BEYOND STRAIGHTNESS: NON-CROSSING FLOW MATCHING VIA QUANTILE ALIGNTREE COUPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

The performance of Flow Matching largely depends on the quality of the coupling between the source and target distributions. However, independent coupling often leads to path crossings and local velocity ambiguity, while OT-based couplings typically incur high construction costs. To address this challenge, we propose **Quantile AlignTree Flow Matching (QAT-FM)**, an efficient structured coupling strategy that constructs a hierarchical coupling between a Gaussian prior and the target data distribution via a quantile-aligned tree structure. QAT-FM constructs the coupling in $\mathcal{O}(Nd \log N)$ time and supports per-pair source sampling with $\mathcal{O}(d)$ complexity, enabling scalable training for large-scale high-dimensional generative tasks. Theoretically, we prove that the QAT coupling satisfies marginal consistency, induces non-crossing linear interpolation paths, and consistently improves path separation at intermediate times compared with independent coupling, thereby alleviating local velocity ambiguity. QAT-FM further extends naturally to conditional generation, enabling structured conditional coupling while preserving global Gaussian alignment. Experiments across diverse benchmark datasets demonstrate that QAT-FM achieves competitive generative performance while substantially reducing coupling construction cost.

1 INTRODUCTION

In recent years, continuous-time flow-based generative models have become increasingly prominent in generative modeling (Chen et al., 2018; Grathwohl et al., 2019; Lipman et al., 2023; Liu et al., 2023). A representative class is continuous normalizing flows (CNFs), which define continuous transformations as the solution flows of ordinary differential equations (ODEs) whose time-dependent velocity fields are parameterized by neural networks. During sampling, new data are generated by numerically integrating the learned ODE dynamics. To improve training efficiency, Flow Matching (Lipman et al., 2023) directly supervises the velocity field with targets derived from pre-specified conditional probability paths, avoiding the repeated numerical ODE integration required by conventional CNF training, enabling scalable training of generative models. Flow Matching has achieved competitive performance across diverse domains, including image generation (Ma et al., 2024; Esser et al., 2024), speech synthesis (Le et al., 2023; Liu et al., 2024), and text generation (Hu et al., 2024; 2026).

Although Flow Matching improves the scalability of training continuous-time generative models, its performance remains sensitive to the conditional probability paths and source–target couplings constructed during training (Lipman et al., 2023; Tong et al., 2024; Pooladian et al., 2023). In common linear-path formulations, Flow Matching trains the velocity field by regressing target velocities along linear interpolation paths between noise and data samples; different coupling choices therefore directly alter the structure of the supervision signal. When training paths cross, or pass too closely through the same intermediate region with substantially different target velocities, *velocity ambiguity* arises locally; under mean-squared-error training, these inconsistent velocity directions are averaged, increasing the difficulty of fitting the velocity field (Guo & Schwing, 2025). By contrast, training paths with the *non-crossing* property or good *path separation* reduce this local ambiguity, yielding a more deterministic mapping from intermediate states to target velocities and thereby improving training stability. In conditional generation, coupling strategies must also preserve the alignment between samples and conditions, since couplings that ignore the conditioning information may produce geometrically favorable but conditionally inconsistent supervision signals (Chemseddine et al., 2025; Cheng & Schwing, 2025).

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.

Several existing approaches seek to alleviate path crossing and velocity ambiguity by improving coupling quality. The reflow procedure in Rectified Flow (Liu et al., 2023) reconstructs pairings between noise and data using a learned model, yielding more deterministic and straighter generation trajectories; however, this requires additional sampling and re-training steps, imposing substantial computational overhead. OT-based methods (Villani et al., 2009) construct geometrically optimal pairings by minimizing transport cost; under a quadratic cost, the induced displacement interpolation exhibits a favorable non-crossing structure that alleviates velocity ambiguity. Nonetheless, exact empirical OT is computationally expensive and difficult to scale to large datasets; in conditional generation, maintaining conditioning consistency typically requires condition-aware constraints or condition-wise couplings, further increasing algorithmic complexity. Mini-batch OT (Pooladian et al., 2023; Tong et al., 2024) substantially reduces this cost by solving OT within each mini-batch, but the resulting coupling is primarily local and cannot strictly preserve the non-crossing structure at the full-dataset level. Most of the above methods construct couplings only over finite Gaussian samples; SD-FM (Mousavi-Hosseini et al., 2026) further improves the alignment of continuous Gaussian information via gradient-based optimization, but its iterative optimization incurs substantial sampling or optimization cost and does not yield an explicit low-cost global coupling construction.

Table 1: Comparison of Flow Matching coupling strategies.

Method	**Build Cost**[1]	**Sample Cost**	**Non-Cross**	**Path Sep.**	**Gaussian Align.**
FM	—	$\mathcal{O}(Nd)$	✗	✗	✓✓
RecFlow	$\mathcal{O}(NE\Theta)$	$\mathcal{O}(N\Theta)$	✓	✗	✗
OT-FM$_{\text{(Full)}}$	$\mathcal{O}(N^3 + N^2 d)$	$\mathcal{O}(N^2 + Nd)$	✓	✓	✗
OT-FM$_{\text{(Minibatch)}}$	$\mathcal{O}(NB + NBd)$	$\mathcal{O}(NB + Nd)$	✗	✗	✗
SD-FM	$\mathcal{O}(NKd)$	$\mathcal{O}(N^2 d)$	✓	✓	✓
QAT-FM$_{\text{(Ours)}}$	$\mathcal{O}(Nd \log N)$	$\mathcal{O}(Nd)$	✓	✓✓[2]	✓✓

[1] **Notation.** N: dataset size; B: mini-batch size; d: data dimension; K: SD optimization steps; E: reflow rounds; Θ: one network forward pass. Complexities are reported up to constant factors.
[2] ✓✓indicates that the property is also preserved in conditional generation.

Table 1 compares representative Flow Matching coupling strategies along five axes: coupling construction cost (**Build Cost**), sampling cost (**Sample Cost**), whether the coupling induces non-crossing paths at the full-dataset level (**Non-Cross**), whether it improves path separation (**Path Sep.**), and whether the source marginal is aligned with a continuous Gaussian prior (**Gaussian Align.**). As shown, prior methods generally trade off computational efficiency, path geometry, and Gaussian-prior alignment. To address these limitations, we propose **Quantile AlignTree Flow Matching (QAT-FM)**, a low-cost structured coupling strategy that achieves non-crossing paths, improved path separation, and continuous Gaussian-prior alignment, with several of these properties naturally extending to conditional generation.

Contributions. Our main contributions are as follows:

1) We propose **Quantile AlignTree (QAT)**, an efficient semi-discrete transport construction based on binary tree-structured quantile alignment. QAT induces an exact coupling between an empirical data distribution and a continuous Gaussian prior. With construction complexity $\mathcal{O}(Nd \log N)$ and per-pair sampling complexity $\mathcal{O}(d)$, QAT scales to large-scale, high-dimensional generative modeling tasks.
2) The QAT coupling supports multiple tree-splitting strategies and can naturally incorporate conditioning information into both tree construction and coupling.
3) We prove that QAT induces favorable path structure: the resulting linear interpolation paths satisfy the *non-crossing* property, and path separation at intermediate times is uniformly greater than that of standard independent-coupling flow matching, alleviating local velocity ambiguity.
4) We conduct systematic experiments on low-dimensional synthetic data, unconditional generation, discrete class-conditional generation, continuous text-conditional generation, and large-scale image synthesis. Results show that QAT-FM achieves competitive generation quality with substantially lower coupling construction complexity than OT-based methods.

2 Preliminaries

Notation. Let $\mathcal{P}(\mathbb{R}^d)$ denote the set of Borel probability measures on $\mathbb{R}^d$. For $\alpha \in \mathcal{P}(\mathbb{R}^d)$ and $\beta \in \mathcal{P}(\mathbb{R}^{d'})$, a joint distribution $\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^{d'})$ is called a *coupling* of α and β if its marginals are α and β; we write $\Pi(\alpha, \beta)$ for the collection of all such couplings. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, let

$\boldsymbol{x}^{(j)}$ denote its j-th coordinate. For $\rho \in \mathcal{P}(\mathbb{R}^d)$, let $\rho^{(j)}$ denote its j-th marginal distribution, $F_\rho^{(j)}$ the cumulative distribution function (CDF) of $\rho^{(j)}$, and $\left(F_\rho^{(j)}\right)^{-1}$ its generalized inverse. For a measurable map $T : \mathbb{R}^d \to \mathbb{R}^k$, we write $T_\# \rho$ for the pushforward of ρ under T.

Flow-based Generative Modeling. Given a time-dependent velocity field $v : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, write $v_t(x) := v(t,x)$. The *flow map* $\phi_t : \mathbb{R}^d \to \mathbb{R}^d$ induced by v_t is defined as the solution to the ODE

$$\frac{d}{dt}\phi_t(x) = v_t(\phi_t(x)), \qquad \phi_0(x) = x.$$

Setting $\rho_t = (\phi_t)_\# \gamma$, where γ denotes the source distribution, the pair (ρ_t, v_t) satisfies the continuity equation

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0, \tag{1}$$

with $\rho_0 = \gamma$. When $(\phi_1)_\# \gamma = \mu$, the flow transports the source γ to the target distribution μ. The goal of flow-based generative modeling is therefore to learn a parameterized velocity field v_θ whose induced flow map satisfies $(\phi_1)_\# \gamma = \mu$.

Flow Matching (Lipman et al., 2023) provides a tractable training procedure for this velocity field. It specifies a probability path $(\rho_t)_{t\in[0,1]}$ connecting γ to μ together with a compatible target velocity field u_t. The parameterized velocity field is then learned via the regression objective

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{\substack{t\sim \mathrm{Unif}[0,1] \\ x_t \sim \rho_t}} \left\| v_\theta(t, x_t) - u_t(x_t) \right\|^2. \tag{2}$$

This formulation avoids backpropagating through numerical ODE solvers during training.

Conditional Flow Matching. In practice, the marginal velocity field u_t is often intractable to construct directly. Conditional Flow Matching[1] (CFM) (Lipman et al., 2023; Tong et al., 2024) addresses this by conditioning on paired endpoints: given $\pi \in \Pi(\gamma, \mu)$, draw $(x_0, x_1) \sim \pi$, specify a conditional path $p_t(\cdot \mid x_0, x_1)$ and its compatible velocity $u_t(\cdot \mid x_0, x_1)$, and train $v_\theta(t, x_t)$ to match $u_t(x_t \mid x_0, x_1)$, where $x_t \sim p_t(\cdot \mid x_0, x_1)$. A canonical and widely used choice is the linear interpolation path

$$x_t = (1-t)x_0 + t x_1, \qquad u_t(x_t \mid x_0, x_1) = x_1 - x_0,$$

under which the training objective reduces to

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{\substack{t\sim \mathrm{Unif}[0,1] \\ (x_0,x_1)\sim\pi}} \left\| v_\theta(t, (1-t)x_0 + t x_1) - (x_1 - x_0) \right\|^2. \tag{3}$$

This reveals that the endpoint pairing is governed by the coupling $\pi \in \Pi(\gamma, \mu)$. The simplest choice is the independent coupling $\pi = \gamma \otimes \mu$. A more principled alternative is the *OT coupling*

$$\pi_{\mathrm{OT}} \in \arg\min_{\pi \in \Pi(\gamma,\mu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x_0 - x_1\|^2 \, d\pi(x_0, x_1), \tag{4}$$

which under the linear path tends to produce shorter, geometrically natural paired trajectories (Pooladian et al., 2023; Tong et al., 2024). However, computing the exact global OT coupling is expensive at large scale; in practice, mini-batch OT is used as an approximation, reducing computational cost at the expense of coupling exactness and cross-batch consistency. For the special case $\gamma = \mathcal{N}(0, I_d)$, Mousavi-Hosseini et al. (2026) proposes solving a semi-discrete OT problem in equation 4 to better align the continuous Gaussian source with the empirical target distribution.

Data Partitioning by Trees. The preceding discussion shows that the choice of coupling π directly governs the geometry of particle trajectories in FM, while exact global OT coupling is prohibitively expensive at large scale (Zhang et al., 2025; Mousavi-Hosseini et al., 2026). This motivates binary tree-based space partitioning, which provides a simple way to organize data hierarchically and will serve as the foundation of our coupling construction. Given a data space $\mathcal{X} \subseteq \mathbb{R}^d$ and a dataset $\mathcal{D} = \{x_i\}_{i=1}^N \subseteq \mathcal{X}$, let $\widehat{\mu}_{\mathcal{D}} = N^{-1} \sum_{i=1}^N \delta_{x_i}$ denote its empirical distribution. A *binary partition tree* T consists of an internal node set $\mathcal{I}$ and a leaf node set $\mathcal{L}$. Each node η is associated with a region $R_\eta \subseteq \mathcal{X}$ and data subset $\mathcal{D}_\eta = \mathcal{D} \cap R_\eta$, where the root r satisfies $R_r = \mathcal{X}$.

[1]Here, "conditional" refers to conditioning on paired endpoints. To avoid ambiguity with conditional generation, we refer to this formulation simply as Flow Matching (FM) throughout the remainder of the paper unless otherwise specified.

