

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 FROM ATTENTION TO DIFFUSION: A UNIFIED EN- TROPIC 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 two tracks—Transformers (core diagnostics) and a compact image diffusion model (parity and first-order weak-error behavior)—with validity conditions and diagnostic abstention protocols detailed in Appendix N.4. 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.

This unification builds on and extends several research streams that have developed in isolation. Attention mechanisms have been interpreted through optimal transport in recent work (Sander et al., 2022; Tay et al., 2020; Xu et al., 2023; Daneshmand, 2024), but prior approaches typically employ balanced optimal transport formulations that are fundamentally incompatible with causal masking in autoregressive language modeling. We resolve this by proving that standard row-softmax attention precisely solves a semi-relaxed entropic optimal transport problem that preserves autoregressive causality. Continuous-depth interpretations of neural networks through neural ordinary differential equations (Chen et al., 2018; Dupont et al., 2019; Bai et al., 2019) have been extended to transformers (Zhang et al., 2021; Chen et al., 2023; Kan et al., 2025), but existing analyses typically lack rigorous finite-depth error control. We strengthen these perspectives by introducing an explicit bounded variation regime with quantitative finite-depth error bounds controlled by a budget parameter, proving that the continuous-depth limit satisfies a well-posed probability flow ordinary differential equation on the probability simplex with the softmax Jacobian acting as a mobility tensor.

Score-based generative models have revealed fundamental connections between stochastic differential equations and deterministic probability flow ordinary differential equations (Song et al., 2021; Huang et al., 2021; Lipman et al., 2022). We demonstrate that transformer attention implements discretizations of probability flow ordinary differential equations in the same geometric family as those underlying diffusion models but operating under semi-relaxed entropic regularization, explaining why autoregressive-diffusion hybrid architectures work well empirically (Hoogeboom et al., 2022; Ma et al., 2025). Schrödinger Bridge theory provides a dynamic formulation of entropy-regularized optimal transport (Léonard, 2014; De Bortoli et al., 2021; Shi et al., 2023); we operationalize Schrödinger Bridge alignment within transformer attention dynamics by defining a rotational energy quantity that measures deviations from optimality and vanishes if and only if the attention-induced flow satisfies the Schrödinger Bridge characterization, transforming abstract optimality conditions into practical diagnostics estimable from model activations. Comprehensive literature review with detailed comparisons appears in Appendix C.

Contributions. Under mild regularity assumptions (detailed in Section 2.2), our main results are:

1. *Layer-level principle.* Standard row-softmax attention implements a principled Kullback-Leibler proximal transport step in the sense of mirror descent or Jordan-Kinderlehrer-Otto schemes, establishing the foundational connection between neural architecture and optimal transport geometry formalized in Proposition 2.1.
2. *Depth-to-time convergence with explicit rates.* The discrepancy between discrete layer dynamics and continuous probability flow is controlled by a finite-depth budget parameter denoted Ξ_L and defined precisely through bounded variation conditions in Theorem 3.1, providing quantitative error bounds that determine when continuum approximations apply to finite architectures.
3. *Diffusion unification through weak approximation.* Stacked attention layers weakly approximate time-inhomogeneous anisotropic reverse diffusions in probability law, with error that separates discretization effects from logit variation and optional degeneracy regularization, as established through the anisotropic Fokker-Planck analysis in Theorem 4.5.
4. *Schrödinger Bridge alignment certificate.* A rotational energy quantity denoted \mathcal{R} provides a necessary and sufficient condition for exact Schrödinger Bridge alignment, quantifying deviations from gradient flow structure and serving as a practical diagnostic for transport optimality as formalized in Theorem 5.2.

The framework yields three falsifiable predictions: continuum approximation accuracy when Ξ_L is small, diminishing returns from depth when mobility degrades, and correlation between low \mathcal{R} and improved generation quality. Our empirical study tests these predictions across transformer language models and compact image diffusion, with extended protocols in Appendix 3–1. Figure 5 visualizes how these results connect: Proposition 2.1 establishes the foundational single-layer principle; Theorem 3.1 extends this to finite-depth convergence through bounded variation compactness; Theorem 3.7 provides well-posedness infrastructure supporting Theorem 4.5’s diffusion unification; and Theorem 5.2 characterizes transport optimality via rotational energy. The diagram identifies the functional analysis machinery underlying each result, directly addressing proof architecture questions.

2 PRELIMINARIES AND CONCEPTUAL FRAMEWORK

2.1 CONCEPTUAL OVERVIEW

Before establishing formal machinery, we outline the key geometric quantities. The softmax Jacobian $J_{\text{sm}}(z)$ acts as the mobility tensor on the probability simplex, with temperature modulating transport capacity via $J_{\text{sm}}^\tau(z) = \tau^{-1} J_{\text{sm}}(z/\tau)$. The finite-depth budget Ξ_L quantifies how well discrete layers approximate continuous flow; small Ξ_L ensures the probability-flow ordinary differential equation accurately captures layerwise behavior. Rotational energy \mathcal{R} measures deviation from optimal transport; exact Schrödinger Bridge alignment occurs when \mathcal{R} vanishes.

