QUANTIZED ONE-STEP GENERATIVE MODELING VIA CELL-MARGIN QUANTIZATION-AWARE TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-bit quantization and one-step sampling can substantially reduce the cost of flow-matching models, yet existing quantization-aware training methods typically minimize velocity-field reconstruction error, although generation quality is ultimately determined by the terminal distribution after ODE integration. We introduce **CM-QAT**, which trains a quantized student for one-step sampling with one function evaluation (1-NFE) to preserve the teacher's terminal geometry in a frozen feature space. Specifically, CM-QAT constructs a public nested partition and uses a differentiable prototype-margin objective to encourage each source sample to reach the correct terminal cell. Our counterexamples show that small velocity-L_2 error is neither necessary for terminal cell correctness nor sufficient for a distribution-free cell-risk guarantee without boundary-margin control. We derive conditional feature-space W_1 bounds from recursive cell-histogram discrepancies and relate cell-crossing probability to endpoint displacement and boundary margins. Experiments on Fashion-MNIST, CIFAR-10, and class-conditioned ImageNet-256 show that CM-QAT outperforms recent low-bit flow-generation methods and enables generation with models quantized to as low as 2-bit & 3-bit weights and sampled in a single step (1-NFE).

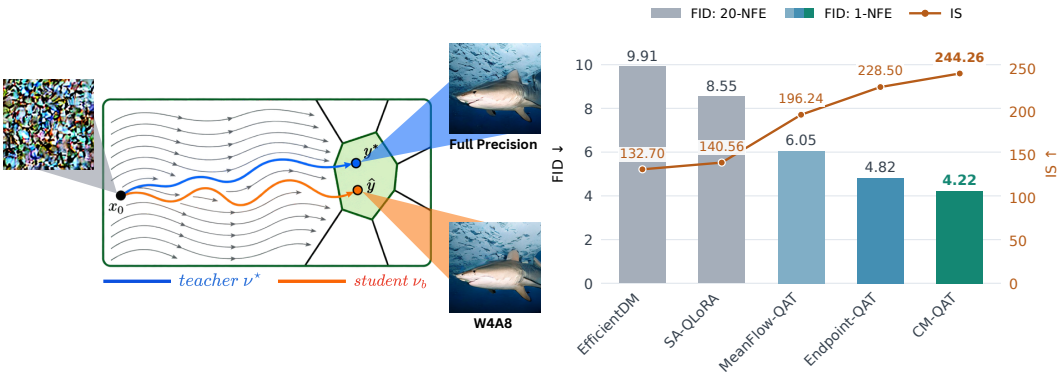


Figure 1: **CM-QAT Overview.** Left: CM-QAT preserves the full-precision teacher's terminal feature-space cell under one-step quantized generation. Right: Result on ImageNet-256 with W4A8 and 1-step generation, CM-QAT achieves the best FID among compared baselines.

1 INTRODUCTION

Recent years have witnessed rapid progress in generative modeling, driven by advances in diffusion (Ho et al., 2020) and flow-matching (Lipman et al., 2022) models, with applications spanning in various fields: image generation (Rombach et al., 2022), diffusion language models (Nie et al., 2026), speech synthesis (Popov et al., 2021; Shen et al., 2024), and robot learning (Chi et al., 2024; Black et al., 2025). However, their deployment is still constrained by two major computational bottlenecks: long iterative sampling chains and the repeated evaluation of large denoising or velocity networks, which is particularly costly for real-time, resource-constrained, and large-scale applications. The first bottleneck has been addressed through advanced samplers and distillation methods

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.

(Chen et al., 2018; Geng et al., 2025; Holderrieth et al., 2026) that construct efficient trajectories by approximating ODEs or SDEs; in flow matching, a learned velocity field defines an ODE that transports a simple reference distribution to the data distribution without ODE simulation during training. The second bottleneck can be reduced through compression techniques such as quantization (Jacob et al., 2018), which lower model size, memory, and per-evaluation cost. A natural approach for flow matching is therefore quantization-aware training, where the quantized velocity field v_b is trained to approximate the full-precision field $v_\star$ in L_2: $\mathcal{L}_{\mathrm{vel}} = \mathbb{E}_{t,x}\left[\|v_b(t,x) - v_\star(t,x)\|_2^2\right]$.

Our counterexamples show that, however, velocity-L_2 error alone does not provide a distribution-free guarantee of terminal cell correctness. A continuous normalizing flow (CNF) (Rezende & Mohamed, 2015; Onken et al., 2021) generates samples by integrating an ODE, while downstream metrics such as FID, precision, and mode coverage depend on the terminal distribution $\rho_b = (\Phi_b^1)_{\#}\rho_0$. Thus, large field error can still preserve the correct semantic region, whereas small field error can move substantial mass across a decision boundary and alter the cell histogram, particularly when mass lies near the boundary. This motivates studying where integrated endpoints land relative to a partition of feature space, rather than relying solely on pointwise velocity matching. We therefore formulate low-bit, one-step flow matching from a terminal-distribution perspective. Mai et al. (2026) shows that partitioning the target space into local cells yields a bit-efficient, certifiable transport representation, where Wasserstein error is controlled by cell diameter rather than raw field error alone. Related cell structures also arise in semi-discrete flow matching and optimal transport (Pierret et al., 2026; Kong et al., 2026). Motivated by this view, we introduce Cell-Margin Quantization-Aware Training for Continuous Normalizing Flows (CM-QAT-CNF), or **CM-QAT**, which partitions a frozen feature space into semantically meaningful cells and trains the quantized flow so that its integrated endpoints remain in the corresponding target cells.

Contributions. (1) We derive conditional feature-space terminal-distribution bounds that combine multi-scale cell-histogram control with a boundary-crossing analysis based on endpoint displacement, boundary-tube mass, and one-sided Lipschitz stability. (2) We give counterexamples showing that small velocity-L_2 error is neither necessary for terminal cell correctness nor sufficient for a distribution-free cell-risk guarantee without boundary-margin control. (3) We develop CM-QAT-CNF, a quantization-aware training algorithm for one-step generation at 1-NFE with hierarchical feature-space cells, prototype-margin supervision, and a sampled log-norm regularizer motivated by the conditional stability bound. (4) We evaluate the method on Fashion-MNIST, CIFAR-10, and ImageNet-256 against prior methods.

2 Preliminary

2.1 Definitions & Notations

State space & feature map. Let $\mathcal{X} \subset \mathbb{R}^d$ be the ODE state space, and let $\phi : \mathcal{X} \to \mathcal{Z} \subset \mathbb{R}^m$ be a fixed Borel-measurable feature map. Let $\mathcal{P}(\mathcal{X})$ denote the set of all probability measures on $\mathcal{X}$, and let $\mu \in \mathcal{P}(\mathcal{X})$ be a distribution over the state space. The feature map ϕ induces the pushforward distribution $\mu^\phi = \phi_{\#}\mu$ on $\mathcal{Z}$. In other words, if $X \sim \mu$, then $\phi(X) \sim \mu^\phi$. Equivalently, for every measurable set $A \subseteq \mathcal{Z}$, $\mu^\phi(A) = \mu\left(\phi^{-1}(A)\right)$. Thus, μ^ϕ is the distribution obtained by mapping samples from the state space $\mathcal{X}$ into the feature space $\mathcal{Z}$ through ϕ. For image or latent generative models, ϕ is taken to be a frozen encoder or latent feature map.

Feature-space partition & measurements. Assume $\mathcal{Z}$ is compact with diameter $D_{\mathcal{Z}} = \sup_{z,z' \in \mathcal{Z}} \|z - z'\|_2$, and let $\mathcal{C} = \{C_1, \ldots, C_M\}$ be a finite partition of $\mathcal{Z}$ into nonempty Borel cells (e.g., Voronoi or polyhedral). The cell-assignment map is $c(z) = j$ iff $z \in C_j$, with fixed boundary tie-breaking, and the decision-rule margin is $m_{\mathcal{C}}(z) = \inf_{z' \in \mathcal{Z} \setminus C_{c(z)}} \|z - z'\|_2$, i.e., the distance to the nearest competing cell rather than the ambient topological boundary, which may be degenerate on lower-dimensional supports. Each cell has a representative $z_j \in C_j \cap \mathcal{Z}$ and $\mathrm{diam}(C_j) \le \Delta$. On constrained spaces such as the unit sphere, z_j may be a medoid or a centroid projected onto $\mathcal{Z}$, provided it remains in C_j. At the finest level, Δ_K is a common cell-diameter bound and thus also bounds $\sup_{z \in C_j} \|z - z_j\|_2$. For a distribution μ, the induced histogram is $p^\mu \in \Delta_M$ with $p_j^\mu = \mu^\phi(C_j)$, and the collapsed measure is $\bar{\mu}_{\mathcal{C}} = \sum_{j=1}^M p_j^\mu \delta_{z_j}$, where δ_{z_j} is the Dirac measure at z_j. For $p, q \in \Delta_M$, $\mathrm{TV}(p,q) = \frac{1}{2}\|p - q\|_1$.

Quantized CNF. Let $\rho_0 \in \mathcal{P}(\mathcal{X})$ be the source distribution. A velocity field $v : [0,1] \times \mathcal{X} \to \mathbb{R}^d$ induces the flow Φ_v^t through $\dot{X}_t = v(t, X_t)$, $X_0 \sim \rho_0$, with terminal map $T_v = \Phi_v^1$. The full-precision teacher $v_\star$ induces $T_\star = \Phi_{v_\star}^1$ and $\nu_\star = (T_\star)_{\#}\rho_0$, so $Y_\star = T_\star(X_0) \sim \nu_\star$; ν denotes the true data distribution, which may differ from $\nu_\star$. For a b-bit quantizer Q_b applied to student parameters θ, let $v_b = v_{Q_b(\theta)}$, with exact terminal map $T_b = \Phi_{v_b}^1$ and distribution $\rho_b = (T_b)_{\#}\rho_0$, so $Y_b = T_b(X_0) \sim \rho_b$. At inference, a numerical solver of step size h and order p implements $\widetilde{T}_b$, producing $\widetilde{\rho}_b = (\widetilde{T}_b)_{\#}\rho_0$. Thus, $\nu_\star$, ρ_b, and $\widetilde{\rho}_b$ denote the teacher, exact quantized-student, and deployed-solver terminal distributions, respectively.

2.2 Problem Formulation

We evaluate the deployed quantized sampler through the feature-space Wasserstein distance $W_{1,\mathcal{Z}}(\phi_{\#}\widetilde{\rho}_b, \phi_{\#}\nu)$ which measures transport cost in $\mathcal{Z}$ and reduces to the state-space W_1 only when $\phi = \mathrm{id}$. Given a coupling $\pi \in \Pi(\rho_0, \nu)$, we define the terminal cell risk as $R_{\mathcal{C}}(v_b; \pi) = \Pr_{(X_0,X_1)\sim\pi}[c(\phi(T_b(X_0))) \neq c(\phi(X_1))]$. When source-data couplings are unavailable, we use teacher pairs $(X_0, T_\star(X_0))$ to provide the target cells. Our goal is to learn a b-bit velocity field that minimizes $W_{1,\mathcal{Z}}(\phi_{\#}\widetilde{\rho}_b, \phi_{\#}\nu)$ under a fixed inference budget, measured by the target number of function evaluations (NFE), with one-step generation at 1-NFE as the primary setting.

3 Theoretical Analysis: The Cell Margin

3.1 Velocity error and terminal cell correctness

First, we show that velocity-L_2 error is neither necessary for terminal cell correctness nor, without boundary-margin control, sufficient for a distribution-free cell-risk guarantee, motivating terminal cell supervision; see Fig. 2a and section B.1. In part (B), ρ_0 varies with ε and concentrates near the boundary: the collapsed W_1 remains bounded below while the uncollapsed terminal W_1 vanishes. Since both constructions use $v_\star = 0$, the same velocity errors hold along teacher trajectories.

Proposition 1. *Let the training invariant be $\mathcal{E}_2(v, v_\star) = \int_0^1 \mathbb{E}_{X\sim\rho_0}\left[\|v(t,X) - v_\star(t,X)\|_2^2\right] dt$. The constructions below have finite first moments, take $\phi = \mathrm{id}$ and $\nu = \nu_\star$, and use the shorthand $R_{\mathcal{C}}(T, T_\star) = \Pr[c(T(X_0)) \neq c(T_\star(X_0))]$ for $X_0 \sim \rho_0$. Part (A) is stated on $\mathbb{R}^d$ independently of the compact feature-space bounds below. Part (B) uses a compact interval. There are two constructions:*

(A) Large velocity error, zero terminal cell error. *For every $A > 0$ there exist $v_A, v_\star$ with $\mathcal{E}_2(v_A, v_\star) = A^2/2$, yet $T_A = T_\star$, $R_{\mathcal{C}}(T_A, T_\star) = 0$, and $W_{1,\mathcal{Z}}(\phi_{\#}(T_A)_{\#}\rho_0, \phi_{\#}(T_\star)_{\#}\rho_0) = 0$.*

(B) Arbitrarily small velocity error, constant cell error. *For every $\varepsilon > 0$ and $\alpha \in (0,1)$ there exist $\rho_0, \nu, v_\star, v_\varepsilon$ and a two-cell partition with $\mathcal{E}_2(v_\varepsilon, v_\star) \leq \varepsilon$ but $R_{\mathcal{C}}(T_\varepsilon, T_\star) \geq \alpha$. With representatives $z_- = -1, z_+ = 1$, the collapsed terminal cost obeys $W_1(\bar{\rho}_\varepsilon, \bar{\nu}_\star) \geq 2\alpha$.*

3.2 Multi-scale feature-space cell bound

Our goal is to connect terminal cell preservation to a distributional guarantee without requiring exact teacher–student endpoint agreement. Theorem 2 shows that feature-space Wasserstein error is controlled by two quantities: a within-cell resolution term $2\Delta_K$ and a cross-cell mass discrepancy given by the finest-level histogram TV. For nested partitions, this TV is further bounded by discrepancies along the partition tree, and paired teacher–student cell errors provide trainable upper bounds through differentiable surrogates. Thus, preserving multi-scale cell assignments controls terminal $W_{1,\mathcal{Z}}$ up to the finest cell resolution, providing the theoretical motivation for CM-QAT's multi-scale endpoint cell loss. These are population inequalities; the proof is given in Section B.2.

