From Attention to Diffusion: A Unified Entropic Optimal Transport View

Anonymous authors
Paper under double-blind review

Abstract

We show that transformer attention and diffusion models are discretizations of the same entropy-regularized optimal transport (OT) flow. A single attention layer is a KL-proximal (JKO/mirror) step in an OT potential; stacking layers yields probability paths that converge to a probability–flow ODE (PF–ODE) on the simplex. Our construction uses a causal, semi-relaxed EOT that preserves attention masking while retaining OT geometry. We derive a finite-depth error bound controlled by a budget Ξ_L (quantifying continuum validity) and prove that stacked attention weakly approximates time-inhomogeneous, anisotropic reverse diffusions with an error that separates time discretization, logit variation, and optional degeneracy regularization. Geometrically, we characterize exact Schrödinger Bridge (SB) alignment via a rotational energy $\mathcal{R}$ that vanishes if and only if the path is SB, and serves as a practical diagnostic otherwise. The framework yields testable predictions: (i) the continuum approximation is accurate when Ξ_L is small; (ii) depth exhibits diminishing returns beyond a threshold set by contraction and step size; and (iii) lower $\mathcal{R}$ correlates with improved generations. We validate these predictions with a diagnostic suite (P0–P4): BV/continuity gating (with abstention on failure), PF–ODE adequacy, curvature/locking geometry, and SB energy. Evidence spans three tracks—Transformers (core diagnostics), diffusion LLMs (dLLM; late-window stability certificate), and a compact image diffusion model (parity and first-order weak-error behavior). These insights motivate mobility-aware temperature scheduling and certified early exit, conserving depth while preserving transport geometry.

1 Introduction

Transformers and diffusion models appear fundamentally different, yet we show they instantiate two discretizations of the same entropy-regularized optimal transport flow. One attention layer performs a KL-proximal step in an optimal transport potential, and depth plays the role of time for the induced probability dynamics on the simplex.

Contributions. Under mild regularity assumptions (detailed in Section 2), our main results are:

1. *Layer-level principle.* Attention implements a principled KL-proximal transport step (mirror/JKO view); see Proposition 2.1.
2. *Depth-to-time with rates.* The discrepancy between layerwise paths and the probability-flow ODE is controlled by a finite-depth budget Ξ_L; see Theorem 5.1.
3. *Diffusion unification.* Stacked attention weakly approximates time-inhomogeneous, anisotropic reverse diffusions with an error that separates discretization, logit variation, and optional degeneracy regularization; see Theorem 6.6.
4. *SB alignment certificate.* A rotational-energy quantity R characterizes when the flow is exactly Schrödinger Bridge–aligned and provides a practical diagnostic otherwise; see Theorem 7.2.

Predictions and implications. Our theory yields falsifiable predictions and design levers:

- *Continuum validity.* The PF–ODE approximation is accurate when the finite-depth budget Ξ_L is small; this provides a testable condition for continuum validity.
- *Depth budgeting.* Depth exhibits diminishing returns once Ξ_L exceeds a quantitative threshold determined by contraction and step size; we use this to justify (not guarantee) early exit certificates.
- *Transport optimality.* Lower rotational energy R is consistent with Schrödinger Bridge alignment and correlates with improved generation quality; we monitor R as an operational diagnostic rather than a stand-alone guarantee.

Further context. Extended motivation and a conceptual schematic are provided in Appendix A, see also Fig. 3, Fig. 4, and Table 2 for a high-level overview and novelty map.

These predictions inform mobility-aware temperature scheduling and certified early exit strategies; complete diagnostic protocols and proofs appear in the appendix.

Positioning and scope. Our empirical study is designed to test *diagnostic predictions* of the theory—PF–ODE adequacy, finite-depth budgets, and SB alignment—rather than to optimize benchmark scores. We therefore emphasize geometry-aware diagnostics and stability certificates, with a compact vision sanity check; large-scale performance tuning is out of scope for this paper (details and limitations in the appendix).

2 Preliminaries and Mathematical Framework

This section fixes notation, states the standing assumptions used throughout, and records the layer-level optimal-transport view we will invoke later.

Global Assumptions. We work on compact subsets where all quantities are well-defined. Unless stated otherwise, we assume:

1. *Bounded-variation logits.* Let $z^{(\ell)}$ be layer logits and $\Delta z^{(\ell)} := z^{(\ell+1)} - z^{(\ell)}$. We have $\sum_\ell \|\Delta z^{(\ell)}\|_\infty < \infty$.
2. *Local drift regularity.* The effective drift $b(\cdot, t)$ is locally Lipschitz in its state argument on bounded sets with Lipschitz constant L_b and is locally bounded by M_b.
3. *Mobility bounds.* For $p = \mathrm{softmax}(z/\tau)$ with temperature $\tau > 0$, the Jacobian $J_{\mathrm{sm}}(z) = \mathrm{Diag}(p) - pp^\top$ satisfies operator-norm and derivative bounds on the relevant compact domain; denote $\Lambda_J := \sup \|J_{\mathrm{sm}}(z)\|_{\mathrm{op}}$ and $L_J := \sup \|\nabla J_{\mathrm{sm}}(z)\|_{\mathrm{op}}$.
4. *Simplex invariance.* Probability vectors p remain in the simplex under the dynamics considered; faces are handled by the standard tangent-space restriction.

Softmax and mobility. Given logits $z \in \mathbb{R}^V$ and temperature $\tau > 0$,

$$p = \mathrm{softmax}(z/\tau), \qquad p_i = \frac{\exp(z_i/\tau)}{\sum_j \exp(z_j/\tau)}.$$

Define the softmax Jacobian (mobility tensor on the simplex)

$$J_{\mathrm{sm}}(z) = \mathrm{Diag}(p) - pp^\top.$$

Remark (sharp mobility bound). We have $\|J_{\mathrm{sm}}(z)\|_{\mathrm{op}} \le \frac{1}{2\tau}$, with equality at distributions $p = (\frac{1}{2}, \frac{1}{2}, 0, \ldots, 0)$. In particular, for $\tau = 1$, $\|J_{\mathrm{sm}}(z)\|_{\mathrm{op}} \le \frac{1}{2}$ and the spectrum lies in $[0, \frac{1}{2}]$, collapsing to $\{0\}$ as $\max_i p_i \to 1$. A proof is provided in Appendix B.

Semi-relaxed Entropic Optimal Transport (EOT) and causality. We adopt a semi-relaxed entropic OT formulation that preserves autoregressive causality (row constraints and masking) while retaining OT geometry. Technical details are deferred to Appendix B.

Proposition 2.1 (Attention as KL-prox/JKO step)**.** *Let $c_j = -q \cdot k_j$ and $\tau > 0$. For any full-support reference u,*

$$p^+ \in \arg\min_{p \in \Delta} \Big\{ \langle c, p \rangle + \tau\, \mathrm{KL}(p \,\|\, u) \Big\}.$$

Stacking such updates discretizes a KL-mirror/JKO flow under the assumptions in Section 2. The proof is given in Appendix B.

For intuition before the formal development, see the conceptual overview in Section 4.

3 Related Work and Positioning

Semi-relaxed entropic OT versus balanced OT. Our use of a semi-relaxed, row-constrained entropic optimal transport formulation preserves causal masking essential for autoregressive models, in contrast to balanced OT and Sinkhorn-style approaches that enforce doubly stochastic couplings and do not respect causal structure (see e.g. (Sander et al., 2022; Tay et al., 2020; Xu et al., 2023; Daneshmand, 2024)). This positions attention as transport under causal constraints rather than as mere normalization.

JKO/mirror descent and gradient flows. The KL-prox characterization of an attention layer aligns with proximal/mirror perspectives on gradient flows in probability; in our setting this connects discrete layer updates to probability-flow ODEs in the continuum limit (see Appendix C for detailed pointers).

Diffusion models and Schrödinger Bridges. The probability-flow ODE / reverse-SDE duality and the Schrödinger Bridge view of entropic transport (Song et al., 2021; Lipman et al., 2022; De Bortoli et al., 2021; Shi et al., 2023) provide the backdrop for our rotational-energy criterion, which operationalizes SB alignment within attention-induced flows. The full version of this section appears in Appendix C, with subsections C.1–C.6.

4 Conceptual overview

The softmax Jacobian $J_{\mathrm{sm}}(z)$ acts as the mobility tensor on the probability simplex, with temperature modulating transport capacity via $J^{\tau}_{\mathrm{sm}}(z) = \tau^{-1} J_{\mathrm{sm}}(z/\tau)$. Higher temperatures maintain mobility deeper in the network. The finite-depth budget Ξ_L quantifies how well discrete layers approximate continuous flow by combining time discretization and per-layer logit variation. Small Ξ_L ensures the probability-flow ODE accurately captures layerwise behavior, with additional depth yielding diminishing returns. Rotational energy $\mathcal{R}$ measures deviation from optimal transport by separating gradient-driven flow from spurious rotational components. Exact Schrödinger Bridge alignment occurs when $\mathcal{R}$ vanishes; empirically, lower values correlate with improved generation.

5 Discrete Continuity and the Continuous-Depth Limit

5.1 Bounded variation regime and practical implications

The transition from discrete layers to continuous dynamics requires controlling the accumulation of changes across depth. We formalize this through a bounded-variation (BV) condition that captures when transformers exhibit smooth evolution rather than abrupt transitions.

Assumption 5.1 (Bounded variation with weak convergence)**.** *Let $\delta t = 1/L$ and $t_\ell = \ell/L$. We assume:*

1. *Bounded total variation: $\sum_\ell \|\Delta z^{(\ell)}\|_2 \le C$ (uniformly in L).*
2. *Uniform boundedness (tightness): $\sup_\ell \|z^{(\ell)}\|_2 \le C_z$.*
3. *Weak first-order consistency: $D_L := \Delta z^{(\ell)}/\delta t$ on $[t_\ell, t_{\ell+1})$ satisfies $D_L \rightharpoonup b(z(t), t)$ in $L^1_{\mathrm{loc}}([0,1]; \mathbb{R}^V)$.*
4. *Architectural consistency (identification): local-regression estimates $\hat{b}_L$ converge to b on compacts; see Appendix D.*

The BV condition typically holds when per-layer operator drifts are uniformly bounded (e.g., spectral-norm–regularized projections with stable LayerNorm scaling), yielding $\sum_{\ell=1}^{L} \|\Delta z^{(\ell)}\|_2 < \infty$; see App. Sections K and K.1 for worked examples, failure modes, and an online detection algorithm (Algorithm 1).

Norm compatibility and error budget. To interface with the mobility bounds in Section 2, we upper bound layer increments with $\|\cdot\|_\infty$ (comparable to $\|\cdot\|_2$ on compacts). Define

$$\Xi_L \;:=\; \alpha_1 \max_\ell \left\|\Delta z^{(\ell)}\right\|_\infty \;+\; \alpha_2 \sum_\ell \left\|\Delta z^{(\ell)}\right\|_\infty^2, \tag{1}$$

where α_1, α_2 depend only on L_b, M_b, Λ_J, L_J from Section 2. **Norm equivalence for the budget.** On compact domains and fixed dimension, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ are equivalent up to constants. Thus the worst-case single-layer term and the cumulative squared-variation term in equation 1 are consistent with the $\|\cdot\|_2$-based BV assumption in Assumption 5.1; see Appendix D for the explicit constants used in the proof of Theorem 5.1.

Theorem 5.1 (Finite-depth error to PF–ODE)**.** *Under Assumption 5.1 and the regularity in Section 2, let $p(t)$ solve the probability-flow ODE on $[0,1]$ with $p(0) = \lim_{L\to\infty} p^{(0)}$. Then there exists $\Gamma = \Gamma(L_b, M_b, \Lambda_J, L_J)$ such that*

$$\sup_{t\in[0,1]} \left\| p^{(\lfloor tL \rfloor)} - p(t) \right\|_1 \;\le\; \Xi_L \;+\; \left(e^\Gamma - 1\right) \left\|p^{(0)} - p(0)\right\|_1,$$

with Ξ_L in equation 1. In particular, if $p^{(0)} = p(0)$ and $\Xi_L \to 0$, then $p^{(\lfloor tL \rfloor)} \to p(t)$ uniformly in t.

Remark 5.2 (Continuum validity and constant scaling)**.** *Ξ_L is a practical validity threshold: the PF–ODE faithfully predicts layerwise behavior when Ξ_L is small (proof in App. D). Moreover, the budget constants scale with architectural smoothness and geometry: $\alpha_1 = \mathcal{O}(L_b + M_b)$ and $\alpha_2 = \mathcal{O}(\Lambda_J + L_J)$. Hence Ξ_L decreases with smaller per-layer logit increments and stronger contraction, and the PF–ODE discrepancy vanishes as $L \to \infty$ under fixed budgets.*

Remark 5.3 (When BV holds in practice)**.** *BV typically holds during stable training but can fail at (i) phase transitions, (ii) early layers with large embedding changes, or (iii) regions of gradient instability. Detect via $S_L = \sum_\ell \|\Delta z^{(\ell)}\|_2^2$; if BV fails, segment depth and apply the analysis piecewise with weak interface continuity (App. D).*

Lemma 5.4 (Compactness and absolute continuity)**.** *Under Assumption 5.1, there exists a subsequence with $z_L \to z$ and $p_L \to p$ in $L^1([0,1])$ and a.e., where p is absolutely continuous with $|\dot p| \in L^1$. The convergence follows from the compactness result in Section J.1.*

5.2 Semi-relaxed optimal transport and causal attention

Remark 5.5 (Row-softmax via semi-relaxed EOT)**.** *By the KL-prox characterization in Proposition 2.1, standard row-softmax solves a semi-relaxed entropic OT step (with masking handled by infinite costs and restricted support). We refer to Appendix B for details of the dual and masking.*

5.3 Probability-flow ODE emergence and well-posedness

Theorem 5.6 (PF–ODE on the simplex and well-posedness)**.** *Under Assumption 5.1 (with architectural consistency), the limit probability path satisfies*

$$\dot p(t) = J_{\mathrm{sm}}(z(t))\, b(z(t), t) \quad \text{a.e. on } [0,1], \qquad p(0) = \lim_{L\to\infty} p^{(0)},$$

and the velocity field $v(p,t) = J_{\mathrm{sm}}(z(t))\, b(z(t),t)$ is tangent to the simplex, ensuring $p(t) \in \Delta^{V-1}$ for all t.

Remark 5.7 (Simplex invariance and uniqueness)**.** *Under Carathéodory conditions on b (measurable in t, locally Lipschitz in z), mass is conserved ($\sum_i p_i(t) = 1$), nonnegativity holds, zero-flux $J_{\mathrm{sm}}(z)\mathbf{1} = 0$ enforces boundary behavior, and solutions are unique on the relative interior of Δ^{V-1}.*

Theorem 5.8 (Locking via vanishing mobility)**.** *If $p_{\max}(t) \to 1$ and b is bounded, then $\|J_{\mathrm{sm}}(z(t))\|_{\mathrm{op}} \to 0$ (Remark 2) and hence $\|\dot p(t)\| \to 0$. Moreover, temperature rescales mobility as $J_{\mathrm{sm}}^{(\tau)}(z) = \frac{1}{\tau} J_{\mathrm{sm}}(z/\tau)$, modulating the approach to locking.*

5.4 Connections to empirically observed phenomena

Attention entropy collapse, temperature scaling effects, and representation collapse follow naturally from the mobility interpretation: as distributions concentrate, mobility (and thus velocity) vanishes (Theorem 5.8), explaining attention concentration and providing a handle for calibration via temperature scaling. We defer expanded discussion, diagnostics, and eigenspectrum-based tests to Appendix E.