108 2.2 MATHEMATICAL PRELIMINARIES AND NOTATION
109110 This subsection establishes notation, states the standing assumptions used throughout, and
111 records the layer-level optimal transport view we invoke in subsequent analysis.
112113 **Global Assumptions.** We collect here the global assumptions used throughout. Assump-
114 tions 2.1 and 3.1 provide the bounded-variation and regularity conditions, with the latter
115 adding architectural consistency for the continuum-limit results.
116117 **Assumption 2.1** (Bounded variation and architectural consistency). *We work on compact
subsets where all quantities are well-defined. Unless stated otherwise, we assume:*118 1. *Bounded-variation logits with uniform mean: Let $z^{(\ell)}$ denote the layer logits and $\Delta z^{(\ell)} :=$
119 $z^{(\ell+1)} - z^{(\ell)}$. We assume that the averaged per-layer logit variation*
120

121
$$C_{\text{BV}} := \frac{1}{L} \sum_{\ell=0}^{L-1} \|\Delta z^{(\ell)}\|_{\infty}$$

122

123 *remains bounded by a constant C_{BV}^* uniformly across all layer counts L . Equivalently,
124 the total variation $\sum_{\ell=0}^{L-1} \|\Delta z^{(\ell)}\|_{\infty}$ grows at most linearly in L , so that typical layer-
125 to-layer changes, rather than accumulated variation, control the quality of continuum
126 approximation.*
127128 2. *Local drift regularity. The effective drift $b(\cdot, t)$ is locally Lipschitz in its state argument
129 on bounded sets with Lipschitz constant L_b and is locally bounded by M_b .*
130 3. *Mobility bounds. For $p = \text{softmax}(z/\tau)$ with temperature $\tau > 0$, the Jacobian $J_{\text{sm}}(z) =$
131 $\text{Diag}(p) - pp^{\top}$ satisfies operator-norm and derivative bounds on the relevant compact
132 domain; denote $\Lambda_J := \sup \|J_{\text{sm}}(z)\|_{\text{op}}$ and $L_J := \sup \|\nabla J_{\text{sm}}(z)\|_{\text{op}}$.*
133 4. *Simplex invariance. Probability vectors p remain in the simplex under the dynamics
134 considered; faces are handled by the standard tangent-space restriction.*
135136 **Softmax and Mobility.** Given logits $z \in \mathbb{R}^V$ and temperature $\tau > 0$, the softmax
137 operation and its induced mobility tensor are defined by
138

139
$$p = \text{softmax}(z/\tau), \quad p_i = \frac{\exp(z_i/\tau)}{\sum_j \exp(z_j/\tau)}, \quad J_{\text{sm}}(z) = \text{Diag}(p) - pp^{\top}.$$

140 The Jacobian $J_{\text{sm}}(z)$ characterizes how probability mass flows under logit perturbations,
141 acting as the mobility tensor that governs transport dynamics on the probability simplex.
142143 **Remark (Sharp Mobility Bound).** We have $\|J_{\text{sm}}(z)\|_{\text{op}} \leq \frac{1}{2\tau}$, with equality at dis-
144 tributions $p = (\frac{1}{2}, \frac{1}{2}, 0, \dots, 0)$. In particular, for $\tau = 1$, $\|J_{\text{sm}}(z)\|_{\text{op}} \leq \frac{1}{2}$ and the spectrum
145 lies in $[0, \frac{1}{2}]$, collapsing to $\{0\}$ as $\max_i p_i \rightarrow 1$. A proof is provided in Appendix B. This
146 sharp bound explains why temperature scheduling proves essential for maintaining mobility
147 in deep networks as distributions become increasingly peaked.
148149 2.3 SEMI-RELAXED ENTROPIC OPTIMAL TRANSPORT FOR ATTENTION
150151 Standard attention with row-softmax normalization solves a semi-relaxed entropic optimal
152 transport problem that preserves autoregressive causality. For a query vector $q \in \mathbb{R}^{d_k}$ and
153 key vectors $k_j \in \mathbb{R}^{d_k}$, define the cost $c_j = -q \cdot k_j$, so that high similarity corresponds
154 to low transport cost. Given a reference distribution $u \in \Delta^{V-1}$ (typically uniform) and
155 temperature $\tau > 0$, the semi-relaxed entropic OT problem is

156
$$\min_{p \in \Delta^{V-1}} \left\{ \sum_{j=1}^V p_j c_j + \tau \text{KL}(p \| u) \right\}. \quad (1)$$

157

158 The first-order optimality conditions yield the softmax solution
159

160
$$p_j = \frac{\exp(q \cdot k_j / \tau)}{\sum_{k=1}^V \exp(q \cdot k_k / \tau)} = \text{softmax}(q K^{\top} / \tau)_j,$$

so that each attention row solves a semi-relaxed entropic OT problem where the row-stochastic constraint is enforced but column marginals are unconstrained, preserving autoregressive structure. Causal masking is implemented by assigning infinite cost $c_j = +\infty$ to masked positions. The complete derivation including the Lagrangian formulation, uniqueness, and masked formulation appears in Appendix B.

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

$$p^+ \in \arg \min_{p \in \Delta} \left\{ \langle c, p \rangle + \tau \text{KL}(p \parallel u) \right\}.$$

Stacking such updates discretizes a Kullback-Leibler mirror descent or Jordan-Kinderlehrer-Otto flow under the assumptions in Section 2.2. The proof is given in Appendix B.

This proposition establishes that each attention layer implements a principled optimal transport step, providing the foundation for our continuous-depth analysis in subsequent sections.

3 DISCRETE CONTINUITY AND THE CONTINUOUS-DEPTH LIMIT

3.1 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 3.1 (Bounded variation and architectural consistency). *Let $\delta t = 1/L$ and $t_\ell = \ell/L$. We assume:*

1. *Bounded total variation: $\sum_\ell \|\Delta z^{(\ell)}\|_2 \leq C$ (uniformly in L).*
2. *Uniform boundedness (tightness): $\sup_\ell \|z^{(\ell)}\|_2 \leq C_z$.*
3. *Architectural consistency (identification): local-regression estimates \hat{b}_L converge to b on compacts; see Appendix D.*

Note. Weak L^1 convergence of D_L to b is not assumed here; it follows from Lemma 3.5 via the calibration-generalization argument.