Theorem 2. *Let $\mathcal{C}^{(0)} \succ \mathcal{C}^{(1)} \succ \cdots \succ \mathcal{C}^{(K)}$ be nested finite Borel partitions of the compact feature space $\mathcal{Z}$, with level K finest of cell diameter $\leq \Delta_K$. Let $p^{(k)}, q^{(k)}$ be the level-k histograms of ν^ϕ and ρ^ϕ. For a parent cell $a \in \mathcal{C}^{(k-1)}$, let $p^{(k)}(\cdot \mid a), q^{(k)}(\cdot \mid a)$ be the conditionals over the children of a (with the convention that the weighted conditional TV term is 0 when $\min(p_a^{(k-1)}, q_a^{(k-1)}) = 0$). Then for any $\rho, \nu \in \mathcal{P}(\mathcal{X})$,*

$$W_{1,\mathcal{Z}}(\rho^\phi, \nu^\phi) \;\le\; 2\Delta_K + D_{\mathcal{Z}}\ \mathrm{TV}(p^{(K)}, q^{(K)}) \tag{1}$$

and we have a recursive TV refinement inequality over the partition tree:

$$\begin{aligned}\mathrm{TV}(p^{(K)}, q^{(K)}) \le \mathrm{TV}(p^{(0)}, q^{(0)}) \\ + \sum_{k=1}^{K} \sum_{a \in \mathcal{C}^{(k-1)}} \min(p_a^{(k-1)}, q_a^{(k-1)})\ \mathrm{TV}\big(p^{(k)}(\cdot \mid a), q^{(k)}(\cdot \mid a)\big).\end{aligned} \tag{2}$$

This is not an exact chain rule: the per-level terms upper-bound the finest TV but need not equal it (see the proof). Moreover, if $(\widehat{Y}, Y)$ is any coupling with $\widehat{Y} \sim \rho$, $Y \sim \nu$, then for every level k, $\mathrm{TV}(p^{(k)}, q^{(k)}) \le \Pr[c_k(\phi(\widehat{Y})) \ne c_k(\phi(Y))]$ (the cell map c_k acts on feature space $\mathcal{Z}$). Hence any measurable nonnegative pointwise surrogate ℓ_k with $\mathbf{1}\{c_k(\phi(\widehat{y})) \ne c_k(\phi(y))\} \le \ell_k(\widehat{y}, c_k(\phi(y)))$ gives the trainable bound

$$W_{1,\mathcal{Z}}(\rho^\phi, \nu^\phi) \;\le\; 2\Delta_K + D_{\mathcal{Z}} \sum_{k=0}^{K} \mathbb{E}\big[\ell_k(\widehat{Y}, c_k(\phi(Y)))\big]. \tag{3}$$

3.3 Boundary-crossing bound, split into a sampler-agnostic part and a smooth-ODE estimate

We separate boundary-crossing analysis into two steps. Proposition 3 relates cell error to endpoint displacement and the teacher boundary margin, defined by $m_{\mathcal{C}}(z) = \inf_{z' \in \mathcal{Z} \setminus C_{c(z)}} \|z - z'\|_2$ and $\beta^{\phi}_{\mathcal{C},\nu_\star}(r) = \Pr[m_{\mathcal{C}}(\phi(Y)) \le r]$ for $Y \sim \nu_\star$; see Section C.2 for geometric details. This result also applies to hard integer-quantized inference, with robustness determined by boundary margins and B_ϕ or its held-out upper bound. Under additional smoothness assumptions, Lemma 4 further bounds the displacement through velocity mismatch, flow expansivity e^{M_+}, and solver error. Thus, the two results connect quantization error to cell risk through terminal displacement and boundary margins, while the smooth ODE bound need not apply to discontinuous activation quantizers; in particular, a sampled log-norm penalty does not establish its required uniform one-sided Lipschitz bound.

Proposition 3 (Cell error from endpoint displacement)**.** *Assume ϕ is Borel measurable with values in compact $\mathcal{Z}$, and $\mathcal{C}$ is a finite Borel partition with cell diameters at most Δ and the relative margin above. Let $\widehat{Y} = \widetilde{T}(X_0)$ be the endpoint of any measurable implemented sampler $\widetilde{T}$ (including a true integer-quantized solver), $Y = T_\star(X_0)$, where $X_0 \sim \rho_0$, $T_\star$ is measurable, and $\widetilde{\rho} = \widetilde{T}_{\#}\rho_0$, and set the realized feature displacement*

$$B_\phi = \mathbb{E}\big[\|\phi(\widehat{Y}) - \phi(Y)\|_2\big]. \tag{4}$$

Then for every $r > 0$, $\Pr[c(\phi(\widehat{Y})) \ne c(\phi(Y))] \le \min\{1,\ \beta^{\phi}_{\mathcal{C},\nu_\star}(r) + B_\phi/r\}$, and the direct displacement bound gives $W_{1,\mathcal{Z}}(\phi_{\#}\widetilde{\rho}, \phi_{\#}\nu_\star) \le B_\phi$. Combining via the teacher gap $\varepsilon^{\phi}_{\mathrm{teacher}} \ge W_{1,\mathcal{Z}}(\phi_{\#}\nu_\star, \phi_{\#}\nu)$,

$$W_{1,\mathcal{Z}}(\phi_{\#}\widetilde{\rho}, \phi_{\#}\nu) \;\le\; \varepsilon^{\phi}_{\mathrm{teacher}} + \min\Big\{B_\phi,\ 2\Delta + D_{\mathcal{Z}}\ \min\big(1,\ \beta^{\phi}_{\mathcal{C},\nu_\star}(r) + \tfrac{B_\phi}{r}\big)\Big\}. \tag{5}$$

If $\beta^{\phi}_{\mathcal{C},\nu_\star}(r) \le A r^{\kappa}$ for all $r \in (0, r_0]$, where $A, \kappa, r_0 > 0$, $B_\phi > 0$, and the minimizer $r_\star = (B_\phi/\kappa A)^{1/(\kappa+1)} \le r_0$ (otherwise take the infimum over $r \in (0, r_0]$), the uncapped wrong-cell upper bound optimizes to $C_\kappa A^{1/(\kappa+1)} B_\phi^{\kappa/(\kappa+1)}$ with $C_\kappa = (\kappa+1)\kappa^{-\kappa/(\kappa+1)}$, so

$$W_{1,\mathcal{Z}}(\phi_{\#}\widetilde{\rho}, \phi_{\#}\nu) \;\le\; \varepsilon^{\phi}_{\mathrm{teacher}} + \min\Big\{B_\phi,\ 2\Delta + D_{\mathcal{Z}}\, C_\kappa\, A^{1/(\kappa+1)}\, B_\phi^{\kappa/(\kappa+1)}\Big\}. \tag{6}$$

If $B_\phi = 0$, the paired feature endpoints agree almost surely and the cell error is zero. For fixed positive A, κ and $D_{\mathcal{Z}} > 0$, the direct bound is asymptotically stronger as $B_\phi \downarrow 0$. The exponent $\kappa/(\kappa+1)$ is sharp up to constants for wrong-cell probability and histogram TV, as shown in Section B.3. It is not a sharp exponent for feature-space W_1, which the direct displacement bound controls linearly.

Lemma 4 (Smooth-ODE perturbation estimate for B_ϕ)**.** *In the setting of Proposition 3, assume that ϕ is L_ϕ-Lipschitz on the teacher and deployed-student endpoint set. Assume that the teacher and student ODEs have unique measurable flows on $[0,1]$ for ρ_0-almost every initial condition. The student field v_b is assumed continuous in (t,x) and locally Lipschitz in x, with a time-dependent one-sided Lipschitz constant $m_b \in L^1([0,1])$: $\langle v_b(t,x) - v_b(t,y), x-y\rangle \le m_b(t)\|x-y\|_2^2$ for almost every t and all x, y in a common domain containing both trajectories. Set $M_+ = \int_0^1 \max(m_b(t),0)\,dt$. Let*

$$a_b = \int_0^1 \mathbb{E}\big[\|v_b(t,X_t^\star) - v_\star(t,X_t^\star)\|_2\big]\,dt, \qquad X_t^\star = \Phi_{v_\star}^t(X_0), \tag{7}$$

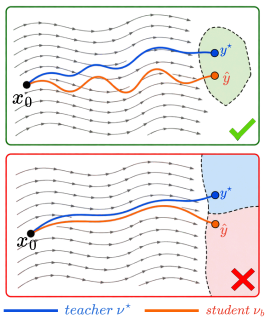
with measurable integrand and $a_b < \infty$, and let the solver satisfy $\mathbb{E}\big[\|\widetilde{T}_b(X_0) - T_b(X_0)\|_2\big] \le C_{\mathrm{sol}} h^p$ for $0 < h \le h_0$, with $p > 0$ and $C_{\mathrm{sol}} < \infty$ independent of h. Then the displacement of Proposition 3, with $\widehat{Y} = \widetilde{T}_b(X_0)$, obeys

$$B_\phi = \mathbb{E}\big[\|\phi(\widehat{Y}) - \phi(Y)\|_2\big] \;\le\; L_\phi\big(e^{M_+} a_b + C_{\mathrm{sol}} h^p\big). \tag{8}$$

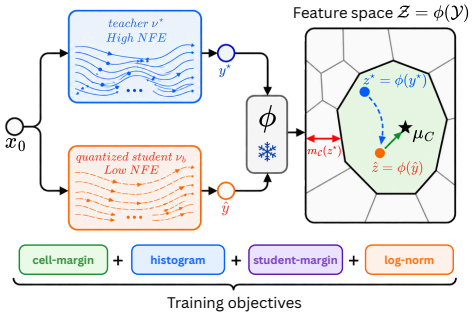
For normalized features $\phi(x) = f(x)/\|f(x)\|_2$, a sufficient condition for the Lipschitz assumption in Lemma 4 is that f is L_f-Lipschitz and $\|f(x)\|_2 \ge \mu > 0$ on the relevant endpoint set. The triangle inequality then gives $\|\phi(x) - \phi(y)\|_2 \le (2/\mu)\|f(x) - f(y)\|_2$, so $L_\phi \le 2L_f/\mu$. The sampler bound in Proposition 3 requires only measurable features and does not need this Lipschitz assumption.

4 CM-QAT: Cell Margin Quantization-Aware Training for Continuous Normalizing Flows

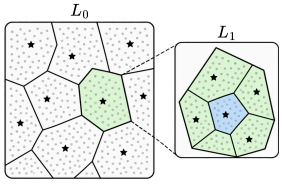
In this section, we propose the CM-QAT-CNF algorithm (Fig. 2b), or CM-QAT for short. The primary deployment setting is one-step generation with 1-NFE. The target-NFE notation in Algorithm 1 also accommodates the budget variations in Fig. 3c. From theoretical analysis, we aim to train the student exactly as it will be sampled: $\widehat{y} = \widetilde{T}_{b,h}(x_0) = \mathrm{ODESolve}_h(v_b, x_0)$, backpropagating through the target solver rather than training only on random intermediate t. The training objectives are motivated by the conditional bounds in Section 3.



(a) Field error vs. terminal cell error.



(b) CM-QAT-CNF trains the quantized student for one-step generation at 1-NFE using terminal feature-cell objectives.



(c) Hierarchical construction of genuinely nested feature-space cells.

Figure 2: **CM-QAT Approach.** (a) Velocity-field error alone does not determine terminal quality: a large trajectory deviation can preserve the terminal cell, while a small deviation near a cell boundary can change the terminal assignment. (b) CM-QAT-CNF integrates the quantized student at the one-step deployment budget of 1-NFE and aligns its terminal endpoint with the teacher in a frozen feature space using 4 objectives. (c) The feature partition is constructed hierarchically, yielding genuinely nested cells for multi-scale supervision.