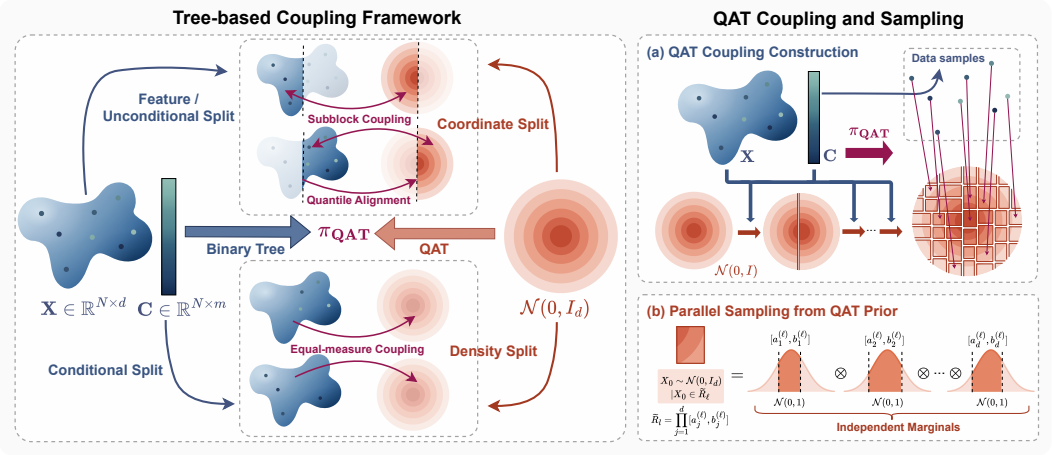


Figure 1: Overview of the QAT coupling and sampling procedure. (*Left*) QAT builds a binary partition tree over the data using feature-based or condition-aware splits, and transfers the same tree topology to the Gaussian source by quantile alignment, yielding the structured coupling π_{QAT}. (*Right*) each data leaf is paired with an aligned Gaussian source region. Since these source regions are axis-aligned hyperrectangles, sampling from the corresponding QAT source conditional reduces to parallel per-coordinate truncated normal sampling.

For each internal node $\eta \in \mathcal{I}$, let $\widehat{\mu}_\eta = |\mathcal{D}_\eta|^{-1} \sum_{x\in\mathcal{D}_\eta} \delta_x$ denote the conditional empirical measure on $\mathcal{D}_\eta$. A *split strategy* selects a split coordinate $\kappa_\eta \in [d]$ and threshold $s_\eta \in \mathbb{R}$ based on $\widehat{\mu}_\eta$, partitioning R_η into

$$R_\eta^- = \{x \in R_\eta : x^{(\kappa_\eta)} \le s_\eta\}, \qquad R_\eta^+ = \{x \in R_\eta : x^{(\kappa_\eta)} > s_\eta\},$$

and correspondingly partitioning $\mathcal{D}_\eta$ into $\mathcal{D}_\eta^- = \mathcal{D}_\eta \cap R_\eta^-$ and $\mathcal{D}_\eta^+ = \mathcal{D}_\eta \cap R_\eta^+$. After recursive splitting, the leaf regions form a partition of $\mathcal{X}$, denoted $\mathcal{R}(T) := \{R_\ell : \ell \in \mathcal{L}\}$.

3 Quantile AlignTree Flow Matching

3.1 Quantile AlignTree Coupling

The preceding analysis shows that in the linear FM objective, the coupling π governs how source samples are paired with target samples, directly shaping the geometry of flow paths. The key idea of QAT is to build a binary partition tree over the empirical data distribution and transfer its topology to the source distribution by matching the split mass ratios at each internal node. Figure 1 summarizes the resulting coupling and sampling procedure.

Let $\widehat{\mu}_\mathcal{D}$ denote the empirical measure of dataset $\mathcal{D} = \{x_i\}_{i=1}^N \subseteq \mathbb{R}^d$, and let $\gamma \in \mathcal{P}(\mathbb{R}^d)$ be the source distribution, e.g., $\gamma = \mathcal{N}(0, I_d)$. Following the binary partition tree construction of Section 2, we build a data tree $T(\widehat{\mu}_\mathcal{D})$ on $\widehat{\mu}_\mathcal{D}$. For each internal node $\eta \in \mathcal{I}$, let (κ_η, s_η) denote its split coordinate and threshold, and let $\widehat{\mu}_\eta$ be the conditional empirical measure at that node. Based on the data tree, we define the Quantile AlignTree as follows.

Definition 1 (Quantile AlignTree (QAT)) *Given the data tree $T(\widehat{\mu}_\mathcal{D})$ and source distribution γ, the* Quantile AlignTree *of γ with respect to the data tree, denoted $\widetilde{T}_\gamma(T)$, is constructed recursively with the same topology as $T(\widehat{\mu}_\mathcal{D})$. At each internal node η, the split coordinate κ_η is inherited from the data tree, but the data-side threshold s_η is replaced by the source-side aligned threshold*

$$\widetilde{s}_\eta = \left(F_{\gamma_\eta}^{(\kappa_\eta)}\right)^{-1}\left(F_{\widehat{\mu}_\eta}^{(\kappa_\eta)}(s_\eta)\right), \tag{5}$$

where γ_η denotes the conditional distribution of γ on the current source-side region $\widetilde{R}_\eta$. The source-side region is then partitioned as

$$\widetilde{R}_\eta^- = \{z \in \widetilde{R}_\eta : z^{(\kappa_\eta)} \le \widetilde{s}_\eta\}, \qquad \widetilde{R}_\eta^+ = \{z \in \widetilde{R}_\eta : z^{(\kappa_\eta)} > \widetilde{s}_\eta\},$$

and the left and right child nodes are constructed recursively.

The quantile-based aligned threshold equation 5 guarantees that the source side and the data side carry equal mass ratios at every split. Specifically, when the source-side conditional marginal is continuous,

$$\gamma_\eta(\widetilde{R}_\eta^-) = \widehat{\mu}_\eta(R_\eta^-), \qquad \gamma_\eta(\widetilde{R}_\eta^+) = \widehat{\mu}_\eta(R_\eta^+).$$

Consequently, the data tree and the aligned source tree carry equal mass at corresponding leaf nodes, naturally inducing a coupling from γ to $\widehat{\mu}_{\mathcal{D}}$.

Definition 2 (QAT Coupling) *Let $\mathcal{L}$ be the nonempty leaf node set of data tree $T(\widehat{\mu}_{\mathcal{D}})$. For each leaf $\ell \in \mathcal{L}$, let $\mathcal{D}_\ell = \mathcal{D} \cap R_\ell$ and $\alpha_\ell := \frac{1}{N}|\mathcal{D}_\ell|$. Let $\widehat{\mu}_\ell = |\mathcal{D}_\ell|^{-1} \sum_{x_i \in \mathcal{D}_\ell} \delta_{x_i}$ denote the conditional empirical measure on $\mathcal{D}_\ell$, and let γ_ℓ denote the conditional source measure associated with leaf ℓ in the Quantile AlignTree. Define the* QAT coupling *as*

$$\pi_{\mathrm{QAT}} := \sum_{\ell \in \mathcal{L}} \alpha_\ell \, \gamma_\ell \otimes \widehat{\mu}_\ell. \tag{6}$$

Then $\pi_{\mathrm{QAT}} \in \Pi(\gamma, \widehat{\mu}_{\mathcal{D}})$.

Intuitively, the QAT coupling matches each data leaf $\mathcal{D}_\ell$ with an aligned source region. To draw a training pair $(x_0, x_1) \sim \pi_{\mathrm{QAT}}$, one samples a leaf ℓ with probability α_ℓ, draws $x_0 \sim \gamma_\ell$ from the corresponding conditional source distribution, and draws $x_1 \sim \widehat{\mu}_\ell$ from the conditional empirical distribution on that leaf.

Remark 1 (Gaussian Sampling) *When $\gamma = \mathcal{N}(0, I_d)$, sampling from γ_ℓ is especially convenient. Since all splits are axis-aligned, every leaf source region is an axis-aligned hyperrectangle. If leaf ℓ corresponds to source region $\widetilde{R}_\ell = \prod_{j=1}^{d} [a_j^{(\ell)}, b_j^{(\ell)}]$, sampling from γ_ℓ reduces to drawing each coordinate independently from a truncated standard normal:*

$$x_0^{(j)} \sim \mathcal{N}(0,1)\big|_{[a_j^{(\ell)}, b_j^{(\ell)}]}, \qquad j = 1, \ldots, d.$$

The interval endpoints are determined by all split constraints along the root-to-leaf path, and the sampling is fully parallelizable across coordinates.

3.2 QAT Flow Matching

Substituting the QAT coupling into the linear FM objective equation 3 yields the QAT-FM training objective. Let $\widehat{\mu}_{\mathcal{D}}$ be the empirical data distribution, γ the source distribution, and $\pi_{\mathrm{QAT}} \in \Pi(\gamma, \widehat{\mu}_{\mathcal{D}})$ the QAT coupling of Definition 2. QAT-FM learns the velocity field via

$$\mathcal{L}_{\text{QAT-FM}}(\theta) := \mathbb{E}_{\substack{t \sim \mathrm{Unif}[0,1] \\ (x_0, x_1) \sim \pi_{\mathrm{QAT}}}} \left\| v_\theta\big(t, (1-t)x_0 + t x_1\big) - (x_1 - x_0) \right\|^2. \tag{7}$$

Compared to standard FM, QAT-FM differs only in the sample pairing: π_{QAT} replaces the independent coupling, while the linear interpolation path and velocity regression target remain unchanged. This structured pairing yields the following geometric properties.

Theorem 1 (Non-Crossing Property) *Let $\pi_{\mathrm{QAT}} \in \Pi(\gamma, \widehat{\mu}_{\mathcal{D}})$ be the QAT coupling. Consider two paired samples $(x_0, x_1), (x_0', x_1') \in \mathrm{supp}(\pi_{\mathrm{QAT}})$. If x_1 and x_1' belong to two distinct leaf nodes $\ell \neq \ell'$ of the data tree, then the corresponding linear interpolation paths*

$$x_t = (1-t)x_0 + t x_1, \qquad x_t' = (1-t)x_0' + t x_1'$$

do not intersect, i.e., $x_t \neq x_t'$ for all $t \in [0,1]$, almost surely.

Theorem 1 establishes that paths associated with distinct target leaves are strictly non-crossing. This property alleviates the velocity ambiguity that can arise from overlapping training paths: since the FM objective fits v_θ pointwise at each $t \in [0,1]$, two training paths satisfying $x_t = x_t'$ but $x_1 - x_0 \neq x_1' - x_0'$ impose conflicting supervision signals at the same spatial location, increasing the difficulty of fitting v_θ. Paths that remain well-separated at intermediate times are therefore desirable. Beyond preventing cross-leaf path crossings, the QAT coupling also consistently improves the expected inter-path distance, as the following theorem quantifies.

Theorem 2 (Improved Path Separation) *Let $\gamma = \mathcal{N}(0, I_d)$, let $\pi_{\mathrm{QAT}} \in \Pi(\gamma, \widehat{\mu}_{\mathcal{D}})$ be the QAT coupling, and let $\pi_{\mathrm{ind}} := \gamma \otimes \widehat{\mu}_{\mathcal{D}}$ be the independent coupling. For any coupling π, let $\mathbb{E}_{\pi^{\otimes 2}}$ denote expectation over two independent pairs $(x_0, x_1), (x_0', x_1') \sim \pi$, and define $x_t = (1-t)x_0 + t x_1$ and $x_t' = (1-t)x_0' + t x_1'$. Then for every $t \in (0,1)$,*

$$\mathbb{E}_{\pi_{\mathrm{QAT}}^{\otimes 2}}\big[\|x_t - x_t'\|_2^2\big] - \mathbb{E}_{\pi_{\mathrm{ind}}^{\otimes 2}}\big[\|x_t - x_t'\|_2^2\big] = 4t(1-t)\,\Delta_{\mathrm{QAT}},$$

where Δ_{QAT} is the node-level path-separation gain induced by the QAT tree:

$$\Delta_{\mathrm{QAT}} := \sum_{\eta \in \mathcal{I}} \alpha_\eta q_\eta (1 - q_\eta) \Big(\mathbb{E}_{\gamma_{\eta^+}}\big[X^{(\kappa_\eta)}\big] - \mathbb{E}_{\gamma_{\eta^-}}\big[X^{(\kappa_\eta)}\big] \Big) \Big(\mathbb{E}_{\widehat{\mu}_{\eta^+}}\big[X^{(\kappa_\eta)}\big] - \mathbb{E}_{\widehat{\mu}_{\eta^-}}\big[X^{(\kappa_\eta)}\big] \Big). \tag{8}$$

Here $\alpha_\eta := \widehat{\mu}_{\mathcal{D}}(R_\eta)$ is the data mass at node η, $q_\eta := F^{(\kappa_\eta)}_{\widehat{\mu}_\eta}(s_\eta)$ is the conditional left-child mass fraction at split coordinate κ_η, and $\gamma_{\eta^\pm}$, $\widehat{\mu}_{\eta^\pm}$ denote the conditional source and empirical distributions on the left and right child regions of η. Under non-degenerate splits, $\Delta_{\mathrm{QAT}} > 0$, so the QAT coupling strictly dominates the independent coupling in expected inter-path separation.

3.3 Conditional QAT Strategy

In conditional generation tasks, data samples carry conditioning variables. Let $\{(x_i, c_i)\}_{i=1}^N \subseteq \mathbb{R}^d \times \mathbb{R}^m$ be the joint dataset, $\widehat{\mu}_{\mathcal{D}_C} := N^{-1}\sum_{i=1}^N \delta_{(x_i,c_i)}$ the joint empirical distribution, and $\widehat{\mu}_C := N^{-1}\sum_{i=1}^N \delta_{c_i}$ the marginal over condition space. A standard OT coupling based only on data-space distances can ignore the alignment between samples and their conditioning variables. Alternatively, solving OT directly between the joint source $\gamma \otimes \widehat{\mu}_C$ and the joint empirical distribution $\widehat{\mu}_{\mathcal{D}_C}$ increases the optimization dimension and computational cost (Cheng & Schwing, 2025; Mousavi-Hosseini et al., 2026). We therefore propose the following Conditional QAT strategy.