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 2.2, we upper bound layer increments with $\|\cdot\|_\infty$ (comparable to $\|\cdot\|_2$ on compacts). Define

$$\Xi_L := \alpha_1 \max_\ell \|\Delta z^{(\ell)}\|_\infty + \alpha_2 \sum_\ell \|\Delta z^{(\ell)}\|_\infty^2, \quad (2)$$

where α_1, α_2 depend only on L_b, M_b, Λ_J, L_J from Section 2.2. **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 2 are consistent with the $\|\cdot\|_2$ -based BV assumption in Assumption 3.1; see Appendix D for the explicit constants used in the proof of Theorem 3.1.

Theorem 3.1 (Finite-depth error to PF-ODE). *Under Assumption 3.1 and the regularity in Section 2.2, let $p(t)$ solve the probability-flow ODE on $[0, 1]$ with $p(0) = \lim_{L \rightarrow \infty} p^{(0)}$. Then there exists $\Gamma = \Gamma(L_b, M_b, \Lambda_J, L_J)$ such that*

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

with Ξ_L in equation 2. In particular, if $p^{(0)} = p(0)$ and $\Xi_L \rightarrow 0$, then $p^{(\lfloor tL \rfloor)} \rightarrow p(t)$ uniformly in t .

216 *Proof sketch.* The proof controls the per-layer error $\Delta_\ell = \|p^{(\ell)} - p(t_\ell)\|_1$ between the discrete
 217 stack and the PF-ODE at times $t_\ell = \ell/L$.
 218

219 **Step 1: Local truncation.** A Taylor expansion of the PF-ODE solution around t_ℓ ,
 220 combined with stability estimates for the simplex Jacobian J_{sm} and drift b , yields a one-
 221 step inequality

$$222 \quad \Delta_{\ell+1} \leq (1 + C_1 \delta t) \Delta_\ell + C_2 \|\Delta z^{(\ell)}\|_\infty \delta t + C_3 \delta t^2,$$

223 with $C_1 = L_b \Lambda_J + M_b L_J$ capturing Lipschitz and mobility bounds.

224 **Step 2: Global accumulation.** Iterating over all layers and applying a discrete Grönwall
 225 lemma (Lemma D.1) produces

$$226 \quad \Delta_L \leq e^{C_1} \Delta_0 + e^{C_1} (C_2 C_{\text{BV}} + C_3 L^{-1}),$$

228 where $C_{\text{BV}} = \frac{1}{L} \sum_\ell \|\Delta z^{(\ell)}\|_\infty$ is the averaged bounded-variation constant.

229 **Step 3: Budget definition.** The finite-depth budget Ξ_L in equation 2 combines worst-
 230 case jumps and cumulative squared variation, with explicit constants depending only on
 231 $(L_b, M_b, \Lambda_J, L_J)$. This yields the stated bound with $\Gamma = C_1$.

232 Complete details including explicit constant derivations and norm equivalence appear in
 233 Appendix D. \square
 234

235 **Remark 3.2** (Continuum validity and constant scaling). Ξ_L is a practical validity thresh-
 236 old: the PF-ODE faithfully predicts layerwise behavior when Ξ_L is small (see proof above;
 237 additional technical details in App. D). Moreover, the budget constants scale with archi-
 238 tectural smoothness and geometry: $\alpha_1 = \mathcal{O}(L_b + M_b)$ and $\alpha_2 = \mathcal{O}(\Lambda_J + L_J)$. Hence Ξ_L
 239 decreases with smaller per-layer logit increments and stronger contraction, and the PF-ODE
 240 discrepancy vanishes as $L \rightarrow \infty$ under fixed budgets.

241 **Remark 3.3** (When BV holds in practice). BV typically holds during stable training but can
 242 fail at phase transitions, early layers, or gradient instability. Detect via $S_L = \sum_\ell \|\Delta z^{(\ell)}\|_2^2$;
 243 if BV fails, apply piecewise analysis (App. D).

244 **Lemma 3.4** (Compactness and absolute continuity). Under Assumption 3.1, there exists a
 245 subsequence with $z_L \rightarrow z$ and $p_L \rightarrow p$ in $L^1([0, 1])$ and a.e., where p is absolutely continuous
 246 with $|p| \in L^1$. The compactness and identification statements follow from Theorem J.1 in
 247 Appendix J.1.

248 **Lemma 3.5** (Drift identification via architectural consistency). Under Assumption 3.1
 249 with (i) bounded total variation and uniform boundedness and (ii) architectural consistency,
 250 define $D_L(t) := \Delta z^{(\ell)}/\delta t$ on $[t_\ell, t_{\ell+1})$, and let \hat{b}_L be the local regression estimator fit to the
 251 same layer transitions. Then

$$252 \quad \|D_L - \hat{b}_L\|_{L^1([0,1])} \rightarrow 0 \quad \text{and} \quad \|\hat{b}_L - b\|_{L^1([0,1])} \rightarrow 0,$$

253 hence $\|D_L - b\|_{L^1([0,1])} \rightarrow 0$ and, in particular, $D_L \rightharpoonup b$ in $L^1([0, 1])$. Proof sketch. **Calib-
 254 ration:** D_L and \hat{b}_L are computed from identical transitions, so regression residuals control
 255 $\|D_L - \hat{b}_L\|_{L^1}$. **Generalization:** architectural consistency yields $\|\hat{b}_L - b\|_{L^1} \rightarrow 0$ on compacts.
 256 Triangle inequality concludes. (Complete details appear in Appendix J.1.)
 257

258 3.2 SEMI-RELAXED OPTIMAL TRANSPORT AND CAUSAL ATTENTION

259 **Remark 3.6** (Row-softmax via semi-relaxed EOT). By the KL-prox characterization in
 260 Proposition 2.1, standard row-softmax solves a semi-relaxed entropic OT step (with masking
 261 handled by infinite costs and restricted support). We refer to Appendix B for details of the
 262 dual and masking.