4.1 PROTOTYPE CELL-MARGIN ENDPOINT HEAD

CM-QAT first constructs a nested feature-space partition $\mathcal{C}$ from teacher endpoints (Fig. 2c). This corresponds to the partition-construction stage of Algorithm 1: build a frozen multi-scale prototype bank $\mathcal{P}^{(k)} = \{z_1^{(k)}, \dots, z_{M_k}^{(k)}\}$ in feature space. The multi-scale TV refinement inequality of Theorem 2 requires genuinely nested partitions $\mathcal{C}^{(0)} \succ \cdots \succ \mathcal{C}^{(K)}$. Independent k-means at each M_k does not produce nested cells (a fine cell can straddle a coarse boundary). We therefore build the tree hierarchically: $\mathcal{C}^{(0)}$ is a coarse k-means, and each finer level is obtained by running k-means within each parent cell, so $C_{a,b}^{(k)} = C_a^{(k-1)} \cap \mathrm{Vor}_{a,b}$ with prototypes seeded inside $C_a^{(k-1)}$. The level-k cell map is then tree-structured: descend the parent path, then pick the nearest child prototype. Because the cells are a tree rather than a flat Voronoi diagram, we use the path-wise surrogate in Section C.1: at each level s, the loss dominates the event that the student picks the wrong child among the siblings under the teacher's level-$(s-1)$ parent $a_{s-1}^\star(x_0)$, and any level-k tree-cell error implies a wrong child somewhere on the path, so the sum of path losses dominates the level-k hard error. Teacher labels are the path nodes $a_s^\star(x_0)$ from the tree map applied to $\phi(T_\star(x_0))$. Restricting the logit competition to $\mathrm{ch}(a_{s-1}^\star)$ – the children of the teacher's parent at level s, define:

$$
\begin{aligned}
&s_j^{(s)}(\widehat{y}) = -\|\phi(\widehat{y}) - z_j^{(s)}\|_2^2/\tau_p, && \ell_s^{\mathrm{tree}}(\widehat{y}, y) = \frac{\mathrm{softplus}\big((\gamma - s_{a_s^\star}^{(s)}(\widehat{y}) + g_s(\widehat{y}))/\tau\big)}{\mathrm{softplus}(\gamma/\tau)},\\
&g_s(\widehat{y}) = \alpha \log \sum_{j \in \mathrm{ch}(a_{s-1}^\star),\ j \neq a_s^\star} \exp\big(s_j^{(s)}(\widehat{y})/\alpha\big), && \mathcal{L}_{\mathrm{cell}}(\widehat{y}, y) = \sum_{s=0}^{K} \lambda_s \ell_s^{\mathrm{tree}}(\widehat{y}, y).
\end{aligned}
\tag{9}
$$

Here $y = T_\star(x_0)$ determines the teacher path, $a_{-1}^\star$ is the root, $K \geq 0$ is an integer, $\tau_p, \alpha, \tau, \gamma > 0$, and $\lambda_s \geq 0$. A singleton child set contributes zero loss, without evaluating its empty log-sum. The displayed loss is per teacher–student pair.

The unweighted path-loss sum dominates the hard error pointwise, so its expectation bounds the cell-error probability. For fixed strictly positive weights, set $\lambda_{\min} = \min_{0 \leq s \leq K} \lambda_s > 0$. Then Corollary 5 gives

$$
\Pr\Big[c_K^{\mathrm{tree}}(\phi(\widehat{Y})) \neq c_K^{\mathrm{tree}}(\phi(Y))\Big] \leq \mathbb{E}\left[\sum_{s=0}^{K} \ell_s^{\mathrm{tree}}(\widehat{Y}, Y)\right] \leq \frac{\mathbb{E}[\mathcal{L}_{\mathrm{cell}}(\widehat{Y}, Y)]}{\lambda_{\min}}. \tag{10}
$$

Weights at least one are sufficient for direct domination by the weighted loss. Strict positivity suffices for the rescaled bound in equation 10. If any weight is zero, this conversion does not apply and the omitted path losses must be controlled separately. These are population bounds for a fixed sampler. A small training average alone does not establish small population cell risk.

The surrogate replaces nonsmooth set-distance computations and the explicit $\min_{k \neq c}$ operation with a smooth prototype-based objective whenever ϕ is smooth. If independent multi-scale k-means is used instead, the resulting partitions are generally not nested. In that case, the global-Voronoi surrogate at each level yields only the finest-level guarantee $W_{1,\mathcal{Z}} \leq 2\Delta_K + D_{\mathcal{Z}}\,\mathrm{TV}\big(p^{(K)}, q^{(K)}\big)$, which does not require nested partitions.

4.2 LOG-NORM & STUDENT-MARGIN REGULARIZER.

Log-norm / expansivity-regularized velocity blocks. Motivated by Lemma 4, CM-QAT penalizes local log norms along sampled trajectories in the differentiable regime. Let $J_{v_b}(t, x) = D_x v_b(t, x)$ be the state Jacobian, and let $\eta_{\log}$ denote the sampling law of the time–state pairs used for this term. The formal objective is

$$
\mathcal{L}_{\mathrm{lognorm}} = \mathbb{E}_{(t,x) \sim \eta_{\log}}\left[\lambda_{\max}\left(\frac{J_{v_b}(t, x) + J_{v_b}(t, x)^\top}{2}\right)_+^2\right], \tag{11}
$$

where $a_+ = \max\{a, 0\}$. In practice, we use a sampled approximation of this objective only as an optimization surrogate, not as a certified estimate of the largest eigenvalue. Reducing this sampled objective does not establish the uniform one-sided Lipschitz bound $m_b(t)$ or its integral M_+ on the common trajectory domain. For hard activation quantization, the applicable population result is Proposition 3, which uses the realized endpoint map and does not require a differentiable velocity field.

Student-margin regularizer. During student training, we additionally retain the boundary penalty

$$\mathcal{L}_{\text{student-margin}} = \mathbb{E}\left[\exp\left(-\frac{m_{\mathcal{C}}(\phi(\widehat{y}))}{\gamma}\right)\right]$$

only as a student-margin regularizer: it encourages student endpoints to remain farther from cell boundaries, but does not modify the teacher-defined boundary-tube term in Proposition 3.

Algorithm 1 Cell Margin Quantization-Aware Training for Continuous Normalizing Flows

Require: teacher $v_\star$, source ρ_0, feature map ϕ, target NFE, bit budget
1: Generate teacher endpoints $y_i^\star = T_\star(x_{0,i})$.
2: Build genuinely-nested prototype partitions: hierarchical k-means in $\phi(y_\star)$ (coarse level, then k-means within each parent cell so $C_{a,b}^{(k)} = C_a^{(k-1)} \cap \mathrm{Vor}_{a,b}$).
3: Initialize quantized student v_b from $v_\star$.
4: **if** target precision is extreme low-bit **then**
5: Warm-start from a higher-bit QAT checkpoint ▷ e.g., W4A8 → W4A4/W2A8
6: **end if**
7: **for** each QAT step **do**
8: Sample x_0; compute teacher endpoint $y^\star = T_\star(x_0)$.
9: Compute teacher cell labels $c_k(y^\star)$.
10: Integrate quantized student at target NFE: $\widehat{y} = \mathrm{ODESolve}_h(v_b, x_0)$.
11: Compute $\mathcal{L}_{\text{Total}}$.
12: Update with STE fake quantization.
13: **end for**
14: Export integer model and evaluate at the same NFE.

4.3 Training Procedure

The CM-QAT training procedure is summarized in Algorithm 1. Its final objective combines the cell-based CM-QAT losses with lightweight velocity stabilization and endpoint alignment:

$$\begin{aligned} \mathcal{L}_{\text{CM}} &= \mathcal{L}_{\text{cell}} + \lambda_{\text{hist}}\,\mathcal{L}_{\text{hist}} + \lambda_{\text{margin}}\,\mathcal{L}_{\text{student-margin}} + \lambda_{\text{log}}\,\mathcal{L}_{\text{lognorm}}, \\ \mathcal{L}_{\text{Total}} &= \mathcal{L}_{\text{CM}} + \lambda_{\text{vel}}\,\mathcal{L}_{\text{vel}} + \lambda_{\text{endpoint}}\,\mathcal{L}_{\text{endpoint}}, \qquad \lambda_{\text{vel}} \ll 1. \end{aligned} \tag{12}$$

Here, $\mathcal{L}_{\text{hist}} = \mathrm{Sinkhorn}_\varepsilon\left(\sum_j \widehat{q}_j \delta_{z_j}, \sum_j \widehat{p}_j \delta_{z_j}\right)$, $\mathcal{L}_{\text{vel}} = \int_0^1 \mathbb{E}\left[\|v_b(t, X_t^\star) - v_\star(t, X_t^\star)\|_2^2\right] dt$, and $\mathcal{L}_{\text{endpoint}} = \mathbb{E}\left[\frac{1}{d}\left\|\widehat{Y} - Y^\star\right\|_2^2\right]$, where d is the state-space dimension. The histogram and endpoint losses are optimization surrogates: the TV certificate in Theorem 2 uses hard cell probabilities, while relating endpoint loss to B_ϕ requires a corresponding feature-displacement bound. The velocity term acts only as a weak stabilizer, with cell-based supervision as the primary signal. Teacher-cell labels $c_k^\star(x_0)$ are obtained from the same teacher pass used for $\mathcal{L}_{\text{vel}}$, avoiding additional minibatch-OT coupling and introducing only the additive teacher discrepancy $\varepsilon_{\text{teacher}}^{\phi} = W_{1,\mathcal{Z}}\left(\phi_\# \nu_\star, \phi_\# \nu\right)$.

Complexity. Each training iteration uses one teacher pass, NFE student evaluations, cell assignment costing $O(M)$ or $O(\log M)$ with a hierarchical tree, and an $O(M^2)$ Sinkhorn computation over a moderate number of cells. With checkpointed or adjoint backpropagation, endpoint optimization requires $O(\sqrt{\mathrm{NFE}})$ activation memory. Overall, CM-QAT adds only modest overhead over standard Quantization-Aware Training.

5 Experiment

5.1 Implementation

We evaluate CM-QAT on unconditional generation for two datasets: Fashion-MNIST, and CIFAR-10; and on ImageNet-256 for class-conditional generation. For each dataset, we evaluate four quantization configurations, W8A8, W4A8, W4A4, and W2A8, all under 1-step generation. For Fashion-MNIST, we compute FID using features extracted by the classification encoder ϕ. For CIFAR-10

and ImageNet-256, we use OpenAI's evaluation suite (Dhariwal & Nichol, 2021) to report Fréchet Inception Distance (FID), Inception Score (IS), precision, and recall for comprehensive evaluation. We consider three standard baselines: OT-PTQ (Varam et al., 2025), Endpoint-QAT, and Velocity-QAT. We further adapt and reimplement Q-Sched (Frumkin & Marculescu, 2026) and Q-Drift (Ryu et al., 2026) for quantized flow-matching models. Detailed implementation configurations is in Section D.1, and algorithms for all baselines are provided in section E.

5.2 Unconditional Generation

For the unconditional tasks, we adopt the U-Net architecture from TorchCFM (Tong et al., 2024) with reduced-capacity variants: a Tiny U-Net for Fashion-MNIST, and a Small U-Net for CIFAR-10, with generation performed directly in pixel space. We report Fashion-MNIST and CIFAR-10 here, using the corresponding trained teachers to initialize the quantized students. As shown in Table 1, CM-QAT consistently achieves the best FID across all quantization settings, with particularly large gains over Endpoint-QAT under the more aggressive W4A4 and W2A8 regimes. Notably, Endpoint-QAT consistently outperforms Velocity-QAT in this setting, providing empirical support for Proposition 1 that low velocity-field error alone does not guarantee terminal distribution fidelity. The full-precision TorchCFM teachers evaluated with a 128-NFE Dopri5 solver achieve FID-ϕ of 0.0078 on Fashion-MNIST and Clean-FID of 3.64 on CIFAR-10; the larger remaining gap on CIFAR-10 is mainly due to distilling a many-step teacher into a 1-step student, whereas this gap is substantially reduced for the native 1-step ImageNet teacher considered next.

Table 1: **FID for 1-step generation under different quantization settings.** Fashion-MNIST reports FID-ϕ, while CIFAR-10 reports Clean-FID. The FP OT-CFM teacher uses 128-step sampling.

	Fashion-MNIST (FID-ϕ $\times 10^{-1}\downarrow$)				CIFAR-10 (Clean-FID $\downarrow$)			
Method	W8A8	W4A8	W4A4	W2A8	W8A8	W4A8	W4A4	W2A8
OT-PTQ	4.08	4.68	6.09	5.99	382.83	322.60	471.26	413.36
Q-Sched	2.90	4.06	5.57	5.18	279.43	323.69	462.85	465.36
Q-Drift	4.67	5.57	5.71	4.18	389.91	361.46	460.28	385.11
Velocity-QAT	4.60	4.96	4.56	4.93	384.02	404.61	340.09	358.64
Endpoint-QAT	0.16	0.32	1.43	1.18	13.76	32.40	51.52	169.94
CM-QAT	**0.11**	**0.14**	**0.21**	**0.58**	**10.17**	**20.48**	**32.67**	**108.38**
OT-CFM (FP, 128-NFE)	0.078				3.637			

5.3 Class-conditional Generation

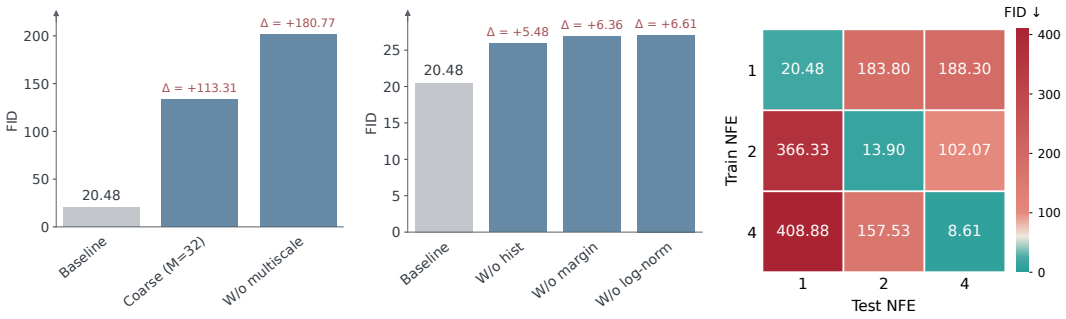
Table 2: Performance evaluation of class-conditioned image generation on ImageNet-256×256.

