; \big(\nabla_x v(x, t)\big)\, v(x, t)$$

Theorem 7 (Straightness if and only if vanishing time derivative along X)**.** *Assume X is rectifiable and Z satisfies Definition 4. Assume moreover that v^X is continuously differentiable in (x, t). Then $V\big((X_0, X_1)\big) = 0$, if and only if $D_t v^X(X_t, t) = 0$, for $t \in [0, 1]$.*

Proof. $\Rightarrow$: If $V = 0$, Theorem 5 gives $Z \equiv X$ and $v^X(X_t, t) = \Delta^X$, which is constant in t along each path. The chain rule implies $D_t v^X(X_t, t) = \partial_t v^X(X_t, t) + \nabla v^X(X_t, t)\, \dot{X}_t = 0$.

$\Leftarrow$: If $D_t v^X(X_t, t) = 0$ for $t \in [0, 1]$, then $v^X(X_t, t)$ is constant in t along each path. With $Z_0 = X_0$, the solution has the linear form $X_t = (1 - t)X_0 + tX_1$ and distinct trajectories cannot intersect because the ordinary differential equation with a well-posed vector field has unique solutions. By Theorem 3, $\mathrm{Law}(Z_t) = \mathrm{Law}(X_t)$ for all t. Applying Theorem 5 then yields $V((X_0, X_1)) = 0$. □

2. Implementation

2.1. Model Setup

For the CIFAR-10 experiments, we construct the generative model v_θ using the EDM framework [1] with the DDPM++ architecture [4]. For the ImageNet 256×256 experiments, we adopt the same architecture as SiT-XL/2 [3] for the generative model v_θ.

Across both settings, the primary difference lies in how the input latent code z is incorporated. As discussed previously, we identify two effective conditioning mechanisms: *adaptive normalization*, where z is added to the time embedding prior to computing the shift and offset parameters, and *bottleneck sum*, where the latent is fused with intermediate activations at the lowest resolution using a weighted sum before upsampling.

The latent code z is produced by two MLP layers and serves as a conditioning signal for the velocity network v_θ. For the variational autoencoder q_ϕ, we use the same backbone architecture as the generative model v_θ, but with reduced block depth and modified output layers to produce the latent code z.

2.2. Model Configurations

For the CIFAR-10 experiments, the configuration of the model is shown in Table 1. For the ImageNet 256×256 experiments, the configuration of the model is shown in Table 2.

2.3. Hardware Specifications

For the CIFAR-10 experiments, we use 8× A100 (80G) GPU for all settings. For the ImageNet 256×256 experiments, we use 16× A100 (80G) GPU for all settings.

3. Ablation Study

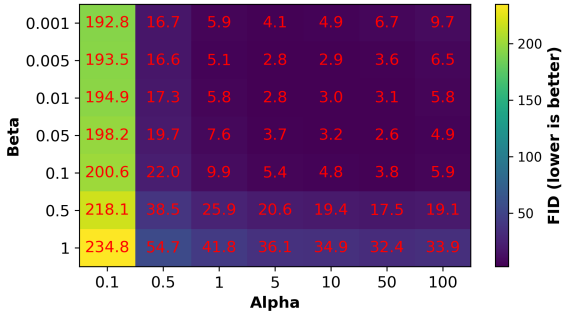

Figure 1. **Ablation Study Results by Adjusting α and β on CIFAR-10 Dataset.** The value in the grid represents the FID score of one-step generation (NFE = 1), lower is better.

To evaluate the contribution of each hyperparameter to S-VFM, we conduct an ablation study by evaluating the generation performance while varying the hyperparameters α and β. Figure 1 presents the one-step generation FID scores for different combinations of α and β on the CIFAR-10 dataset in a heatmap format, where lower values indicate better performance. The results show that S-VFM exhibits a stable performance plateau around the optimal region, suggesting the method is relatively insensitive to hyperparameter variations and demonstrating its adaptive robustness.

4. Discussion on Broader Applications

The theoretical and empirical advantages of *Straight Variational Flow Matching* (S-VFM) suggest potential extensions across several frontier generative tasks. First, our focus on straight, non-intersecting trajectories provides a natural synergy with Dataset Distillation (DD). The efficiency of S-VFM in single-step synthesis can significantly accelerate the generation of synthetic training sets, building upon the representativeness improvements in diffusion-based distillation [8] and the hierarchical amplification strategies of HIERAMP [9]. Second, the inference efficiency of S-VFM makes it a strong candidate for **mobile and streaming applications**, such as the video generation paradigms explored in S2DiT [10], where reducing Function Evaluations (NFE) is critical for real-time deployment.

Furthermore, the variational latent code z in our framework, which captures a global "generation overview", could be adapted for Test-Time Adaptation (TTA) and Visual Prompting. By leveraging the connection between straight flows and Optimal Transport, S-VFM could enhance OT-guided prompting methods like OT-VP [6] or facilitate dynamic adaptation through prompt coresets [7]. Finally, the structural consistency achieved by penalizing velocity field variations is vital for medically accurate image synthesis. S-VFM's stable generation paths could be integrated with expert feedback loops, similar to the Doctor Approved [5] framework, to ensure that the generated content maintains high fidelity to domain-specific constraints.