263 3.3 PROBABILITY-FLOW ODE EMERGENCE AND WELL-POSEDNESS

264 **Theorem 3.7** (PF-ODE on the simplex and well-posedness). Under Assumption 3.1 and
 265 the regularity in Section 2.2, the limit probability path satisfies

$$266 \quad \dot{p}(t) = J_{\text{sm}}(z(t)) b(z(t), t) \quad \text{a.e. on } [0, 1], \quad p(0) = \lim_{L \rightarrow \infty} p^{(0)},$$

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

273 Complete proof in Appendix Q.1.

274 **Remark 3.8** (Simplex invariance and uniqueness). Under Carathéodory conditions on b
 275 (measurable in t , locally Lipschitz in z), mass is conserved ($\sum_i p_i(t) = 1$), nonnegativity
 276 holds, zero-flux $J_{\text{sm}}(z)\mathbf{1} = 0$ enforces boundary behavior, and solutions are unique on the
 277 relative interior of Δ^{V-1} .

278 **Theorem 3.9** (Locking via vanishing mobility). If $p_{\max}(t) \rightarrow 1$ and b is bounded, then
 279 $\|J_{\text{sm}}(z(t))\|_{\text{op}} \rightarrow 0$ (Remark 2.2) and hence $\|\dot{p}(t)\| \rightarrow 0$. Moreover, temperature rescales
 280 mobility as $J_{\text{sm}}^{(\tau)}(z) = \frac{1}{\tau} J_{\text{sm}}(z/\tau)$, modulating the approach to locking.

281 **Example 3.10** (Two-token mobility collapse (summary)). Consider a minimal attention
 282 layer with two tokens. For logits $z = (z_1, z_2)$, the softmax Jacobian reduces to

$$284 \quad J_{\text{sm}}(z) = p(1-p) \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},$$

285 where $p = p_1 = \text{softmax}(z)_1$. The operator norm is $\|J_{\text{sm}}(z)\|_{\text{op}} = 2p(1-p)$, which attains
 286 its maximum $1/2$ at the uniform distribution $p = 1/2$ and collapses to zero as $p \rightarrow 0$ or
 287 $p \rightarrow 1$. With temperature τ , the effective mobility scales as $\tau^{-1}p(1-p)$.

288 As attention mass locks onto one token, the mobility eigenvalue vanishes, forcing $\dot{p} =$
 289 $J_{\text{sm}}(z) b(z, t)$ to approach zero even if the drift b remains nonzero. Temperature rescaling
 290 modulates this: larger τ maintains nontrivial mobility deeper into the network. This
 291 illustrates the mechanism behind Theorem 3.9: as distributions concentrate, the mobility
 292 tensor loses rank and dynamics freeze for geometric reasons, not because the drift dis-
 293 ppears. Complete eigenvalue calculations appear in Appendix L.3.

296 3.4 CONNECTIONS TO EMPIRICALLY OBSERVED PHENOMENA

297 Attention entropy collapse, temperature scaling effects, and representation collapse follow
 298 naturally from the mobility interpretation: as distributions concentrate, mobility (and thus
 299 velocity) vanishes (Theorem 3.9), explaining attention concentration and providing a handle
 300 for calibration via temperature scaling. We defer expanded discussion, diagnostics, and
 301 eigenspectrum-based tests to Appendix E.

304 4 DIFFUSION DUALITY WITH ANISOTROPIC NOISE

306 4.1 STOCHASTIC DYNAMICS AND WEAK FOKKER-PLANCK FORMULATION

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

$$310 \quad dH_t = F(H_t, t) dt + \Sigma(H_t, t) dW_t, \quad (3)$$

312 with diffusion tensor $a = \Sigma\Sigma^\top$. Our analysis accommodates minimal regularity (F locally
 313 integrable with weak derivatives, a measurable and locally bounded), anisotropy (a may be
 314 degenerate or near-singular), and time-inhomogeneity.

315 **Lemma 4.1** (Distributional calculus in weak FP regime). Under local Fisher-information
 316 conditions ($p_H > 0$ a.e., $p_H \nabla \log p_H \in L^1_{\text{loc}}$), the product rule holds distributionally:

$$317 \quad \nabla \cdot \nabla \cdot (a p_H) = \nabla \cdot ((\nabla \cdot a) p_H + a \nabla p_H) \quad \text{in } \mathcal{D}'.$$

319 Proof via mollification and weak convergence in Appendix F.

320 **Theorem 4.2** (PF-ODE / reverse-SDE duality). Let $a(x, t) = \sigma(x, t)\sigma(x, t)^\top$ and suppose
 321 $p_H(\cdot, t) > 0$ solves the Fokker-Planck equation

$$323 \quad \partial_t p_H = -\nabla \cdot (F p_H) + \frac{1}{2} \sum_{i,j} \partial_{x_i x_j} (a_{ij} p_H)$$

324 with suitable decay/no-flux boundary conditions. Define the deterministic flow
 325

$$326 \quad u(x, t) = F(x, t) - \frac{1}{2} \left(a(x, t) \nabla_x \log p_H(x, t) + (\nabla \cdot a)(x, t) \right), \quad (4)$$

328 where $(\nabla \cdot a)_i := \sum_j \partial_{x_j} a_{ij}$. Then the PF-ODE with velocity u shares identical marginals
 329 with the Itô SDE for all t . If $a \equiv 2\beta I$ is spatially constant, then $u = F - \beta \nabla \log p_H$ recovers
 330 the standard probability flow drift. Complete proof in Appendix F.

331 **Corollary 4.3** (Simplex marginal preservation). For the softmax projection $\varphi(h) =$
 332 $\text{softmax}(W^\top h)$, the pushforward measures satisfy $\varphi_\# p_H(\cdot, t) = \varphi_\# \rho(\cdot, t)$ a.e. in time, ex-
 333 tending the duality to simplex-valued processes. Proof in Appendix F.