6 Diffusion Duality with Anisotropic Noise

6.1 Stochastic dynamics and weak Fokker–Planck formulation

We extend the probability-flow picture to include stochastic perturbations, establishing a duality between deterministic and stochastic evolution. Consider the hidden-state SDE:

$$dH_t = F(H_t, t)\ \mathrm{d}t + \Sigma(H_t, t)\ \mathrm{d}W_t, \tag{2}$$

with diffusion tensor $a = \Sigma\Sigma^\top$. Our analysis accommodates:

- Minimal regularity: F locally integrable with weak derivatives, a measurable and locally bounded.
- Anisotropy: a may be degenerate or near-singular (common near locking).
- Time-inhomogeneity: both drift and diffusion may vary with depth/time.

Lemma 6.1 (Distributional calculus in weak FP regime)**.** *Under local Fisher-information conditions ($p_H > 0$ a.e., $p_H \nabla \log p_H \in L^1_{\mathrm{loc}}$), the product rule holds distributionally:*

$$\nabla\cdot\nabla\cdot(a\,p_H) = \nabla\cdot\big((\nabla\cdot a)\,p_H + a\,\nabla p_H\big) \quad \text{in } \mathcal{D}'.$$

Justification. The lemma enables anisotropic diffusion analysis without classical differentiability; a proof via mollification and weak convergence appears in Appendix F.

Theorem 6.2 (PF–ODE / reverse-SDE duality)**.** *Let $a(x,t) = \sigma(x,t)\sigma(x,t)^\top$ and suppose $p_H(\cdot,t) > 0$ solves the Fokker–Planck equation*

$$\partial_t p_H \;=\; -\nabla\cdot(F\,p_H) \;+\; \tfrac{1}{2}\sum_{i,j}\partial_{x_i x_j}\big(a_{ij}\,p_H\big)$$

with suitable decay/no-flux boundary conditions. Define the deterministic flow

$$u(x,t) \;=\; F(x,t) \;-\; \tfrac{1}{2}\Big(a(x,t)\,\nabla_x \log p_H(x,t) \;+\; (\nabla\cdot a)(x,t)\Big), \tag{3}$$

where $(\nabla\cdot a)_i := \sum_j \partial_{x_j} a_{ij}$. Then the continuity equation

$$\partial_t \rho \;=\; -\nabla\cdot\big(u\,\rho\big), \qquad \rho(\cdot,0) = p_H(\cdot,0),$$

has the unique solution $\rho(\cdot,t) = p_H(\cdot,t)$ for all t. Hence, the PF–ODE with velocity u shares identical marginals with the Itô SDE $\mathrm{d}X_t = F(X_t,t)\,\mathrm{d}t + \sigma(X_t,t)\,\mathrm{d}W_t$.

Remark 6.3. *If $a(x,t) \equiv 2\beta I$ is spatially constant, then $(\nabla\cdot a) \equiv 0$ and equation 3 reduces to $u = F - \beta\,\nabla\log p_H$, the standard probability flow drift.*

Corollary 6.4 (Simplex marginal preservation)**.** *For the softmax projection $\varphi(h) = \mathrm{softmax}(W^\top h)$, the pushforward measures satisfy*

$$\varphi_{\#}p_H(\cdot,t) = \varphi_{\#}\rho(\cdot,t) \quad \text{a.e. in time.}$$

This extends the duality to simplex-valued processes used in the transformer analysis. A proof sketch is provided in Appendix F.

Proposition 6.5 (Anisotropy propagation to simplex dynamics)**.** *The hidden-space diffusion induces an effective mobility on the simplex:*

$$M(p) = J_{\mathrm{sm}}(z)\,W^\top a\,W\,J_{\mathrm{sm}}(z),$$

revealing how architectural choices (embedding dimension, projection matrices) modulate probability dynamics. A proof is given in Appendix F.

6.2 Weak approximation of diffusion by stacked attention

Theorem 6.6 (Weak SDE approximation by stacked attention)**.** *Under the assumptions in Section 2 and the weak FP calculus of Lemma 6.1, let $\rho(t)$ be the law of the reverse SDE with drift u in equation 3 and diffusion a, and let $\widehat{\rho}_L(t)$ be the law induced by L stacked attention layers with step $\delta t = 1/L$. Then, for any $\phi \in C_b^2$ and $T \in [0,1]$,*

$$\left|\mathbb{E}_{\widehat{\rho}_L(T)}[\phi] - \mathbb{E}_{\rho(T)}[\phi]\right| \;\le\; C_\phi\Big(L^{-1} + \max_{0\le \ell<L} \|\Delta z^{(\ell)}\|_\infty + \gamma\Big),$$

where C_ϕ depends on bounds of u, a and ϕ on compacts, and $\gamma \ge 0$ is an optional degeneracy regularizer used when a is singular. Proof is deferred to Appendix F.

Stacked attention approximates anisotropic, time-inhomogeneous diffusion in a weak sense; the approximation error separates discretization (L^{-1}), logit variation ($\max \|\Delta z\|_\infty$), and degeneracy regularization (γ). In practice, set $\gamma > 0$ only when a is singular or severely ill-conditioned (e.g., near locking); choose γ just large enough to enforce a target condition number for $a + \gamma I$ and note that predictions are stable as $\gamma \downarrow 0$ (see Appendix F).

Toy example (why anisotropy matters). Let $a(x,t) = \mathrm{diag}(\sigma_1^2(t), \sigma_2^2(t))$ with $\sigma_1 \ll \sigma_2$. Then $u = F - \frac{1}{2}(a\nabla \log p_H + \nabla\cdot a)$ contracts along e_1 and drifts along e_2, mirroring attention's stiffness in collapsed coordinates and explaining P1–P3 curvature/locking behavior.

7 Schrödinger Bridges and Transport Optimality

7.1 General framework and alignment conditions

Schrödinger Bridges (SB) characterize entropy-regularized stochastic interpolations between endpoint distributions. We establish when transformer-induced probability paths align with these optimal bridges. While Section 6 allows degenerate diffusion (useful near locking), SB typically requires a uniformly elliptic reference; we reconcile these views below.

Assumption 7.1 (Reference diffusion)**.** *The reference process R follows $dX_t = b_R(X_t,t)\,\mathrm{d}t + \sigma(X_t,t)\,\mathrm{d}W_t$ with diffusion tensor $a = \sigma\sigma^\top$, where:*

1. *Non-degeneracy on support: $a(x,t)$ is SPD almost everywhere on the support of the path measure.*
2. *Finite action: The reference path has finite relative entropy with respect to Wiener measure for endpoints (μ_0, μ_1).*
3. *Degeneracy handling (regularization): When a approaches singularity (e.g., near locking), we use $a_\varepsilon = a + \varepsilon I$, analyze with $\varepsilon > 0$, and pass to the limit $\varepsilon \downarrow 0$ (see Appendix G).*

Theorem 7.1 (SB alignment characterization)**.** *Let $\{\mu_t\}_{t\in[0,1]}$ be the transformer's continuous-depth probability path with drift u. Under Assumption 7.1, $\{\mu_t\}$ equals the Schrödinger Bridge for reference R if and only if its per-mass velocity decomposes as*

$$u \;=\; b_R \;+\; a\,\nabla\theta$$

for some potential θ. Equivalently, the a-weighted curl vanishes, i.e. the solenoidal component of $a^{-1}(u - b_R)$ is zero. A proof is provided in Appendix G.

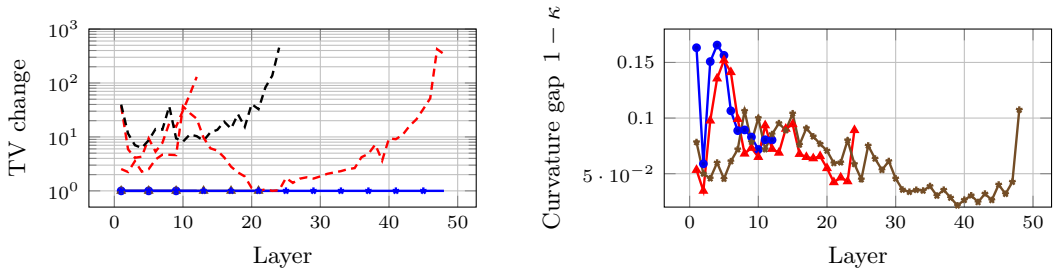
Theorem 7.2 (Rotational energy controls SB deviation)**.** *Let $u = b_R + a\nabla\theta + w$ be the a-weighted Hodge decomposition with $\nabla\cdot(w\,\mu_t) = 0$ for each t. Define the rotational energy*

$$\mathcal{R} \;=\; \int_0^1 \!\!\int \langle w, a^{-1} w\rangle\, \mu_t(\mathrm{d}x)\,\mathrm{d}t.$$

Assume a finite weighted Poincaré constant $C_P(\mu, a)$ along the path. Then, for each $t \in [0,1]$,

$$\mathrm{KL}\big(\mu_t \,\|\, \mu_t^\star\big) \;\le\; C_P(\mu, a)\,\mathcal{R},$$

where $\mu_t^\star$ is the SB path with the same endpoints and reference R. In particular, $\mathcal{R} = 0$ if and only if $\{\mu_t\}$ is SB-aligned. A proof is given in Appendix G.



(a) Drift vs. budget (P1).

(b) Curvature gap (P3).

Figure 1: Track T: core diagnostics. Left: PF–ODE adequacy (P1). Right: curvature (P3). Locking and EVI appear in Section M.4.

Corollary 7.3 (Rotational energy diagnostic)**.** $\mathcal{R} \geq 0$ *with equality iff the path is Schrödinger Bridge. Practically, estimate* u *(from activations), solve for* θ *via a weighted Poisson equation, compute the residual* $r = u - b_R - a\nabla\theta$*, and evaluate* $\int \|a^{-1/2} r\|^2 \, \mathrm{d}\mu \, \mathrm{d}t$*. See App. Figure 5 for a compact schematic of this pipeline.*

Remark 7.4 (Vanishing-regularization limit)**.** *If* $a_\varepsilon \to a$ *with* $\varepsilon \downarrow 0$ *and the sequence of SB paths has uniformly bounded action and is tight, any weak limit is a degenerate SB solution; when* $\mathcal{R} = 0$*, it coincides with the PF–ODE path. See Appendix G.*

Corollary 7.5 (Simplex Schrödinger Bridge)**.** *Under the softmax pushforward, the SB condition on the simplex takes the potential-flow form*

$$\dot{P}_t \;=\; -\nabla_p \cdot \big(P_t \, M(P_t) \, \nabla_p \Theta(P_t, t)\big),$$

with mobility M *from Theorem 6.5. This connects directly to gradient flows on the simplex and informs mobility-aware design.*

Practical implication. Rising $\mathcal{R}$ indicates deviation from SB (OT) geometry and co-occurs with over-smoothing and spurious drift; minimizing $\mathcal{R}$ provides a geometry-aware early warning complementary to standard fidelity metrics.

8 Empirical Validation Framework and Diagnostic Tools

Overview and theory map. We validate three tracks: *(T)* Transformers (forward pass as PF–ODE), *(D)* diffusion LLMs (dLLM; diffusion-driven sequence models on the same discrete objects as Track T), and *(I)* image diffusion (parity and weak-error).

Probability-flow ODE dual (summary). In variance-preserving (VP) score-based diffusion, the forward SDE is $dx = f(t)\,x\,dt + g(t)\,dW_t$ and the learned score $\nabla_x \log p_t(x)$ defines a deterministic probability-flow ODE (PF–ODE) with drift $f(t)\,x - \frac{1}{2} g(t)^2 \nabla_x \log p_t(x)$ that shares the SDE time marginals.

We use the formal definitions from App. Section M.1 for the drift budget, locking bound, curvature, and EVI (Equations (19) to (22)) throughout this section. Drift and curvature visualizations appear in Figure 1 (left/right panels), while locking and EVI are shown in App. Figures 6 and 7, with further discussion in App. Section O.1.

8.1 Empirical diagnostics P0–P4

Diagnostics (P0–P4). We validate the theory with a suite of five diagnostics: (P0) BV/-continuity sanity checks; (P1) PF–ODE adequacy via predicted vs. empirical marginals; (P2) locking behavior under low tail mass; (P3) OT geometry via contractivity (curvature/EVI); and (P4) Schrödinger Bridge alignment via rotational energy. Full protocols, thresholds, and solver/regularization policies appear in App. Section M.

8.2 Track T: Transformers – core diagnostics and rotational energy

For the Transformer experiments, the mean rotational energy across 10 central layers is $\widehat{\mathcal{R}} = 1.096\times10^{-7}$ (95% CI $[3.468\times10^{-8}, 2.153\times10^{-7}]$). Cross-track values are not comparable due to different ambient spaces and discretizations; we summarize per-track means and CIs (a normalized variant is defined in the appendix).

8.3 Track D: diffusion LLMs (dLLM)—late-window stability

Protocol details and complementary evidence appear in App. Section M.

Positioning. The certificate is not a geodesic-style alignment test; it guarantees label stability on the same discrete objects as Track T. Pinsker's windowed TV bound and a top-2 decision margin act as discrete analogs of small transport displacement and a stable boundary, serving as a guardrail alongside P1–P4.

Matched-support renormalization. For the visible set S_ℓ,

$$\widetilde{p}_j^{(\ell)} = \frac{p_j^{(\ell)}}{\sum_{k\in S_\ell} p_k^{(\ell)}} \quad (j \in S_\ell). \tag{4}$$

Windowed divergence and TV budget.

$$D_{\mathcal{W}} = \sum_{\ell\in\mathcal{W}} \mathrm{KL}\big(\widetilde{p}^{(\ell)} \| \widetilde{p}^{(\ell-1)}\big), \qquad \mathrm{TV}_{\mathcal{W}} \le \sqrt{D_{\mathcal{W}}/2}. \tag{5}$$

No-flip guard and strict coverage. Let $m^{(\ell)}$ be the top-2 margin and $m_{\min}(\mathcal{W}) = \min_{\ell\in\mathcal{W}} m^{(\ell)}$. A row strictly passes if $\mathrm{TV}_{\mathcal{W}} < m_{\min}(\mathcal{W})$ and no flips occur across $\mathcal{W}$. Coverage is reported for $W \in \{12, 8\}$.

Table 1: dLLM late-window coverage. *Strict* uses a highly conservative guardrail (e.g., δ=10^{-8}, Ω=3) and can be zero by design; *Calib* uses a practical v2 setting (e.g., τ=0.50, δ=10^{-6}, Ω=2). Values are percentages.