Definition 3 (Conditional QAT) *Let $\{c_i\}_{i=1}^N \subseteq \mathbb{R}^m$ be conditioning variables. Conditional QAT constructs a data tree on the joint samples $\{(x_i, c_i)\}_{i=1}^N \subseteq \mathbb{R}^{d+m}$. For each internal node η, let $\kappa_\eta \in [d+m]$ denote its split coordinate. When building the source-side AlignTree, if $\kappa_\eta \le d$ (a data coordinate), the source side is partitioned using the standard QAT aligned threshold in equation 5; if $\kappa_\eta > d$ (a condition coordinate), no split is applied to the source and both child nodes inherit the parent source region, the split is used only to refine the target-side grouping: $\widetilde{R}_{\eta^-} = \widetilde{R}_{\eta^+} = \widetilde{R}_\eta$.*

The Conditional QAT coupling takes the same form as Definition 2, with the target empirical distribution replaced by the joint distribution $\widehat{\mu}_{\mathcal{D}_C}$:

$$\pi_{\mathrm{CQAT}} = \sum_{\ell \in \mathcal{L}_C} \alpha_\ell \, \gamma_\ell \otimes \widehat{\mu}_{\mathcal{D}_C,\ell} \in \Pi(\gamma, \widehat{\mu}_{\mathcal{D}_C}),$$

where $\mathcal{L}_C$ is the set of nonempty leaves of the joint data tree, and $\widehat{\mu}_{\mathcal{D}_C,\ell}$ is the conditional joint empirical distribution at leaf ℓ. A sample from π_{CQAT} can be written as $(x_0, (x_1, c))$, where the linear FM path is applied only in the data space between x_0 and x_1, while c is provided as conditioning input to the velocity model. By encoding conditional structure through leaf-level matching in the joint tree, this construction avoids explicitly solving the product-space OT problem between $\gamma \otimes \widehat{\mu}_C$ and $\widehat{\mu}_{\mathcal{D}_C}$, while retaining the efficient sampling form of unconditional QAT.

Remark 2 (Theoretical extensions) *Conditional QAT inherits the theoretical properties of QAT under appropriate conditions. Specifically, if the root-to-ℓ and root-to-ℓ' paths first diverge at a data-coordinate split node (i.e., $\kappa_\eta \le d$), then the linear interpolation paths drawn from the corresponding leaf couplings satisfy the non-crossing property of Theorem 1. Under the same non-degeneracy conditions, Conditional QAT also inherits the path-separation bound of Theorem 2. Proofs and further details are given in Appendix A.3.*

3.4 Efficient QAT Construction and Sampling

When $\gamma = \mathcal{N}(0, I_d)$, each source-side leaf region $\widetilde{R}_\ell$ is fully characterized by its coordinate-wise bounds (a_ℓ, b_ℓ) (Remark 1). Regardless of whether the data tree is built from unconditional or conditional samples, each leaf ℓ stores only the data index set $\mathcal{S}_\ell$ and the source-side bounds (a_ℓ, b_ℓ), which fully specify the leaf-wise product coupling $\gamma_\ell \otimes \widehat{\mu}_\ell$. Training pairs are drawn by sampling a leaf ℓ proportionally to its mass, then independently drawing a source point from the corresponding truncated Gaussian region and a target point from $\mathcal{S}_\ell$. The full construction, sampling algorithm, and complexity analysis are given in Appendix B.

Split Strategy. QAT is compatible with a variety of split strategies. Throughout this work we adopt a maximum-variance rule: for each node η, the split coordinate is the dimension of highest empirical variance and the split threshold is the empirical mean along that coordinate,

$$\kappa_\eta = \arg\max_j \mathrm{Var}_{\widehat{\mu}_\eta}\big(Y^{(j)}\big), \qquad s_\eta = \mathbb{E}_{\widehat{\mu}_\eta}\big[Y^{(\kappa_\eta)}\big],$$

where $Y = X$ in the unconditional case and $Y = (X, C)$ in the conditional case. For discrete empirical data, we recurse until each leaf contains at most one sample; in the conditional setting, the terminal splitting phase prioritizes data coordinates over condition coordinates, maximizing the number of source-side splits and thereby increasing the path-separation gain of Theorem 2.

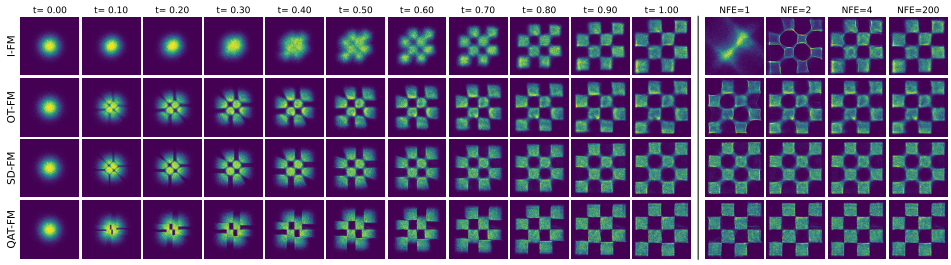


Figure 2: **2D checkerboard results.** (*Left*) sample evolution induced by different coupling methods from $t = 0$ to $t = 1$. (*Right*) generated samples with varying numbers of function evaluations (NFE).

High-Dimensional Data Handling. For high-dimensional data, axis-aligned partitioning in the raw data space may fail to capture the principal directions of variation. We therefore first rotate the data via an orthogonal transform $P \in \mathbb{R}^{d\times d}$, setting $z_i = Px_i$ and constructing the QAT coupling on the top-k subspace $z_i^{(1:k)}$. By the rotational invariance of the standard Gaussian, $P_{\#}\mathcal{N}(0, I_d) = \mathcal{N}(0, I_d)$ for any orthogonal P. Training then proceeds by sampling $(\epsilon^{(1:k)}, z_i^{(1:k)}) \sim \pi_{\mathrm{QAT}}$, independently drawing $\epsilon^{(-k)} \sim \mathcal{N}(0, I_{d-k})$, and setting

$$x_0 = P^\top\big(\epsilon^{(1:k)}, \epsilon^{(-k)}\big).$$

The marginal distribution of x_0 remains $\mathcal{N}(0, I_d)$, while the QAT tree operates entirely within the low-dimensional subspace. In practice, P is composed of a Patch-Hadamard transform and a PCA rotation at the mean-pooling scale, concentrating the principal data variation into a small number of coordinates. Details are given in Appendix B.3.

4 Experiments

We evaluate QAT-FM through a series of experiments: we begin with path visualizations on two-dimensional synthetic data, then proceed to high-dimensional image generation benchmarks covering unconditional generation (CIFAR-10), class-conditional generation (ImageNet-32), text-conditional generation (CelebA-64), and large-scale latent-space generation (ImageNet-256). Additional implementation details are provided in Appendix C; extended experimental results appear in Appendix D.

4.1 2D Synthetic Experiments

Following Lipman et al. (2023); Pooladian et al. (2023), we compare different coupling methods on two-dimensional checkerboard data. Four methods are evaluated: I-FM, OT-FM (Tong et al., 2024), SD-FM (Mousavi-Hosseini et al., 2026), and the proposed QAT-FM. All methods share the same network architecture and training configuration.

Figure 2 (left) shows sample evolution from $t = 0$ to $t = 1$. OT-FM and SD-FM capture the global structure of the target distribution earlier than I-FM, while QAT-FM yields better local separation and more organized intermediate transport. Figure 2 (right) shows generated samples with varying NFE. QAT-FM recovers a clear checkerboard structure at small NFE, indicating that its structured coupling improves few-step sample quality.

4.2 Unconditional Generation on CIFAR-10

We evaluate QAT-FM on unconditional generation using CIFAR-10 (Krizhevsky et al., 2009), which contains $50{,}000$ RGB training images of size 32×32 across 10 categories. The velocity-field network follows the architecture of Tong et al. (2024); training details are provided in Appendix C. We generate $50{,}000$ samples using several numerical solvers and compute FID (Heusel et al., 2017). Results are averaged over 5 random seeds and reported with standard deviations.

Table 2 reports CIFAR-10 results at 100k training steps. Under few-step Euler sampling, OT-FM performs best, while QAT-FM substantially improves over I-FM and SD-FM, indicating that structured coupling can improve generation quality under coarse numerical integration. Under the adaptive Dopri5 solver, QAT-FM achieves the lowest FID, showing competitive final generation quality in the unconditional setting. Figure 3 further shows that QAT-FM maintains strong performance during late-stage training.

Table 2: FID results on unconditional CIFAR-10 generation at 100k training steps.

Method	Euler-2	Euler-100	Dopri5	NFE↓
I-FM	$164.18_{\pm 0.13}$	$\mathbf{3.79}_{\pm \mathbf{0.04}}$	$3.38_{\pm 0.03}$	$137.30_{\pm 0.30}$
OT-FM	$\mathbf{77.26}_{\pm \mathbf{0.12}}$	$3.96_{\pm 0.04}$	$3.49_{\pm 0.03}$	$\mathbf{129.52}_{\pm \mathbf{0.86}}$
SD-FM	$162.66_{\pm 0.23}$	$4.03_{\pm 0.02}$	$3.61_{\pm 0.01}$	$135.78_{\pm 0.36}$
QAT-FM	$98.68_{\pm 0.17}$	$3.91_{\pm 0.04}$	$\mathbf{3.26}_{\pm \mathbf{0.02}}$	$145.55_{\pm 1.05}$

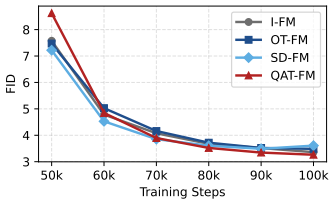


Figure 3: FID (Dopri5) vs. steps.

Ablation Studies. Our main experiments use patch size $p = 4$, PCA dimension $k = 32$, and mean-based tree splitting. To assess the sensitivity of these choices, we conduct three ablation studies: (1) varying the PCA dimension $k \in \{32, 64, 128\}$; (2) varying the patch size $p \in \{1, 2, 4\}$; and (3) comparing three split strategies—mean split, median split, and random split with ratio $\sim \mathcal{U}(0.25, 0.75)$. Results are summarized in Table 3.

Table 3: Ablation studies on PCA dimension, patch size, and tree splitting strategy.

PCA Dim.	FID↓	NFE↓
$k = 32$	3.26	145.48
$k = 64$	3.27	143.50
$k = 128$	3.24	143.76

Patch Size	FID↓	NFE↓
$p = 1$	3.24	145.36
$p = 2$	3.24	145.55
$p = 4$	3.26	145.48

Split Rule	FID↓	NFE↓
Mean	3.26	145.48
Median	3.22	143.89
Random	3.24	144.87

As shown in Table 3, QAT-FM is insensitive to the choice of PCA dimension, patch size, and tree splitting strategy; FID and NFE vary only slightly across all configurations. These results suggest that the performance gains of QAT-FM do not rely on carefully tuned tree construction hyperparameters, but stem primarily from the global tree-structured quantile alignment mechanism.

4.3 Conditional Generation Experiments

Class-conditional generation. We evaluate QAT-FM on class-conditional generation using ImageNet-1k (Deng et al., 2009), which contains 1,000 categories and over 1M training images. All images are resized to 32×32, with random horizontal flipping used for augmentation. We build the Conditional QAT coupling using both original and horizontally flipped training images. At evaluation, we generate 50K samples and compute FID against the training set. Results are averaged over 5 random seeds and reported with standard deviations.

Table 4: FID results on class-conditional ImageNet-32×32 generation.

Method	Euler-2	Euler-5	Euler-10	Euler-20	Euler-50	Euler-100	Dopri5	NFE↓
I-FM	$122.40_{\pm 0.08}$	$23.04_{\pm 0.07}$	$10.12_{\pm 0.07}$	$6.48_{\pm 0.07}$	$4.77_{\pm 0.07}$	$4.25_{\pm 0.06}$	$3.82_{\pm 0.04}$	$\mathbf{140.27}_{\pm \mathbf{1.99}}$
OT-FM	$115.63_{\pm 0.42}$	$31.82_{\pm 0.11}$	$15.79_{\pm 0.02}$	$10.59_{\pm 0.00}$	$8.14_{\pm 0.02}$	$7.37_{\pm 0.00}$	$6.65_{\pm 0.02}$	$142.51_{\pm 0.45}$
C2OT	$\mathbf{100.41}_{\pm \mathbf{0.05}}$	$\mathbf{20.94}_{\pm \mathbf{0.08}}$	$\mathbf{10.10}_{\pm \mathbf{0.06}}$	$6.80_{\pm 0.04}$	$5.15_{\pm 0.04}$	$4.62_{\pm 0.04}$	$4.16_{\pm 0.02}$	$143.97_{\pm 0.75}$
QAT-FM	$105.16_{\pm 0.30}$	$22.88_{\pm \mathbf{0.05}}$	$10.30_{\pm 0.08}$	$\mathbf{6.40}_{\pm \mathbf{0.07}}$	$\mathbf{4.54}_{\pm \mathbf{0.07}}$	$\mathbf{3.99}_{\pm \mathbf{0.07}}$	$\mathbf{3.56}_{\pm \mathbf{0.06}}$	$143.35_{\pm 2.22}$

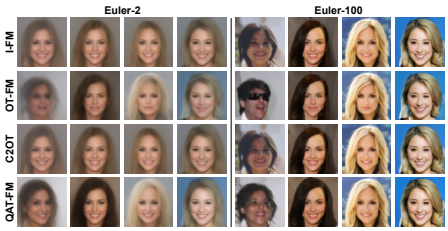
Table 4 reports FID on class-conditional ImageNet-32×32 generation. Both C2OT and QAT-FM improve few-step generation over I-FM, demonstrating the effectiveness of condition-aware coupling. While C2OT performs best under small Euler step counts, its advantage reverses as the solver becomes more accurate. By contrast, QAT-FM remains competitive in few-step sampling and achieves the best performance under larger-step Euler solvers and Dopri5. This indicates that Conditional QAT provides more stable supervision for high-quality conditional FM training.

Text-conditional generation. We further evaluate QAT-FM on text-conditional CelebA-64×64 generation. CelebA (Liu et al., 2015) contains over 200K celebrity face images with attribute annotations, and CelebA-Dialog (Jiang et al., 2021) provides fine-grained natural-language descriptions for these images. We use the text captions as conditioning inputs. Following C2OT (Cheng & Schwing, 2025), each caption is encoded by a CLIP (Radford et al., 2021)-like DFN text encoder (Fang et al., 2024), and the resulting embedding is used as the condition variable. We compare QAT-FM with I-FM, OT-FM, and C2OT under the same text-conditioning setting. At evaluation, we generate 50K samples and report FID for image fidelity and SigLIP-2 (Tschannen et al., 2025) CLIP score for text-image alignment.