334 **Proposition 4.4** (Anisotropy propagation to simplex dynamics). The hidden-space diffu-
 335 sion induces an effective mobility on the simplex: $M(p) = J_{\text{sm}}(z) W^\top a W J_{\text{sm}}(z)$, revealing
 336 how architectural choices modulate probability dynamics. Proof in Appendix F.

338 4.2 WEAK APPROXIMATION OF DIFFUSION BY STACKED ATTENTION

340 **Theorem 4.5** (Weak SDE approximation by stacked attention). Under the assumptions in
 341 Section 2.2 and the weak FP calculus of Lemma 4.1, let $\rho(t)$ be the law of the reverse SDE
 342 with drift u in equation 4 and diffusion a , and let $\widehat{\rho}_L(t)$ be the law induced by L stacked
 343 attention layers with step $\delta t = 1/L$. Then, for any $\phi \in C_b^2$ and $T \in [0, 1]$,

$$344 \quad \left| \mathbb{E}_{\widehat{\rho}_L(T)}[\phi] - \mathbb{E}_{\rho(T)}[\phi] \right| \leq C_\phi \left(L^{-1} + \max_{0 \leq \ell < L} \|\Delta z^{(\ell)}\|_\infty + \gamma \right),$$

346 where C_ϕ depends on bounds of u, a and ϕ on compacts, and $\gamma \geq 0$ is an optional degeneracy
 347 regularizer. Proof in Appendix F.

349 Stacked attention approximates anisotropic, time-inhomogeneous diffusion in a weak sense;
 350 the approximation error separates discretization, logit variation, and degeneracy regular-
 351 ization. Anisotropic diffusion with widely varying eigenvalues induces directional stiffness
 352 mirroring attention’s collapsed-coordinate behavior, explaining curvature and locking diag-
 353 nóstics. Toy example and degeneracy guidelines in Appendix F.

355 5 SCHRÖDINGER BRIDGES AND TRANSPORT OPTIMALITY

357 5.1 GENERAL FRAMEWORK AND ALIGNMENT CONDITIONS

359 Schrödinger Bridges (SB) characterize entropy-regularized stochastic interpolations between
 360 endpoint distributions. We establish when transformer-induced probability paths align with
 361 these optimal bridges. While Section 4 allows degenerate diffusion (useful near locking), SB
 362 typically requires a uniformly elliptic reference; we reconcile these views below.

363 **Assumption 5.1** (Reference diffusion). The reference process R follows $dX_t =$
 364 $b_R(X_t, t) dt + \sigma(X_t, t) dW_t$ with diffusion tensor $a = \sigma\sigma^\top$, where:

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

371 **Theorem 5.1** (SB alignment characterization). Let $\{\mu_t\}_{t \in [0, 1]}$ be the transformer’s
 372 continuous-depth probability path with drift u . Under Assumption 5.1, $\{\mu_t\}$ equals the
 373 Schrödinger Bridge for reference R if and only if its per-mass velocity decomposes as

$$374 \quad u = b_R + a \nabla \theta$$

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

378 **Theorem 5.2** (Rotational energy controls SB deviation). *Let $u = b_R + a\nabla\theta + w$ be the*
 379 *a -weighted Hodge decomposition with $\nabla \cdot (w\mu_t) = 0$ for each t . Define the rotational energy*
 380

$$381 \quad \mathcal{R} = \int_0^1 \int \langle w, a^{-1}w \rangle \mu_t(dx) dt.$$

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

$$384 \quad \text{KL}(\mu_t \| \mu_t^*) \leq C_P(\mu, a) \mathcal{R},$$

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

387 **Corollary 5.3** (Rotational energy diagnostic). *$\mathcal{R} \geq 0$ with equality iff the path is*
 388 *Schrödinger Bridge. Practically, estimate u (from activations), solve for θ via a weighted*
 389 *Poisson equation, compute the residual $r = u - b_R - a\nabla\theta$, and evaluate $\int \|a^{-1/2}r\|^2 d\mu dt$.*
 390 *See App. Figure 6 for a compact schematic of this pipeline.*

391 **Example 5.4** (Rotational energy on the two-simplex). *Consider a toy flow on $\Delta^2 =$*
 392 *$\{(p_1, p_2, p_3) : p_i \geq 0, \sum_i p_i = 1\}$ with coordinates (p_1, p_2) and $p_3 = 1 - p_1 - p_2$. Let*
 393 *$a = \sigma^2 I$ and $b_R = 0$. A gradient flow $u = a\nabla\theta$ with $\theta(p_1, p_2) = -\frac{\alpha}{2}(p_1^2 + p_2^2)$ yields*
 394 *$u = (-\alpha\sigma^2 p_1, -\alpha\sigma^2 p_2)$. The curl vanishes:*

$$395 \quad \partial_{p_1} u_2 - \partial_{p_2} u_1 = 0,$$

396 *confirming zero rotational energy and pure Schrödinger Bridge structure. Conversely, the*
 397 *rotational flow $u = (\beta p_2, -\beta p_1)$ circulating around the simplex center yields*

$$398 \quad \partial_{p_1}(-\beta p_1) - \partial_{p_2}(\beta p_2) = -2\beta \neq 0,$$

400 *indicating rotational energy proportional to β^2 . This flow cannot arise as a Schrödinger*
 401 *Bridge because it lacks potential structure; the rotational component represents spurious*
 402 *circulation that wastes transport capacity on cycles rather than moving mass toward the*
 403 *terminal distribution. Our diagnostic computes an empirical analogue by discretizing the*
 404 *flow field from layer activations, approximating spatial derivatives via finite differences, and*
 405 *integrating across the trajectory. Small \mathcal{R} indicates approximate gradient structure and*
 406 *Schrödinger Bridge alignment; large \mathcal{R} reveals spurious rotational components deviating*
 407 *from optimal transport.*