Parameter	Value	Description
— Model Architecture (S-VFM) —		
img_resolution	32	Input/output spatial size inherited from CIFAR-10
in_channels / out_channels	3 / 3	RGB inputs and velocity outputs share channels
embedding_type	positional	Sinusoidal noise embeddings for EDM schedule
encoder_type / decoder_type	standard / standard	Plain UNet backbone without auxiliary paths
model_channels	128	Base channel width before multipliers
channel_mult	[2, 2, 2]	Per-resolution width multipliers across three scales
num_blocks	3	Residual blocks per resolution stage
dropout	0.1	Activation dropout inside UNet blocks
resample_filter	[1, 1]	FIR filter for up-/downsampling convolutions
latent_dim	768	Latent conditioning width injected via map_latent
phi_hidden_channels	128	Width of latent encoder CNN backbone
phi_num_layers	2	Number of convolutional layers in latent encoder
phi_time_embedding_dim	512	Timestamp embedding dimension for latent encoder
sigma_min / sigma_max	0.002 / 80.0	EDM noise schedule bounds
α	10.0	α in the proposed S-VFM
β	1e-2	β in the proposed S-VFM
— Training / Optimization —		
batch_size	512	Global batch size
use_ema	True	Whether to maintain an exponential moving average of weights
ema_rate	0.9999	EMA decay factor
ema_type	traditional	EMA update style
class_conditional	True	Enables training with class labels
total_training_steps	400000	Number of optimization iterations
velocity_learning_rate	0.0002	Learning rate for velocity (UNet) parameters
phi_learning_rate	0.0002	Learning rate for latent encoder
velocity_weight_decay	0.0	Weight decay applied to velocity network
phi_weight_decay	0.0	Weight decay applied to latent encoder

Table 1. **Configuration of S-VFM on CIFAR-10.** Architecture follows the EDM [1] of the DDPM++ [4] setting.

Parameter	Value	Description
— Model Architecture (S-VFM) —		
`img_resolution`	256×256	Input/output resolution on ImageNet-1k
`params` (M)	677	Number of learnable parameters
`depth`	28	Transformer block layers
`hidden_dim`	1152	Token embedding / MLP width
`num_heads`	16	Attention heads per block
`patch_size`	2×2	Image patch size for tokenization
`latent_dim`	1152	Latent conditioning width injected via `map_latent`
`phi_num_layers`	3	Number of depth
α	10.0	α in the proposed S-VFM
β	1e-2	β in the proposed S-VFM
— Training / Optimization —		
`epochs`	240	Total training epochs on ImageNet-1k
`batch_size`	256	Global batch size
`optimizer`	Adam	Optimizer for all parameters
`lr_schedule`	constant	Learning rate schedule
`learning_rate`	0.0002	Base learning rate
`Adam` (β_1, β_2)	(0.9, 0.95)	Adam momentum parameters
`weight_decay`	0.0	Weight decay
`use_ema`	True	Maintain EMA of model weights
`ema_decay`	0.9999	EMA decay rate
`ema_type`	traditional	EMA update style
`class_conditional`	True	Trained with class labels
`total_training_steps`	400000	Number of optimization iterations

Table 2. **Configuration of S-VFM on ImageNet-1k/256.** Architecture follows the SiT-XL/2 [3] setting.

References

[1] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 3, 4

[2] Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[3] 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*, pages 23–40. Springer, 2024. 3, 5

[4] 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. 3, 4

[5] Janet Wang, Yunbei Zhang, Zhengming Ding, and Jihun Hamm. Doctor approved: Generating medically accurate skin disease images through ai-expert feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 3

[6] Yunbei Zhang, Akshay Mehra, and Jihun Hamm. Ot-vp: Optimal transport-guided visual prompting for test-time adaptation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1122–1132. IEEE, 2025. Oral Presentation. 3

[7] Yunbei Zhang, Akshay Mehra, Shuaicheng Niu, and Jihun Hamm. Dpcore: Dynamic prompt coreset for continual test-time adaptation. In *Forty-second International Conference on Machine Learning (ICML)*, 2025. 3

[8] Lin Zhao, Yushu Wu, Xinru Jiang, Jianyang Gu, Yanzhi Wang, Xiaolin Xu, Pu Zhao, and Xue Lin. Taming diffusion for dataset distillation with high representativeness. In *International Conference on Machine Learning*, pages 77760–77780. PMLR, 2025. 3

[9] Lin Zhao, Xinru Jiang, Xi Xiao, Qihui Fan, Lei Lu, Yanzhi Wang, Xue Lin, Octavia Camps, Pu Zhao, and Jianyang Gu. Hieramp: Coarse-to-fine autoregressive amplification for generative dataset distillation. *arXiv preprint arXiv:2603.06932*, 2026. 3

[10] Lin Zhao, Yushu Wu, Aleksei Lebedev, Dishani Lahiri, Meng Dong, Arpit Sahni, Michael Vasilkovsky, Hao Chen, Ju Hu, Aliaksandr Siarohin, et al. S2dit: Sandwich diffusion transformer for mobile streaming video generation. *arXiv preprint arXiv:2601.12719*, 2026. 3