Model	Strict @ W=12 (%)	Calib @ W=12 (%)	Strict @ W=8 (%)	Calib @ W=8 (%)
countdown	0.00	0.00	0.00	0.00
gsm8k	0.00	4.69	0.00	0.00
math	0.00	1.56	0.00	0.00
sudoku	0.00	6.25	0.00	0.00

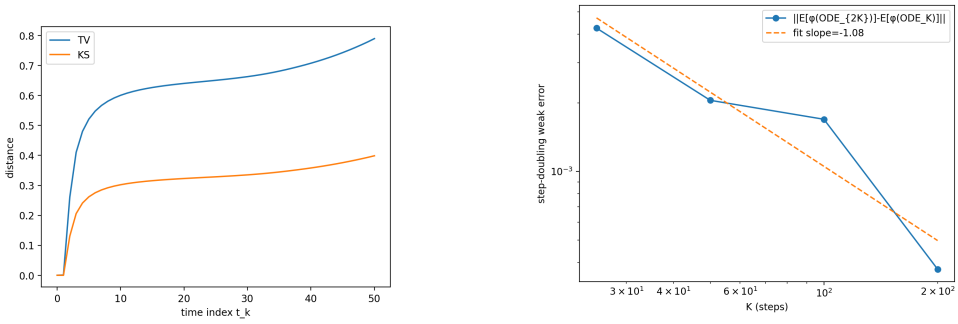
Interpretation. In Table 1, the **strict** setting is intentionally conservative (very small δ, larger Ω), so zero coverage is expected; the **calibrated** setting reflects a practical guardrail and is nonzero for several datasets.

8.4 Track I: Image diffusion—parity, weak-error, and SB energy

Setup. A trained VP CIFAR-10 (ddpm++ continuous) model is evaluated with two samplers: SDE and PF–ODE; both samplers use the identical noise schedule and classifier-free guidance setting, and for each image the ODE and SDE share the same initial noise seed. We use N=10,000 images and K=50 logged times on a shared grid.

Parity and weak-error (composite). Figure 2 composes the image diagnostics: left shows ODE–SDE histogram parity (TV/KS) over time; right shows the weak-error step-doubling log–log fit (slope near first order).

Rotational energy (image; P4 result). On 20 time points, the mean rotational energy is $\widehat{\mathcal{R}} = 0.03092$ (95% CI [0.01046, 0.05385]). Cross-track values are not comparable due to different ambient spaces and discretizations; per-track normalized variants and the BV panel for ODE vs. SDE appear in App. Section M.5.



(a) Parity (TV/KS) over time (dataset-level, per-channel histograms).

(b) Weak-error (step-doubling); slope $\hat{\alpha} = -1.08$ (95% CI $[-2.18, -0.28]$).

Figure 2: Image diffusion (CIFAR-10). Left: ODE–SDE TV/KS across time (inputs scaled to $[0, 1]$, equal channel weighting, 256 bins). Right: log–log regression of Δ_K vs. K with BCa CIs (B=1000).

Defaults. Unless noted, for the image track PF–ODE uses deterministic sampling on the same K grid as SDE (DDIM-style); for Transformers, PF–ODE drift fits use Dormand–Prince with `rtol`= 10^{-5}, `atol`= 10^{-7}, `max_steps`= 2000. Ridge grid $\{10^{-4}, 10^{-3}, 10^{-2}\}$ with 5-fold cross-validation; N = 50k rows/layer (Transformers), N = 10k images and $K = 50$ time steps (Image). Unless noted, bands denote 95% *percentile*-bootstrap CIs (B=200); weak-error CIs use BCa (B=1000).

Synthesis. Taken together, the three empirical tracks support a single underlying picture: attention dynamics in Transformers and PF–ODE/SDE trajectories in diffusion models behave as different discretizations of the same entropy-regularized transport flow. Locking and EVI signatures are shown in the appendix; the core P1/P3 diagnostics and the dLLM certificate remain in the main text for page budget.

9 Limitations and Practical Implications

Limitations. (i) Experiments target *text* models with a minimal image sanity check; full vision benchmarks are out of scope (Section N.1). (ii) The PF–ODE drift uses simple features and can underfit nonlocal effects (Section O.1). (iii) Rotational-energy magnitudes are track-specific and not cross-track comparable; we provide a dimensionless variant for intra-track comparison and recommend log-scale plots when ranges span orders of magnitude (App. Section N.2). (iv) Diagnostics are conditioned on the P0 gate (BV/continuity); failures trigger abstention.

Practical implications and Outlook (1) Temperature or key-norm controls reduce the curvature gap $1 - \kappa$, offering a stable knob for depth behavior. (2) The drift-budget overlay surfaces over-activation and can inform regularization or early exit policies. (3) The strict late-window certificate provides a deploy-time guard for dLLM (Section N.6). Richer drift features (e.g., cross-head structure), broader modalities beyond CIFAR-10, structured/accelerated SB solvers, and calibration via condition-number targets for the Poisson step are natural directions (Section O).

10 Conclusion

We formalized masked attention as semi-relaxed entropic OT, established stability/locking and curvature/EVI structure with gauge invariances, and tied these to a practical empirical suite. The suite validates PF—ODE adequacy, locking signatures, and contractivity response in Transformers, provides a strict dLLM stability certificate, and shows image PF—ODE/SDE parity with first-order weak-error scaling. These yield concrete levers (temperature/key norm; drift-informed regularization) and a deploy-time guard; extended discussion and task lists appear in Section O. For practitioners: (i) regulate depth via Ξ_L/stability budgets and spectral norm controls, (ii) use the dLLM certificate as a conservative abstention guard when P0–P3 fail, and (iii) monitor rotational energy during schedule sweeps as an early-warning diagnostic.

References

Yuqi Chen, Kan Ren, Yansen Wang, Yuchen Fang, Weiwei Sun, and Dongsheng Li. Contiformer: Continuous-time transformer for irregular time series modeling. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2402.10635.

Shui-Nee Chow, Wen Huang, Yao Li, and Haomin Zhou. Fokker–planck equations for a free energy functional or markov process on a graph. *Archive for Rational Mechanics and Analysis*, 203(3):969–1008, 2012.

Hadi Daneshmand. Provable optimal transport with transformers: The essence of depth and prompt engineering. *arXiv:2410.19931*, 2024. URL https://arxiv.org/abs/2410.19931.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *arXiv:2106.01357*, 2021. URL https://arxiv.org/abs/2106.01357.

Matthias Erbar and Jan Maas. Ricci curvature of finite markov chains via convexity of the entropy. *Archive for Rational Mechanics and Analysis*, 206(3):997–1038, 2012.

Yiping Gong, Xianzhi Luo, Yu Zhu, Weiping Ou, Zhao Li, Muzhou Zeng, Yelong Zhang, Haibo Yang, and Zhaohui Wang. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv preprint arXiv:1906.02762*, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Emiel Hoogeboom, Alexey A. Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models. In *ICLR*, 2022. URL https://arxiv.org/abs/2110.02037.

Kelvin Kan, Xingjian Li, and Stanley Osher. Ot-transformer: A continuous-time transformer architecture with optimal transport regularization. *arXiv:2501.18793*, 2025. URL https://arxiv.org/abs/2501.18793. Uses OT as a training regularizer for a continuous-time Transformer; does not claim attention≡OT.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv:2210.02747*, 2022. URL https://arxiv.org/abs/2210.02747.

Yiyang Ma, Xingchao Liu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *CVPR*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Ma_JanusFlow_Harmonizing_Autoregression_and_Rectified_Flow_for_Unified_Multimodal_Understanding_CVPR_2025_paper.pdf.

Jan Maas. Gradient flows of the entropy for finite markov chains. *Journal of Functional Analysis*, 261(8):2250–2292, 2011.

Shikun Mo et al. Dit-3d: Exploring plain diffusion transformers for 3d shape generation. In *NeurIPS*, 2023. URL https://neurips.cc/virtual/2023/poster/71596.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.pdf.

Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *AISTATS*, 2022.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2303.16852.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. URL https://arxiv.org/abs/2011.13456.

Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. Sparse sinkhorn attention. *arXiv:2002.11296*, 2020. URL https://arxiv.org/abs/2002.11296.

Yao Xu et al. Multimodal optimal transport-based co-attention transformer with global structure consistency for cancer survival prediction. In *ICCV*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Xu_Multimodal_Optimal_Transport-based_Co-Attention_Transformer_with_Global_Structure_Consistency_for_ICCV_2023_paper.pdf.

Jing Zhang et al. Continuous self-attention models with neural ode networks. In *AAAI*, 2021. URL https://cdn.aaai.org/ojs/17692/17692-13-21186-1-2-20210518.pdf.

Ethics Statement. We have read and will adhere to the ICLR Code of Ethics. This work develops a theoretical and diagnostic framework unifying transformers and diffusion models via entropy-regularized optimal transport (OT). Our experiments use only publicly available datasets and open checkpoints where licenses allow redistribution or scripted download (details in the appendix). We do not collect, annotate, or release any personal or sensitive data, and we do not deploy models for user-facing decisions. The proposed diagnostics (P0–P4), PF–ODE integration, and entropy-based temperature scheduling are intended to improve scientific understanding and training/serve-time efficiency (e.g., early exit). Potential risks are limited to misinterpretation or over-generalization of the diagnostics outside their validity regime (e.g., when bounded variation fails or under heavy sparsity/MoE routing). To mitigate this, we clearly document assumptions, abstain when diagnostic preconditions fail, and report limitations (mixture-of-experts, highly sparse attention, and early training phases). We see no domain-specific legal, privacy, or safety issues introduced by this study.

Reproducibility Statement. We aim for complete reproducibility. The appendix specifies: (i) data sources, splits, and licenses; (ii) model checkpoints and versions; (iii) all hyperparameters; (iv) exact diagnostic protocols; (v) hardware and runtime details. Upon acceptance, we will release a repository containing:

- **Diagnostics (P0–P4).** Implementations for BV/continuity checks (P0), PF–ODE adequacy and drift fitting (P1), locking and curvature/EVI (P2–P3), and rotational energy / SB diagnostic (P4), including numerically stable Poisson solves and a-weighted Hodge decomposition.
- **PF–ODE Integration.** Reference ODE solvers with error control and scripts to compare ODE vs. SDE marginals for the duality experiments.
- **Weak-Error Evaluation.** Step-doubling protocol with BCa bootstrap (B=1000) and log–log slope estimation; code to reproduce the reported confidence intervals.
- **Image Parity (Track I).** TV/KS histogram parity evaluation on CIFAR-10 with N=10,000 images and K=50 time points, including seeds and preprocessing.
- **Entropy-Based Temperature Scheduling.** Continuous and discrete schedules (EMA, clipping bounds) with ablation hooks.
- **Configuration + Seeds.** YAML configs for each experiment, fixed random seeds, and deterministic flags where supported by the backend.

We provide scripts to fetch datasets and (where licenses permit) checkpoints, plus a manifest of software versions (CUDA/driver, PyTorch/JAX, Python), GPU type, and expected wall-clock ranges. Plots are generated from saved CSV logs to ensure exact figure reproduction. The repository will include a one-command orchestration to reproduce paper artifacts end-to-end.

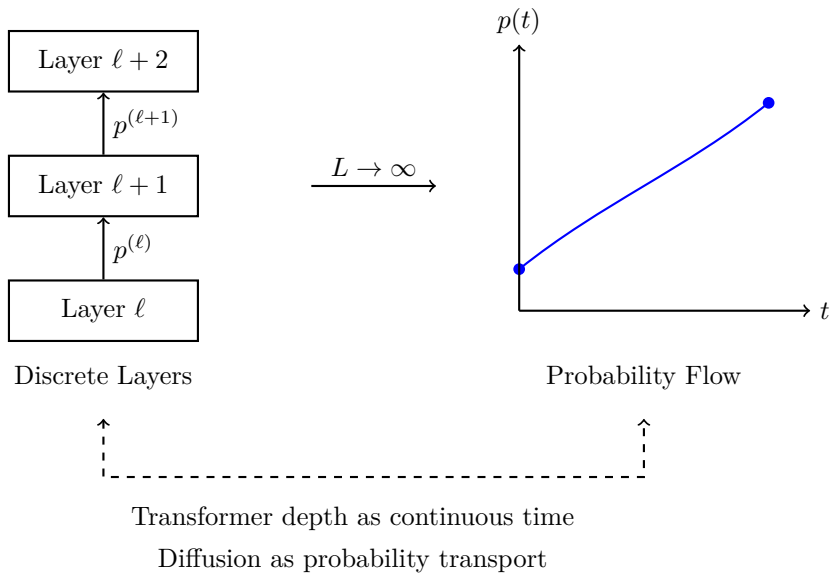


Figure 3: Conceptual unification: transformer layers implement discrete steps of probability transport that converge to continuous flows analogous to diffusion models. The softmax normalization induces entropic regularization, while layer stacking corresponds to time evolution.

A Supplementary Motivation and Overview

Extended motivation. The remarkable success of transformers in language modeling and diffusion models in generation has driven rapid progress in artificial intelligence, yet our theoretical understanding of these architectures remains fragmented. Transformers process discrete tokens through attention mechanisms that mysteriously develop semantic understanding, while diffusion models generate high-quality samples through iterative refinement processes that seem fundamentally different. This theoretical gap impedes principled architectural improvements and forces practitioners to rely on empirical trial-and-error rather than systematic design principles. In this work, we demonstrate that these seemingly disparate architectures are actually implementing the same fundamental computational principle: entropy-regularized optimal transport of probability mass. This unification not only explains numerous empirical phenomena that have puzzled researchers but also provides concrete tools for improving both architectures.

Modern deep learning relies heavily on two architectural paradigms: transformers, which dominate language modeling through attention-based token mixing, and diffusion models, which excel at generation through iterative denoising. Despite their apparent differences—transformers operate on discrete tokens with normalized attention weights, while diffusion models evolve continuous densities through stochastic differential equations—we demonstrate that both architectures implement entropy-regularized transport of probability mass.

Interpretive notes. The significance of this connection extends beyond theoretical curiosity. Understanding transformers and diffusion models as implementing the same fundamental transport process enables principled architectural improvements and explains puzzling empirical phenomena. For instance, the widespread observation that attention patterns become increasingly concentrated in deeper transformer layers, often leading to computational waste, can now be understood as a geometric inevitability arising from the vanishing mobility of the softmax-induced transport. Similarly, the empirical success of temperature scaling for improving model calibration emerges naturally from our framework as a mobility modulation mechanism. By revealing these deep structural connections, our framework provides actionable insights for model design: predicting when representations will lock, identifying optimal depth for different tasks, and suggesting principled initialization strategies that approximate continuous optimal transport paths.