408 **Remark 5.5** (Vanishing-regularization limit). *If $a_\varepsilon \rightarrow a$ with $\varepsilon \downarrow 0$ and the sequence of SB*
 409 *paths has uniformly bounded action and is tight, any weak limit is a degenerate SB solution;*
 410 *when $\mathcal{R} = 0$, it coincides with the PF-ODE path. See Appendix G.*

411 **Corollary 5.6** (Simplex Schrödinger Bridge). *Under the softmax pushforward, the SB con-*
 412 *dition on the simplex takes the potential-flow form*

$$413 \quad \dot{P}_t = -\nabla_p \cdot (P_t M(P_t) \nabla_p \Theta(P_t, t)),$$

414 *with mobility M from Theorem 4.4. This connects directly to gradient flows on the simplex*
 415 *and informs mobility-aware design.*

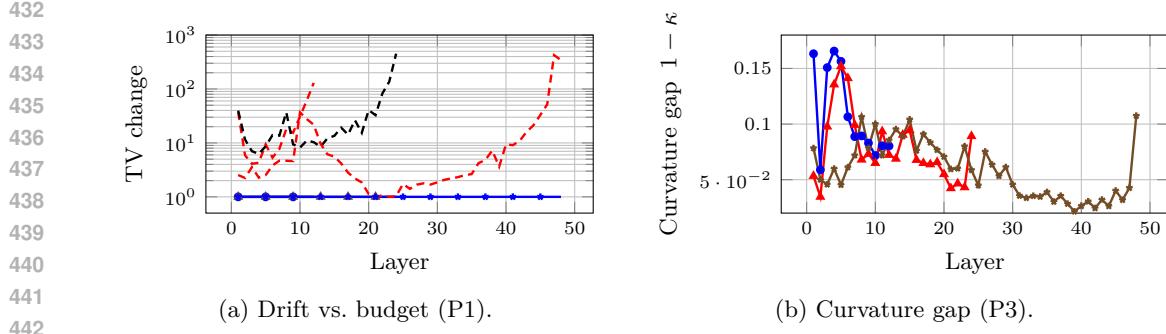
416 **Practical implication.** Rising \mathcal{R} indicates deviation from SB (OT) geometry and co-
 417 occurs with over-smoothing and spurious drift; minimizing \mathcal{R} provides a geometry-aware
 418 early warning complementary to standard fidelity metrics.

420 6 EMPIRICAL VALIDATION FRAMEWORK AND DIAGNOSTIC TOOLS

422 **Overview and theory map.** We validate two tracks: *(T)* Transformers (forward pass as
 423 PF-ODE) and *(I)* image diffusion (parity and weak-error).

425 **Probability-flow ODE dual (summary).** In variance-preserving (VP) score-based diffu-
 426 sion, the forward SDE is $dx = f(t) x dt + g(t) dW_t$ and the learned score $\nabla_x \log p_t(x)$ defines
 427 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
 428 shares the SDE time marginals.

429 We use the formal definitions from App. Section M.1 for the drift budget, locking bound,
 430 curvature, and EVI (Equations (22) to (25)) throughout this section. Drift and curvature
 431 visualizations appear in Figure 1 (left/right panels), while locking and EVI are shown in
 App. Figures 7 and 8.



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

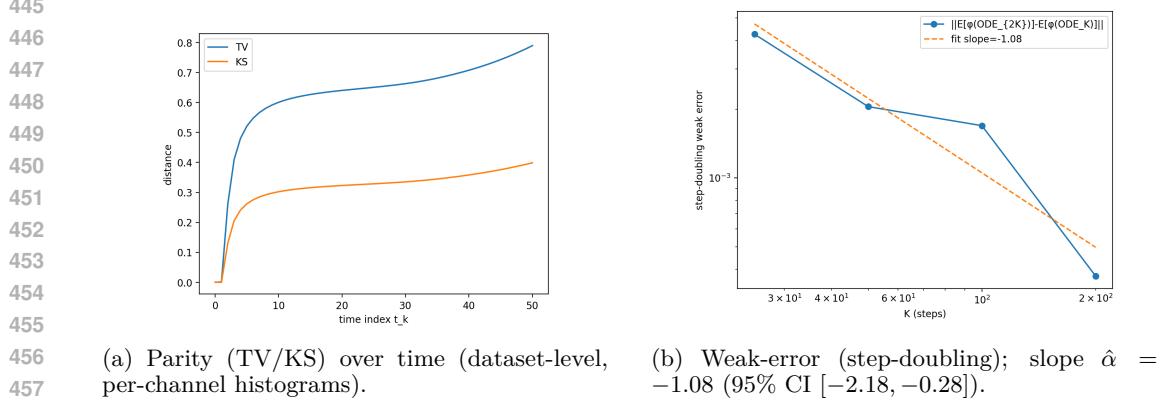


Figure 2: Image diffusion (CIFAR-10). Left: ODE-SDE TV/KS across time (inputs scaled to $[0, 1]$, equal channel weighting, 256 bins). Right: log-log regression of Δ_K vs. K with BCa CIs ($B=1000$).

6.1 EMPIRICAL DIAGNOSTICS P0–P4

Diagnostics (P0–P4). Five diagnostics validate the theory: (P0) BV/continuity checks; (P1) PF-ODE adequacy; (P2) locking behavior; (P3) OT contractivity; (P4) SB alignment via rotational energy. Full protocols appear in App. Section M.

6.2 TRACK T: TRANSFORMERS – CORE DIAGNOSTICS AND ROTATIONAL ENERGY

For the Transformer experiments, the mean rotational energy across 10 central layers is $\hat{\mathcal{R}} = 1.096 \times 10^{-7}$ (95% CI $[3.468 \times 10^{-8}, 2.153 \times 10^{-7}]$). 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).