Method	**Bits (W/A)**	**NFE**	**Size (MB)**	**BOPs (T)**	**IS** $\uparrow$	**FID** $\downarrow$	**Precision** $\uparrow$	**Recall** $\uparrow$
MeanFlow	FP32	1	1837.70	79.33	246.58	3.86	81.67%	48.65%
SA-PTQ	8/8	20	436.79	8.76	120.71	10.16	65.39%	56.97%
MeanFlow-QAT	8/8	1	451.23	4.96	210.72	4.99	75.37%	52.84%
Endpoint-QAT	8/8	1	451.23	4.96	237.61	4.05	80.59%	48.95%
CM-QAT (ours)	8/8	1	451.23	4.96	**246.06**	**3.96**	**81.60%**	49.18%
EfficientDM	4/8	20	219.12	4.38	132.70	9.91	71.05%	53.62%
SA-QLoRA	4/8	20	219.12	4.38	140.56	8.55	73.20%	54.49%
MeanFlow-QAT	4/8	1	234.22	2.48	196.24	6.05	73.68%	52.29%
Endpoint-QAT	4/8	1	234.22	2.48	228.50	4.82	79.18%	48.86%
CM-QAT (ours)	4/8	1	234.22	2.48	**244.26**	**4.22**	**80.92%**	48.59%
EfficientDM	4/4	20	219.12	2.19	225.20	17.28	60.33%	52.82%
SA-QLoRA	4/4	20	219.12	2.19	242.03	13.73	58.90%	55.38%
MeanFlow-QAT	4/4	1	234.22	1.24	1.96	366.84	0.03%	0.00%
Endpoint-QAT	4/4	1	234.22	1.24	96.22	29.07	51.89%	47.79%
CM-QAT (ours)	4/4	1	234.22	1.24	**242.13**	**5.99**	**78.06%**	43.84%
Endpoint-QAT	3/8	1	189.29	1.86	146.17	11.09	70.87%	43.14%
CM-QAT (ours)	3/8	1	189.29	1.86	**250.77**	**5.04**	**82.11%**	41.89%

We implement MeanFlow-QAT (replacing Velocity-QAT), Endpoint-QAT, and CM-QAT on ImageNet-256, while adopting the remaining baselines from SA-QLoRA (Zeng et al., 2026). Our

implementations use a pretrained SiT-L/2 MeanFlow checkpoint (Geng et al., 2025), comparable in scale to the roughly 400M-parameter LDM-4 backbone used by SA-QLoRA. As shown in Table 2, CM-QAT achieves the best IS, FID, and precision among the quantized methods, with slightly lower recall. At W8A8, it remains competitive with full-precision MeanFlow, while at W4A8 and W4A4 it maintains FID in the 4–6 range versus roughly 9–14 for prior methods. MeanFlow-QAT collapses at W4A4, whereas endpoint-based training remains viable, highlighting the importance of terminal supervision at low precision. CM-QAT remains effective down to W3A8, the lowest bit-width that produced meaningful generation; despite multiple attempts, W2A8 failed to do so and is therefore omitted. The lower recall may stem from the MeanFlow backbone or endpoint-oriented objective, which we leave for future study.

5.4 Ablations



(a) Multi-scale partition ablation. (b) Training-objective ablation. (c) Train–test NFE mismatch.

Figure 3: **Ablation studies on CIFAR-10 with W4A8 quantization.** (a) Effect of the hierarchical multi-scale partition. (b) Contribution of the histogram, margin, and log-norm objectives. (c) Effect of training and inference NFE, showing that matching the target NFE during training is critical for few-step generation.

Figure 3 confirms the importance of each CM-QAT component. Multi-scale supervision is critical: using only coarse or fine cells substantially degrades FID, showing that both global structure and local fidelity matter. Removing the histogram, margin, or log-norm objective also worsens FID, with the log-norm ablation causing the largest drop. Finally, the NFE ablation shows a clear train–test mismatch penalty: models perform best at their training NFE, highlighting the importance of matching the deployment NFE during training.

6 Conclusion, Limitation & Future Work

We introduced CM-QAT, a quantization-aware training framework that reframes low-bit, one-step flow generation as a terminal cell-preservation problem rather than velocity-field reconstruction. Our theory connects hierarchical cell consistency, endpoint displacement, and boundary margins to conditional feature-space distributional bounds. The training objective uses surrogates motivated by these bounds at the one-step deployment budget of 1-NFE. Experiments on Fashion-MNIST, CIFAR-10, and ImageNet-256 show that CM-QAT consistently improves over velocity- and endpoint-based quantization baselines and remains effective under aggressive low-bit, 1-step generation. These results suggest that preserving terminal geometry provides a useful principle for jointly addressing quantization and one-step sampling. A current limitation is that CM-QAT can exhibit reduced recall on ImageNet, indicating a remaining precision-diversity trade-off. Moreover, its effectiveness depends on the chosen frozen feature representation and partition. The diameter term in Theorem 2 can be loose for normalized high-dimensional features, so informative numerical guarantees may require coarser semantic partitions, which we leave for future work. Future work can also study adaptive feature spaces and objectives that better preserve mode coverage under extreme quantization.

AI USE STATEMENT

In this work, we used generative AI tools for assist in implementing the method, assist in the writing of proofs. We have not used generative AI tools for help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments; and generate synthetic data sets, clean and reformat dataset, support qualitative and thematic data analysis, assist with translation are not applicable to this work. Additionally, we used generative AI tools for draft parts of this research paper and edit this research paper to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the main training configurations and implementation details in the appendix to facilitate reproducibility. We will further polish and release the CM-QAT source code and trained checkpoints upon acceptance.

REFERENCES

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y Galliker, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *9th Annual Conference on Robot Learning*, 2025.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Ricky T. Q. Chen, Brandon Amos, and Maximilian Nickel. Semi-discrete normalizing flows through differentiable tessellation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 2024.

Alexander Davydov and Francesco Bullo. Perspectives on contractivity in control, optimization, and learning. *IEEE Control Systems Letters*, 8:2087–2098, 2024.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Natalia Frumkin and Diana Marculescu. Q-sched: Pushing the boundaries of few-step diffusion models with quantization-aware scheduling. In *Catch, Adapt, and Operate: Monitoring ML Models Under Drift Workshop*, 2026.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yefei He, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. EfficientDM: Efficient quantization-aware fine-tuning of low-bit diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Peter Holderrieth, Uriel Singer, Tommi Jaakkola, Ricky T. Q. Chen, Yaron Lipman, and Brian Karrer. GLASS flows: Efficient inference for reward alignment of flow and diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yushi Huang, Ruihao Gong, Xianglong Liu, Jing Liu, Yuhang Li, Jiwen Lu, and Dacheng Tao. Temporal feature matters: A framework for diffusion model quantization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10):8823–8837, 2025.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2704–2713, 2018.

Lingkai Kong, Molei Tao, Yang Liu, Bryan Wang, Jinmiao Fu, Chien-Chih Wang, and Huidong Liu. Alignflow: Improving flow-based generative models with semi-discrete optimal transport. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The eleventh international conference on learning representations*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International conference on learning representations (ICLR)*, 2023.

Hung Mai, Hai Nguyen, Luong Doan, Ngoc Vu, Khanh Nguyen, Nhung Duong, and Tuan Do. Field codes for distributed coupling samplers and certified empirical transport, 2026.

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.

Bao Nguyen, Binh Nguyen, and Viet Anh Nguyen. Bellman optimal stepsize straightening of flow-matching models. In *The Twelfth International Conference on Learning Representations*, 2024.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *Advances in Neural Information Processing Systems*, 38:50608–50646, 2026.

Derek Onken, Samy Wu Fung, Xingjian Li, and Lars Ruthotto. Ot-flow: Fast and accurate continuous normalizing flows via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9223–9232, 2021.

Mingue Park, Jisung Hwang, Seungwoo Yoo, Kyeongmin Yeo, and Minhyuk Sung. Pairflow: Closed-form source-target coupling for few-step generation in discrete flow models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Emile Pierret, Johannes Hertrich, Samuel Hurault, and Julie Delon. Tessellations of semi-discrete flow matching, 2026.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International conference on machine learning*, pp. 8599–8608. PMLR, 2021.

Danyal Rehman, Tara Akhound-Sadegh, Artem Gazizov, Yoshua Bengio, and Alexander Tong. FALCON: Few-step accurate likelihoods for continuous flows. In *The Fourteenth International Conference on Learning Representations*, 2026.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1530–1538. JMLR.org, 2015.

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*, pp. 10684–10695, 2022.

H Ryu, N Park, and H Shim. Dgq: Distribution-aware group quantization for text-to-image diffusion models. arxiv 2025. *arXiv preprint arXiv:2501.04304*.

Sooyoung Ryu, Mathieu Salzmann, and Saqib Javed. Q-drift: Quantization-aware drift correction for diffusion model sampling, 2026.

Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. Post-training quantization on diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1972–1981.

Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yichong Leng, Lei He, Tao Qin, Jiang Bian, et al. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. In *International conference on learning representations*, volume 2024, pp. 698–722, 2024.

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. Expert Certification.

Dara Varam, Diaa A. Abuhani, Imran Zualkernan, Raghad AlDamani, and Lujain Khalil. Low-bit, high-fidelity: Optimal transport quantization for flow matching, 2025.

Junyi Wu, Haoxuan Wang, Yuzhang Shang, Mubarak Shah, and Yan Yan. PTQ4dit: Post-training quantization for diffusion transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Binxing Xu, Hao Gu, Lujun Li, Hao Wang, Bei Liu, Jiacheng Liu, Qiyuan Zhu, Xintong Yang, Chao Li, Sirui Han, and Yike Guo. Bit-by-bit: Progressive QAT strategy with outlier channel splitting for stable low-bit LLMs. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22284–22299, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6.

Qian Zeng, Jie Song, Yuanyu Wan, Huiqiong Wang, and Mingli Song. Sampling-aware quantization for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35831–35840, 2026.

Jiaji Zhang, Ruichao Sun, Hailiang Zhao, Jiaju Wu, Peng Chen, Hao Li, Yuying Liu, Kingsum Chow, Gang Xiong, and Shuiguang Deng. Segquant: A semantics-aware and generalizable quantization framework for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43717–43726, June 2026.

A Related work

Flow Matching (Lipman et al., 2022) learns a time-dependent velocity field whose ODE transports a simple prior to the data distribution, while OT-CFM (Tong et al., 2024) and rectified flow (Liu et al., 2023) further improve FM by using transport-informed coupling or iterative re-pairing to straighten trajectories and improve coarse-step sampling. Related geometric perspectives use cell structures in flow-based transport: semi-discrete FM induces terminal assignment cells in the source space (Pierret et al., 2026), AlignFlow uses SDOT Laguerre cells for deterministic noise-data alignment (Kong et al., 2026), and semi-discrete normalizing flows employ differentiable Voronoi tessellations to connect discrete and continuous variables (Chen et al., 2022). Recent semi-discrete FM further leverages semi-discrete OT to construct scalable noise–data couplings beyond minibatch OT (Mousavi-Hosseini et al., 2026). Such straightening facilitates numerical integration but does not by itself imply contractivity; flow expansivity is instead characterized through the Jacobian logarithmic norm in neural-ODE stability theory (Davydov & Bullo, 2024; Chen et al., 2018). Because standard FM still requires multi-step ODE integration, reducing the NFE becomes a natural next

objective. Geng et al. (2025) introduces MeanFlow, which learns interval-averaged velocities to enable one-step generation without pretraining or teacher distillation. Other few-step methods optimize solver schedules, straighten pretrained flows, or learn integrated flow maps over large time intervals (Nguyen et al., 2024; Rehman et al., 2026; Park et al., 2026). For compression, diffusion quantization (Wu et al., 2024; Shang et al.) is predominantly post-training or distribution-aware, using timestep-aware calibration, activation reconstruction, attention preservation, or temporal feature alignment to accommodate the strongly time-dependent distributions encountered during denoising (Ryu et al.; Huang et al., 2025). QAT-based methods such as EfficientDM (He et al., 2024) further fine-tune quantized models by matching full-precision per-step predictions or intermediate outputs. SA-QLoRA (Zeng et al., 2026) jointly considers fast sampling and low-bit compression by aligning mixed-order trajectories to reduce quantization-induced directional errors that accumulate across sparse denoising steps. Meanwhile, semantics-aware approaches such as SegQuant (Zhang et al., 2026) preserve segment-level relationships through graph-structured quantization objectives. In FM, OT-based quantization aligns full-precision and quantized weight distributions before ODE integration (Varam et al., 2025), while Q-Sched (Frumkin & Marculescu, 2026) and Q-Drift (Ryu et al., 2026) address low-bit few-step sampling through scheduler preconditioning and post-hoc drift correction, respectively.

B PROOF

B.1 PROOF OF PROPOSITION 1

Restatement of Proposition 1. Let the training invariant be $\mathcal{E}_2(v, v_\star) = \int_0^1 \mathbb{E}_{X\sim\rho_0}\big[\|v(t,X) - v_\star(t,X)\|_2^2\big]\,dt$. As in Proposition 1, the examples have finite first moments and use $\phi = \mathrm{id}$, $\nu = \nu_\star$, and $R_{\mathcal{C}}(T, T_\star) = \Pr[c(T(X_0)) \neq c(T_\star(X_0))]$. Part (A) is stated on $\mathbb{R}^d$. Part (B) uses a compact interval. There are two constructions:

(A) Large velocity error, zero terminal cell error. For every $A > 0$ there exist $v_A, v_\star$ with $\mathcal{E}_2(v_A, v_\star) = A^2/2$, yet $T_A = T_\star$, $R_{\mathcal{C}}(T_A, T_\star) = 0$, and $W_{1,\mathcal{Z}}(\phi_\#(T_A)_\#\rho_0, \phi_\#(T_\star)_\#\rho_0) = 0$.