Figure 4 shows that QAT-FM consistently outperforms I-FM, OT-FM, and C2OT in FID across different solvers, while also requiring the lowest average NFE. Meanwhile, QAT-FM attains the same CLIP scores as I-FM and C2OT, and higher CLIP scores than OT-FM, indicating that the improvement

	Method	Euler-2	Euler-100	Dopri5	NFE↓
FID↓	I-FM	82.40	3.60	3.13	149.97
	OT-FM	70.24	3.42	2.98	146.53
	C2OT	81.39	3.50	3.08	148.75
	QAT-FM	**61.84**	**3.17**	**2.68**	**144.08**
CLIP↑	I-FM	**0.11**	**0.10**	**0.10**	149.97
	OT-FM	0.09	0.08	0.08	146.53
	C2OT	**0.11**	**0.10**	**0.10**	148.75
	QAT-FM	**0.11**	**0.10**	**0.10**	**144.08**

(a) Quantitative results.



(b) Qualitative samples.

Figure 4: Text-conditional generation results on CelebA-64 × 64. (a) Quantitative comparison in terms of FID and CLIP score under different solvers. (b) Qualitative samples generated by Euler-2 and Euler-100 using the caption: *"The full face of this female is beamed with happiness. This person looks very young and has no bangs. There is not any glasses on her face."*



Figure 5: Qualitative comparison between I-FM and QAT-FM on ImageNet-256 × 256 class-conditional generation using Euler-250 sampling. Each column uses the same input condition.

in image fidelity does not come at the cost of weaker text-image alignment. The qualitative samples further show that QAT-FM produces more coherent faces under Euler-2 and sharper details under Euler-100, consistent with the quantitative results.

Large-scale latent generation. To assess scalability, we further evaluate QAT-FM on class-conditional ImageNet-256 × 256 latent generation (Rombach et al., 2022; Ma et al., 2024). Following Yao et al. (2025), images are encoded into VA-VAE latents, and the flow model is trained in the latent space. We construct the Conditional QAT directly over the VAE latents, while following the network architecture, training hyperparameters, and evaluation protocol of Yao et al. (2025). Due to computational constraints, we train for 64 epochs using the official recommended setting. The implementation and evaluation details are provided in Appendix C. Table 5 shows that QAT-FM consistently improves FID and Inception Score over I-FM across all solver step counts, while maintaining competitive sFID, precision, and recall. The improvements hold in both few-step and high-accuracy sampling regimes, indicating that the QAT coupling benefits latent-space flow training beyond a particular solver choice. Figure 5 provides qualitative examples under Euler-250 sampling. These results suggest that QAT-FM can be integrated into modern latent-space flow backbones with minimal architectural changes while improving generation quality.

Table 5: Class-conditional generation results on ImageNet-256 × 256 latent generation.

Method	Steps	IS↑	FID↓	sFID↓	Pre.↑	Rec.↑
I-FM	10	152.19	11.83	**20.73**	0.64	**0.49**
QAT-FM	10	**162.15**	**10.67**	20.74	**0.66**	0.48
I-FM	50	242.65	2.57	**5.62**	**0.80**	0.57
QAT-FM	50	**250.36**	**2.46**	5.69	**0.80**	**0.58**
I-FM	250	248.94	2.08	4.29	**0.81**	0.59
QAT-FM	250	**256.96**	**1.99**	**4.25**	0.80	**0.60**

5 Conclusion

We study the prior-data coupling problem in Flow Matching and propose QAT-FM, a simple yet effective coupling strategy based on tree-structured quantile alignment. QAT-FM avoids the high cost of OT-based coupling while preserving favorable path geometry, including non-crossing paths and improved path separation. It also extends naturally to conditional generation through condition-aware tree construction. Extensive experiments verify the effectiveness and scalability of our method across synthetic data, image generation, and large-scale latent-space generation. Overall, QAT-FM provides an efficient and structured coupling mechanism for improving Flow Matching training.

AI USE STATEMENT

In this work, we used generative AI tools for grammatical correction, language polishing, and operational assistance with running experiments. The research methodology, theoretical development, mathematical claims and proofs, algorithms, experimental design, and experimental settings were independently developed and determined by the authors. All AI-assisted work and experimental outputs were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The datasets used in this paper are all publicly available and do not involve any ethical issues.

REPRODUCIBILITY STATEMENT

The assumptions and complete proofs of all theoretical results are provided in Appendix A. Detailed construction and sampling algorithms are given in Appendix B. Appendix C reports the datasets, preprocessing procedures, model architectures, training hyperparameters, sampling configurations, and evaluation protocols used in the experiments. Additional quantitative and qualitative results are provided in Appendix D. All reported results were obtained using the experimental settings described in the paper and appendices.

REFERENCES

Jannis Chemseddine, Paul Hagemann, Gabriele Steidl, and Christian Wald. Conditional wasserstein distances with applications in bayesian ot flow matching. *Journal of Machine Learning Research*, 26(141):1–47, 2025. URL http://jmlr.org/papers/v26/24-0586.html.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

Ho Kei Cheng and Alexander Schwing. The curse of conditions: Analyzing and improving optimal transport for conditional flow-based generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15875–15884, October 2025.

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. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=FPnUhsQJ5B.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T Toshev, and Vaishaal Shankar. Data filtering networks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KAk6ngZ09F.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet,

Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/2a27b8144ac02f67687f76782a3b5d8f-Paper.pdf.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, and David Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJxgknCcK7.

Pengsheng Guo and Alex Schwing. Variational rectified flow matching. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Rk18ZikrFI.

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 *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Kathy Horadam. *Hadamard matrices and their applications*. Princeton university press, 2012.

Keya Hu, Linlu Qiu, Yiyang Lu, Hanhong Zhao, Tianhong Li, Yoon Kim, Jacob Andreas, and Kaiming He. Elf: Embedded language flows, 2026. URL https://arxiv.org/abs/2605.10938.

Vincent Hu, Di Wu, Yuki Asano, Pascal Mettes, Basura Fernando, Björn Ommer, and Cees Snoek. Flow matching for conditional text generation in a few sampling steps. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 380–392, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-short.33. URL https://aclanthology.org/2024.eacl-short.33/.

Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13799–13808, October 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=gzCS252hCO.

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

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code, 2024. URL https://arxiv.org/abs/2412.06264.

Alexander H. Liu, Matthew Le, Apoorv Vyas, Bowen Shi, Andros Tjandra, and Wei-Ning Hsu. Generative pre-training for speech with flow matching. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KpoQSgxbKH.

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. URL https://openreview.net/forum?id=XVjTT1nw5z.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72980-5.

Alireza Mousavi-Hosseini, Stephen Y. Zhang, Michal Klein, and marco cuturi. Flow matching with semidiscrete couplings. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4EGjzT6w80.

Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input perturbation reduces exposure bias in diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26245–26265. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/ning23a.html.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11410–11420, June 2022.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28100–28127. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/pooladian23a.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions, 2017. URL https://arxiv.org/abs/1710.05941.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

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. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW. Expert Certification.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

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

Zhendong Wang, Yifan Jiang, Huangjie Zheng, Peihao Wang, Pengcheng He, Zhangyang Wang, Weizhu Chen, and Mingyuan Zhou. Patch diffusion: Faster and more data-efficient training of diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=iv2sTQtbst.

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

Stephen Zhang, Alireza Mousavi-Hosseini, Michal Klein, and Marco Cuturi. On fitting flow models with large sinkhorn couplings, 2025. URL https://arxiv.org/abs/2506.05526.

A PROOFS

A.1 PROOF OF THEOREM 1

Proof. By the binary tree structure, for any two distinct data leaves $\ell \neq \ell'$, there exists a unique first internal node η at which the root-to-ℓ and root-to-ℓ' paths diverge. Without loss of generality, assume ℓ lies in the left subtree of η and ℓ' lies in the right subtree.

By the construction of the data tree, node η is split along coordinate κ_η with threshold s_η, so its left and right child regions are

$$R_{\eta^-} = \{x \in R_\eta : x^{(\kappa_\eta)} \leq s_\eta\}, \qquad R_{\eta^+} = \{x \in R_\eta : x^{(\kappa_\eta)} > s_\eta\}.$$

Since $x_1 \in R_{\eta^-}$ and $x_1' \in R_{\eta^+}$,

$$x_1^{(\kappa_\eta)} \leq s_\eta < (x_1')^{(\kappa_\eta)}. \tag{9}$$

On the source side, QAT reuses the same split coordinate κ_η and replaces the data-side threshold s_η by the aligned threshold $\widetilde{s}_\eta$. The corresponding source-side child regions are

$$\widetilde{R}_{\eta^-} = \{z \in \widetilde{R}_\eta : z^{(\kappa_\eta)} \leq \widetilde{s}_\eta\}, \qquad \widetilde{R}_{\eta^+} = \{z \in \widetilde{R}_\eta : z^{(\kappa_\eta)} > \widetilde{s}_\eta\}.$$

Since QAT matches each data leaf with the corresponding source leaf region, the source samples paired with leaves ℓ and ℓ' satisfy $x_0 \in \widetilde{R}_{\eta^-}$ and $x_0' \in \widetilde{R}_{\eta^+}$. Hence

$$x_0^{(\kappa_\eta)} \leq \widetilde{s}_\eta < (x_0')^{(\kappa_\eta)}. \tag{10}$$

For any $t \in [0,1]$, the two linear interpolation paths are

$$x_t = (1-t)x_0 + tx_1, \qquad x_t' = (1-t)x_0' + tx_1'.$$

Examining the κ_η-th coordinate,

$$(x_t')^{(\kappa_\eta)} - x_t^{(\kappa_\eta)} = (1-t)\big\{(x_0')^{(\kappa_\eta)} - x_0^{(\kappa_\eta)}\big\} + t\big\{(x_1')^{(\kappa_\eta)} - x_1^{(\kappa_\eta)}\big\}.$$

By equation 9 and equation 10, both differences on the right-hand side are strictly positive. Therefore,

$$(x_t')^{(\kappa_\eta)} - x_t^{(\kappa_\eta)} > 0, \qquad \forall\, t \in [0,1],$$

so $x_t' \neq x_t$ for all $t \in [0,1]$. This proves that linear interpolation paths from two distinct QAT leaves are strictly non-coincident almost surely. □

A.2 PROOF OF THEOREM 2

Proof. For any coupling $\pi \in \Pi(\gamma, \widehat{\mu}_\mathcal{D})$, let (X_0, X_1) and (X_0', X_1') be two independent pairs drawn from π. By the linear interpolation $X_t - X_t' = (1-t)(X_0 - X_0') + t(X_1 - X_1')$,

$$\begin{aligned}\mathbb{E}_{\pi^{\otimes 2}}\|X_t - X_t'\|_2^2 = &(1-t)^2\mathbb{E}\|X_0 - X_0'\|_2^2 + t^2\mathbb{E}\|X_1 - X_1'\|_2^2 \\ &+ 2t(1-t)\mathbb{E}_{\pi^{\otimes 2}}\left[(X_0 - X_0')^\top (X_1 - X_1')\right].\end{aligned} \tag{11}$$

The first two terms depend only on the marginal distributions γ and $\widehat{\mu}_\mathcal{D}$, and are therefore identical under π_{QAT} and π_{ind}. The difference between the two couplings arises entirely from the cross term. Under the independent coupling $\pi_{\mathrm{ind}} = \gamma \otimes \widehat{\mu}_\mathcal{D}$, $X_0 - X_0'$ and $X_1 - X_1'$ are mutually independent with zero mean, giving

$$\mathbb{E}_{\pi_{\mathrm{ind}}^{\otimes 2}}\left[(X_0 - X_0')^\top (X_1 - X_1')\right] = 0. \tag{12}$$

Recursive analysis of the QAT cross term. For each node η, let π_η denote the normalized QAT coupling conditioned on node η's region, and let $\mathbb{E}_\eta$ denote expectation when both pairs are drawn independently from π_η. Define

$$\Gamma_\eta := \mathbb{E}_\eta\left[(X_0 - X_0')^\top (X_1 - X_1')\right].$$

If η is a leaf, the QAT coupling within that leaf is $\gamma_\eta \otimes \widehat{\mu}_\eta$, so $X_0 - X_0'$ and $X_1 - X_1'$ are independent with zero mean, giving $\Gamma_\eta = 0$.