6.3 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 $\hat{\mathcal{R}} = 0.03092$ (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.

486 Defaults. Unless noted, for the image track PF-ODE uses deterministic sampling
 487 on the same K grid as SDE (DDIM-style); for Transformers, PF-ODE drift fits use
 488 Dormand-Prince with $\text{rtol} = 10^{-5}$, $\text{atol} = 10^{-7}$, $\text{max_steps} = 2000$. Ridge grid
 489 $\{10^{-4}, 10^{-3}, 10^{-2}\}$ with 5-fold cross-validation; $N = 50\text{k}$ rows/layer (Transformers), $N =$
 490 10k images and $K = 50$ time steps (Image). Unless noted, bands denote 95% *percentile*-
 491 *bootstrap* CIs ($B=200$); weak-error CIs use BCa ($B=1000$).

492 Synthesis. Taken together, the two empirical tracks support a single underlying picture:
 493 attention dynamics in Transformers and PF-ODE/SDE trajectories in diffusion models
 494 behave as different discretizations of the same entropy-regularized transport flow. Locking
 495 and EVI signatures appear in the appendix; the core P1/P3 diagnostics remain in the main
 496 text.

498 7 LIMITATIONS AND PRACTICAL IMPLICATIONS

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

507 Practical implications and Outlook (1) Temperature or key-norm controls reduce the
 508 curvature gap $1 - \kappa$, offering a stable knob for depth behavior. (2) The drift-budget overlay
 509 surfaces over-activation and can inform regularization or early exit policies. (3) Diagnostic
 510 abstention protocols based on P0 gating conditions provide conservative guardrails when
 511 validity assumptions are violated (Section N.4). Richer drift features (e.g., cross-head struc-
 512 ture), broader modalities beyond CIFAR-10, structured/accelerated SB solvers, and calibra-
 513 tion via condition-number targets for the Poisson step are natural directions (Section O).

514 8 CONCLUSION

We formalized masked attention as semi-relaxed entropic OT, established stability/locking
 516 and curvature/EVI structure with gauge invariances, and tied these to a practical empirical
 517 suite. The suite validates PF-ODE adequacy, locking signatures, and contractivity response
 518 in Transformers, and shows image PF-ODE/SDE parity with first-order weak-error scaling.
 519 These yield concrete levers (temperature/key norm; drift-informed regularization) and
 520 geometry-aware diagnostics; extended discussion appears in Section O. For practitioners: (i)
 521 regulate depth via Ξ_L /stability budgets and spectral norm controls, (ii) monitor rotational
 522 energy during schedule sweeps as an early-warning diagnostic, and (iii) abstain when P0
 523 validity conditions fail.

525 REFERENCES

527 Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances
 528 in Neural Information Processing Systems*, 2019. URL <https://arxiv.org/abs/1909.01377>.

530 Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural or-
 531 dinary differential equations. In *Advances in Neural Information Processing Systems*,
 532 2018.

533 Yuqi Chen, Kan Ren, Yansen Wang, Yuchen Fang, Weiwei Sun, and Dongsheng Li. Con-
 534 tiformer: Continuous-time transformer for irregular time series modeling. In *NeurIPS*,
 535 2023. URL <https://arxiv.org/abs/2402.10635>.

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

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

543

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

547 Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural odes. In *Advances
 548 in Neural Information Processing Systems*, 2019.

549

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

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

555

556 Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neu-
 557 ral networks. In *Proceedings of the 34th International Conference on Machine Learning*,
 558 2017.

559

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

562

563 Chin-Wei Huang, Jae Hyun Lim, and Aaron C. Courville. A variational perspective on
 564 diffusion-based generative models and score matching. In *Advances in Neural Information
 565 Processing Systems*, 2021. URL <https://proceedings.neurips.cc/paper/2021/hash/c11abfd29e4d9b4d4b566b01114d8486-Abstract.html>.

566

567 Kelvin Kan, Xingjian Li, and Stanley Osher. Ot-transformer: A continuous-
 568 time transformer architecture with optimal transport regularization. *arXiv preprint
 569 arXiv:2501.18793*, 2025.

570

571 Christian Léonard. A survey of the schrödinger problem and some of its connections with
 572 optimal transport. *Discrete and Continuous Dynamical Systems A*, 2014.

573

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

576

577 Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng
 578 Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu, and
 579 Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multi-
 580 modal understanding and generation. In *CVPR*, 2025. arXiv preprint arXiv:2411.07975.

581

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

584

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

587

588 William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*,
 589 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.pdf.

590

591 Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Trans-
 592 formers with doubly stochastic attention. In *AISTATS*, 2022.

593

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

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

597
 598 Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. Sparse sinkhorn
 599 attention. *arXiv:2002.11296*, 2020. URL <https://arxiv.org/abs/2002.11296>.

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

605 Jing Zhang et al. Continuous self-attention models with neural ode networks. In *AAAI*,
 606 2021. URL <https://cdn.aaai.org/ojs/17692/17692-13-21186-1-2-20210518.pdf>.

608
 609 **Ethics Statement.** We have read and will adhere to the ICLR Code of Ethics. This work
 610 develops a theoretical and diagnostic framework unifying transformers and diffusion models
 611 via entropy-regularized optimal transport (OT). Our experiments use only publicly available
 612 datasets and open checkpoints where licenses allow redistribution or scripted download (de-
 613 tails in the appendix). We do not collect, annotate, or release any personal or sensitive data,
 614 and we do not deploy models for user-facing decisions. The proposed diagnostics (P0–P4),
 615 PF–ODE integration, and entropy-based temperature scheduling are intended to improve
 616 scientific understanding and training/serve-time efficiency (e.g., early exit). Potential risks
 617 are limited to misinterpretation or over-generalization of the diagnostics outside their val-
 618 idity regime (e.g., when bounded variation fails or under heavy sparsity/MoE routing).
 619 To mitigate this, we clearly document assumptions, abstain when diagnostic preconditions
 620 fail, and report limitations (mixture-of-experts, highly sparse attention, and early training
 621 phases). We see no domain-specific legal, privacy, or safety issues introduced by this study.