(B) Arbitrarily small velocity error, constant cell error. For every $\varepsilon > 0$ and $\alpha \in (0,1)$ there exist $\rho_0, \nu, v_\star, v_\varepsilon$ and a two-cell partition with $\mathcal{E}_2(v_\varepsilon, v_\star) \le \varepsilon$ but $R_{\mathcal{C}}(T_\varepsilon, T_\star) \ge \alpha$. With representatives $z_- = -1, z_+ = 1$, the collapsed terminal cost obeys $W_1(\bar\rho_\varepsilon, \bar\nu_\star) \ge 2\alpha$.

Proof. **Part A.** Take $\mathcal{X} = \mathbb{R}^d$, $\phi = \mathrm{id}$, any source law ρ_0 with finite first moment, and teacher $v_\star(t,x) = 0$, so $T_\star(x) = x$. For $A > 0$ set $v_A(t,x) = A\sin(2\pi t)\,e_1$. The ODE $\dot X_t = A\sin(2\pi t)e_1$ integrates to $X_1 = X_0 + Ae_1\int_0^1 \sin(2\pi t)\,dt = X_0$, so $T_A = T_\star$. Every terminal distribution and cell assignment is identical, giving zero terminal W_1 and zero cell error. Yet $\mathcal{E}_2(v_A, v_\star) = \int_0^1 A^2\sin^2(2\pi t)\,dt = A^2/2$, which is arbitrarily large.

Part B. Take $\mathcal{X} = [-2,2]$, $\phi = \mathrm{id}$, $C_- = [-2,0)$, $C_+ = [0,2]$, teacher $v_\star = 0$, $T_\star(x) = x$. Fix $0 < \eta < 1$ and let $\rho_0 = \nu_\star = \alpha\,\mathrm{Unif}(0,\eta/2) + (1-\alpha)\delta_1$, so all teacher endpoints lie in C_+. To keep the flow a self-map of the compact $\mathcal{X}$ (a constant $-\eta$ field would push $x = -2$ outside $[-2,2]$), use the compactly-supported cutoff field $v_\eta(t,x) = -\eta\,\chi(x)$, where χ is smooth, $0 \le \chi \le 1$, $\chi \equiv 1$ on $[-1,2]$ (which contains the source support $(0,\eta/2)\cup\{1\}$ and all its trajectories, since $\eta < 1$) and $\chi \equiv 0$ on $[-2,-3/2]$. On $[-1,2]$ the field is the constant $-\eta$, so $T_\eta(x) = x - \eta$ for every x in the source support. For $x \in (0,\eta/2)$, $T_\eta(x) \in (-\eta, -\eta/2) \subset C_-$: all α mass crosses the boundary, while $\delta_1 \mapsto 1-\eta > 0$ stays in C_+, and the flow never leaves $[-2,2]$. Hence $R_{\mathcal{C}}(T_\eta, T_\star) = \alpha$. The velocity error is $\mathcal{E}_2(v_\eta, v_\star) = \int_0^1 \mathbb{E}_{X\sim\rho_0}\big[\eta^2\chi(X)^2\big]\,dt = \eta^2$ (since $\chi \equiv 1$ on the source support). Choosing $\eta = \min\{1/2, \sqrt{\varepsilon}\}$ gives $\mathcal{E}_2 \le \varepsilon$. The teacher collapsed histogram places all mass on $z_+ = 1$. The student places α on $z_- = -1$ and $1-\alpha$ on z_+. Moving α mass from -1 to 1 costs 2α, so $W_1(\bar\rho_\eta, \bar\nu_\star) = 2\alpha$. For the uncollapsed laws, the translation coupling gives $W_1(\rho_\eta, \nu_\star) \le \eta$, and their means differ by η, giving the matching lower bound. Thus $W_1(\rho_\eta, \nu_\star) = \eta \to 0$. □

B.2 PROOF OF THEOREM 2

Restatement of Theorem 2. Let $\mathcal{C}^{(0)} \succ \mathcal{C}^{(1)} \succ \cdots \succ \mathcal{C}^{(K)}$ be nested finite Borel partitions of compact $\mathcal{Z}$, with level K finest of cell diameter $\le \Delta_K$. Let $p^{(k)}, q^{(k)}$ be the level-k histograms of ν^ϕ

and ρ^ϕ. For a parent cell $a \in \mathcal{C}^{(k-1)}$, let $p^{(k)}(\cdot \mid a), q^{(k)}(\cdot \mid a)$ be the conditionals over the children of a (with the convention that the weighted conditional TV term is 0 when $\min(p_a^{(k-1)}, q_a^{(k-1)}) = 0$). Then for any $\rho, \nu \in \mathcal{P}(\mathcal{X})$,

$$W_{1,\mathcal{Z}}(\rho^\phi, \nu^\phi) \;\le\; 2\Delta_K + D_{\mathcal{Z}}\ \mathrm{TV}(p^{(K)}, q^{(K)}) \tag{13}$$

and we have a recursive TV refinement inequality over the partition tree:

$$\begin{aligned}\mathrm{TV}(p^{(K)}, q^{(K)}) \le \mathrm{TV}(p^{(0)}, q^{(0)}) \\ + \sum_{k=1}^{K} \sum_{a \in \mathcal{C}^{(k-1)}} \min(p_a^{(k-1)}, q_a^{(k-1)})\ \mathrm{TV}\big(p^{(k)}(\cdot \mid a), q^{(k)}(\cdot \mid a)\big).\end{aligned} \tag{14}$$

This is not an exact chain rule: the per-level terms upper-bound the finest TV but need not equal it (see the proof). Moreover, if $(\widehat{Y}, Y)$ is any coupling with $\widehat{Y} \sim \rho$, $Y \sim \nu$, then for every level k, $\mathrm{TV}(p^{(k)}, q^{(k)}) \le \Pr[c_k(\phi(\widehat{Y})) \neq c_k(\phi(Y))]$ (the cell map c_k acts on feature space $\mathcal{Z}$). Hence any measurable nonnegative pointwise surrogate ℓ_k with $\mathbf{1}\{c_k(\phi(\widehat{y})) \neq c_k(\phi(y))\} \le \ell_k(\widehat{y}, c_k(\phi(y)))$ gives the trainable bound

$$W_{1,\mathcal{Z}}(\rho^\phi, \nu^\phi) \;\le\; 2\Delta_K + D_{\mathcal{Z}} \sum_{k=0}^{K} \mathbb{E}\big[\ell_k(\widehat{Y}, c_k(\phi(Y)))\big]. \tag{15}$$

Proof. Define the finest-level collapse map $\Pi_K(z) = z_{c_K(z)}$, so $(\Pi_K)_\# \rho^\phi = \bar{\rho}_{\mathcal{C}^{(K)}}$ and $(\Pi_K)_\# \nu^\phi = \bar{\nu}_{\mathcal{C}^{(K)}}$. Couple $Z \sim \rho^\phi$ with $\Pi_K(Z)$. Since Z and $\Pi_K(Z)$ lie in the same level-K cell, $\|Z - \Pi_K(Z)\|_2 \le \Delta_K$, so $W_{1,\mathcal{Z}}(\rho^\phi, \bar{\rho}_{\mathcal{C}^{(K)}}) \le \Delta_K$, and likewise $W_{1,\mathcal{Z}}(\nu^\phi, \bar{\nu}_{\mathcal{C}^{(K)}}) \le \Delta_K$. By the triangle inequality $W_{1,\mathcal{Z}}(\rho^\phi, \nu^\phi) \le 2\Delta_K + W_{1,\mathcal{Z}}(\bar{\rho}_{\mathcal{C}^{(K)}}, \bar{\nu}_{\mathcal{C}^{(K)}})$. The collapsed measures are discrete on $\{z_j\}$. Their total surplus mass is $\mathrm{TV}(p^{(K)}, q^{(K)})$, and every unit of surplus moves to a deficit at cost $\le D_{\mathcal{Z}}$, so $W_{1,\mathcal{Z}}(\bar{\rho}_{\mathcal{C}^{(K)}}, \bar{\nu}_{\mathcal{C}^{(K)}}) \le D_{\mathcal{Z}}\, \mathrm{TV}(p^{(K)}, q^{(K)})$. This proves the first claim.

For the multi-scale inequality it suffices to prove the one-step version. Let P, Q be distributions on pairs (A, B) (parent A, child B), with $P_{ab} = P_a P_{b|a}$, $Q_{ab} = Q_a Q_{b|a}$. When a parent has zero mass, choose its conditional distribution arbitrarily. The corresponding weighted term vanishes. Using $\mathrm{TV}(P, Q) = 1 - \sum_{a,b} \min(P_{ab}, Q_{ab})$ and $\min(P_a P_{b|a}, Q_a Q_{b|a}) \ge \min(P_a, Q_a) \min(P_{b|a}, Q_{b|a})$,

$$\begin{aligned}\sum_{a,b} \min(P_{ab}, Q_{ab}) &\ge \sum_a \min(P_a, Q_a) \sum_b \min(P_{b|a}, Q_{b|a}) \\ &= \sum_a \min(P_a, Q_a)\big(1 - \mathrm{TV}(P_{\cdot|a}, Q_{\cdot|a})\big).\end{aligned} \tag{16}$$

With $\sum_a \min(P_a, Q_a) = 1 - \mathrm{TV}(P_A, Q_A)$, rearranging gives

$$\mathrm{TV}(P, Q) \le \mathrm{TV}(P_A, Q_A) + \sum_a \min(P_a, Q_a)\, \mathrm{TV}(P_{\cdot|a}, Q_{\cdot|a}). \tag{17}$$

Applying this recursively over the partition tree yields the stated bound. The inequality is generally strict, since $\min(P_a P_{b|a}, Q_a Q_{b|a}) \ge \min(P_a, Q_a) \min(P_{b|a}, Q_{b|a})$ can be strict. The exact identity is $\mathrm{TV}(P, Q) = \mathrm{TV}(P_A, Q_A) + \sum_a \big[\min(P_a, Q_a) - \sum_b \min(P_a P_{b|a}, Q_a Q_{b|a})\big]$, and the bracketed term is $\le \min(P_a, Q_a)\, \mathrm{TV}(P_{\cdot|a}, Q_{\cdot|a})$. We therefore call this a recursive refinement inequality rather than an exact chain rule.

Finally, let $J = c_k(\phi(\widehat{Y}))$, $L = c_k(\phi(Y))$, so $J \sim q^{(k)}$, $L \sim p^{(k)}$. For any A, $p^{(k)}(A) - q^{(k)}(A) = \Pr[L \in A] - \Pr[J \in A] \le \Pr[J \neq L]$. Taking the sup over A gives $\mathrm{TV}(p^{(k)}, q^{(k)}) \le \Pr[J \neq L]$, and surrogate domination gives $\Pr[J \neq L] \le \mathbb{E}[\ell_k]$. Apply this at $k = K$ in Equation (1) and use $\mathbb{E}[\ell_K] \le \sum_{k=0}^{K} \mathbb{E}[\ell_k]$ to obtain Equation (3). □

B.3 Proof of Proposition 3

Restatement of Proposition 3. Under the measurability, compactness, and cell-diameter assumptions of Proposition 3, let $X_0 \sim \rho_0$, $\widehat{Y} = \widetilde{T}(X_0)$, $Y = T_\star(X_0)$, $\widetilde{\rho} = \widetilde{T}_{\#}\rho_0$, and $B_\phi = \mathbb{E}\big[\|\phi(\widehat{Y}) - \phi(Y)\|_2\big]$. Then for every $r > 0$, $\Pr[c(\phi(\widehat{Y})) \neq c(\phi(Y))] \le \min\{1,\ \beta^{\phi}_{\mathcal{C},\nu_\star}(r) + B_\phi/r\}$, and the direct displacement bound gives $W_{1,\mathcal{Z}}(\phi_{\#}\widetilde{\rho}, \phi_{\#}\nu_\star) \le B_\phi$. Combining via the teacher gap $\varepsilon^{\phi}_{\text{teacher}} \ge W_{1,\mathcal{Z}}(\phi_{\#}\nu_\star, \phi_{\#}\nu)$,

$$W_{1,\mathcal{Z}}(\phi_{\#}\widetilde{\rho}, \phi_{\#}\nu) \ \le\ \varepsilon^{\phi}_{\text{teacher}} + \min\Big\{B_\phi,\ 2\Delta + D_{\mathcal{Z}}\ \min\big(1,\ \beta^{\phi}_{\mathcal{C},\nu_\star}(r) + \tfrac{B_\phi}{r}\big)\Big\}. \tag{18}$$

If $\beta^{\phi}_{\mathcal{C},\nu_\star}(r) \le Ar^\kappa$ for all $r \in (0, r_0]$, where $A, \kappa, r_0, B_\phi > 0$, and the minimizer $r_\star = (B_\phi/\kappa A)^{1/(\kappa+1)} \le r_0$ (otherwise take the infimum over $r \in (0, r_0]$), the uncapped wrong-cell upper bound optimizes to $C_\kappa A^{1/(\kappa+1)} B_\phi^{\kappa/(\kappa+1)}$ with $C_\kappa = (\kappa+1)\kappa^{-\kappa/(\kappa+1)}$, so

$$W_{1,\mathcal{Z}}(\phi_{\#}\widetilde{\rho}, \phi_{\#}\nu) \ \le\ \varepsilon^{\phi}_{\text{teacher}} + \min\Big\{B_\phi,\ 2\Delta + D_{\mathcal{Z}}\, C_\kappa\, A^{1/(\kappa+1)}\, B_\phi^{\kappa/(\kappa+1)}\Big\}. \tag{19}$$

When $B_\phi = 0$, the paired feature endpoints agree almost surely. For fixed $A, \kappa, D_{\mathcal{Z}} > 0$ the direct bound is asymptotically stronger as $B_\phi \downarrow 0$. The fractional exponent concerns cell error and histogram TV, not feature-space W_1.