For any internal node $\eta \in \mathcal{I}$ with left and right children η^- and η^+, let

$$Z_\eta := \mathbf{1}\{X_0 \in \widetilde{R}_{\eta^-},\, X_1 \in R_{\eta^-}\}, \qquad Z'_\eta := \mathbf{1}\{X'_0 \in \widetilde{R}_{\eta^-},\, X'_1 \in R_{\eta^-}\}$$

indicate whether the first and second pairs fall into the left child. Write $p_\eta^- := \widehat{\mu}_\eta(R_{\eta^-})$ and $p_\eta^+ := \widehat{\mu}_\eta(R_{\eta^+})$. By the law of total expectation,

$$\begin{aligned}\Gamma_\eta =& \mathbb{E}\big[\mathbb{E}_\eta\big[(X_0 - X_0')^\top (X_1 - X_1') \,\big|\, Z_\eta, Z'_\eta\big]\big] \\ =& (p_\eta^-)^2 \Gamma_{\eta^-} + (p_\eta^+)^2 \Gamma_{\eta^+} \\ &+ p_\eta^- p_\eta^+ \,\mathbb{E}_\eta\big[(X_0 - X_0')^\top (X_1 - X_1') \,\big|\, Z_\eta = 1,\, Z'_\eta = 0\big] \\ &+ p_\eta^+ p_\eta^- \,\mathbb{E}_\eta\big[(X_0 - X_0')^\top (X_1 - X_1') \,\big|\, Z_\eta = 0,\, Z'_\eta = 1\big]\,.\end{aligned} \tag{13}$$

When both pairs fall into the same child, they contribute $(p_\eta^-)^2\Gamma_{\eta^-}$ and $(p_\eta^+)^2\Gamma_{\eta^+}$ respectively. For the cross-subtree terms, write $m_{0,\eta}^\pm := \mathbb{E}_{\gamma_{\eta^\pm}}[X]$ and $m_{1,\eta}^\pm := \mathbb{E}_{\widehat{\mu}_{\eta^\pm}}[X]$. Taking the case $Z_\eta = 0$, $Z'_\eta = 1$ (first pair in right child, second in left), the cross-subtree contribution is

$$\begin{aligned}&p_\eta^+ p_\eta^- \,\mathbb{E}_\eta\big[(X_0 - X_0')^\top (X_1 - X_1') \,\big|\, Z_\eta = 0,\, Z'_\eta = 1\big] \\ =& p_\eta^+ p_\eta^- \Big\{\mathbb{E}_{\eta^+}[X_0^\top X_1] - (m_{0,\eta}^+)^\top m_{1,\eta}^- - (m_{0,\eta}^-)^\top m_{1,\eta}^+ + \mathbb{E}_{\eta^-}[X_0^\top X_1]\Big\}.\end{aligned}$$

Since $\Gamma_{\eta^\pm} = 2\big\{\mathbb{E}_{\eta^\pm}[X_0^\top X_1] - (m_{0,\eta}^\pm)^\top m_{1,\eta}^\pm\big\}$, the above simplifies to

$$p_\eta^+ p_\eta^- \left\{\frac{1}{2}\Gamma_{\eta^+} + \frac{1}{2}\Gamma_{\eta^-} + \left(m_{0,\eta}^+ - m_{0,\eta}^-\right)^\top \left(m_{1,\eta}^+ - m_{1,\eta}^-\right)\right\}.$$

The case $Z_\eta = 1$, $Z'_\eta = 0$ contributes the same amount. Substituting into equation 13,

$$\Gamma_\eta = p_\eta^- \Gamma_{\eta^-} + p_\eta^+ \Gamma_{\eta^+} + 2p_\eta^- p_\eta^+ \left(m_{0,\eta}^+ - m_{0,\eta}^-\right)^\top \left(m_{1,\eta}^+ - m_{1,\eta}^-\right). \tag{14}$$

Reduction to split-coordinate mean differences. Since the source distribution is $\mathcal{N}(0, I_d)$ and all QAT splits are axis-aligned, the source regions $\widetilde{R}_{\eta^-}$ and $\widetilde{R}_{\eta^+}$ differ only in the κ_η-th coordinate interval. By Gaussian coordinate independence, for any $j \neq \kappa_\eta$,

$$(\mathrm{proj}_j)_\# \gamma_{\eta^+} = (\mathrm{proj}_j)_\# \gamma_{\eta^-},$$

so $(m_{0,\eta}^+)^{(j)} = (m_{0,\eta}^-)^{(j)}$ for all $j \neq \kappa_\eta$. Writing

$$\begin{aligned}\Delta_\eta :=& \left(m_{0,\eta}^+ - m_{0,\eta}^-\right)^\top \left(m_{1,\eta}^+ - m_{1,\eta}^-\right) \\ =& \left[(m_{0,\eta}^+)^{(\kappa_\eta)} - (m_{0,\eta}^-)^{(\kappa_\eta)}\right] \left[(m_{1,\eta}^+)^{(\kappa_\eta)} - (m_{1,\eta}^-)^{(\kappa_\eta)}\right],\end{aligned}$$

equation equation 14 becomes

$$\Gamma_\eta = p_\eta^- \Gamma_{\eta^-} + p_\eta^+ \Gamma_{\eta^+} + 2p_\eta^- p_\eta^+ \Delta_\eta. \tag{15}$$

Telescoping from the root. Let $\alpha_\eta := \widehat{\mu}_{\mathcal{D}}(R_\eta)$, $\alpha_{\eta^-} := \widehat{\mu}_{\mathcal{D}}(R_{\eta^-})$, $\alpha_{\eta^+} := \widehat{\mu}_{\mathcal{D}}(R_{\eta^+})$. Since $p_\eta^- = \alpha_{\eta^-}/\alpha_\eta$ and $p_\eta^+ = \alpha_{\eta^+}/\alpha_\eta$, multiplying equation 15 by α_η gives

$$\alpha_\eta \Gamma_\eta = \alpha_{\eta^-} \Gamma_{\eta^-} + \alpha_{\eta^+} \Gamma_{\eta^+} + 2\frac{\alpha_{\eta^-}\alpha_{\eta^+}}{\alpha_\eta} \Delta_\eta. \tag{16}$$

Telescoping from the root r (where $\alpha_r = 1$) and using $\Gamma_\ell = 0$ at all leaves,

$$\mathbb{E}_{\pi_{\mathrm{QAT}}^{\otimes 2}}\big[(X_0 - X_0')^\top (X_1 - X_1')\big] = \Gamma_r = 2\sum_{\eta \in \mathcal{I}} \frac{\alpha_{\eta^-}\alpha_{\eta^+}}{\alpha_\eta} \Delta_\eta. \tag{17}$$

Setting $q_\eta := \alpha_{\eta^-}/\alpha_\eta$, so that $\frac{\alpha_{\eta^-}\alpha_{\eta^+}}{\alpha_\eta} = \alpha_\eta q_\eta(1-q_\eta)$,

$$\mathbb{E}_{\pi_{\mathrm{QAT}}^{\otimes 2}}\big[(X_0 - X_0')^\top (X_1 - X_1')\big] = 2\Delta_{\mathrm{QAT}}. \tag{18}$$

Combining equation 11, equation 12, and equation 18,

$$\begin{aligned} &\mathbb{E}_{\pi_{\mathrm{QAT}}^{\otimes 2}}\big[\|X_t - X_t'\|_2^2\big] - \mathbb{E}_{\pi_{\mathrm{ind}}^{\otimes 2}}\big[\|X_t - X_t'\|_2^2\big] \\ &= 2t(1-t)\mathbb{E}_{\pi_{\mathrm{QAT}}^{\otimes 2}}\big[(X_0 - X_0')^\top (X_1 - X_1')\big] \\ &= 4t(1-t)\Delta_{\mathrm{QAT}}. \end{aligned} \tag{19}$$

Under non-degenerate splits, each effective split node satisfies $\Delta_\eta \geq 0$; if at least one node has $\Delta_\eta > 0$, then $\Delta_{\mathrm{QAT}} > 0$. This completes the proof. □

A.3 Properties of CQAT

Let $\mathcal{D}_c = \{(x_i, c_i)\}_{i=1}^N \subseteq \mathbb{R}^d \times \mathbb{R}^m$ be the conditional dataset with joint empirical distribution $\widehat{\mu}_{\mathcal{D}_c}$. Denote the Conditional QAT coupling by $\pi_{\mathrm{CQAT}} \in \Pi(\gamma, \widehat{\mu}_{\mathcal{D}_c})$, where $\gamma = \mathcal{N}(0, I_d)$. Throughout, X_1 refers to the data-coordinate component of a joint sample (X_1, C).

Proposition 1 (Non-crossing property of CQAT) *Let $\ell \neq \ell'$ be two leaf nodes of the Conditional QAT tree, and let η be the first internal node at which the root-to-ℓ and root-to-ℓ' paths diverge. If η is a data-coordinate split node, i.e., $\kappa_\eta \leq d$, then the linear interpolation paths drawn from the two leaf couplings do not intersect.*

Proof. Without loss of generality, assume ℓ lies in the left subtree η^- and ℓ' in the right subtree η^+. Since $\kappa_\eta \leq d$, the CQAT tree performs a data-coordinate split at η. By the tree construction, for any $(x_1, c_1) \in R_\ell$ and $(x_1', c_1') \in R_{\ell'}$,

$$x_1^{(\kappa_\eta)} \leq s_\eta < (x_1')^{(\kappa_\eta)}.$$

Moreover, since η is a data-coordinate split node, the source-side region is simultaneously partitioned: for any $x_0 \in \widetilde{R}_\ell$ and $x_0' \in \widetilde{R}_{\ell'}$,

$$x_0^{(\kappa_\eta)} \leq \widetilde{s}_\eta < (x_0')^{(\kappa_\eta)}.$$

Hence, for all $t \in [0,1]$,

$$(x_t')^{(\kappa_\eta)} - x_t^{(\kappa_\eta)} = (1-t)\big[(x_0')^{(\kappa_\eta)} - x_0^{(\kappa_\eta)}\big] + t\big[(x_1')^{(\kappa_\eta)} - x_1^{(\kappa_\eta)}\big] > 0,$$

which implies $x_t' \neq x_t$ for all $t \in [0,1]$. □

Proposition 2 (Path separation property of CQAT) *Let $x_t = (1-t)x_0 + tx_1$ and $x_t' = (1-t)x_0' + tx_1'$, where x_1, x_1' denote the data-coordinate components of joint samples. For any $t \in (0,1)$,*

$$\mathbb{E}_{\pi_{\mathrm{CQAT}}^{\otimes 2}}\big[\|x_t - x_t'\|_2^2\big] - \mathbb{E}_{\pi_{\mathrm{ind}}^{\otimes 2}}\big[\|x_t - x_t'\|_2^2\big] = 4t(1-t)\Delta_{\mathrm{CQAT}},$$

where Δ_{CQAT} is the node-level separation induced by the Conditional QAT tree structure:

$$\begin{aligned} \Delta_{\mathrm{CQAT}} := \sum_{\eta \in \mathcal{I}_c:\, \kappa_\eta \leq d} &\alpha_\eta q_\eta (1-q_\eta)\Big(\mathbb{E}_{\gamma_{\eta^+}}\big[X^{(\kappa_\eta)}\big] - \mathbb{E}_{\gamma_{\eta^-}}\big[X^{(\kappa_\eta)}\big]\Big) \\ &\times \Big(\mathbb{E}_{\widehat{\mu}_{\mathcal{D}_c,\eta^+}}\big[X^{(\kappa_\eta)}\big] - \mathbb{E}_{\widehat{\mu}_{\mathcal{D}_c,\eta^-}}\big[X^{(\kappa_\eta)}\big]\Big). \end{aligned} \tag{20}$$

Here $\mathcal{I}_c$ is the set of internal nodes of the Conditional QAT tree; $\alpha_\eta := \widehat{\mu}_{\mathcal{D}_c}(R_\eta)$ is the joint data mass at node η; and $q_\eta := \widehat{\mu}_{\mathcal{D}_c,\eta}(R_{\eta^-})$ is the conditional mass of the left child.

Proof. The telescoping recursion equation 17 in the proof of Theorem 2 relies only on the left-right child decomposition at each internal node and the product structure of the leaf couplings. Replacing the QAT tree with the Conditional QAT tree and $\widehat{\mu}_{\mathcal{D}}$ with the joint empirical distribution $\widehat{\mu}_{\mathcal{D}_c}$, the same argument gives

$$\mathbb{E}_{\pi_{\mathrm{CQAT}}^{\otimes 2}}\big[(X_0 - X_0')^\top (X_1 - X_1')\big] = 2\sum_{\eta \in \mathcal{I}_c} \frac{\alpha_{\eta^-}\alpha_{\eta^+}}{\alpha_\eta}\Delta_\eta,$$

where $\alpha_\eta := \widehat{\mu}_{\mathcal{D}_c}(R_\eta)$ and Δ_η is the local cross-term contribution at node η.

It remains to show that condition-coordinate split nodes contribute zero. If $\kappa_\eta > d$, the CQAT tree does not partition the source-side region at η, so both child nodes inherit the same source region: $\gamma_{\eta^-} = \gamma_{\eta^+}$. Consequently,

$$\mathbb{E}_{\gamma_{\eta^+}}[X] = \mathbb{E}_{\gamma_{\eta^-}}[X],$$

and the source-side mean difference vanishes, giving $\Delta_\eta = 0$. Therefore, only data-coordinate split nodes ($\kappa_\eta \le d$) contribute, and the sum reduces to

$$\mathbb{E}_{\pi_{\mathrm{CQAT}}^{\otimes 2}}\left[(X_0 - X_0')^\top (X_1 - X_1')\right] = 2\Delta_{\mathrm{CQAT}}.$$

Combining with the linear interpolation expansion equation 11 and the fact that the cross-term vanishes under the independent coupling π_{ind}, we obtain

$$\mathbb{E}_{\pi_{\mathrm{CQAT}}^{\otimes 2}}\left[\|X_t - X_t'\|_2^2\right] - \mathbb{E}_{\pi_{\mathrm{ind}}^{\otimes 2}}\left[\|X_t - X_t'\|_2^2\right] = 4t(1-t)\Delta_{\mathrm{CQAT}}.$$

Under non-degenerate splits, if at least one data-coordinate split node contributes positively, then $\Delta_{\mathrm{CQAT}} > 0$. This completes the proof. □

B Algorithms

B.1 QAT Coupling Construction and Sampling

This section presents pseudocode for QAT coupling construction and training-pair sampling. QAT uses $\gamma = \mathcal{N}(0, I_d)$ as the source distribution and maintains source-side coordinate bounds $(a_\ell, b_\ell) \in (\mathbb{R}\cup\{\pm\infty\})^d \times (\mathbb{R}\cup\{\pm\infty\})^d$ at each leaf node. Throughout, Φ and Φ^{-1} denote the standard-normal CDF and its inverse, respectively.