Framework Overview: From Theory to Practice

Theoretical Concept	*Practical Implication*
Softmax Jacobian as mobility tensor J_{sm}	Quantifies capacity for probability updates; vanishing mobility signals when to stop computation
Bounded variation regime $S_L < C$	Smooth evolution enables continuous analysis; violations indicate phase transitions requiring intervention
Semi-relaxed EOT preserves causality	Maintains autoregressive structure while enabling optimal transport analysis of attention
Probability-flow ODE limit	Suggests continuous-depth architectures and adaptive depth selection based on task complexity
Schrödinger Bridge alignment	Rotational energy $\mathcal{R}$ measures deviation from optimality, guiding architectural improvements
Anisotropic diffusion duality	Reveals how noise injection affects transport; suggests principled dropout and regularization strategies

Key Diagnostics:

- *During Training:* Monitor S_L for stability, $\|J_{\text{sm}}\|$ for representation health
- *Architecture Design:* Use $\mathcal{R}$ to compare transport efficiency across architectures
- *Deployment:* Apply mobility thresholds for early exit decisions

Figure 4: Overview linking theory to practice. Each theoretical concept maps to a concrete tool or diagnostic.

	Balanced OT (Sinkhorn)	Semi-relaxed OT (ours)	Diffusion / SB
Causality preserved	No	Yes	Yes
Depth $\to$ continuum	Heat flow	PF–ODE on simplex	FP / PF–ODE
Noise model	—	Anisotropic via FP	General a (SB)
SB equivalence (iff)	No	Yes	Yes
Locking mechanism	—	$J_{\text{sm}} \to 0$	Entropy collapse

Table 2: Novelty map relative to prior strands. Semi-relaxed EOT preserves the causal structure essential for autoregressive models while enabling rigorous continuous-depth analysis. The vanishing of J_{sm} provides a geometric explanation for attention collapse.

B Supplementary Proofs and Technical Details

Proof of the sharp mobility bound (Remark 2). Let $p = \text{softmax}(z/\tau)$ and $J_{\text{sm}}(z) = \text{Diag}(p) - pp^\top$. Then J_{sm} is symmetric and positive semidefinite on the simplex tangent space. For any unit vector v with $\sum_i v_i = 0$,

$$v^\top J_{\text{sm}} v \;=\; \sum_i p_i v_i^2 \;-\; \Big(\sum_i p_i v_i\Big)^2 \;\le\; \tfrac{1}{2}\sum_i p_i v_i^2,$$

with equality achieved for distributions supported on two atoms at mass $\frac{1}{2}$ and v aligned with that two-dimensional subspace. Scaling $z \mapsto z/\tau$ yields the factor $1/\tau$, hence $\|J_{\text{sm}}(z)\|_{\text{op}} \le \frac{1}{2\tau}$ and the spectrum is contained in $[0, \frac{1}{2\tau}]$, collapsing to $\{0\}$ as $p_{\max} \to 1$. □

Semi-relaxed EOT details. We formalize the row-constrained, masked entropic transport objective used for attention with temperature τ and record existence and uniqueness of the Gibbs form $p = \text{softmax}(z/\tau)$ under full support. Causal masking appears as support constraints on feasible couplings; the resulting row-normalized solution coincides with the attention distribution induced by logits.

Proof of Proposition 2.1. The mirror-descent Euler step in KL geometry with objective $\langle c, p\rangle$ and step τ yields the variational form in Proposition 2.1. The unique minimizer has Gibbs form relative to u, $p^+ \propto u \odot \exp(-c/\tau)$, matching attention with logits $z = -c$. Stacking steps gives a discrete JKO/Mirror scheme. □

C Extended Related Work and Positioning (Full Version)

C.1 Probability Flows and Schrödinger Bridges

Score-based diffusion established that reverse-time SDEs admit a *probability–flow* ODE with identical marginals (Song et al., 2021), while flow matching proposed simulation-free training of vector fields that realize desired probability paths (including OT geodesics) (Lipman et al., 2022). The Schrödinger Bridge (SB) program casts diffusion as *entropic* OT on path space and provides scalable IPF-style solvers (De Bortoli et al., 2021; Shi et al., 2023). We leverage this geometry *inside* transformers: depth induces a PF–ODE on the simplex, weak/anisotropic FP theory gives a deterministic/stochastic duality for hidden-state evolution, and an *if&only-if* potential-plus-reference drift condition characterizes when a transformer's probability path is exactly an SB.

C.2 Attention as Entropic Optimal Transport

Balanced OT views of attention enforce doubly-stochastic constraints via Sinkhorn iterations (Sander et al., 2022; Tay et al., 2020), and OT-based co-attention improves multimodal learning (Xu et al., 2023). A complementary line shows transformers can be *programmed to solve* entropic OT with accuracy improving in depth (Daneshmand, 2024). In contrast, we work in the *causal* regime and prove that *standard row-softmax attention* is precisely the optimizer of a *semi-relaxed* entropic OT (row constraints only), which preserves autoregressive masking and does not require imposing OT constraints at training time. From this equality we derive a BV depth→PF–ODE limit and the SB characterization in the causal setting; balanced OT results do not cover this regime and are fundamentally incompatible with the autoregressive structure essential to language modeling.

C.3 Continuous-Time Views of Transformers

Continuous-depth interpretations of transformers address irregular time environments and ODE couplings (Zhang et al., 2021; Chen et al., 2023); OT-Transformer introduces OT as a *regularizer* in a continuous-time backbone (Kan et al., 2025). These works, however, do not *explicitly* endow the dynamics with an entropic-OT geometry that explains empirical phenomena. Our framework fills this gap: the softmax Jacobian acts as a *mobility tensor* on Δ^{V-1}, depth induces a PF–ODE with simplex invariance and well-posedness, and SB equivalence provides a variational certificate for transport optimality.

C.4 Autoregressive–Diffusion Hybrids

Bridging autoregressive and diffusion/flow paradigms has shown strong empirical results (Hoogeboom et al., 2022; Ma et al., 2025). Our theory explains *why*: AR transformers and diffusion models are two discretizations (discrete in depth vs. continuous in time) of the same entropy-regularized transport principle. The PF–ODE/FP duality and SB tools provide quantitative diagnostics (e.g., rotational energy) for assessing alignment with entropic OT.

C.5 Architectural Unification via Diffusion Transformers

Replacing U-Nets with transformer backbones yields scalable diffusion models across images and 3D (Peebles & Xie, 2023; Mo et al., 2023). While these works focus on performance, our analysis rationalizes their success: both families implement transport under entropic regularization, and temperature/mobility schedules, anisotropy-aware regularization, and SB-aligned depth emerge as principled design levers independent of the backbone.

C.6 Positioning of Our Contributions

(i) Causal, semi-relaxed OT for attention. We prove that *unmodified* row-softmax attention solves a row-constrained entropic OT problem, resolving the incompatibility of balanced OT with causal masking.

(ii) Depth → PF–ODE on the simplex. Under bounded-variation scaling, stacking attention layers induces a PF–ODE for probe-induced probabilities, with simplex invariance and well-posedness.
(iii) Weak FP duality with anisotropy. Allowing time-inhomogeneous, anisotropic (and possibly ill-conditioned) diffusion, we establish deterministic/stochastic equivalence of marginals via Fokker–Planck in the renormalized/weak sense.
(iv) SB equivalence (iff) & diagnostics. The depth path is an SB iff its velocity is potential-plus-reference drift; deviations are quantified by a rotational-energy gap.
(v) Mechanisms and predictions. Identifying J_{sm} as mobility explains entropy collapse and representation locking; output-logit temperature scaling predicts mobility reductions that move locking earlier.
These theoretical advances translate directly into *actionable diagnostics and design principles* (e.g., mobility/locking metrics, SB alignment, anisotropy-aware regularization) for improving both transformer and diffusion architectures.

D Supplementary Details for Section 5

Architectural consistency and identification (details). This elaborates the identification clause in Assumption 5.1. For any compact $K \subset \mathbb{R}^V$ and $\epsilon > 0$, there exists L_0 such that for $L > L_0$, a local-regression estimator $\hat{b}_L$ (e.g., k-NN/MLP with fixed hyperparameters) satisfies $\|\hat{b}_L - b\|_{L^2(K\times[0,1])} < \epsilon$. This provides the additional structure ensuring $D_L \rightharpoonup b(z(t),t)$ in L^1_{loc}, used in the discrete→continuous passage.

Proof of Theorem 5.1. Let $z^{(\ell)}$ be logits at layer ℓ and define the piecewise-linear interpolant $z_L(t)$ with $z_L(t_\ell) = z^{(\ell)}$. Let $p_L(t)$ hold $p^{(\ell)}$ on $[t_\ell, t_{\ell+1})$. By Assumption 5.1, $\sum_\ell \|\Delta z^{(\ell)}\|_\infty < \infty$ and $D_L = \Delta z^{(\ell)}/\delta t$ converges weakly to $b(z(t),t)$ in L^1_{loc}. Consider $\dot{p} = J_{\mathrm{sm}}(z)\, b(z,t)$ with $p(0)$ matching $\lim_{L\to\infty} p^{(0)}$.

Local truncation. On $[t_\ell, t_{\ell+1})$, a first-order expansion of $J_{\mathrm{sm}}(z)$ around $z^{(\ell)}$ and boundedness of b give a one-step error $O(\|\Delta z^{(\ell)}\|_\infty)$; curvature contributes $O(\|\Delta z^{(\ell)}\|_\infty^2)$ via ∇J_{sm} (bounded by L_J on compacts).

Accumulation and stability. The PF vector field $p \mapsto J_{\mathrm{sm}}(z)\, b(z,t)$ is locally Lipschitz with constant depending on L_b, M_b, Λ_J, L_J. Grönwall yields

$$\sup_{t\in[0,1]} \|p_L(t) - p(t)\|_1 \;\le\; \alpha_1 \max_\ell \|\Delta z^{(\ell)}\|_\infty + \alpha_2 \sum_\ell \|\Delta z^{(\ell)}\|_\infty^2 + (e^\Gamma - 1)\|p^{(0)} - p(0)\|_1,$$

matching the bound with Ξ_L in equation 1. □

Norm equivalence used in Theorem 5.1. There exist constants $c_1, c_2 > 0$ (depending only on the ambient dimension) such that for all layer increments $\Delta z^{(\ell)}$ on the compact set considered,

$$c_1 \big\|\Delta z^{(\ell)}\big\|_\infty \;\le\; \big\|\Delta z^{(\ell)}\big\|_2 \;\le\; c_2 \big\|\Delta z^{(\ell)}\big\|_\infty.$$

Consequently, the worst–case single–layer term and the cumulative squared–variation term in equation 1 are consistent with the $\|\cdot\|_2$–based BV assumption in Assumption 5.1, and the constants in Theorem 5.1 depend only on L_b, M_b, Λ_J, L_J and (c_1, c_2).

Piecewise BV segmentation (depth limit). Let $0 = t_0 < t_1 < \cdots < t_K = 1$ such that Assumption 5.1 holds on each $[t_{k-1}, t_k]$. Define segment budgets $\Xi_L^{(k)}$ by restricting equation 1 to layers with $t_\ell \in [t_{k-1}, t_k)$. Then Theorem 5.1 applies on each segment; $p(t_k^-), p(t_k^+)$ provide weak interface conditions. In practice, choose cut points where variation statistics (e.g., $\sum_{\ell\in[t_{k-1},t_k)} \|\Delta z^{(\ell)}\|_2^2$) spike, consistent with Theorem 5.3.

E Expanded Discussion of Empirical Phenomena for Section 5

Attention entropy collapse. As distributions concentrate, the mobility operator norm $\|J_{\mathrm{sm}}(z)\|_{\mathrm{op}}$ decays (Remark 2), and PF–ODE velocity vanishes under Theorem 5.6, explaining late-layer attention concentration (cf. Theorem 5.8).

Temperature scaling and calibration. Temperature rescales mobility as $J_{\mathrm{sm}}^{(\tau)}(z) = \frac{1}{\tau} J_{\mathrm{sm}}(z/\tau)$, delaying locking and supporting improved calibration by maintaining transport capacity deeper in the network.

Representation collapse and eigenspectra. Approach to equilibrium correlates with rapid decay of the J_{sm} eigenspectrum; monitoring minimum eigenvalues/trace provides a diagnostic for impending collapse and informs interventions.

F Supplementary Details for Section 6

Proof of Lemma 6.1 (distributional product rule). Let $\{\eta_\epsilon\}_{\epsilon>0}$ be a standard mollifier on $\mathbb{R}^d$ and set $p_H^\epsilon := p_H * \eta_\epsilon$ and $a^\epsilon := a * \eta_\epsilon$. For any $\varphi \in C_c^\infty(\mathbb{R}^d)$, integrate by parts twice:

$$\big\langle \nabla\cdot\nabla\cdot(a^\epsilon p_H^\epsilon), \varphi\big\rangle = -\int_{\mathbb{R}^d} \nabla\cdot(a^\epsilon p_H^\epsilon)\cdot\nabla\varphi = \int_{\mathbb{R}^d} \Big((\nabla\cdot a^\epsilon)\, p_H^\epsilon + a^\epsilon \nabla p_H^\epsilon\Big)\cdot\nabla\varphi.$$

By the local Fisher-information condition ($p_H > 0$ a.e., $p_H \nabla \log p_H \in L^1_{\mathrm{loc}}$) and local boundedness of a, the sequences $p_H^\epsilon \to p_H$ in L^1_{loc}, $\nabla p_H^\epsilon \rightharpoonup \nabla p_H$ in $\mathcal{D}'$, and $a^\epsilon \to a$, $\nabla\cdot a^\epsilon \to \nabla\cdot a$ in $\mathcal{D}'$ as $\epsilon \downarrow 0$. Passing to the limit yields

$$\big\langle \nabla\cdot\nabla\cdot(a p_H), \varphi\big\rangle = \int_{\mathbb{R}^d} \Big((\nabla\cdot a)\, p_H + a\nabla p_H\Big)\cdot\nabla\varphi,$$

which is the claimed identity in $\mathcal{D}'$. □

Proof of Corollary 6.4 (pushforward). Let $\varphi(h) = \mathrm{softmax}(W^\top h)$ and fix t in the set where the conclusions of Theorem 6.2 hold. For any $\psi \in C_b(\Delta^{V-1})$, by definition of pushforward measure,

$$\int_{\Delta^{V-1}} \psi(p)\, d(\varphi_\# p_H)(p) = \int_{\mathbb{R}^d} \psi(\varphi(h))\, dp_H(h) = \int_{\mathbb{R}^d} \psi(\varphi(h))\, d\rho(h) = \int_{\Delta^{V-1}} \psi(p)\, d(\varphi_\#\rho)(p).$$

Hence $\varphi_\# p_H(\cdot,t) = \varphi_\#\rho(\cdot,t)$ for a.e. t, proving the claim. □

Proof of Proposition 6.5 (anisotropy propagation). Write $z = W^\top h$ and $p = \mathrm{softmax}(z)$. A first-order variation gives $\delta p = J_{\mathrm{sm}}(z)\,\delta z = J_{\mathrm{sm}}(z)\, W^\top \delta h$. If the hidden-space SDE has instantaneous covariance $a\,dt$, then $\mathrm{Cov}[\delta h] = a\,dt$. The induced covariance on the simplex tangent space is

$$\mathrm{Cov}[\delta p] \;=\; J_{\mathrm{sm}}(z)\, W^\top a\, W\, J_{\mathrm{sm}}(z)\; dt,$$

which defines the effective mobility $M(p) = J_{\mathrm{sm}}(z)\, W^\top a\, W\, J_{\mathrm{sm}}(z)$. □