Proof. Let $E_{\text{wrong}} = \{c(\phi(\widehat{Y})) \neq c(\phi(Y))\}$, and for fixed $r > 0$ set $E_{\text{bd}} = \{m_{\mathcal{C}}(\phi(Y)) \le r\}$, $E_{\text{move}} = \{\|\phi(\widehat{Y}) - \phi(Y)\|_2 \ge r\}$. With the Euclidean relative margin, if neither occurs then $\phi(Y)$ is more than r (in $\|\cdot\|_2$) from every other cell and $\phi(\widehat{Y})$ is within Euclidean r of $\phi(Y)$, so $\phi(\widehat{Y})$ stays in $\phi(Y)$'s cell and the label is correct. The margin and the movement use the same metric. Hence $E_{\text{wrong}} \subseteq E_{\text{bd}} \cup E_{\text{move}}$. By Markov, $\Pr(E_{\text{wrong}}) \le \beta^{\phi}_{\mathcal{C},\nu_\star}(r) + \mathbb{E}\big[\|\phi(\widehat{Y}) - \phi(Y)\|_2\big]/r = \beta^{\phi}_{\mathcal{C},\nu_\star}(r) + B_\phi/r$, and it is a probability so the $\min\{1, \cdot\}$ cap holds. The coupling $(\phi(\widehat{Y}), \phi(Y))$ has transport cost equal to B_ϕ, giving $W_{1,\mathcal{Z}}(\phi_{\#}\widetilde{\rho}, \phi_{\#}\nu_\star) \le B_\phi$. Applying Theorem 2 with this coupling gives $W_{1,\mathcal{Z}}(\phi_{\#}\widetilde{\rho}, \phi_{\#}\nu_\star) \le 2\Delta + D_{\mathcal{Z}} \Pr(E_{\text{wrong}})$. Taking the minimum and adding $\varepsilon^{\phi}_{\text{teacher}}$ by the triangle inequality proves Equation (5). For $B_\phi > 0$, $f'(r) = \kappa A r^{\kappa-1} - B_\phi/r^2$ changes sign from negative to positive at $r_\star$, so minimizing $f(r) = Ar^\kappa + B_\phi/r$ gives Equation (6) when $r_\star \le r_0$. Otherwise the minimum over $(0, r_0]$ occurs at r_0. If $B_\phi = 0$, the nonnegative displacement is zero almost surely. No smoothness of any field is used. □

Tightness (for the wrong-cell probability). Take $\mathcal{X} = \mathcal{Z} = [-R, R]$, $\phi = \text{id}$, $C_- = [-R, 0)$, $C_+ = [0, R]$, and $T_\star = \text{id}$. Fix $A, \kappa > 0$, set $R = A^{-1/\kappa}$, and choose the source and teacher law on $[0, R]$ with $\Pr[Y \le r] = Ar^\kappa$ for $0 \le r \le R$. Then $\beta^{\phi}_{\mathcal{C},\nu_\star}(r) = Ar^\kappa$ on this interval. For $0 < u \le R$, define the measurable sampler on the source support by $\widetilde{T}_u(y) = y - (u - y)_+$ and extend it measurably to $[-R, 0)$ with values in $[-R, R]$. Exactly the points $0 \le y < u/2$ cross the boundary, and

$$\begin{aligned} B_\phi = \mathbb{E}\big[|\widetilde{T}_u(Y) - Y|\big] &= \int_0^u (u - y) A\kappa y^{\kappa-1}\, dy = \frac{A}{\kappa+1} u^{\kappa+1}, \\ \Pr[\text{wrong}] &= A(u/2)^\kappa = 2^{-\kappa}(\kappa+1)^{\kappa/(\kappa+1)} A^{1/(\kappa+1)} B_\phi^{\kappa/(\kappa+1)}. \end{aligned} \tag{20}$$

All teacher mass lies in C_+, so the two-cell histogram TV equals this wrong-cell probability. Thus the exponent is sharp for both quantities, even on a fixed compact feature space with a fixed teacher law. The example does not give a fractional-power lower bound for feature-space W_1.

B.4 Proof of Lemma 4

Restatement of Lemma 4. Assume the endpoint Lipschitz, flow-existence, and field-regularity hypotheses of Lemma 4, including an integrable one-sided Lipschitz bound $m_b(t)$ on a common trajectory domain. Set $M_+ = \int_0^1 \max(m_b(t), 0)\, dt$. Let

$$a_b = \int_0^1 \mathbb{E}\big[\|v_b(t, X_t^\star) - v_\star(t, X_t^\star)\|_2\big]\, dt, \qquad X_t^\star = \Phi^t_{v_\star}(X_0), \tag{21}$$

with measurable integrand and $a_b < \infty$, and let the solver satisfy $\mathbb{E}\big[\|\widetilde{T}_b(X_0) - T_b(X_0)\|_2\big] \le C_{\mathrm{sol}} h^p$ for $0 < h \le h_0$, with C_{sol} independent of h. Then the displacement of Proposition 3 obeys

$$B_\phi = \mathbb{E}\big[\|\phi(\widehat{Y}) - \phi(Y)\|_2\big] \;\le\; L_\phi\big(e^{M_+} a_b + C_{\mathrm{sol}} h^p\big). \tag{22}$$

Proof. Let $e_t = X_t^b - X_t^\star$ and $\delta_t = \|v_b(t, X_t^\star) - v_\star(t, X_t^\star)\|_2$. For almost every t, adding and subtracting $v_b(t, X_t^\star)$,

$$\begin{aligned} \frac{d}{dt}\frac{1}{2}\|e_t\|_2^2 &= \langle e_t, v_b(t, X_t^b) - v_b(t, X_t^\star)\rangle + \langle e_t, v_b(t, X_t^\star) - v_\star(t, X_t^\star)\rangle \\ &\le m_b(t)\|e_t\|_2^2 + \|e_t\|_2 \delta_t. \end{aligned} \tag{23}$$

Thus $\frac{d}{dt}\|e_t\|_2 \le m_b(t)\|e_t\|_2 + \delta_t$ almost everywhere: divide by $\|e_t\|_2$ on its positive set, and use that the absolutely continuous norm has derivative zero almost everywhere on its zero set. With $e_0 = 0$, Grönwall gives $\|e_1\|_2 \le \int_0^1 \exp(\int_s^1 m_b(u)\,du)\delta_s\,ds \le e^{M_+}\int_0^1 \delta_s\,ds$ since $\int_s^1 m_b \le M_+$. Taking expectations and applying Tonelli's theorem to the nonnegative discrepancy, then adding solver error, $\mathbb{E}\big[\|\widehat{Y} - Y\|_2\big] \le e^{M_+} a_b + C_{\mathrm{sol}} h^p$. Multiplying by L_ϕ bounds B_ϕ. □

C Extended technical detail

C.1 Path-wise surrogate for hierarchical cells

Differentiable surrogate for Theorem 2 (prototype logits, smooth for smooth ϕ). For the prototype surrogate we assume each level's partition $\mathcal{C}^{(k)}$ is the Voronoi partition induced by its prototypes $z_1^{(k)}, \dots, z_{M_k}^{(k)}$, i.e. $c_k(z) = \arg\min_j \|z - z_j^{(k)}\|_2$, with the same deterministic tie-break used by c_k. Non-Voronoi measurable cells are still covered by Theorem 2. They require a surrogate that dominates error for their actual cell-assignment rule. Fix $\tau_p, \alpha, \tau > 0$ and $\gamma \ge 0$. For $M_k \ge 2$, let $s_j(y) = -\|\phi(y) - z_j^{(k)}\|_2^2/\tau_p$ and $\mathrm{lse}_{\alpha,-c}(y) = \alpha \log \sum_{j \ne c} \exp(s_j(y)/\alpha)$. Define the margin $m_k(y, c) = s_c(y) - \mathrm{lse}_{\alpha,-c}(y)$ and

$$\ell_k(y, c) = \frac{\mathrm{softplus}\big((\gamma - m_k(y, c))/\tau\big)}{\mathrm{softplus}(\gamma/\tau)}. \tag{24}$$

Here $\mathrm{softplus}(u) = \log(1 + e^u)$. For $M_k = 1$, define $\ell_k = 0$, since no classification error is possible. Under the Voronoi assumption, if c is not the nearest-prototype cell of $\phi(y)$ then $m_k(y, c) \le 0$, hence $\ell_k(y, c) \ge 1$, so the loss dominates the hard error. Indeed, even at a tie assigned to another cell, some competitor satisfies $s_j(y) \ge s_c(y)$ and $\mathrm{lse}_{\alpha,-c}(y) \ge s_j(y)$. The loss is nonnegative for correct assignments and is smooth whenever ϕ is smooth.

Path-wise surrogate for hierarchical (tree) cells. The domination above is for a global level-k Voronoi classifier $c_k(z) = \arg\min_j \|z - z_j^{(k)}\|_2$. The hierarchical partition of Section 4.1 instead uses a tree rule c_k^{tree} that descends the parent path, selects the nearest child among the assigned parent's children, and restricts the level-k logits to those children. A loss restricted to the teacher parent's children does not by itself dominate the tree error: if the student endpoint leaves the teacher's parent cell entirely, the restricted margin can still look correct while $c_k^{\mathrm{tree}}(\widehat{z}) \ne c_k^{\mathrm{tree}}(z)$. We therefore use a path-wise surrogate. Fix the root convention $a_{-1}^\star = \mathrm{root}$ with $\mathrm{ch}(a_{-1}^\star) = \mathcal{C}^{(0)}$, so the level-$s = 0$ child decision is well-defined. For level s let $\ell_s^{\mathrm{tree}}(\widehat{y}, y)$ dominate the event that the student selects the wrong child of the teacher's level-$(s-1)$ parent, i.e. $\mathbf{1}\{\text{wrong child at level } s\} \le \ell_s^{\mathrm{tree}}$ (the same prototype-margin softplus, evaluated over the siblings under the teacher's parent). This comparison is evaluated at $\phi(\widehat{y})$ against the children of the teacher parent even if the student's actual path has already diverged. A singleton child set contributes zero loss. At the first divergence, the parent paths still agree, so a tree-cell error implies a wrong decision in one of these comparisons. Hence,

$$\mathbf{1}\{c_k^{\mathrm{tree}}(\phi(\widehat{y})) \ne c_k^{\mathrm{tree}}(\phi(y))\} \le \sum_{s=0}^{k} \mathbf{1}\{\text{wrong child at level } s\} \le \sum_{s=0}^{k} \ell_s^{\mathrm{tree}}(\widehat{y}, y), \tag{25}$$

so $\Pr[c_k^{\mathrm{tree}}\text{ error}] \le \sum_{s=0}^{k}\mathbb{E}[\ell_s^{\mathrm{tree}}]$, and combining with Theorem 2's $\mathrm{TV}(p^{(k)},q^{(k)}) \le \Pr[c_k^{\mathrm{tree}}\text{ error}]$ gives the trainable bound $W_{1,\mathcal{Z}}(\rho^\phi,\nu^\phi) \le 2\Delta_K + D_{\mathcal{Z}}\sum_{s=0}^{K}\mathbb{E}[\ell_s^{\mathrm{tree}}]$. The global-Voronoi surrogate is the special case of a single-level tree. The following consequence makes the role of the level weights explicit.

Corollary 5 (Positive level weights)**.** *Fix a measurable feature map, a hierarchical Voronoi partition satisfying the hypotheses of Theorem 2, and a coupling $(\widehat{Y},Y)$ with $\widehat{Y}\sim\widetilde{\rho}$ and $Y\sim\nu_\star$. Let ℓ_s^{tree} be the nonnegative path losses in equation 9, and let $\mathcal{L}_{\mathrm{cell}}=\sum_{s=0}^{K}\lambda_s\ell_s^{\mathrm{tree}}$ with fixed weights $\lambda_s>0$. Set $\lambda_{\min}=\min_s\lambda_s$. Then equation 10 holds. If $\varepsilon_{\mathrm{teacher}}^{\phi}\ge W_{1,\mathcal{Z}}(\phi_\#\nu_\star,\phi_\#\nu)$, then*

$$W_{1,\mathcal{Z}}(\phi_\#\widetilde{\rho},\phi_\#\nu) \le \varepsilon_{\mathrm{teacher}}^{\phi} + 2\Delta_K + \frac{D_{\mathcal{Z}}}{\lambda_{\min}}\mathbb{E}\Big[\mathcal{L}_{\mathrm{cell}}(\widehat{Y},Y)\Big]. \tag{26}$$

Proof. Nonnegativity gives $\mathcal{L}_{\mathrm{cell}} \ge \lambda_{\min}\sum_s \ell_s^{\mathrm{tree}}$ pointwise. Taking expectations in equation 25 proves equation 10. Applying Theorem 2 to the teacher–student coupling and adding the teacher–data gap by the triangle inequality proves equation 26. □

For positive weights, the unweighted population sum can also be recovered as $\sum_s \lambda_s^{-1}\mathbb{E}[\lambda_s\ell_s^{\mathrm{tree}}]$ if the levelwise expectations are available. A zero-weight level must be controlled separately. The corollary concerns population losses for a fixed sampler and does not turn its training minibatch loss into a held-out certificate.