Algorithm 1 QAT Coupling Construction

Require: Dataset $\mathcal{D} = \{x_i\}_{i=1}^N \subseteq \mathbb{R}^d$
$\mathcal{S}_r \leftarrow \{1, \dots, N\}$, $a_r \leftarrow (-\infty, \dots, -\infty)$, $b_r \leftarrow (+\infty, \dots, +\infty)$
$\mathcal{Q} \leftarrow \{r\}$, $\mathcal{L} \leftarrow \varnothing$
while $\mathcal{Q} \neq \varnothing$ **do**
  Pop one node η from $\mathcal{Q}$
  if $|\mathcal{S}_\eta| \le 1$ **then**
    Add $(\mathcal{S}_\eta,\, a_\eta,\, b_\eta)$ to $\mathcal{L}$
  else
    $\kappa_\eta \leftarrow \arg\max_{j\in[d]} \mathrm{Var}_{\{x_i : i\in\mathcal{S}_\eta\}}(x^{(j)})$, $\quad s_\eta \leftarrow \frac{1}{|\mathcal{S}_\eta|}\sum_{i\in\mathcal{S}_\eta} x_i^{(\kappa_\eta)}$
    $\mathcal{S}_{\eta^-} \leftarrow \{i \in \mathcal{S}_\eta : x_i^{(\kappa_\eta)} \le s_\eta\}$, $\quad \mathcal{S}_{\eta^+} \leftarrow \mathcal{S}_\eta \setminus \mathcal{S}_{\eta^-}$, $\quad q_\eta \leftarrow |\mathcal{S}_{\eta^-}|/|\mathcal{S}_\eta|$
    $\widetilde{s}_\eta \leftarrow \Phi^{-1}\big(\Phi(a_\eta^{(\kappa_\eta)}) + q_\eta \cdot \big(\Phi(b_\eta^{(\kappa_\eta)}) - \Phi(a_\eta^{(\kappa_\eta)})\big)\big)$
    $(a_{\eta^-}, b_{\eta^-}) \leftarrow (a_\eta, b_\eta)$; $\quad (a_{\eta^+}, b_{\eta^+}) \leftarrow (a_\eta, b_\eta)$
    $b_{\eta^-}^{(\kappa_\eta)} \leftarrow \widetilde{s}_\eta$; $\quad a_{\eta^+}^{(\kappa_\eta)} \leftarrow \widetilde{s}_\eta$
    Push η^- and η^+ into $\mathcal{Q}$
  end if
end while
return $\mathcal{L} = \{(\mathcal{S}_\ell,\, a_\ell,\, b_\ell) : \ell \in \mathcal{L}\}$

Because the construction terminates when $|\mathcal{S}_\ell| \le 1$, each non-empty leaf holds exactly one training sample. Every leaf therefore carries mass $1/N$, and sampling a leaf is equivalent to drawing a data index uniformly at random.

In flow matching training, the sampled pair (ϵ, x) serves as (x_0, x_1).

Complexity analysis. Assuming the tree is approximately balanced, it has $\mathcal{O}(\log N)$ levels. At each level, computing coordinate means and variances over all N samples costs $\mathcal{O}(Nd)$, giving a total construction cost of $\mathcal{O}(Nd \log N)$. Sampling a single training pair requires generating d independent truncated-Gaussian variates via the inverse-CDF method, each in $O(1)$, for a per-sample cost of $\mathcal{O}(d)$. Epoch-level sampling therefore costs $\mathcal{O}(Nd)$, matching standard Gaussian sampling in asymptotic complexity.

Algorithm 2 QAT Coupling Training-Pair Sampling

Require: Leaf list $\mathcal{L} = \{(\mathcal{S}_\ell,\, a_\ell,\, b_\ell) : \ell \in \mathcal{L}\}$
Sample $\ell \sim \mathrm{Unif}(\mathcal{L})$; let $\mathcal{S}_\ell = \{i\}$ and set $x \leftarrow x_i$
for $j = 1, \ldots, d$ **do**
    Sample $u_j \sim \mathrm{Unif}\big(\Phi(a_\ell^{(j)}),\, \Phi(b_\ell^{(j)})\big)$; set $\epsilon^{(j)} \leftarrow \Phi^{-1}(u_j)$
end for
return $(\epsilon,\, x)$

Vectorized implementation. In practice, the source-side bounds for each sample's corresponding leaf are precomputed and stacked into two matrices $\mathbf{A}, \mathbf{B} \in (\mathbb{R} \cup \{\pm\infty\})^{N \times d}$ indexed by data order. Given a batch of indices $\mathcal{I} = \{i_b\}_{b=1}^{B}$, one retrieves $\mathbf{A}_{\mathcal{I},:}$ and $\mathbf{B}_{\mathcal{I},:}$ directly and draws $U_{bj} \sim \mathrm{Unif}\big(\Phi(\mathbf{A}_{i_b,j}), \Phi(\mathbf{B}_{i_b,j})\big)$ in parallel, then sets $\epsilon_{bj} = \Phi^{-1}(U_{bj})$. The entire batch sampling step is therefore fully vectorizable.

B.2 CQAT Coupling Construction and Sampling

In CQAT, the condition variable is embedded as a Euclidean vector and concatenated with the data feature to form $y_i = (x_i, c_i) \in \mathbb{R}^{d+m}$; the algorithm then builds a tree directly on the joint sample set $\mathcal{D}_c = \{y_i\}_{i=1}^{N}$. To prevent late-stage splits driven purely by condition coordinates — which partition the data index set without updating the source-side bounds — we restrict the candidate split coordinates to $[d]$ whenever the node sample count falls below a threshold n_{data}. Algorithm 3 gives the full construction.

Algorithm 3 CQAT Coupling Construction

Require: Joint dataset $\mathcal{D}_c = \{y_i\}_{i=1}^{N} \subseteq \mathbb{R}^{d+m}$, data-split threshold n_{data}
$\mathcal{S}_r \leftarrow \{1, \ldots, N\}$, $a_r \leftarrow (-\infty, \ldots, -\infty)$, $b_r \leftarrow (+\infty, \ldots, +\infty)$
$\mathcal{Q} \leftarrow \{r\}$, $\mathcal{L} \leftarrow \varnothing$
while $\mathcal{Q} \neq \varnothing$ **do**
    Pop one node η from $\mathcal{Q}$
    if $|\mathcal{S}_\eta| \leq 1$ **then**
        Add $(\mathcal{S}_\eta,\, a_\eta,\, b_\eta)$ to $\mathcal{L}$
    else
        $\mathcal{J}_\eta \leftarrow [d]$ if $|\mathcal{S}_\eta| \leq n_{\mathrm{data}}$, else $\mathcal{J}_\eta \leftarrow [d+m]$
        $\kappa_\eta \leftarrow \arg\max_{j \in \mathcal{J}_\eta} \mathrm{Var}_{\{y_i : i \in \mathcal{S}_\eta\}}(y^{(j)})$, $s_\eta \leftarrow \frac{1}{|\mathcal{S}_\eta|} \sum_{i \in \mathcal{S}_\eta} y_i^{(\kappa_\eta)}$
        $\mathcal{S}_{\eta^-} \leftarrow \{i \in \mathcal{S}_\eta : y_i^{(\kappa_\eta)} \leq s_\eta\}$, $\mathcal{S}_{\eta^+} \leftarrow \mathcal{S}_\eta \setminus \mathcal{S}_{\eta^-}$
        $(a_{\eta^-}, b_{\eta^-}) \leftarrow (a_\eta, b_\eta)$; $(a_{\eta^+}, b_{\eta^+}) \leftarrow (a_\eta, b_\eta)$
        if $\kappa_\eta \leq d$ **then**
            $q_\eta \leftarrow |\mathcal{S}_{\eta^-}| / |\mathcal{S}_\eta|$
            $\widetilde{s}_\eta \leftarrow \Phi^{-1}\big(\Phi(a_\eta^{(\kappa_\eta)}) + q_\eta \cdot \big(\Phi(b_\eta^{(\kappa_\eta)}) - \Phi(a_\eta^{(\kappa_\eta)})\big)\big)$
            $b_{\eta^-}^{(\kappa_\eta)} \leftarrow \widetilde{s}_\eta$; $a_{\eta^+}^{(\kappa_\eta)} \leftarrow \widetilde{s}_\eta$
        end if
        Push η^- and η^+ into $\mathcal{Q}$
    end if
end while
return $\mathcal{L} = \{(\mathcal{S}_\ell,\, a_\ell,\, b_\ell) : \ell \in \mathcal{L}\}$

Scale balancing. Since CQAT builds on data and condition coordinates jointly, scale imbalance between x_i and c_i may bias split selection toward high-variance condition dimensions. We therefore build the tree on $(X,\, wC)$. For discrete class labels, we set $w = \infty$, which is formally equivalent to first partitioning all samples by class and then constructing an independent QAT within each class subtree — matching the approach of C2OT (Cheng & Schwing, 2025). For continuous conditions,

we adopt the variance-normalized weight

$$w = \left(\frac{\sum_j \mathrm{Var}(X^{(j)})}{\sum_k \mathrm{Var}(C^{(k)})} \right)^{1/2}, \tag{21}$$

so that the rescaled condition variable has aggregate variance comparable to that of the data features.

Complexity analysis. Building the tree over $d + m$ coordinates gives a construction cost of $\mathcal{O}(N(d+m)\log N)$ under a balanced-tree assumption. Condition-coordinate splits ($\kappa_\eta > d$) partition the data index set without touching the source-side bounds; data-coordinate splits ($\kappa_\eta \le d$) perform quantile alignment exactly as in QAT. At training time, Algorithm 2 applies directly: one returns the data sample x_i together with its paired condition variable c_i. Since source sampling remains confined to d-dimensional Gaussian space, the per-sample cost is $\mathcal{O}(d)$, unchanged from QAT.

B.3 High-Dimensional Image Data Processing

For high-dimensional image data, we do not build the QAT directly in raw pixel space. Instead, we first apply an orthogonal transform to concentrate the principal variation into the leading coordinates. Crucially, only a pure rotation is applied — no mean subtraction — so that if the source is $\mathcal{N}(0, I_d)$, it remains standard Gaussian under both the forward and inverse transforms.

Patch-Hadamard transform. Let the input image have spatial resolution $H \times W$ with C channels, flattened to $x_i \in \mathbb{R}^d$ ($d = HWC$). For patch size p, the image is divided into $d' = (H/p)(W/p)C$ non-overlapping patches, each of spatial dimension $r = p^2$. A normalized Hadamard matrix $H_{\mathrm{patch}} \in \mathbb{R}^{r\times r}$ (Horadam, 2012) is applied independently to each patch. Because the normalized Hadamard matrix is orthogonal with constant direction $r^{-1/2}\mathbf{1}_r$, extracting the coefficient along this direction is equivalent to patch-level mean pooling. We denote the resulting mean-pooled representation as $\bar{x}_i \in \mathbb{R}^{d'}$.

Let $H_p \in \mathbb{R}^{d\times d}$ denote the global patch-wise Hadamard transform induced by patch size p. Since each per-patch transform is orthogonal, H_p is itself orthogonal. We order the output coordinates so that the first d' entries of $H_p x_i$ correspond to the constant-direction coefficients across all patches. In practice, if only the low-dimensional tree representation is needed, $\bar{x}_i$ can be obtained directly by patch-level mean pooling, with PCA directions estimated on this d'-dimensional representation.

PCA rotation and low-dimensional QAT. Patch-Hadamard pooling avoids the cost of PCA in the full d-dimensional pixel space. We estimate the PCA rotation matrix $\mathbf{P} \in \mathbb{R}^{d'\times d'}$ from $\{\bar{x}_i\}_{i=1}^N$ without mean shift. Setting $z_i = \mathbf{P}\bar{x}_i \in \mathbb{R}^{d'}$, we use only the leading k coordinates $z_i^{(1:k)}$ to construct the QAT coupling. Algorithm 4 summarizes the full construction pipeline.

Algorithm 4 High-Dimensional QAT Construction

Require: Image dataset $\mathcal{D} = \{x_i\}_{i=1}^N$, patch size p, reduced dimension k
1: $d' \leftarrow (H/p)(W/p)C$
2: $\bar{x}_i \leftarrow (H_p x_i)^{(1:d')}$, for $i = 1, \dots, N$
3: Estimate PCA rotation $\mathbf{P} \in \mathbb{R}^{d'\times d'}$ from $\{\bar{x}_i\}_{i=1}^N$ without mean shift
4: $\widetilde{z}_i \leftarrow (\mathbf{P}\bar{x}_i)^{(1:k)}$, for $i = 1, \dots, N$
5: Run Algorithm 1 on $\{\widetilde{z}_i\}_{i=1}^N$
6: **return** $\mathcal{L}$, $\mathbf{P}$, H_p

Inverse-transform sampling. The low-dimensional QAT constructs a coupling over the leading k PCA-pooled coordinates, whose source marginal is $\mathcal{N}(0, I_k)$. To recover a source sample in the original d-dimensional pixel space, we fill the remaining coordinates with independent standard Gaussian noise: first the $d' - k$ residual PCA-pooled coordinates, then the high-frequency Hadamard coefficients for each patch (all per-patch dimensions except the constant direction). Since $\mathbf{P}$ and H_p are both orthogonal and all unfilled coordinates are drawn as independent standard Gaussians, the resulting source sample has marginal exactly $\mathcal{N}(0, I_d)$.

Algorithm 5 High-Dimensional QAT Sampling

Require: $\mathcal{D} = \{x_i\}_{i=1}^N$, leaf list $\mathcal{L} = \{(\mathcal{S}_\ell,\, a_\ell,\, b_\ell) : \ell \in \mathcal{L}\}$, $\mathbf{P}$, H_p, d', k
Sample one leaf $\ell \sim \mathrm{Unif}(\mathcal{L})$; let $\mathcal{S}_\ell = \{i\}$ and set $x \leftarrow x_i$
for $j = 1, \ldots, k$ **do**
  Sample $u_j \sim \mathrm{Unif}\big(\Phi(a_\ell^{(j)}),\, \Phi(b_\ell^{(j)})\big)$; set $\epsilon^{(j)} \leftarrow \Phi^{-1}(u_j)$
end for
Sample $\epsilon^{(k+1:d')} \sim \mathcal{N}(0, I_{d'-k})$
$\epsilon_{\mathrm{pool}} \leftarrow \big(\epsilon^{(1:k)},\, \epsilon^{(k+1:d')}\big)$
$\bar{\epsilon} \leftarrow \mathbf{P}^\top \epsilon_{\mathrm{pool}}$
Sample $\epsilon_{\mathrm{high}} \sim \mathcal{N}(0, I_{d-d'})$
$\epsilon_H \leftarrow \big(\bar{\epsilon},\, \epsilon_{\mathrm{high}}\big)$
$x_0 \leftarrow H_p^\top \epsilon_H$
return $(x_0,\, x)$

Complexity analysis. Let $d = HWC$, $r = p^2$, and $d' = (H/p)(W/p)C$. Computing patch-level mean pooling costs $\mathcal{O}(Nd)$; the full patch-wise Hadamard transform can be applied via the Fast Walsh-Hadamard Transform (Horadam, 2012) in $\mathcal{O}(Nd \log r)$. PCA is estimated on the d'-dimensional pooled representation, and the low-dimensional QAT is built in k-dimensional space, giving tree construction cost $\mathcal{O}(Nk \log N)$. At sampling time, the inverse PCA rotation costs $\mathcal{O}(d'^2)$ per sample and the inverse patch-wise Hadamard transform costs $\mathcal{O}(d \log r)$, giving a total per-sample cost of $\mathcal{O}(d + d'^2 + d \log r)$; batch sampling is fully vectorizable.