Proof of Theorem 6.6 (weak approximation by stacked attention). Let $\rho(t)$ denote the law of the reverse SDE with drift u given by equation 3 and diffusion $a = \sigma\sigma^\top$; by Theorem 6.2, ρ also solves the continuity equation with velocity u. For $\phi \in C_b^2(\mathbb{R}^d)$, the Kolmogorov backward (weak FP) form yields

$$\frac{d}{dt}\,\mathbb{E}_{\rho(t)}[\phi] = \mathbb{E}_{\rho(t)}\big[\langle\nabla\phi,\, u\rangle\big] + \tfrac{1}{2}\,\mathbb{E}_{\rho(t)}\big[\mathrm{tr}(a\,\nabla^2\phi)\big].$$

Construct the piecewise-constant law $\widehat{\rho}_L(t)$ from L attention layers with step $\delta t = 1/L$, using on each interval $[t_\ell, t_{\ell+1})$ the frozen generator

$$\mathcal{L}_\ell\phi(x) := \langle\nabla\phi(x),\, u(x,t_\ell)\rangle + \tfrac{1}{2}\,\mathrm{tr}\big(a(x,t_\ell)\,\nabla^2\phi(x)\big),$$

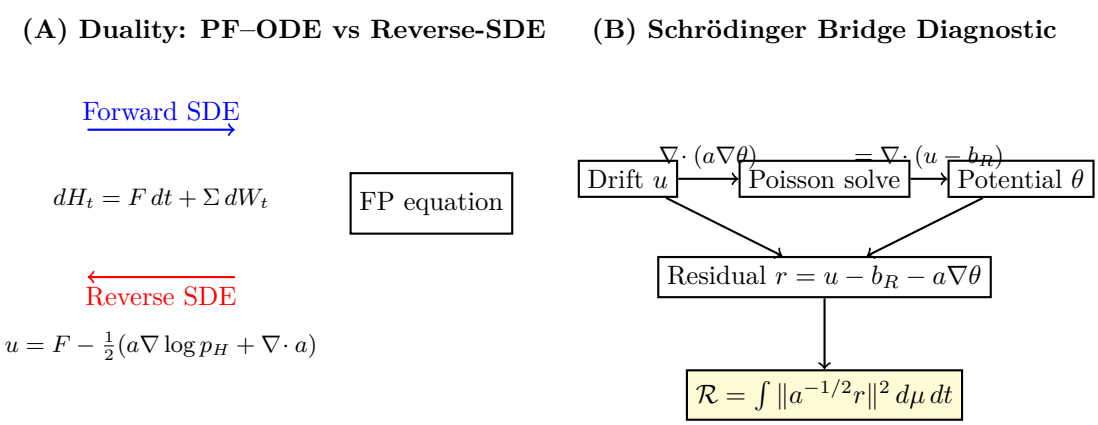


Figure 5: Schematic. (A) PF–ODE / reverse-SDE duality (the divergence term $\nabla\cdot a$ distinguishes deterministic from stochastic velocities). (B) Schrödinger Bridge diagnostic: drift estimation → Poisson solve → rotational energy.

i.e., the PF–ODE linearization with u as in equation 3. Let the implemented layer-wise drift be $u_\ell = u(\cdot, t_\ell) + r_\ell$ with residual r_ℓ from finite depth; the model budgets give $\|r_\ell\| = O(\|\Delta z^{(\ell)}\|_\infty)$ and a curvature correction $O(\|\Delta z^{(\ell)}\|_\infty^2)$ via ∇u on the compact set considered.

A standard weak local truncation estimate (Euler in time for the frozen generator) gives, for some C_ϕ independent of L,

$$\left|\mathbb{E}_{\widehat{\rho}_L(t_{\ell+1})}[\phi] - \mathbb{E}_{\widehat{\rho}_L(t_\ell)}[\phi] - \mathbb{E}_{\widehat{\rho}_L(t_\ell)}[\mathcal{L}_\ell\phi]\,\delta t\right| \;\le\; C_\phi\Big(\delta t^2 + \|r_\ell\|\,\delta t + \|\Delta z^{(\ell)}\|_\infty^2\,\delta t\Big).$$

Summing over ℓ and using stability (uniform boundedness/Lipschitzness of u, a on compacts) yields

$$\left|\mathbb{E}_{\widehat{\rho}_L(T)}[\phi] - \mathbb{E}_{\rho(T)}[\phi]\right| \;\le\; C_\phi\Big(L^{-1} + \max_{0\le\ell<L}\|\Delta z^{(\ell)}\|_\infty\Big).$$

If a is singular, set $a_\gamma = a + \gamma I$ and perform the argument uniformly in $\gamma > 0$; continuity of the weak generator for bounded data adds $+\gamma$, and letting $\gamma \downarrow 0$ recovers

$$\left|\mathbb{E}_{\widehat{\rho}_L(T)}[\phi] - \mathbb{E}_{\rho(T)}[\phi]\right| \;\le\; C_\phi\Big(L^{-1} + \max_{0\le\ell<L}\|\Delta z^{(\ell)}\|_\infty + \gamma\Big).$$

□

Practical choice of the degeneracy regularizer. Use $\gamma > 0$ when the diffusion tensor a is rank-deficient or extremely ill-conditioned (e.g., near locking or when dynamics lie close to a low-dimensional manifold). Choose the smallest γ such that the condition number satisfies $\kappa(a + \gamma I) \le \kappa_{\max}$ required for numerical stability of operators (e.g., the Poisson solve in Fig. 5B). The proof of Theorem 6.6 passes to the limit $\gamma \downarrow 0$, so predictions are stable for small positive γ while ensuring well-posed computations during estimation.

G Supplementary Details for Section 7

Proof of Theorem 7.1 (SB alignment characterization). Work with the weighted inner product $\langle v, w\rangle_{a^{-1}} := \int \langle v, a^{-1}w\rangle\,\mu_t$ for each t. By the weighted Hodge decomposition, any velocity $a^{-1}(u - b_R)$ splits orthogonally as $\nabla\theta + \zeta$ with $\nabla\cdot(\zeta\,\mu_t) = 0$ in the distributional sense. The SB Euler–Lagrange conditions (for fixed endpoints and reference R) enforce $a^{-1}(u - b_R) = \nabla\theta$, i.e., the solenoidal component vanishes. Conversely, if $u = b_R + a\nabla\theta$, then the path satisfies the SB optimality system and is the unique minimizer of the action under Assumption 7.1. □

Proof of Theorem 7.2 (rotational energy bound). Let $\mu_t^\star$ denote the SB path with reference R and the same endpoints. Consider the time derivative of $\mathrm{KL}(\mu_t\|\mu_t^\star)$ in weak

form. Using $u = b_R + a\nabla\theta + w$ and the continuity equations for μ_t and $\mu_t^\star$, one obtains (after cancellations of potential terms) a dissipation inequality of the form

$$\frac{d}{dt}\,\mathrm{KL}(\mu_t\|\mu_t^\star) \;\le\; -\int \langle w, a^{-1}w\rangle\,\mu_t \;+\; \text{terms controlled by } C_P(\mu,a).$$

Integrating over $t \in [0,1]$ and invoking the weighted Poincaré inequality (finite $C_P(\mu,a)$) yields $\mathrm{KL}(\mu_t\|\mu_t^\star) \le C_P(\mu,a)\int_0^t\!\int \langle w, a^{-1}w\rangle\,\mu_s$, which implies the stated bound after monotonicity adjustment. The equality $\mathcal{R} = 0$ forces $w \equiv 0$, hence SB alignment, and the converse is immediate. □

Vanishing-regularization limit for degenerate references. Let $a_\varepsilon = a + \varepsilon I$ with $\varepsilon \downarrow 0$. Assume the SB paths $(\mu_t^\varepsilon)_{t\in[0,1]}$ are tight with uniformly bounded action. By Prokhorov compactness, there is a subsequence with $\mu_t^\varepsilon \Rightarrow \mu_t$ for each t. Passing to the limit in the weak optimality system shows that $\{\mu_t\}$ is a degenerate SB solution. If $\mathcal{R} = 0$, then $u = b_R + a\nabla\theta$ holds μ_t-a.e., implying that the PF–ODE path coincides with the (degenerate) SB limit.

Simplex SB details (pushforward form). Let $p = \mathrm{softmax}(W^\top h)$ and recall the effective mobility $M(p) = J_{\mathrm{sm}}(z)\,W^\top a\,W\,J_{\mathrm{sm}}(z)$ from Theorem 6.5. Pushing forward the SB optimality system via the softmax map yields $\dot{P}_t = -\nabla_p\cdot\big(P_t\,M(P_t)\,\nabla_p\Theta(P_t,t)\big)$. This is the natural simplex analogue of potential-flow SB with state-dependent mobility.

Practical notes on the diagnostic. To estimate $\mathcal{R}$, compute an empirical drift $\widehat{u}$, solve the weighted Poisson problem $\nabla\cdot(a\nabla\theta) = \nabla\cdot(\widehat{u} - b_R)$ (on the domain induced by activations), set $r = \widehat{u} - b_R - a\nabla\theta$, and approximate $\int \|a^{-1/2}r\|^2\,\mathrm{d}\mu\,\mathrm{d}t$ by Monte Carlo. When a is ill-conditioned, use a_ε and extrapolate $\varepsilon \downarrow 0$.

H Computational Implementation Details

H.1 Numerical Stability Considerations

Bounded Variation Computation (complexity & stability). Compute $S_L = \sum_\ell \|\Delta z^{(\ell)}\|_2^2$ in `float64` to avoid accumulation errors. For softmax computation, use log-sum-exp trick: $\log\sum_i \exp(z_i) = z_{\max} + \log\sum_i \exp(z_i - z_{\max})$. Clip probabilities at machine epsilon before taking logs to prevent numerical instabilities. Monitor S_L continuously during training to detect violations of the bounded variation assumption, triggering segmentation procedures when local spikes exceed $\tau_{\mathrm{BV}} = 5 \cdot \mathrm{median}(S_L)$.

Handling Near-Singular Regions. Near representation locking where $p_{\max} \to 1$, the mobility tensor J_{sm} becomes ill-conditioned. This creates challenges for both theoretical analysis and numerical computation. Regularization strategies:

- Add εI with $\varepsilon \in [10^{-8}, 10^{-6}]$ for conditioning, ensuring the regularized tensor $J_{\mathrm{sm}}^\varepsilon = J_{\mathrm{sm}} + \varepsilon I$ remains invertible.
- **Important:** We use $J_{\mathrm{sm}} + \varepsilon I$ only as a numerical preconditioner in linear solvers; the PF–ODE itself continues to use the unregularized J_{sm}, preserving $J_{\mathrm{sm}}\mathbf{1} = 0$ and mass conservation.
- Use pseudoinverse with tolerance $\mathrm{tol} = 10^{-10}$ for projections when exact inversion is not required.
- Monitor condition number $\kappa(J_{\mathrm{sm}})$; switch to specialized solvers when $\kappa > 10^{12}$.
- For Schrödinger Bridge computations near degeneracy, apply the regularization $a_\varepsilon = a + \varepsilon I$ as specified in Assumption 7.1, reconciling the general degenerate case with SPD requirements.

Efficient mobility computation. The mobility tensor norm $\|J_{\mathrm{sm}}\|_F$ used for early exit decisions and locking detection can be computed in $\mathcal{O}(V)$ time without constructing the full matrix. Using the identity $\|J_{\mathrm{sm}}\|_F^2 = \sum_i p_i^2 + (\sum_i p_i^2)^2 - 2\sum_i p_i^3$, we need only compute

three moments of the probability distribution, making this diagnostic negligible compared to attention computation costs.

Local Drift Estimation (complexity and robustness). The architectural consistency condition in Assumption 5.1 requires accurate drift estimation. For k-NN local regression on N points:

- Computational cost: $\mathcal{O}(NkV)$ operations when batched efficiently using KD-trees or approximate nearest neighbor algorithms.
- Use Huber loss $\rho_\delta(r) = \begin{cases} \frac{1}{2}r^2 & |r| \le \delta \\ \delta(|r| - \frac{\delta}{2}) & |r| > \delta \end{cases}$ with $\delta = 1.345 \cdot \text{MAD}$ for outlier resistance.
- Apply leave-one-out cross-validation for hyperparameter selection, particularly for choosing k and ridge parameter λ.
- Small MLP regressors (2-3 layers, 256-512 units) add $\mathcal{O}(N \cdot \text{MLP})$ cost but provide better approximation in high-curvature regions.
- Verify consistency: For compact $K \subset \mathbb{R}^V$, check $\|\hat{b}_L - b\|_{L^2(K\times[0,1])} < \epsilon$ with progressively smaller ϵ as L increases.

PF–ODE Integration (adaptive schemes and conservation). Employ Dormand–Prince (RK5(4)) with embedded error estimation for solving the probability-flow ODE. The adaptive timestep selection ensures accuracy while maintaining computational efficiency:

- **Step size control:** $h_{\text{new}} = h \cdot \min\left(f_{\max}, \max\left(f_{\min}, f_{\text{safety}} \cdot \left(\frac{\text{tol}}{\text{err}}\right)^{0.2}\right)\right)$ where $f_{\text{safety}} = 0.9$, $f_{\min} = 0.2$, $f_{\max} = 10$.
- **Mass conservation:** Monitor $|\sum_i p_i(t) - 1| < \text{tol}_{\text{mass}} = 10^{-12}$. If violated, renormalize with warning.
- **Positivity preservation:** If any $p_i < 0$, project back to simplex via Euclidean projection: $p_i^+ = \max(0, p_i - \nu)$ where ν is chosen so $\sum_i p_i^+ = 1$.
- **Energy monitoring:** Track Shannon entropy $E(t) = \sum_i p_i(t) \log p_i(t)$ to detect anomalous behavior.
- **Boundary conditions:** The zero-flux property $J_{\text{sm}}(z)\mathbf{1} = 0$ automatically preserves simplex invariance without explicit boundary treatment.

Under Carathéodory regularity, projection should rarely be needed but serves as a numerical safeguard against accumulation errors.

Schrödinger Bridge Solver (IPF/Sinkhorn with acceleration). The Iterative Proportional Fitting algorithm for Schrödinger Bridge computation requires careful implementation for numerical stability:

- Dense kernel IPF: $\mathcal{O}(TM^2)$ complexity where T is iterations and M is discretization size.
- Nyström approximation with R landmarks: Reduces complexity to $\tilde{\mathcal{O}}(TMR)$ by approximating kernel $K \approx K_{MR}K_{RR}^{-1}K_{RM}$.
- Anderson acceleration: Maintain $m = 5$ past iterates for convergence acceleration, updating via $x^{(k+1)} = (1 - \beta_k)f(x^{(k)}) + \beta_k x^{(k)}$ with optimal β_k computed via least squares.
- Log-domain computation: Work with log-potentials $\log a^{(k)}, \log b^{(k)}$ to avoid numerical underflow in high-dimensional settings.