C.2 Decision-rule cell margin and boundary-tube mass

To relate the endpoint displacement B_ϕ to cell-crossing probability, we need a margin that measures how far a teacher endpoint is from changing its cell assignment. A standard ambient-boundary distance is unsuitable when $\mathcal{Z}$ lies on a low-dimensional manifold (e.g., the unit sphere induced by L_2-normalized features), since the ambient boundary $\partial_{\mathbb{R}^m}C$ may contain the entire cell and hence give zero margin. We instead define the decision-rule margin

$$m_{\mathcal{C}}(z) = \inf_{z'\in\mathcal{Z}\setminus C_{c(z)}} \|z-z'\|_2, \tag{27}$$

the Euclidean distance from z to the nearest point assigned to another cell. Importantly, this uses the same feature-space Euclidean metric as $B_\phi=\mathbb{E}\big[\|\phi(\widehat{Y})-\phi(Y)\|_2\big]$, so $\|\widehat{z}-z\|_2 < m_{\mathcal{C}}(z)$ guarantees that the cell label cannot change. The convention $\inf\varnothing=+\infty$ covers a one-cell partition. This margin is measurable: on each of finitely many Borel cells it is the distance to a fixed set, which is a continuous function when that set is nonempty.

For Voronoi cells with pairwise distinct prototypes $\{p_j\}$, a conservative computable lower bound is the half-space barrier

$$m_{\mathrm{half}}(z) = \min_{j\ne c(z)} \frac{\|z-p_j\|_2^2 - \|z-p_{c(z)}\|_2^2}{2\|p_j-p_{c(z)}\|_2}. \tag{28}$$

Thus, $\|\widehat{z}-z\|_2 < m_{\mathrm{half}}(z)$ prevents crossing any Voronoi bisector and preserves the cell label. To see this, each squared-distance difference is affine in z, with gradient $2(p_{c(z)}-p_j)$, so its value remains positive under a displacement smaller than the displayed ratio. Consequently $m_{\mathrm{half}}(z)\le m_{\mathcal{C}}(z)$, including a zero barrier at ties. Here m_{half} is an ambient Euclidean barrier, not an intrinsic geodesic distance.

For a hierarchical Voronoi tree, apply the same construction to every sibling comparison along the path of z. If $a_s(z)$ is its level-s node and $p_j^{(s)}$ are the child prototypes, define

$$m_{\mathrm{half}}^{\mathrm{tree}}(z) = \min_{0\le s\le K}\ \min_{\substack{j\in\mathrm{ch}(a_{s-1}(z))\\ j\ne a_s(z)}} \frac{\|z-p_j^{(s)}\|_2^2 - \|z-p_{a_s(z)}^{(s)}\|_2^2}{2\|p_j^{(s)}-p_{a_s(z)}^{(s)}\|_2}, \tag{29}$$

with distinct prototypes within each sibling set and empty minima interpreted as $+\infty$. A displacement smaller than this barrier preserves every path decision by induction, so it lower-bounds the finest tree-cell margin. Using only the final sibling comparison does not control crossings of ancestor boundaries.

Finally, for the teacher endpoint $Y = T_\star(X_0)$, define the boundary-tube mass

$$\beta^{\phi}_{\mathcal{C},\nu_\star}(r) = \Pr[m_{\mathcal{C}}(\phi(Y)) \le r]\,, \tag{30}$$

which measures the probability mass of teacher endpoints lying within distance r of a cell change. When a computable lower margin $\underline{m} \le m_{\mathcal{C}}$ is used, its tube probability satisfies $\beta^{\phi}_{\mathcal{C},\nu_\star}(r) \le \Pr[\underline{m}(\phi(Y)) \le r]$. Thus a held-out upper bound on this latter probability gives a conservative certificate.

C.3 Scope and Technical Assumptions

Scope of Lemma 4. The ODE perturbation estimate applies to smooth students, weight-only quantization with smooth activations, or quantizers with smoothed forward maps, provided the flow and integrable one-sided Lipschitz assumptions hold. In particular,

$$M_+ = \int_0^1 \max\{m_b(t), 0\}\, dt \tag{31}$$

must be finite. A straight-through gradient estimator does not make a hard-quantized forward map continuous. For hard activation-quantized integer inference, the induced field is discontinuous and generally does not admit the finite one-sided Lipschitz bound required by Lemma 4. We therefore do not apply Lemma 4 to the deployed integer model. For such samplers, the applicable result is Proposition 3. It could be evaluated using an independent upper confidence bound on the realized feature displacement B_ϕ and the teacher boundary-tube mass. Existence on the entire time interval and uniqueness of both flows, together with the required measurability and integrability of the field discrepancy, are assumed throughout Lemma 4.

The one-sided Lipschitz bound must hold throughout a common trajectory domain. On a convex domain for a continuously differentiable field, it suffices to bound the largest eigenvalue of the symmetric Jacobian uniformly over that domain, as follows by integrating the Jacobian along the segment joining x and y. Values sampled along finitely many trajectories do not by themselves certify this uniform bound or M_+. Likewise, the solver-error estimate is an explicit hypothesis. The nominal order p alone does not establish it.

All teacher-relative certificates refer to the map used to define $\nu_\star$. If a numerical teacher replaces the exact map $T_\star$, one must either define the reference law using that numerical map or additionally bound its feature-space error relative to the exact teacher.

D Extended Experiment detail

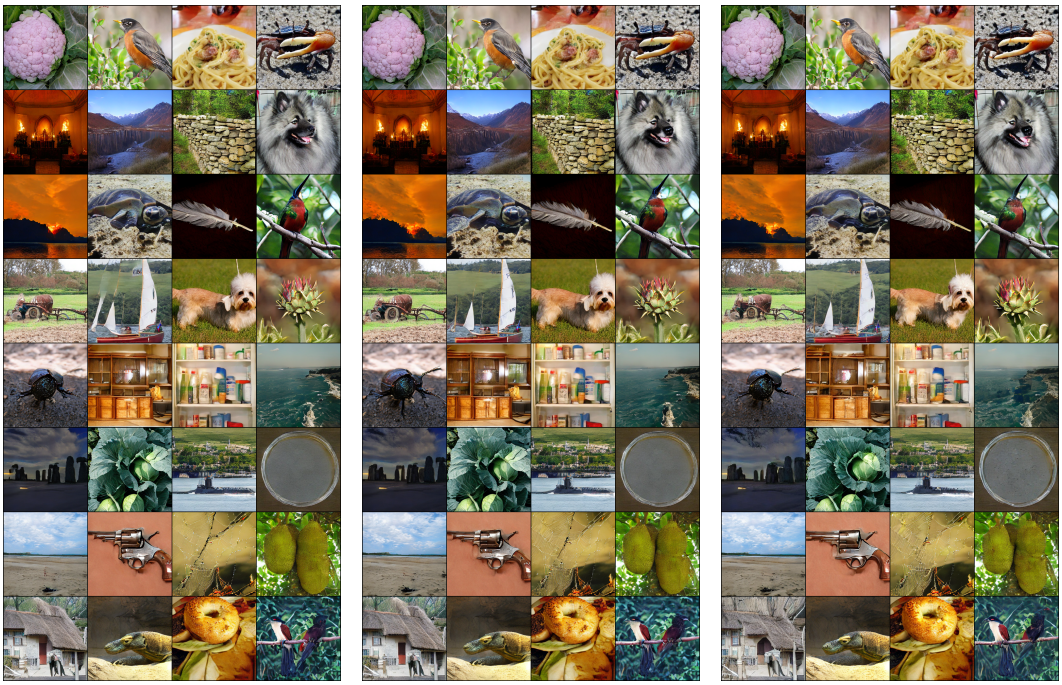
D.1 Implementation

Extreme low-bit initialization note. For extreme low-bit settings (see example in Algorithm 1), we use progressive initialization following Bit-by-Bit (Xu et al., 2026): W4A4, W3A8, W2A8 are warm-started from a W4A8 checkpoint rather than directly from full precision. This simple yet effective strategy improves optimization stability while keeping training and evaluation at the target precision.

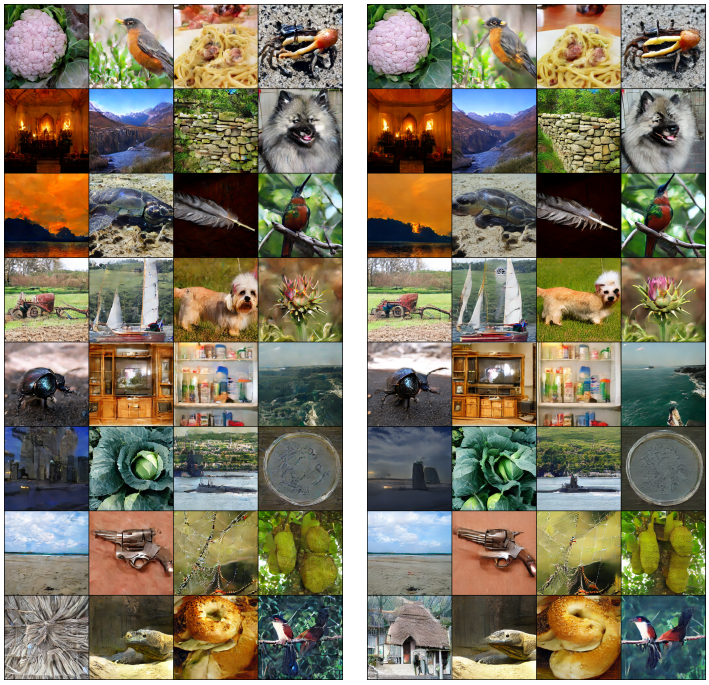
Training configurations. Table 3 summarizes the main CM-QAT training configurations used for Fashion-MNIST, CIFAR-10, and ImageNet-256, including the hierarchical partition, loss coefficients, histogram/stability objectives, and optimization settings.

D.2 Qualitative Visualization

Figure 4 provides a qualitative comparison between the full-precision MeanFlow teacher and 1-NFE CM-QAT models on ImageNet-256 across W8A8, W4A8, W4A4, and W3A8 settings. The samples complement the quantitative evaluation by illustrating the visual quality retained under progressively more aggressive quantization.



(a) MeanFlow FP (b) CM-QAT W8A8 (c) CM-QAT W4A8



(d) CM-QAT W4A4 (e) CM-QAT W3A8

Figure 4: **Qualitative comparison on ImageNet-256 1-NFE Generation.** Generated samples from the full-precision MeanFlow teacher and CM-QAT models under different weight/activation quantization settings.

Table 3: **CM-QAT training configurations.** Representative configurations used for the main experiments.

Configuration	Fashion-MNIST	CIFAR-10	ImageNet-256
Hierarchical partition			
Feature dimension	128	256	4096
Number of levels $K+1$	3	3	4
Prototype counts $\{M_s\}$	$(10, 100, 500)$	$(16, 64, 256)$	$(128, 512, 2048, 8192)$
Children per parent	$10, 10, 5$	$16, 4, 4$	$128, \sim 4, \sim 4, \sim 4$
Partition samples	50k	50k	200k
Partition construction	Hierarchical MiniBatchKMeans		
Level weights $\{\lambda_s\}$	$(1, 1, 1)$	$(1, 1, 1)$	$(1, 1, 1, 1)$
Cell-margin objective			
τ_p	0.07		
α	0.05		
γ	0.1		
τ	$0.5 \to 0.05$ over the first 70% of training		
λ_{CM}	1	1	1
λ_{cell}	1	1	1
λ_{hist}	0.5	0.005	0.7
λ_{margin}	0.05	0.001	0.05
λ_{log}	10^{-4}	10^{-4}	10^{-4}
$\lambda_{\text{velocity}}$	0.02	0	0.02
$\lambda_{\text{endpoint}}$	0	0.1	0.2
Histogram and stability objectives			
Histogram assignment	Finest level; teacher hard, student soft		
Sinkhorn ε / iterations	0.05 / 8		
Log-norm states	16 teacher-pair interpolation states, $t \sim \mathcal{U}[0, 1)$		
Log-norm estimator	Single random-direction finite difference		
Optimization			
Optimizer	AdamW, $\beta = (0.9, 0.999)$		
Learning rate	5×10^{-5}	5×10^{-6}	10^{-5}
Weight decay	0.01	0.01	0
Training steps	40k	80k	15k

E Baseline Algorithms

OT-weight PTQ. We use a faithful frozen-network OT-weight baseline that performs per-output-channel equal-mass weight quantization only. A Lloyd/nearest-centroid codebook variant is treated only as an ablation and not as the cited OT-PTQ method.

Algorithm 2 OT-Weight PTQ for Flow Matching

Require: FP32 teacher weights $\theta_\star$, weight bit-width b_w
1: Freeze the network; do not retrain weights or quantizer parameters.
2: Set the number of quantization levels $K = 2^{b_w}$.
3: **for** each weight tensor W **do**
4: **for** each output channel c **do**
5: Flatten $W[c]$ and sort its entries.
6: Split the sorted values into K equal-mass chunks $\{B_k\}_{k=1}^{K}$.
7: **for** $k = 1, \ldots, K$ **do**
8: Compute centroid $q_k = \frac{1}{|B_k|} \sum_{w \in B_k} w$.
9: Replace all values in B_k by q_k.
10: **end for**

Restore the original element order.
end for
end for
Evaluate at target NFE.

LSQ velocity-QAT. This baseline directly minimizes velocity-field reconstruction,

$$\mathcal{L}_{\text{LSQ-vel}} = \mathbb{E}_{t,x}\left[\|v_b(t,x) - v_\star(t,x)\|_2^2\right].$$

It provides a strong field-L_2 baseline, but does not explicitly account for terminal cell assignments or boundary margins.