C Implementation Details

This section details the per-experiment training configurations. Unless stated otherwise, all methods share the same network architecture, training configuration, and sampling protocol, differing only in the coupling strategy used to form training pairs: I-FM uses independent coupling, OT-FM uses mini-batch OT coupling, and QAT-FM uses the QAT or Conditional QAT coupling proposed in this work.

C.1 Two-Dimensional Experiment

Data. Following Lipman et al. (2023); Pooladian et al. (2023), we consider transport from a two-dimensional standard Gaussian source distribution to a checkerboard target distribution. Training samples are drawn independently from the source and target; the flow paths learned by I-FM, OT-FM, SD-FM, and QAT-FM are then compared under the same network architecture and optimization configuration.

Network and training. Following Lipman et al. (2024), the velocity field is parameterized by a five-layer MLP with hidden width 512 and Swish activations (Ramachandran et al., 2017). All methods are trained for 20K iterations with batch size 256 and Adam (Kingma & Ba, 2015) at learning rate 10^{-3}. For QAT-FM, 50K samples are drawn from the checkerboard target to build the QAT tree, and training pairs are sampled from the induced coupling. For OT-FM, exact EMD is solved within each mini-batch using POT (Flamary et al., 2021). For SD-FM, a discrete reference measure is built from 50K target samples and the corresponding regularized problem is solved with entropy parameter 0.05; we optimize with Adam, as the SGD-based strategy of Genevay et al. (2016); Mousavi-Hosseini et al. (2026) proved unstable and slow to converge in this setting.

C.2 CIFAR-10

We conduct unconditional image generation on CIFAR-10 (Krizhevsky et al., 2009), using the U-Net architecture of Tong et al. (2024). Main network and training hyperparameters are listed in Table 6. We train with Adam, a linear warmup to 2×10^{-4} over 5k steps (constant thereafter), gradient norm clipping at 1.0, and an EMA decay of 0.9999.

Table 6: Hyperparameter settings for image-space experiments.

	CIFAR-10	ImageNet-32	CelebA-64
Channels	192	256	192
Depth	2	3	3
Channel multiplier	1,2,2,2	1,2,2,2	1,2,3,4
Heads	4	4	4
Head channels	64	64	64
Attention resolution	16	4	8
Dropout	0.0	0.0	0.0
Batch size / GPU	256	128	64
GPUs	1	4	4
Effective batch size	256	512	256
Iterations	100k	200k	100k
Learning rate	2e-4	1e-4	1e-4
LR scheduler	Constant	Linear decay	Constant
Warmup steps	5k	20k	10k

For QAT-FM, the QAT coupling is built from the $50{,}000$ CIFAR-10 training images together with their horizontally flipped augmentations, giving $n = 100{,}000$ tree samples. Images are first processed by patch-level mean pooling with patch size $p = 4$; PCA rotation is then estimated on the pooled representation, and the leading $k = 32$ coordinates are retained for tree construction. The tree splits on the maximum-variance coordinate at the empirical mean threshold and recurses until each leaf contains at most one sample. At evaluation, 50K images are generated, and FID is computed using `clean-fid` (Parmar et al., 2022), following the protocol of Tong et al. (2024).

C.3 IMAGENET-32

We evaluate class-conditional generation on ImageNet-1k (Deng et al., 2009) at 32×32 resolution. All images are resized to 32×32 with random horizontal flipping, and class labels are provided as condition inputs to the velocity-field network. We follow the class-conditional U-Net architecture of Tong et al. (2024); main hyperparameters are listed in Table 6. Optimizer, gradient clipping, and EMA settings match the CIFAR-10 setup. The learning rate warms up linearly to 1×10^{-4} over 20k steps and then decays linearly to 3×10^{-5}.

For QAT-FM, the Conditional QAT coupling is built from ImageNet-1k training images and their horizontal-flip augmentations, yielding over 2M tree samples. Patch-level mean pooling ($p = 4$) and PCA rotation are applied as in the CIFAR-10 setup, retaining the leading $k = 32$ coordinates. Since the condition is a discrete class label, training samples are first partitioned by class and an independent QAT coupling is constructed within each class subset. For C2OT (Cheng & Schwing, 2025), we follow its oversampling strategy with ratio 50, yielding a local OT batch size of $128 \times 50 = 6400$ per GPU. Following the C2OT setting for class-conditional generation, we use its conditional weighting scheme and set the condition weight to 10^8.

At evaluation, 50K image samples are generated to estimate Inception feature statistics of the generated distribution; reference statistics are computed from all $\sim$1.28M ImageNet-1k training images, following Dhariwal & Nichol (2021); Ma et al. (2024).

C.4 CELEBA-64

Data. We evaluate text-conditional generation on CelebA (Liu et al., 2015) at 64×64 resolution. Since CelebA does not provide text annotations by default, we use the natural-language descriptions from CelebA-Dialog (Jiang et al., 2021) as text conditions.

Network and Training. We use the U-Net architecture with main hyperparameters in Table 6. Following Cheng & Schwing (2025), scale-shift normalization is applied in the U-Net residual blocks, and text embeddings are projected to the conditioning dimension via a linear layer before being fed to the network. Optimizer, gradient clipping, and EMA settings match the CIFAR-10 experiment; the learning rate warms up linearly to 1×10^{-4} over 10k steps and remains constant thereafter.

For QAT-FM, a Conditional QAT coupling is constructed in the joint space of low-dimensional image representations and text embeddings. The image-side representation is obtained via patch-level mean pooling ($p = 4$) and PCA rotation, retaining the leading $k = 32$ coordinates. Text conditions are encoded by a `DFN5B-CLIP-ViT-H-14` text encoder (Fang et al., 2024); the resulting embeddings are projected to 32 dimensions via PCA and used as condition-side features in the joint tree. To prevent late-stage splits driven purely by condition coordinates, candidate split dimensions are restricted to data coordinates when a node's sample count falls below $n_{\text{data}} = 128$. For C2OT (Cheng & Schwing, 2025), we follow its oversampling strategy with ratio 50, a local OT batch size of $64 \times 50 = 3200$ per GPU, which substantially increases the coupling cost; see Appendix D.1. For a fair conditional comparison, we use the Euclidean distance between text embeddings as the conditional distance and apply the weighting scheme in equation 21.

Evaluation. At evaluation, 50K images are generated and assessed along two axes: image fidelity via FID against training-set reference statistics (Ning et al., 2023; Wang et al., 2023), and text-image alignment via CLIP score computed with the `siglip2-base-patch16-224` checkpoint of SigLIP-2 (Tschannen et al., 2025).

C.5 IMAGENET-256

We conduct class-conditional latent-space generation on ImageNet-256×256 (Deng et al., 2009). Images are encoded by VA-VAE (Yao et al., 2025) into latent representations, and a flow-based LightningDiT (Yao et al., 2025) is trained in the resulting latent space with class labels as condition inputs. All network architecture, optimizer, learning-rate schedule, EMA, precision, and sampling configurations follow the official implementation of Yao et al. (2025); due to computational constraints, we adopt the recommended lightweight training setting of 64 epochs.

For QAT-FM, the Conditional QAT coupling is built directly over VA-VAE latent representations. PCA projection is applied, retaining the leading $k = 256$ coordinates for tree construction; patch-level mean pooling is omitted since VA-VAE already maps images to a compact latent space. The tree splits on the maximum-variance coordinate at the empirical mean and recurses until each leaf contains at most one sample. Classifier-free guidance is applied during inference with a scale of 10.

At evaluation, we follow the protocol of Yao et al. (2025) and compute FID against the ImageNet-256×256 reference statistics from guided-diffusion (Dhariwal & Nichol, 2021). We additionally report IS, sFID, Precision, and Recall.

D EXTENDED EXPERIMENTAL RESULTS

This section provides supplementary results comprising a runtime comparison of coupling strategies and additional visualization of generated samples. Appendix D.1 benchmarks the wall-clock time of each coupling approach at both construction and sampling phases; the subsequent four subsections present additional generated samples from CIFAR-10, ImageNet-32×32, CelebA-64×64, and ImageNet-256×256.

D.1 RUNTIME COMPARISON

We compare the wall-clock time of different coupling strategies at both construction and sampling phases. All experiments fix the feature dimension at $d = 32$ and vary the dataset size n. To reflect the repeated pairing demand of FM training, we report cumulative times over 100 training epochs. For OT-based methods, source points are resampled from a standard Gaussian and the coupling is recomputed from scratch at every epoch; by contrast, the QAT and SD-FM couplings are constructed once offline and reused throughout training. Full OT is solved with the `EMD` solver from POT (Flamary et al., 2021); mini-batch OT (OTBatch) uses `ot.dist_batch` to compute batch pairwise distances, fully exploiting GPU parallelism. Entropy regularization with $\varepsilon = 0.05$ is applied for both SD-FM and batch OT. SD-FM optimizes a semi-discrete coupling with SGD (Mousavi-Hosseini et al., 2026), running until convergence or 100 optimization epochs.

Figure 6 reports cumulative construction and sampling times over 100 epochs. Full OT becomes rapidly infeasible as n grows; batch OT reduces the per-epoch cost but must recompute the coupling

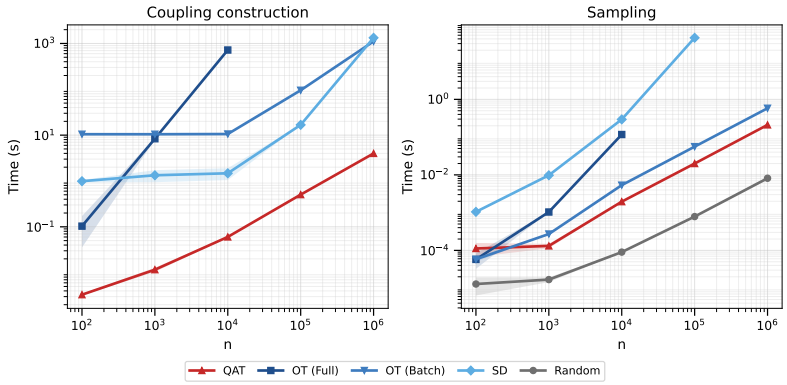


Figure 6: Runtime comparison of different coupling strategies. Left: total coupling construction time over 100 epochs. Right: total sampling time over 100 epochs.

every epoch, keeping its total time high. SD-FM requires only a single construction pass, but the semi-discrete optimization itself is expensive. By contrast, QAT constructs the tree and records the Gaussian bounds at each leaf in a single offline pass, achieving substantially lower construction time at all scales. At sampling time, QAT draws source points by querying leaf bounds and sampling from truncated Gaussians, placing its per-sample cost close to that of independent (random) sampling. OT-based methods must sample from the coupling's joint distribution, and SD-FM requires a nearest-neighbor lookup over all training points at each draw—both become prohibitive at large n. Overall, QAT achieves structured coupling at a sampling efficiency comparable to independent coupling.

D.2 CIFAR-10

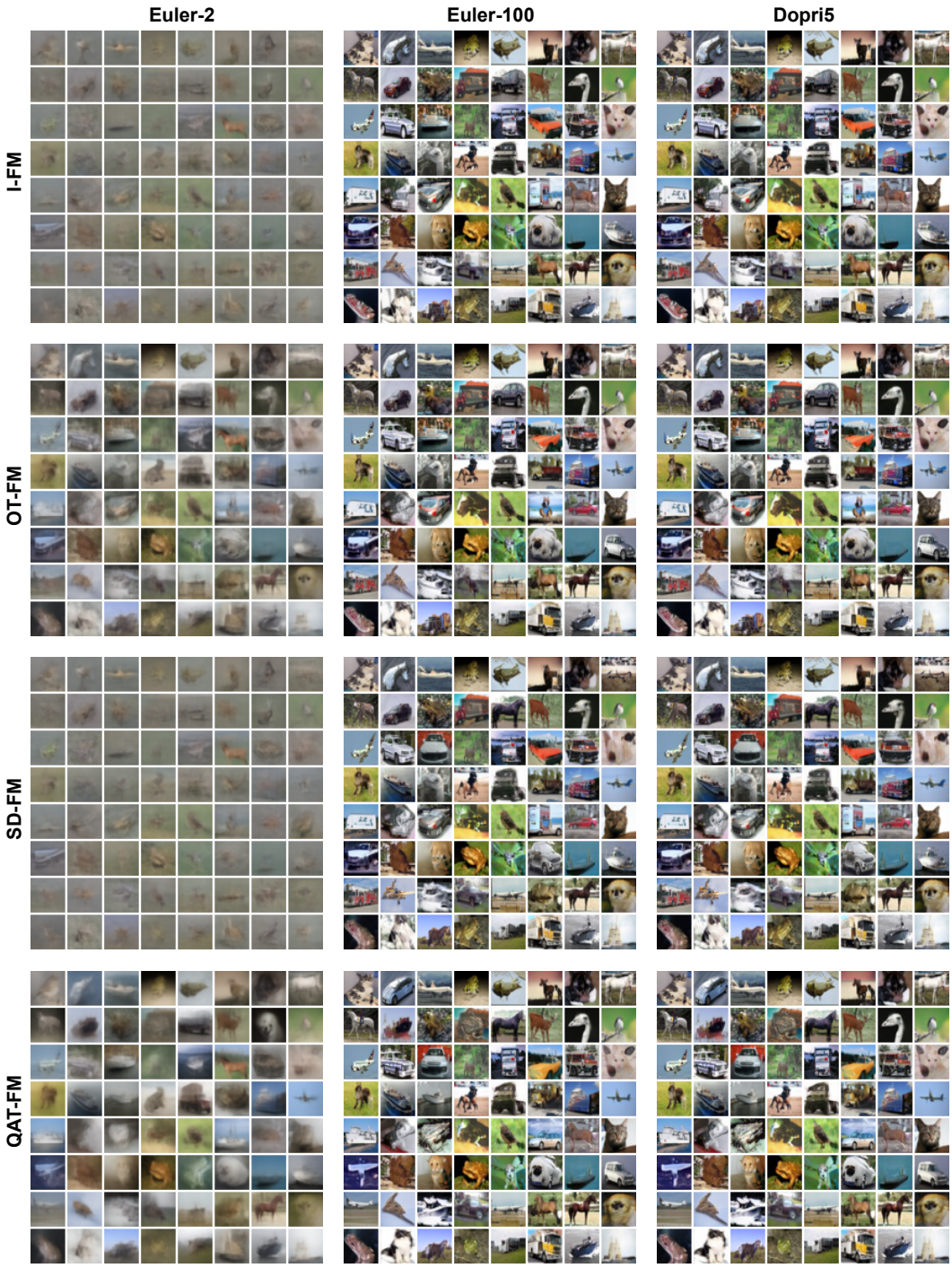


Figure 7: Additional generated samples for unconditional CIFAR-10 generation. Rows correspond to I-FM, OT-FM, SD-FM, and QAT-FM; columns compare three sampling configurations: `Euler-2`, `Euler-100`, and `dopri5`. All images are randomly generated without cherry-picking.