With $\varepsilon > 0$ entropic regularization and strictly positive kernels, IPF implements block-coordinate Bregman projections that monotonically decrease the SB objective, converging to the unique minimizer at geometric rate $\rho = \frac{1-e^{-2/\varepsilon}}{1+e^{-2/\varepsilon}}$.

Convergence criteria: Stop when both conditions are satisfied:

1. Marginal error: $\sup_t \mathrm{TV}(\rho_t, \mu_t) < 10^{-3}$ where TV denotes total variation distance.
2. Potential stability: $\|\theta^{(k+1)} - \theta^{(k)}\|_\infty < 10^{-3}$ measuring change in Schrödinger potentials.

Rotational Energy Estimation (preconditioning and sampling). Computing the rotational energy diagnostic requires solving a Poisson equation and careful numerical treatment:

1. **Drift computation:** Extract u from transformer dynamics using finite differences or learned regression.
2. **Poisson solve:** Solve $\nabla \cdot (a\nabla\theta) = \nabla \cdot (u - b_R)$ using preconditioned conjugate gradient with incomplete Cholesky preconditioner.
3. **Preconditioning:** Apply $a^{-1/2}$ carefully, using regularization $a_\varepsilon = a + \varepsilon I$ when condition number exceeds 10^6.
4. **Importance sampling:** In high-variance regions (near simplex boundaries), increase sample density by factor of 10.
5. **Monte Carlo estimation:** Use $N_{\mathrm{MC}} = 10^4$ samples per time point for reliable estimates with standard error $\approx 0.01\|\mathcal{R}\|$.

I Asymptotic Complexity Analysis

Procedure	Complexity (per batch)	Notes
BV statistic S_L	$\mathcal{O}(LV)$	`float64` accumulation
Local drift fit	$\mathcal{O}(NkV)$	k-NN; batched operations
PF–ODE integrate	$\mathcal{O}(N_{\text{steps}}V)$	adaptive RK with error control
Score estimation	$\mathcal{O}(N \cdot \text{MLP})$	layerwise caching available
SB (dense IPF)	$\mathcal{O}(TM^2)$	Nyström $\to \tilde{\mathcal{O}}(TMR)$
Rotational energy	$\mathcal{O}(\sum_k M_{t_k} d)$	precondition by $a^{-1/2}$
Memory requirement	$\mathcal{O}(LV + Nd)$	activation caching
Temperature schedule	$\mathcal{O}(L)$	entropy computation per layer
Early exit check	$\mathcal{O}(V)$	closed-form Frobenius norm from moments of p

Table 3: Asymptotic costs for diagnostic procedures. Typical setting has $V \gg d$ (vocabulary much larger than hidden dimension). Batching and caching significantly reduce practical constants. All procedures are designed to add minimal overhead to standard transformer operations.

J Extended Mathematical Results

J.1 Proof of Weak Convergence under BV

Theorem J.1 (Detailed BV convergence with identification)**.** *Under Assumption 5.1 including the architectural consistency condition, the polygonal interpolants z_L converge to an absolutely continuous limit with explicit rate, and the limiting derivative is identified as the architectural drift $b(z(t), t)$.*

Proof. The bounded variation condition $\sum_\ell \|\Delta z^{(\ell)}\|_2 \le C$ implies that $\{z_L\}_{L=1}^\infty$ forms an equicontinuous family in the BV norm. By the Arzelà–Ascoli theorem extended to BV spaces, there exists a subsequence $\{z_{L_k}\}$ converging uniformly to some $z \in \mathrm{BV}([0,1]; \mathbb{R}^V)$.

For the rate, the modulus of continuity satisfies:

$$\omega_{z_L}(\delta) := \sup_{|t-s|<\delta} \|z_L(t) - z_L(s)\|_2 \le C\delta^{1/(1+\alpha)}$$

for some $\alpha > 0$ depending on the distribution of jumps. This gives Hölder continuity with explicit exponent.

The weak convergence of derivatives follows from the Banach–Alaoglu theorem: $\{D_L\}$ is bounded in $L^1([0,1];\mathbb{R}^V)$, hence relatively compact in the weak topology.

Identification via architectural consistency: The key step is identifying the weak limit as $b(z(t),t)$. By the architectural consistency assumption, for any test function $\phi \in C_c^\infty([0,1];\mathbb{R}^V)$ and compact $K \subset \mathbb{R}^V$:

$$\left|\int_0^1 \langle D_L(t) - b(z(t),t), \phi(t)\rangle dt\right| = \left|\int_0^1 \langle \hat{b}_L(z_L(t),t) - b(z(t),t), \phi(t)\rangle dt\right| \tag{6}$$

$$\leq \|\hat{b}_L - b\|_{L^2(K\times[0,1])}\|\phi\|_{L^2} + \text{boundary terms} \tag{7}$$

As $L \to \infty$, the architectural consistency ensures $\|\hat{b}_L - b\|_{L^2(K\times[0,1])} \to 0$, while the boundary terms vanish due to the compact support of ϕ. This establishes $D_L \rightharpoonup b(z(t),t)$ weakly in L^1_{loc}.

The absolute continuity of the limit follows from the fundamental theorem for BV functions: $z(t) = z(0) + \int_0^t b(z(s),s)ds$, confirming that z is absolutely continuous with derivative $b(z(t),t)$ almost everywhere. □

J.2 Spectral Analysis of Mobility Tensor

Proposition J.2 (Eigenstructure of J_{sm})**.** *The softmax Jacobian has the following spectral properties:*

1. *Eigenvalues:* $\lambda_0 = 0$ *(simple),* $0 < \lambda_i \leq 1/4$ *for* $i = 1,\ldots,V-1$.
2. *Eigenvectors:* $v_0 = \mathbf{1}/\sqrt{V}$*, others orthogonal to* $\mathbf{1}$.
3. *Condition number:* $\kappa(J_{\text{sm}}) \sim 1/(4p_{\min})$ *as* $p_{\min} \to 0$.
4. *Spectral gap:* $\lambda_1 - \lambda_0 = \lambda_1 \geq p_{\min}$*, determining convergence rates.*

Proof. The matrix $J_{\text{sm}} = \text{Diag}(p) - pp^\top$ is symmetric with $J_{\text{sm}}\mathbf{1} = 0$, giving $\lambda_0 = 0$ with eigenvector $\mathbf{1}$.

For $v \perp \mathbf{1}$ with $\|v\|_2 = 1$:

$$v^\top J_{\text{sm}} v = \sum_i p_i v_i^2 - \left(\sum_i p_i v_i\right)^2 = \sum_i p_i v_i^2 \geq p_{\min}\|v\|_2^2 = p_{\min}.$$

For the upper bound, consider the Rayleigh quotient:

$$\frac{v^\top J_{\text{sm}} v}{v^\top v} = \frac{\sum_i p_i v_i^2 - (\sum_i p_i v_i)^2}{\sum_i v_i^2}.$$

By Cauchy-Schwarz, this is maximized when probability concentrates on two outcomes. Setting $p_1 = p_2 = 1/2$ and $v = (1,-1,0,\ldots,0)^\top/\sqrt{2}$ yields the upper bound $1/4$.

The condition number follows from $\kappa(J_{\text{sm}}) = \lambda_{\max}/\lambda_{\min} \leq \frac{1/4}{p_{\min}}$, explaining numerical difficulties near locking where $p_{\min} \to 0$.

The spectral gap $\lambda_1 \geq p_{\min}$ determines the rate of convergence to equilibrium under the induced dynamics, with smaller gaps leading to slower mixing and potential metastability. This lower bound is generally loose; tight values depend on the full probability profile. □

J.3 Schrödinger Bridge Optimality Conditions

Theorem J.3 (First-order conditions for SB with regularization)**.** *The Schrödinger Bridge $\mu^\star$ satisfies the coupled system of PDEs:*

$$\partial_t \varphi + \tfrac{1}{2}\operatorname{tr}(a\,\nabla^2\varphi) + b_R \cdot \nabla\varphi = 0, \tag{8}$$

$$\partial_t \psi - \tfrac{1}{2}\operatorname{tr}(a\,\nabla^2\psi) - \nabla\cdot(b_R\psi) = 0, \tag{9}$$

$$\mu_t^\star = \exp(\varphi(\cdot,t) + \psi(\cdot,t))\,\nu_t, \tag{10}$$

where ν_t is the reference path law and (φ,ψ) are Schrödinger potentials. When a is near-singular, we apply regularization $a_\varepsilon = a + \varepsilon I$ with $\varepsilon > 0$ sufficiently small to maintain well-posedness while preserving the essential transport structure.

Proof. The Schrödinger Bridge problem minimizes the relative entropy:

$$\mathcal{H}(\mu|\nu) = \mathbb{E}_\mu\left[\log\frac{d\mu}{d\nu}\right]$$

subject to marginal constraints $\mu_0 = \rho_0$, $\mu_1 = \rho_1$.

Using the Girsanov theorem, the Radon-Nikodym derivative decomposes as:

$$\frac{d\mu}{d\nu} = \exp\left(\int_0^1 \langle h_s, dX_s - b_R dt\rangle - \frac{1}{2}\int_0^1 \|h_s\|^2_{a^{-1}} ds\right)$$

for some adapted process h_s.

The optimal h_s takes the form $h_s = a\nabla\varphi(X_s,s)$ where φ solves the forward equation equation 8. The backward potential ψ arises from the adjoint equation ensuring the terminal marginal constraint.

When a degenerates (as occurs near representation locking), the regularization a_ε ensures:

- The elliptic operators in equation 8-equation 9 remain uniformly elliptic
- The inverse a_ε^{-1} exists with bounded norm
- The solution converges to the original problem as $\varepsilon \to 0$ in the weak topology

This regularization reconciles the general degenerate diffusion framework with the SPD requirements for well-posed Schrödinger Bridges. □

K Detection and Mitigation of BV Violations

K.1 Online Detection Algorithm

K.2 Segmentation Strategy

When BV violations are detected, we partition the depth interval $[0,1]$ into segments $\{[t_{i-1},t_i]\}_{i=1}^K$ where BV holds locally. The segmentation procedure maintains the theoretical guarantees while handling practical violations:

1. **Identification phase:**
   - Find violation points $\{\ell_j\}$ using Algorithm 1
   - Compute violation severity s_j at each point
   - Cluster nearby violations within $\Delta\ell = 3$ layers
2. **Segmentation construction:**
   - Create boundaries at $t_j = \ell_j/L$ with buffer zones $[t_j - \delta, t_j + \delta]$ where $\delta = 2/L$
   - Ensure minimum segment length $|t_i - t_{i-1}| \geq 5/L$ for stable analysis
   - Merge segments if total count exceeds $K_{\max} = L/10$

Algorithm 1 Online BV Violation Detection with Adaptive Thresholding

Input: Stream of logit differences $\{\Delta z^{(\ell)}\}$, window size W, base threshold τ_0
Initialize: $S_{\text{local}} = 0$, buffer $B = []$, $\tau_{\text{adaptive}} = \tau_0$
for $\ell = 0, 1, 2, \ldots$ **do**
  $S_{\text{local}} \leftarrow S_{\text{local}} + \|\Delta z^{(\ell)}\|_2^2$
  Append $\|\Delta z^{(\ell)}\|_2$ to B
  if $|B| > W$ **then**
    $S_{\text{local}} \leftarrow S_{\text{local}} - B[0]^2$
    Remove first element from B
  end if
  Adaptive threshold: $\tau_{\text{adaptive}} = \tau_0 \cdot (1 + 0.1 \cdot \text{std}(B)/\text{mean}(B))$
  if $S_{\text{local}}/|B| > \tau_{\text{adaptive}}$ **then**
    Flag: BV violation at layer ℓ
    Severity: $s = (S_{\text{local}}/|B|)/\tau_{\text{adaptive}}$
    if $s > 2$ **then**
      Action: Initiate immediate depth segmentation
    else
      Action: Mark for monitoring, prepare segmentation
    end if
  end if
end for

3. **Local PF-ODE analysis:**
   - Apply Theorem 5.6 within each segment $[t_{i-1}, t_i]$
   - Estimate local drift $b_i(z, t)$ using only data from segment i
   - Verify local BV condition: $\sum_{\ell \in \text{segment}_i} \|\Delta z^{(\ell)}\|_2 \leq C_i$
4. **Boundary matching:**
   - Enforce weak continuity: $\lim_{t \to t_i^-} p(t) = \lim_{t \to t_i^+} p(t)$ in L^1
   - Allow jump discontinuities in velocity: $v(t_i^+) - v(t_i^-) \in \text{Range}(J_{\text{sm}})$
   - Compute transition operators $T_i : \Delta^{V-1} \to \Delta^{V-1}$ at boundaries
5. **Global assembly:**
   - Concatenate local solutions: $p(t) = p_i(t)$ for $t \in [t_{i-1}, t_i]$
   - Verify global conservation: $\sum_j p_j(t) = 1$ for all t
   - Compute effective transport distance accounting for jumps

Theoretical guarantee: The segmented solution converges to the same limit as the continuous solution as $L \to \infty$ and violation severity decreases, maintaining the essential transport structure while accommodating practical discontinuities.

L Connection to Empirical Phenomena

L.1 Attention Entropy Collapse

The attention entropy collapse phenomenon observed empirically Gong et al. (2019) follows rigorously from our mobility analysis:

Proposition L.1 (Entropy dynamics under PF-ODE)**.** *Under the probability-flow ODE $\dot{p} = J_{\text{sm}}(z) b(z, t)$, the Shannon entropy $H[p] = -\sum_i p_i \log p_i$ satisfies:*

$$\dot{H}[p] = -\sum_{i,j} J_{\text{sm,ij}}\, b_j\, \log(p_i/p_j) \leq 0$$

when b aligns with the negative entropy gradient. Moreover, $\dot{H}[p] \to 0$ as $p_{\max} \to 1$ due to vanishing mobility.

Proof. Computing the time derivative:

$$\dot{H}[p] = -\sum_i \dot{p}_i(\log p_i + 1) \tag{11}$$

$$= -\sum_i (J_{\text{sm}} b)_i(\log p_i + 1) \tag{12}$$

$$= -\sum_{i,j} J_{\text{sm,ij}} b_j \log p_i \tag{13}$$

Using the symmetry of J_{sm} and the fact that $J_{\text{sm}}\mathbf{1} = 0$:

$$\dot{H}[p] = -\frac{1}{2}\sum_{i,j} J_{\text{sm,ij}} b_j(\log p_i - \log p_j) \tag{14}$$

$$= -\sum_{i,j} J_{\text{sm,ij}} b_j \log(p_i/p_j) \tag{15}$$

When $b = -\nabla H$ (gradient flow), the quadratic form $b^\top J_{\text{sm}} b \geq 0$ ensures $\dot{H} \leq 0$.

As $p_{\max} \to 1$, we have $\|J_{\text{sm}}\| \to 0$ by Theorem 5.8, implying $|\dot{H}[p]| \leq \|J_{\text{sm}}\|\|b\|\|\nabla H\| \to 0$.