Algorithm 3 LSQ Velocity-QAT for Flow Matching

Require: teacher $v_\star$, student v_b, source ρ_0, target NFE
Initialize v_b from teacher parameters $\theta_\star$.
Initialize weight and activation scales by PTQ calibration.
Enable LSQ fake quantization for weights and activations.
for each QAT step **do**
  Sample $x_0 \sim \rho_0$ and $t \sim \mathcal{U}(0,1)$.
  Compute teacher endpoint $y^\star = T_\star(x_0)$.
  Obtain teacher trajectory state $x_t^\star$ ▷ or linear FM bridge
  Compute

$$\mathcal{L}_{\text{LSQ-vel}} = \|v_b(t,x_t^\star) - v_\star(t,x_t^\star)\|_2^2.$$

  Update student weights and LSQ scales through STE.
end for
Freeze quantization parameters and evaluate at target NFE.

Endpoint-L_2 QAT. We consider both pixel-space endpoint matching, $\|\widehat{y}-y^\star\|_2^2$, and feature-space endpoint matching, $\|\phi(\widehat{y}) - \phi(y^\star)\|_2^2$. The feature-space variant is used as the stronger endpoint-matching baseline.

Algorithm 4 Endpoint-L_2 Quantization-Aware Training

Require: teacher $v_\star$, student v_b, source ρ_0, target NFE
Initialize v_b from the teacher and enable fake quantization.
for each QAT step **do**
  Sample $x_0 \sim \rho_0$.
  Compute high-NFE teacher endpoint $y^\star = T_\star(x_0)$.
  Integrate the quantized student at target NFE:

$$\widehat{y} = \text{ODESolve}_h(v_b, x_0).$$

  Compute endpoint loss

$$\mathcal{L}_{\text{end}} = \|\widehat{y} - y^\star\|_2^2.$$

  Optionally add velocity regularization:

$$\mathcal{L} = \mathcal{L}_{\text{end}} + 0.02\,\mathcal{L}_{\text{vel}}.$$

  Update through STE fake quantization.
end for
Evaluate at the same target NFE.

Q-Sched-CNF. We adapt Q-Sched to continuous normalizing flows by interpreting c_v as the model-output coefficient and c_x as the state coefficient. The quantized velocity field and all quantization parameters remain frozen; only four scalar schedule endpoints are selected on a calibration split. Unlike CM-QAT, Q-Sched-CNF performs no weight updates and uses neither paired teacher-cell labels nor cell-margin objectives. We store the resulting per-step $\{c_x, c_v\}$ coefficients using the `qsched_cnf_v2` format; legacy `alpha/beta/delta` scheduler caches are not used.

Algorithm 5 Q-Sched-CNF: Scheduler-Only Quantization Correction

Require: frozen PTQ field v_q, target NFE K, calibration sources $\mathcal{X}_{\mathrm{cal}}$, target features $\mathcal{Z}_{\mathrm{cal}}$
Freeze v_q and all quantization parameters.
for each candidate $(c_x^s, c_x^e, c_v^s, c_v^e)$ **do**
  Construct

$$c_x = \mathrm{linspace}(c_x^s, c_x^e, K), \qquad c_v = \mathrm{linspace}(c_v^s, c_v^e, K).$$

  for each calibration source x_0 **do**
    Set $x_0^q = x_0$.
    for $i = 0, \dots, K-1$ **do**
      if Euler solver **then**

$$x_{i+1}^q = c_x[i] x_i^q + h_i c_v[i] v_q(t_i, x_i^q).$$

      else
        $k_1 = v_q(t_i, x_i^q)$.

$$x_{\mathrm{pred}} = c_x[i] x_i^q + h_i c_v[i] k_1.$$

        $k_2 = v_q(t_{i+1}, x_{\mathrm{pred}})$.

$$x_{i+1}^q = c_x[i] x_i^q + h_i c_v[i] \frac{k_1 + k_2}{2}.$$

      end if
    end for
    Store final sample $\widehat{y} = x_K^q$.
  end for
  Compute calibration objective

$$\mathcal{J}_{\mathrm{sched}} = \mathrm{Sinkhorn}_\phi\big(\phi(\widehat{\mathcal{Y}}), \mathcal{Z}_{\mathrm{cal}}\big) + \lambda_{\mathrm{reg}} \frac{1}{K} \sum_{i=0}^{K-1} \left[(c_x[i] - 1)^2 + (c_v[i] - 1)^2 \right].$$

end for
Select the schedule minimizing $\mathcal{J}_{\mathrm{sched}}$.
Evaluate the frozen quantized model with the selected schedule.

Q-Drift-CNF. We construct a frozen-network CNF analogue of Q-Drift using only an offline calibration split. The diffusion noise-level index is replaced by the ODE solver step, and quantization drift is estimated as a step-wise, channel-wise affine residual correction. For Heun integration, the same step-i correction is applied to both predictor and corrector field evaluations. Unless otherwise stated, we use 512 calibration pairs. No network weights or quantizer scales are updated.

Algorithm 6 Q-Drift-CNF: Calibration-Based Drift Correction

Require: PTQ field v_q, teacher $v_\star$, calibration set $\mathcal{C}_{\mathrm{cal}}$, target NFE K
Freeze v_q and all quantization parameters.
for each $x_0 \in \mathcal{C}_{\mathrm{cal}}$ **do**
  Roll out the teacher trajectory $\{x_i^\star\}_{i=0}^K$ on the target solver grid.
  for $i = 0, \dots, K-1$ **do**
    Compute

$$v_i^\star = v_\star(t_i, x_i^\star), \qquad v_i^q = v_q(t_i, x_i^\star).$$

    Compute residual $r_i = v_i^\star - v_i^q$.
    Store paired samples (v_i^q, r_i).
  end for
end for
for each solver step i and channel c **do**
  Compute

$$g_i[c] = \mathrm{clip}\left(\frac{\mathrm{Cov}(r_i[c], v_i^q[c])}{\mathrm{Var}(v_i^q[c]) + \lambda_{\mathrm{ridge}}}, -g_{\max}, g_{\max} \right).$$

  Compute

$$b_i[c] = \mathbb{E}[r_i[c]] - g_i[c] \mathbb{E}[v_i^q[c]].$$

Estimate residual variance

$$\sigma^2_{i,\mathrm{res}}[c] = \mathrm{Var}\big(r_i[c] - g_i[c]v_i^q[c] - b_i[c]\big).$$

end for

Define the corrected field

$$v_{\mathrm{corr}}(t_i, x) = (1 + g_i) \odot v_q(t_i, x) + b_i.$$

Use v_{corr} for all target-NFE solver steps.

Evaluate without updating network weights or quantizer scales.

F Summary of Notation

Table 4: Summary of notation.

Notation	Definition
State and feature spaces	
$\mathcal{X} \subseteq \mathbb{R}^d$	ODE state space of dimension d.
$\mathcal{Z} \subseteq \mathbb{R}^m$	Feature space of dimension m.
$\phi : \mathcal{X} \to \mathcal{Z}$	Fixed feature map, such as a frozen encoder or latent feature map.
id	Identity map. When $\phi = \mathrm{id}$, feature-space distances coincide with state-space distances.
$\mathcal{P}(\mathcal{X})$	Set of probability measures on $\mathcal{X}$.
$\mu \in \mathcal{P}(\mathcal{X})$	Generic probability distribution on the state space.
$\phi_{\#}\mu = \mu^\phi$	Pushforward of μ through ϕ. Equivalently, $\mu^\phi(A) = \mu(\phi^{-1}(A))$ for every measurable $A \subseteq \mathcal{Z}$.
$D_{\mathcal{Z}}$	Diameter of the feature space, $D_{\mathcal{Z}} = \sup_{z,z' \in \mathcal{Z}} \|z - z'\|_2$.
Feature-space partition	
$\mathcal{C} = \{C_1, \ldots, C_M\}$	Partition of $\mathcal{Z}$ into M Borel cells.
M	Number of cells in the partition $\mathcal{C}$.
C_j	The jth cell of the partition.
$c : \mathcal{Z} \to [M]$	Cell-assignment map defined by $c(z) = j$ when $z \in C_j$, with a fixed tie-breaking rule on cell boundaries.
$[M]$	Index set $\{1, \ldots, M\}$.
$m_{\mathcal{C}}(z)$	Cell margin of z, defined as its Euclidean distance to the nearest competing cell under the cell-assignment rule.
$z_j \in C_j \cap \mathcal{Z}$	Representative of cell C_j, such as a medoid or a centroid projected onto $\mathcal{Z}$.
$\mathrm{diam}(C_j)$	Diameter of cell C_j, $\mathrm{diam}(C_j) = \sup_{z,z' \in C_j} \|z - z'\|_2$.
Δ	Uniform upper bound on cell diameters: $\mathrm{diam}(C_j) \le \Delta$ for all j.
$\Delta_{\mathcal{C}}$	Maximum representative-based cell radius, $\Delta_{\mathcal{C}} = \max_{1 \le j \le M} \sup_{z \in C_j} \|z - z_j\|_2$.
Cell histograms and collapsed measures	
Δ_M	Probability simplex $\Delta_M = \{p \in \mathbb{R}^M_+ : \sum_{j=1}^M p_j = 1\}$.
$p^\mu \in \Delta_M$	Cell histogram induced by μ, with entries $p^\mu_j = \mu^\phi(C_j)$.
p^μ_j	Probability mass assigned by μ^ϕ to cell C_j.
δ_{z_j}	Dirac probability measure concentrated at z_j.
$\bar{\mu}_{\mathcal{C}}$	Collapsed feature-space measure $\bar{\mu}_{\mathcal{C}} = \sum_{j=1}^M p^\mu_j \delta_{z_j}$.
$\mathrm{TV}(p, q)$	Total variation distance between $p, q \in \Delta_M$: $\mathrm{TV}(p, q) = \frac{1}{2}\|p - q\|_1$.
Continuous normalizing flow	

Table 4: Summary of notation (continued)

Notation	Definition
$\rho_0 \in \mathcal{P}(\mathcal{X})$	Source distribution from which the initial state X_0 is sampled.
$X_0 \sim \rho_0$	Initial state of the ODE.
$v : [0,1] \times \mathcal{X} \to \mathbb{R}^d$	Generic time-dependent velocity field.
X_t	ODE state at time t, satisfying $\frac{dX_t}{dt} = v(t, X_t)$.
Φ_v^t	Flow map induced by v from time 0 to time t.
$T_v = \Phi_v^1$	Terminal map induced by v at time $t = 1$.
Teacher model and data distribution	
$v_\star$	Full-precision teacher velocity field.
$T_\star = \Phi_{v_\star}^1$	Exact terminal map of the full-precision teacher.
$Y_\star = T_\star(X_0)$	Teacher endpoint corresponding to source sample X_0.
$\nu_\star = (T_\star)_{\#}\rho_0$	Terminal distribution generated by the full-precision teacher.
$\nu \in \mathcal{P}(\mathcal{X})$	True data distribution, which need not equal $\nu_\star$.
Quantized student model	
θ	Full-precision parameters of the student velocity field.
b	Bit width used to quantize the student parameters.
Q_b	A b-bit parameter quantizer.
$Q_b(\theta)$	Quantized student parameters.
$v_b = v_{Q_b(\theta)}$	Velocity field induced by the b-bit quantized parameters.
$T_b = \Phi_{v_b}^1$	Exact terminal map of the quantized student.
$Y_b = T_b(X_0)$	Exact quantized-student endpoint corresponding to X_0.
$\rho_b = (T_b)_{\#}\rho_0$	Exact terminal distribution of the quantized student.
Numerical solver and deployed sampler	
h	Step size of the numerical ODE solver.
p	Order of the numerical ODE solver.
$\widetilde{T}_b$	Numerically implemented terminal map of the quantized student.
$\widetilde{\rho}_b = (\widetilde{T}_b)_{\#}\rho_0$	Distribution generated by the deployed quantized sampler.
NFE	Number of velocity-field evaluations used by the numerical solver.
Evaluation objective and cell risk	
$W_{1,\mathcal{Z}}$	The 1-Wasserstein distance computed using the Euclidean transport cost in feature space $\mathcal{Z}$.
$W_{1,\mathcal{Z}}(\phi_{\#}\widetilde{\rho}_b, \phi_{\#}\nu)$	Feature-space Wasserstein distance between the deployed student distribution and the true data distribution.
$\Pi(\rho_0, \nu)$	Set of couplings with marginals ρ_0 and ν.
$\pi \in \Pi(\rho_0, \nu)$	A joint distribution of source–target pairs (X_0, X_1) whose marginals are ρ_0 and ν, respectively.
$(X_0, X_1) \sim \pi$	A source–target pair sampled from the coupling π.
B_ϕ	Expected teacher–student endpoint displacement in feature space, $$B_\phi = \mathbb{E}_{X_0 \sim \rho_0}\left[\left\|\phi(\widetilde{T}_b(X_0)) - \phi(T_\star(X_0))\right\|_2\right].$$
$\beta^\phi_{\mathcal{C},\nu_\star}(r)$	Teacher boundary-tube probability at radius r, $$\beta^\phi_{\mathcal{C},\nu_\star}(r) = \Pr_{Y \sim \nu_\star}\left[m_{\mathcal{C}}(\phi(Y)) \le r\right],$$ i.e., the probability that a teacher endpoint lies within distance r of a cell boundary in feature space.
$R_{\mathcal{C}}(v_b; \pi)$	Terminal cell-mismatch risk, $$R_{\mathcal{C}}(v_b; \pi) = \Pr_{(X_0, X_1) \sim \pi}[c(\phi(T_b(X_0))) \ne c(\phi(X_1))].$$

Table 4: Summary of notation (continued)

Notation	Definition
$(X_0, T_\star(X_0))$	Teacher-generated source–endpoint pair used when direct source–data couplings are unavailable.