D.3 IMAGENET-32

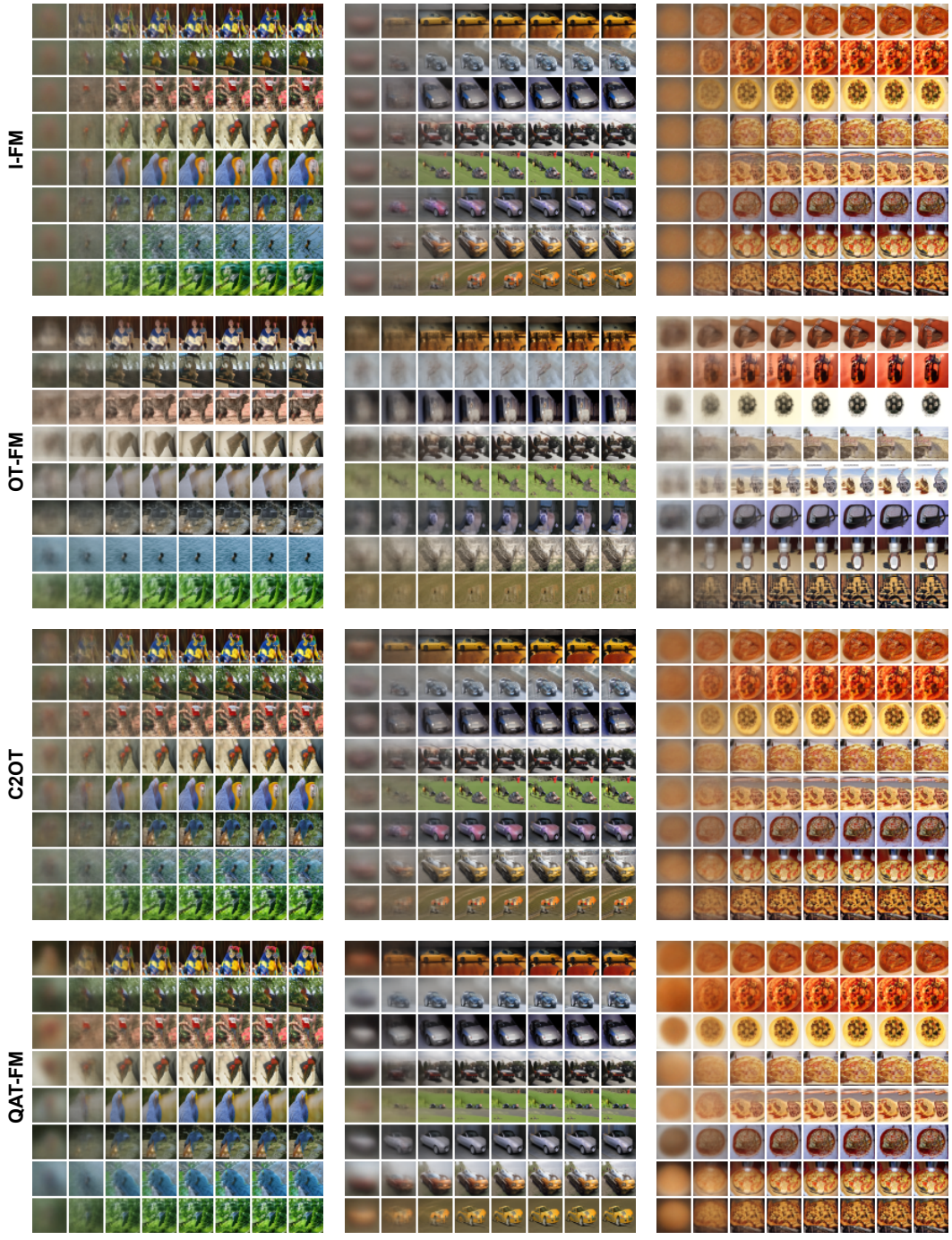


Figure 8: Additional generated samples for class-conditional ImageNet-32 $\times$ 32 generation. The four row blocks correspond to I-FM, OT-FM, C2OT, and QAT-FM, respectively. The three column blocks correspond to macaw, sports car, and pizza class conditions. Within each class block, the eight columns show samples generated with `Euler-1`, `Euler-2`, `Euler-5`, `Euler-10`, `Euler-20`, `Euler-50`, `Euler-100`, and `Dopri5`, respectively. All images are randomly generated without cherry-picking.

D.4 CELEBA-64

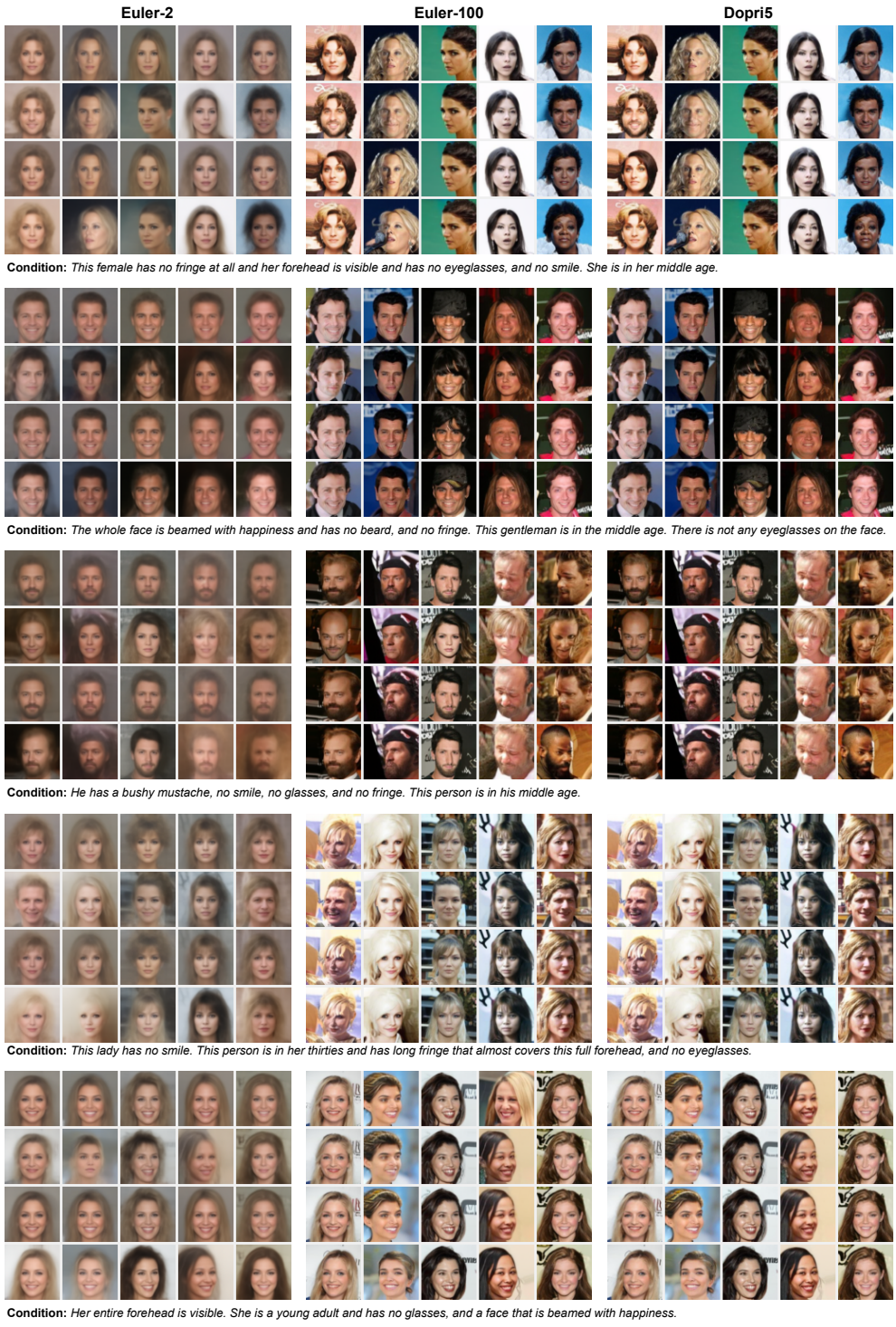


Figure 9: Additional generated samples for text-conditional CelebA-64 $\times$ 64 generation. Each horizontal block corresponds to one CelebA-Dialog text condition; the three columns show results under `Euler-2`, `Euler-100`, and `Dopri5`; within each block, the four rows correspond to I-FM, OT-FM, C2OT and QAT-FM. All images are randomly generated without cherry-picking.

D.5 IMAGENET-256

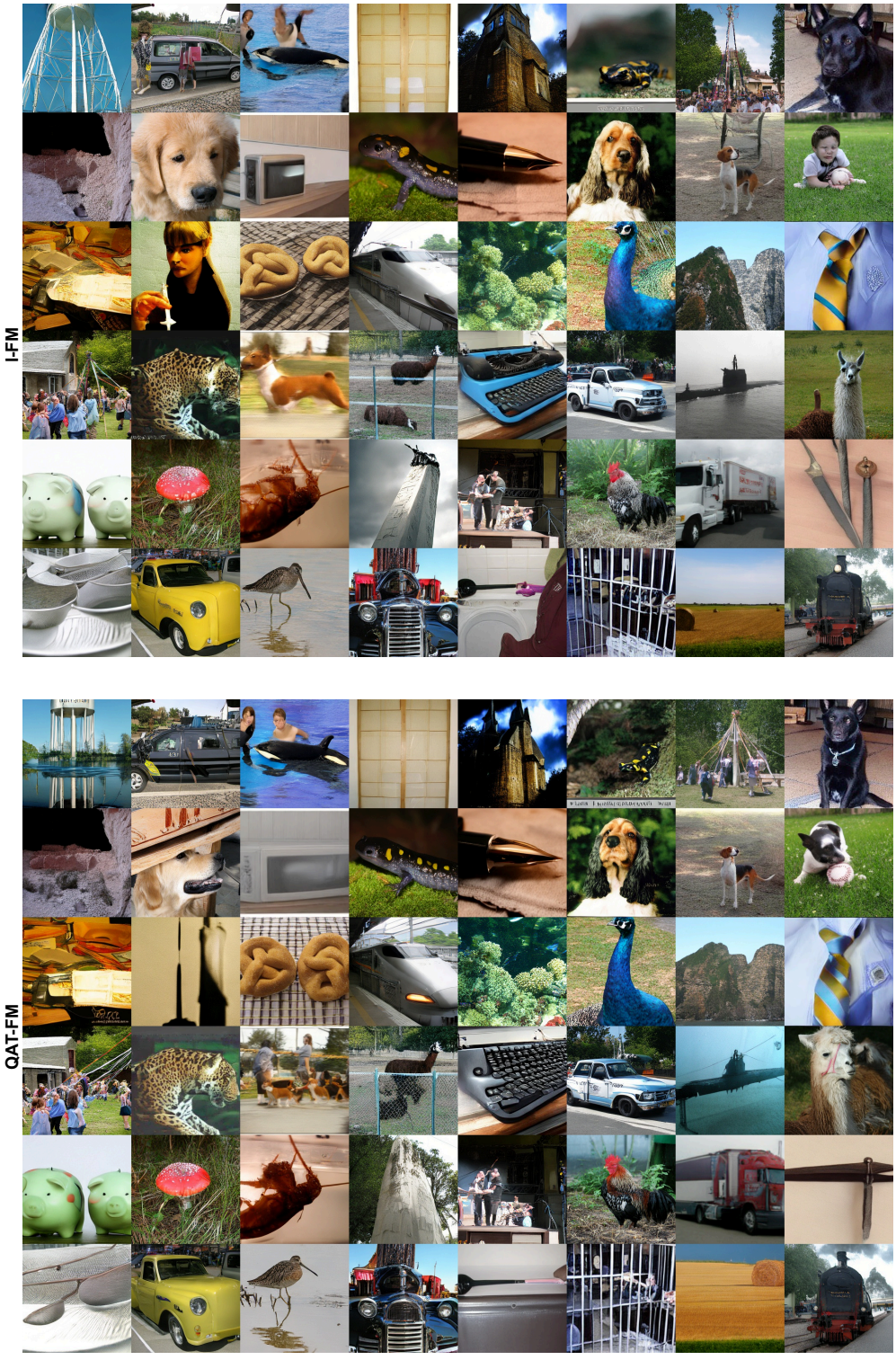


Figure 10: Additional generated samples for class-conditional latent-space generation on ImageNet-256 × 256. The top and bottom groups show samples from I-FM and QAT-FM, respectively, generated under the same random seed and class conditions. All images are randomly generated without cherry-picking.