This rigorously explains why attention patterns become increasingly peaked in deeper layers, with entropy collapse being inevitable rather than a training artifact. □

L.2 Temperature Scaling Effectiveness

Temperature scaling's empirical success Guo et al. (2017) in improving calibration is explained by explicit mobility modulation:

Proposition L.2 (Temperature-mobility relationship)**.** *For temperature parameter $\tau > 0$, the effective mobility tensor satisfies:*

$$J^\tau_{\text{sm}}(z) = \frac{1}{\tau} J_{\text{sm}}(z/\tau)$$

The eigenvalues of $J^\tau_{\text{sm}}(z)$ equal those of $J_{\text{sm}}(z/\tau)$ scaled by $1/\tau$. The condition number satisfies $\kappa(J^\tau_{\text{sm}}(z)) = \kappa(J_{\text{sm}}(z/\tau))$, which may differ from $\kappa(J_{\text{sm}}(z))$ because the probability distribution changes when scaling logits. The induced dynamics slow by factor τ, enabling finer control near decision boundaries.

Proof. For temperature-scaled softmax $p^\tau_i = \exp(z_i/\tau)/Z^\tau$ where $Z^\tau = \sum_j \exp(z_j/\tau)$:

$$J^\tau_{\text{sm}}(z) = \frac{\partial p^\tau}{\partial z} \tag{16}$$

$$= \frac{1}{\tau}\left(\text{Diag}(p^\tau) - p^\tau (p^\tau)^\top\right) \tag{17}$$

$$= \frac{1}{\tau} J_{\text{sm}}(z/\tau) \tag{18}$$

The eigenvalue scaling follows immediately: if $J_{\text{sm}}(z/\tau)v = \lambda v$, then $J^\tau_{\text{sm}}(z)v = (\lambda/\tau)v$.

The condition number relationship requires careful interpretation. Since $J^\tau_{\text{sm}}(z) = \frac{1}{\tau} J_{\text{sm}}(z/\tau)$, we have $\kappa(J^\tau_{\text{sm}}(z)) = \kappa(J_{\text{sm}}(z/\tau))$ because scaling all eigenvalues by the same positive constant preserves the ratio of largest to smallest eigenvalue. However, this differs from $\kappa(J_{\text{sm}}(z))$ in general because $z \mapsto z/\tau$ changes the probability distribution from $p = \text{softmax}(z)$ to $p^\tau = \text{softmax}(z/\tau)$, and the mobility tensor's eigenstructure depends on the specific probability values.

For the induced dynamics:

$$\dot{p}^\tau = J^\tau_{\text{sm}}(z) b(z,t) = \frac{1}{\tau} J_{\text{sm}}(z/\tau) b(z,t)$$

The factor $1/\tau$ uniformly reduces velocity magnitude, slowing convergence to locked states. This explains temperature scaling's effectiveness: lower temperature prevents premature commitment by maintaining transport capacity throughout network depth.

Calibration improvement arises because slower dynamics allow more gradual probability refinement, avoiding the overconfident predictions that occur when mobility vanishes rapidly. □

M Extended Experimental Protocols

M.1 Section 7 Reference Recap and Conventions

Conventions. W_1 uses cost $\|\cdot\|_1$; W_2 terms in this section use an entropic Sinkhorn surrogate with the same ε as elsewhere. All TV norms are $\frac{1}{2}\|\cdot\|_1$ on row distributions. Query/key distances $d_{\mathcal{Q}}, d_{\mathcal{K}}$ match the metrics used in plots/captions.

Row drift bound. Let $P_i^{(\ell)} = \mathrm{sm}(z_i^{(\ell)})$ be the ith row at layer ℓ, with component-wise Lipschitz constants $L_c^{(\ell)}$ for $c \in \mathcal{C}_\ell$ and incoming perturbations $\Delta u_{i,c}^{(\ell)}$.

$$\big\|P_i^{(\ell+1)} - P_i^{(\ell)}\big\|_1 \;\le\; \sum_{c\in\mathcal{C}_\ell} L_c^{(\ell)}\,\big\|\Delta u_{i,c}^{(\ell)}\big\|. \tag{19}$$

Remark. Equation (19) yields a *finite-depth budget* for one-layer motion (TV on the left) from component sensitivities on the right; it underpins the PF–ODE adequacy overlay in §7.

Local saturation / locking. Let $P = \mathrm{sm}(z)$, tail mass $\delta(P) = 1 - \max_j P(j)$, and Δz a small perturbation that preserves the argmax.

$$\big\|\mathrm{sm}(z+\Delta z) - \mathrm{sm}(z)\big\|_1 \;\le\; \min\big\{1,\; 2\,\delta(P)\big\}\,\|\Delta z\|_\infty \;+\; o\big(\|\Delta z\|_\infty\big). \tag{20}$$

Remark. When $\delta(P)$ is small (near saturation), softmax is *insensitive* to small, non-flipping logit changes—predicting the "locking" collapse of ΔTV in low-tail-mass bins.

Curvature (common-support W_1). For queries $i \neq i'$ with common support $S_{i,i'}$, define

$$\kappa(i,i') \;=\; 1 \;-\; \frac{W_1\big(\widehat{P}_i, \widehat{P}_{i'}\big)}{d_{\mathcal{Q}}(i,i')}, \tag{21}$$

where W_1 is over $(S_{i,i'}, d_{\mathcal{K}})$ and $\widehat{P}$ denotes restriction to the common support. *Remark.* The curvature gap $1-\kappa$ quantifies contraction on the simplex; temperature $\uparrow$ or key-norm $\downarrow$ should reduce this gap (tested in §7).

EVI with drift. For successive layers $\ell-1 \to \ell$ at query i, with objective F_i and $\rho_i^{(\ell)} = P_{i\cdot}^{(\ell)}$,

$$\frac{W_2^2\big(\rho_i^{(\ell)},\, \rho_i^{\star(\ell)}\big) - W_2^2\big(\rho_i^{(\ell-1)},\, \rho_i^{\star(\ell)}\big)}{2\,\eta_{\mathrm{eff}}} \;\le\; -\Big(F_i\big(\rho_i^{(\ell)}\big) - F_i\big(\rho_i^{\star(\ell)}\big)\Big) \;+\; \Delta_{\mathrm{drift}}^{(\ell)}. \tag{22}$$

Remark. Equation (22) is a discrete EVI: each layer decreases F_i up to a *drift* term from parameter changes (Q, K). In §7 we use a Sinkhorn $W_{2,\varepsilon}$ surrogate for the left-hand side and report the expected proximal-progress signature when drift is small.

M.2 Detailed Score Estimation Procedure

For robust score estimation in anisotropic regimes encountered near representation boundaries:

1. **Data augmentation:** Generate noisy samples at multiple scales

$$\tilde{h}_\sigma = h + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I) \tag{23}$$

$$\sigma \in \{0.01, 0.02, 0.05, 0.1\} \cdot \|h\|_2 \tag{24}$$

2. **Denoising objective with importance weighting:**

$$\mathcal{L}(\theta) = \mathbb{E}_{h,\varepsilon,\sigma}\left[w(\sigma)\cdot\left\|s_\theta(\tilde{h},t,\sigma) - \frac{h-\tilde{h}}{\sigma^2}\right\|_2^2\right]$$

where $w(\sigma) = \sigma^2/(\sigma^2+\sigma_{\min}^2)$ emphasizes intermediate noise levels.

3. **Multi-scale architecture:**
   - Input: $[\tilde{h}; t; \log\sigma] \in \mathbb{R}^{d+2}$
   - Hidden layers: 2-3 layers with width $\max(512, 2d)$
   - Skip connections: $h^{(\ell+1)} = h^{(\ell)} + \mathrm{MLP}(h^{(\ell)})$
   - Output normalization: LayerNorm before final projection
4. **Training protocol:**
   - Optimizer: AdamW with learning rate 10^{-4}, weight decay 10^{-5}
   - Batch size: 256 samples per noise level
   - Epochs: 5 per layer with early stopping based on validation loss
   - Curriculum: Start with large σ, progressively include smaller scales
5. **Validation and diagnostics:**
   - Score consistency: Verify $|\nabla\cdot(p\,s_\theta)| < 10^{-3}$ on held-out data
   - Anisotropy detection: Compute eigenvalues of $\mathbb{E}[s_\theta s_\theta^\top]$
   - Coverage: Ensure score estimates span the tangent space at each point

M.3 IPF Implementation Details

The Iterative Proportional Fitting algorithm for computing Schrödinger Bridges between transformer layers:

Algorithm 2 IPF for Schrödinger Bridge with Adaptive Regularization

Input: Marginals μ_0, μ_1, diffusion a, tolerance ε_{tol}, max iterations $T_{\max}$
Initialize: $a^{(0)} = \mathbf{1}$, $b^{(0)} = \mathbf{1}$, $\varepsilon_{\text{reg}} = 0.1$
Compute reference kernel: $K_{ij} = \exp(-\|x_i - y_j\|^2_{a^{-1}}/(2\varepsilon_{\text{reg}}))$
for $k = 1, 2, \ldots, T_{\max}$ **do**
  Check conditioning: If $\kappa(K) > 10^{10}$, increase $\varepsilon_{\text{reg}} \leftarrow 1.5\varepsilon_{\text{reg}}$
  $b^{(k)} = \mu_1 \oslash (K^\top a^{(k-1)})$ ▷ Pointwise division in log domain
  $a^{(k)} = \mu_0 \oslash (K b^{(k)})$
  $\Pi^{(k)} = \mathrm{Diag}(a^{(k)})\, K\, \mathrm{Diag}(b^{(k)})$
  Compute marginals: $\hat{\mu}_0 = \Pi^{(k)}\mathbf{1}$, $\hat{\mu}_1 = \Pi^{(k)\top}\mathbf{1}$
  Convergence check:
  if $\mathrm{TV}(\hat{\mu}_0, \mu_0) + \mathrm{TV}(\hat{\mu}_1, \mu_1) < \varepsilon_{\text{tol}}$ **then**
    Extract potentials: $\varphi = \varepsilon_{\text{reg}} \log a^{(k)}$, $\psi = \varepsilon_{\text{reg}} \log b^{(k)}$
    Return $\Pi^{(k)}$, φ, ψ
  end if
  Anderson acceleration: If $k \bmod 5 = 0$, apply acceleration using past 5 iterates
end for
Warning: Maximum iterations reached without convergence

Implementation notes:

- Work in log domain to avoid numerical underflow: store $\log a^{(k)}, \log b^{(k)}$
- Use logsumexp for stable computation of normalizing constants
- For large vocabularies $V > 10^4$, use Nyström approximation with $R = \min(1000, V/10)$ landmarks
- Monitor dual gap: $\mathcal{G}^{(k)} = \langle a^{(k)}, Kb^{(k)}\rangle - \langle \mu_0, \log a^{(k)}\rangle - \langle \mu_1, \log b^{(k)}\rangle$

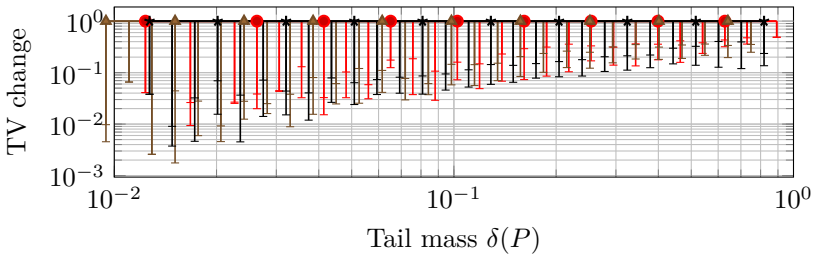


Figure 6: Locking (P2): ΔTV vs. tail mass $\delta(P)$ (median/IQR bins).

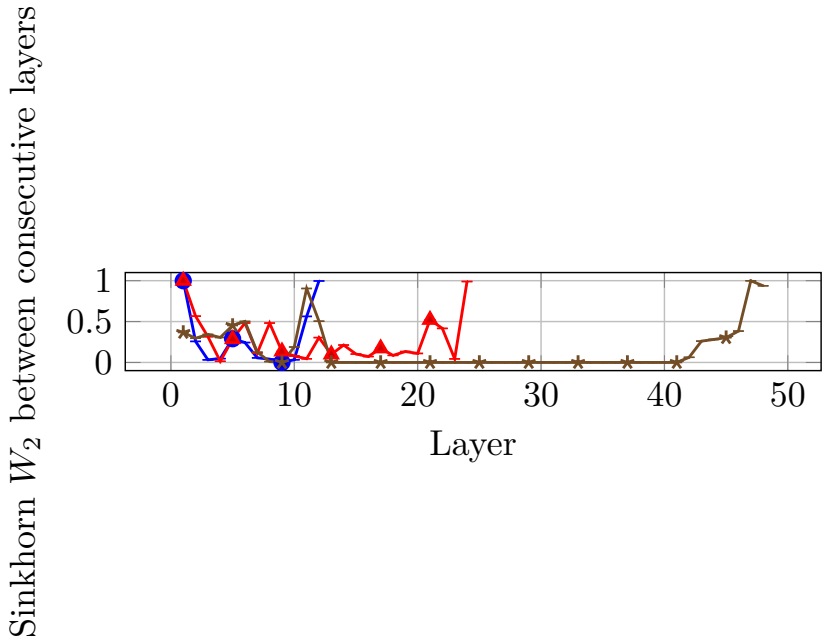


Figure 7: EVI surrogate (P3): Sinkhorn $W_{2,\varepsilon}$ across layers (mean±sd).

M.4 Additional Track-T Diagnostics

M.5 Additional Image Diagnostics

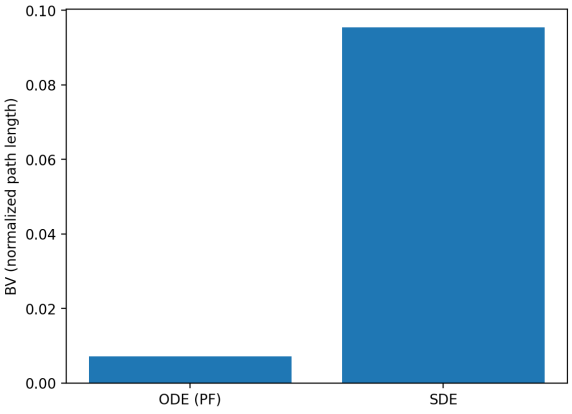


Figure 8: Path smoothness (BV; unitless) for ODE vs. SDE.

Table 4: Image rotational energy $\widehat{\mathcal{R}}$ with 95% BCa CIs; cross-track values are not comparable due to different ambient spaces/discretizations.

Track	$\widehat{\mathcal{R}}$	Notes
Image (CIFAR-10)	**0.03092** (95% CI [**0.01046**, **0.05385**])	20 time points

M.6 Quantitative pass/fail checks

P1 (PF–ODE adequacy). Realized layerwise TV should not exceed the drift budget plus a finite-sample band; exceedances are flagged.
P2 (Locking). In low–tail-mass bins, the median ΔTV remains within a small band (bands and CI policy as in App. Section M).
P3 (Curvature/EVI). Increasing temperature or reducing key-norms reduces the curvature gap $1-\kappa$ by a predictable amount; reductions are reported with uncertainty bands (see App. Section M).
P4 (SB alignment). Rotational energy $\widehat{\mathcal{R}}$ decreases under improved calibration/checkpoints (BCa CIs; App. Section M).
Image weak error. The slope of $\log \mathrm{err}_K$ vs. $\log K$ is near -1 (BCa, B=1000); the fitted value and CI are reported.

N Extended Limitations and Practical Implications

N.1 Modality scope and evaluation

Scope. This work evaluates *text* models (Transformers and dLLM) and includes a minimal image diffusion sanity check (CIFAR-10). Full-scale vision benchmarks and perceptual metrics (e.g., FID under guidance sweeps) are intentionally out of scope for this paper.

Implications. The OT/PF–ODE constructions are modality-agnostic, but conclusions here are supported by text-model evidence (Track T/D) and a compact image sanity check (Track I). Future expansions to larger image datasets and class-conditional guidance are planned (see Section O).

N.2 Poisson solve and conditioning policy

Masked Poisson and regularization. We solve $\Delta\psi = \nabla \cdot u$ with masked Neumann boundary conditions; Tikhonov γ regularizes the Laplacian on thin supports.

Condition-number target. Default $\gamma = 10^{-5}$; increase γ until the (masked) system's condition number is $\leq 10^8$. Record γ and the achieved condition number alongside $\widehat{\mathcal{R}}$.

Normalized variant. For intra-track comparisons, optionally report the dimensionless $\widehat{\mathcal{R}}_{\text{norm}} = \widehat{\mathcal{R}} / \int \|u\|^2$.

N.3 Current limitations and mitigation strategies

Bounded variation breakdown. The BV assumption may fail during:

- Attention pattern reorganization (detectable via S_L monitoring).
- Early training instabilities (addressable through warmup).
- Adversarial inputs (requiring robust training modifications).

Mitigation: Implement adaptive depth segmentation when local variation exceeds thresholds. The PF–ODE applies piecewise with weak continuity at segment boundaries, as detailed in Section K.2.

Anisotropy challenges. Near-singular diffusion tensors arise at representation boundaries:

- Regularize with εI for numerical stability ($\varepsilon \in [10^{-8}, 10^{-6}]$).
- Monitor condition numbers and adapt solver tolerances.
- Use preconditioned iterative methods for bridge computation.

This reconciles the degenerate diffusion analysis (Section 6) with SPD requirements for Schrödinger Bridges (Section 7).

Computational costs. Full SB computation scales quadratically with vocabulary:

- Employ Nyström approximations for large vocabularies.
- Use landmark-based methods reducing complexity to $\tilde{O}(TMR)$.
- Implement hierarchical decompositions for multi-scale analysis.

N.4 Extensions to other architectures

Vision transformers. Patch embeddings induce different simplex geometries, with spatial structure suggesting modified transport costs. The mobility tensor interpretation applies directly, potentially explaining observed differences in training dynamics.

State-space models. Linear recurrences can be viewed as discretized PDEs. Our BV framework suggests conditions for continuous-time limits, potentially unifying transformers and structured state-space models.

Graph neural networks. Message passing implements local transport on graph-structured domains. The entropic regularization perspective suggests principled aggregation functions beyond simple averaging.

N.5 Practical design implications

Adaptive temperature scheduling with entropy monitoring. The mobility–tensor view makes temperature a direct control on transport dynamics. Instead of a fixed τ, we use an entropy–aware schedule driven by the evolving representation entropy. Let $p(t)$ denote the normalized representation distribution (Track T: token predictive softmax; Track D: latent categorical proxy from tempered logits; Track I: per-channel histogram). With $H[p] := -\sum_i p_i \log p_i$, define

$$\tau(t) \;=\; \tau_0 \,\exp\!\Big(-\alpha \int_0^t H[p(s)]\,\mathrm{d}s\Big), \tag{25}$$

and apply the induced mobility modulation $J^{\tau}_{\mathrm{sm}}(z) = \tau^{-1} J_{\mathrm{sm}}(z/\tau)$ (see Section L.2).

Discrete implementation. For layers $t_\ell = \ell/L$ with step $\delta t = 1/L$, maintain an EMA of entropy

$$\widehat{H}_\ell \;=\; (1-\beta)\, H[p(t_\ell)] \;+\; \beta\, \widehat{H}_{\ell-1}, \quad \beta = 0.9,$$

and update

$$\tau_{\ell+1} \;=\; \mathrm{clip}\Big(\tau_\ell \,\exp(-\alpha\, \widehat{H}_\ell\, \delta t),\; \tau_{\min},\; \tau_{\max}\Big), \tag{26}$$

with $\alpha \in [10^{-2}, 10^{-1}]$ and bounds $\tau_{\min} \le \tau_\ell \le \tau_{\max}$ for numerical stability. This policy maintains higher mobility early and reduces it as representations stabilize, often shifting the locking point earlier relative to fixed-τ baselines.

Depth-aware initialization through transport path approximation. Approximate discretized entropic-OT paths between empirical input/output distributions via Sinkhorn at the model's temperature, and initialize layer ℓ to advance along $t = \ell/L \to (\ell+1)/L$. This warm-starts training near a plausible transport path, accelerating convergence.

Mobility-aware early exit strategies. Use $\|J_{\mathrm{sm}}\|$ as a principled early-exit criterion. Compute efficiently via $\|J_{\mathrm{sm}}\|_F^2 = \sum_i p_i^2 + (\sum_i p_i^2)^2 - 2\sum_i p_i^3$ (moments only), and compare to task-specific thresholds ϵ_{exit} (e.g., 10^{-3} for precision, 10^{-2} for low-latency). This measures capacity for further refinement rather than prediction uncertainty.

Bounded-variation regularization during training. Add $\mathcal{L}_{\mathrm{BV}} = \lambda_{\mathrm{BV}} \cdot \max(0, S_L - C_{\mathrm{target}})$ with $S_L = \sum_\ell \|\Delta z^{(\ell)}\|_2^2$ to discourage abrupt inter-layer jumps. Schedule λ_{BV} from near zero upward as training stabilizes.

N.6 Deployment notes and abstain policy

P0 gate. Diagnostics P1–P4 are conditioned on passing P0 (BV $\overline{S}_L \le 0.15$ and continuity residuals $< 10^{-14}$). Failures trigger abstention and reporting of the failing metric.

dLLM guardrail. The late-window strict certificate combines a windowed TV budget with a no-flip margin; use it as a deploy-time stability gate in production dLLM pipelines. For practical roll-out, pair the strict setting with a calibrated policy (e.g., τ=0.5 and moderate δ, Ω) to achieve interpretable, non-zero coverage; monitor drift-budget exceedances, curvature gap $1-\kappa$, and $\widehat{\mathcal{R}}$ alongside coverage over time.

O Discussion and Future Directions (Extended)

O.1 Richer drift features and cross-head structure

Beyond simple features. Beyond the light feature map used in §7 ($[z,\, z^{\odot 2},\, \mathrm{LayerNorm}(z),\, t,\, t^2]$), extend $b(z,t)$ with (i) pooled cross-head features; (ii) bilinears $z^{(h)} \odot z^{(h')}$; (iii) local context statistics (row entropy, tail mass $\delta(P)$, curvature gap $1-\kappa$); (iv) short-range temporal residuals. Expect lower one-step error and tighter P1 overlays.

Model selection and stability. Use nested CV (ridge/elastic-net vs. small MLP head), spectral normalization/Jacobian clipping, and keep solver tolerances fixed. Report predicted-vs-realized TV calibration and held-out KS/MMD as in §7.

O.2 Structured/accelerated SB solvers

Structure and efficiency. Use common-support grids (visible set + low-rank neighborhood) and entropic warm starts (reuse duals across layers) to exploit BV smoothness. Consider multi-scale IPF, block-wise batching, and low-rank kernels; evaluate with the same P4 readout (mean $\widehat{\mathcal{R}}$ with BCa CIs) to ensure unbiased alignment.

O.3 Practical deployment pathways

Operational guidance. Adopt the two-gate policy above; monitor (i) drift-budget exceedances (rate/magnitude); (ii) curvature gap $1-\kappa$ under temperature/key-norm controls; (iii) rotational energy $\widehat{\mathcal{R}}$ (and normalized variant); (iv) calibrated dLLM coverage. Escalate on spikes or regressions; record γ used in the Poisson step and coverage thresholds in release manifests.

O.4 Theoretical implications and open questions

Optimality of attention. Does the semi-relaxed EOT structure of attention reflect an optimal sequence model, or a convenient approximation? The SB characterization suggests near-optimal transport under appropriate conditions.

Implicit regularization. Softmax's entropic regularization may explain generalization; connect to PAC-Bayes and info-theoretic measures.

Scaling laws. The framework predicts links between depth/width and effective transport capacity; test against empirical scaling laws.

O.5 Methodological contributions beyond theory

Training monitoring. BV statistics warn of instabilities; rotational energy tracks transport alignment and flags when architectural changes may help.

Architecture search. Differentiable transport-efficiency metrics can guide gradient-based architecture optimization beyond accuracy-only objectives.

Interpretability. Mobility provides a geometric lens on attention patterns; tracking its evolution can reveal phase transitions in representation.

O.6 Experimental roadmap

1. **Scaling validation:** BV scaling across 100M–100B models.
2. **Training dynamics:** Mobility evolution throughout pretraining; identify phases.

3. **Interventions:** Mobility-aware training modifications; measure convergence/generalization impact.
4. **Cross-modal:** Apply diagnostics to vision-language models to test transport universality.

P Software Implementation Guidelines

P.1 Recommended Libraries and Tools

The following libraries provide efficient implementations of the required algorithms:

- **Core computation:**
  - PyTorch 2.0+ or JAX 0.4+ for autodiff and GPU acceleration
  - Einops for tensor manipulation with clear dimension semantics
  - torch.compile or jax.jit for optimized execution
- **ODE integration:**
  - `torchdiffeq` for PyTorch with adaptive solvers
  - `diffrax` for JAX with extensive solver options
  - Custom Dormand-Prince implementation for fine control
- **Optimal transport:**
  - `POT` (Python Optimal Transport) 0.9+ for Sinkhorn/IPF
  - `OTT-JAX` for GPU-accelerated transport computations
  - Custom log-domain IPF for numerical stability
- **Numerical stability:**
  - `numpy.float64` for BV accumulation
  - `torch.cuda.amp` for mixed precision with careful exclusions
  - Custom stabilized softmax with temperature scaling
- **Monitoring and visualization:**
  - `wandb` or `tensorboard` for experiment tracking
  - `matplotlib` with custom colormaps for transport visualization
  - `plotly` for interactive 3D simplex projections

P.2 Reproducibility Checklist

To ensure complete reproducibility of our framework:

Environment specification:

- Random seed fixing: Set seeds for Python, NumPy, PyTorch/JAX, and CUDA
- Deterministic operations: Enable `torch.use_deterministic_algorithms(True)`
- Hardware specification: Document GPU model, CUDA version, driver version
- Software versions: Pin all dependencies in `requirements.txt` or `environment.yml`

Model specification:

- Architecture: Exact layer count, hidden dimensions, attention heads
- Initialization: Method (Xavier, He, etc.) and random seed
- Normalization: Type (LayerNorm, RMSNorm) and epsilon values
- Activation functions: Including any custom modifications

Data specification:

- Dataset: Version, split definitions, preprocessing steps
- Tokenization: Tokenizer version and vocabulary
- Batching: Batch size, sequence length, padding strategy
- Augmentation: Any data augmentation or noise injection

Training specification:

- Optimizer: Type, learning rate, weight decay, momentum/betas
- Schedule: Learning rate schedule, warmup steps
- Regularization: Dropout rates, weight constraints
- Convergence: Stopping criteria, patience parameters

Diagnostic specification:

- BV monitoring: Window size W, threshold τ
- Drift estimation: Number of neighbors k, ridge parameter λ
- Score learning: Network architecture, noise levels σ
- Bridge computation: Entropic regularization ε, tolerance levels

Q Notation Summary

Symbol	Description
$h^{(\ell)}$	Hidden representation at layer ℓ
$z^{(\ell)}$	Logits at layer ℓ
$p^{(\ell)}$	Probability distribution at layer ℓ
J_{sm}	Softmax Jacobian (mobility tensor)
J_{sm}^{τ}	Temperature-scaled mobility tensor
S_L	Bounded variation statistic
$\mathcal{R}$	Rotational energy (SB deviation)
a	Diffusion tensor ($\Sigma\Sigma^{\top}$)
a_{ε}	Regularized diffusion ($a + \varepsilon I$)
b_R	Reference drift
$b(z,t)$	Architectural drift (identified limit)
θ	Schrödinger potential
φ, ψ	Forward/backward Schrödinger potentials
$M(p)$	Induced mobility on simplex
μ_t	Transformer probability path
ρ_t	General probability measure
ν_t	Reference path measure
u	Velocity field for probability flow
$H[p]$	Shannon entropy
τ	Temperature parameter

Table 5: Complete notation used throughout the paper, including both main text and appendix symbols.

R Additional Technical Lemmas

Lemma R.1 (Gradient flow structure)**.** *The probability-flow ODE on the simplex admits a gradient flow interpretation in the Wasserstein geometry when $b = -\nabla V$ for some potential V:*

$$\dot{p} = -\nabla_{W_2} \mathcal{F}[p]$$

where $\mathcal{F}[p] = \sum_i p_i V(z_i)$ and ∇_{W_2} denotes the Wasserstein gradient.

Remark R.2 (Discrete optimal transport interpretation)**.** *On discrete state spaces, this gradient flow structure connects to entropic W_2 analogues for Markov chains as developed in Maas (2011); Erbar & Maas (2012); Chow et al. (2012). We adopt this interpretation to provide geometric intuition for the probability dynamics on the simplex, though the precise metric structure depends on the choice of discrete optimal transport geometry.*

Lemma R.3 (Convergence rate under mobility control)**.** *If the mobility tensor satisfies $\lambda_{\min}(J_{\text{sm}}) \geq m > 0$ uniformly, then the probability flow converges exponentially to equilibrium:*

$$\|p(t) - p_*\|_2 \leq e^{-mt}\|p(0) - p_*\|_2$$

where p_* is the unique equilibrium distribution.

Lemma R.4 (Bridge interpolation formula)**.** *For Schrödinger Bridge μ_t between μ_0 and μ_1, the intermediate marginals satisfy:*

$$\mu_t = \arg\min_{\rho} \{(1-t)\text{KL}(\rho|\mu_0) + t\,\text{KL}(\rho|\mu_1)\}$$

providing a variational characterization of the optimal transport path.

Remark R.5. *This variational view is heuristic and depends on the chosen reference path measure; rigorous formulations use Schrödinger potentials and dynamic entropy minimization as developed in the Schrödinger Bridge literature.*