622
 623 **Reproducibility Statement.** We aim for complete reproducibility. The appendix spec-
 624 ifies: (i) data sources, splits, and licenses; (ii) model checkpoints and versions; (iii) all
 625 hyperparameters; (iv) exact diagnostic protocols; (v) hardware and runtime details. Upon
 626 acceptance, we will release a repository containing:

- 627 • **Diagnostics (P0–P4).** Implementations for BV/continuity checks (P0), PF–ODE
 628 adequacy and drift fitting (P1), locking and curvature/EVI (P2–P3), and rotational
 629 energy / SB diagnostic (P4), including numerically stable Poisson solves and a -
 weighted Hodge decomposition.
- 630 • **PF–ODE Integration.** Reference ODE solvers with error control and scripts to
 631 compare ODE vs. SDE marginals for the duality experiments.
- 632 • **Weak-Error Evaluation.** Step-doubling protocol with BCa bootstrap ($B=1000$)
 633 and log–log slope estimation; code to reproduce the reported confidence intervals.
- 635 • **Image Parity (Track I).** TV/KS histogram parity evaluation on CIFAR-10 with
 636 $N=10,000$ images and $K=50$ time points, including seeds and preprocessing.
- 637 • **Entropy-Based Temperature Scheduling.** Continuous and discrete schedules
 638 (EMA, clipping bounds) with ablation hooks.
- 639 • **Configuration + Seeds.** YAML configs for each experiment, fixed random seeds,
 640 and deterministic flags where supported by the backend.

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

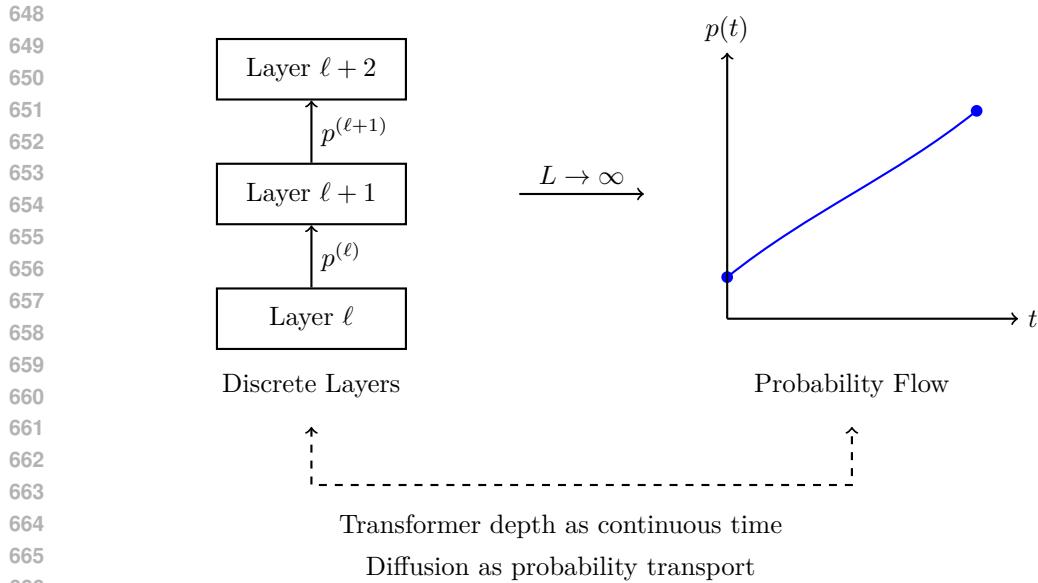


Figure 3: Conceptual unification: transformer layers implement discrete steps of probability transport that converge to continuous flows analogous to diffusion models. The softmax normalization induces entropic regularization, while layer stacking corresponds to time evolution.

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	
704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755	703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755
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
<i>Key Diagnostics:</i>	
<ul style="list-style-type: none"> • <i>During Training:</i> Monitor S_L for stability, $\ J_{\text{sm}}\$ for representation health • <i>Architecture Design:</i> Use \mathcal{R} to compare transport efficiency across architectures • <i>Deployment:</i> 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 \rightarrow 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	—	$J_{\text{sm}} \rightarrow 0$	Entropy collapse

Table 1: 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 2.2). 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 $\sum_i v_i = 0$,

$$v^\top J_{\text{sm}} v = \sum_i p_i v_i^2 - \left(\sum_i p_i v_i \right)^2 \leq \frac{1}{2} \sum_i p_i v_i^2,$$

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 $\|J_{\text{sm}}(z)\|_{\text{op}} \leq \frac{1}{2\tau}$ and the spectrum is contained in $[0, \frac{1}{2\tau}]$, collapsing to $\{0\}$ as $p_{\max} \rightarrow 1$. \square

Semi-relaxed EOT details. We provide the complete derivation of the semi-relaxed entropic optimal transport characterization of attention.

Let $q \in \mathbb{R}^{d_k}$ be the query vector and $k_j \in \mathbb{R}^{d_k}$ the j -th key vector in a shared key-query space. Define the cost $c_j = -q \cdot k_j$, so that high similarity corresponds to low transport cost. Given a reference distribution $u \in \Delta^{V-1}$ (typically uniform), we consider

$$\min_{p \in \Delta^{V-1}} \left\{ \sum_{j=1}^V p_j c_j + \tau \text{KL}(p \| u) \right\}, \quad (2.1)$$

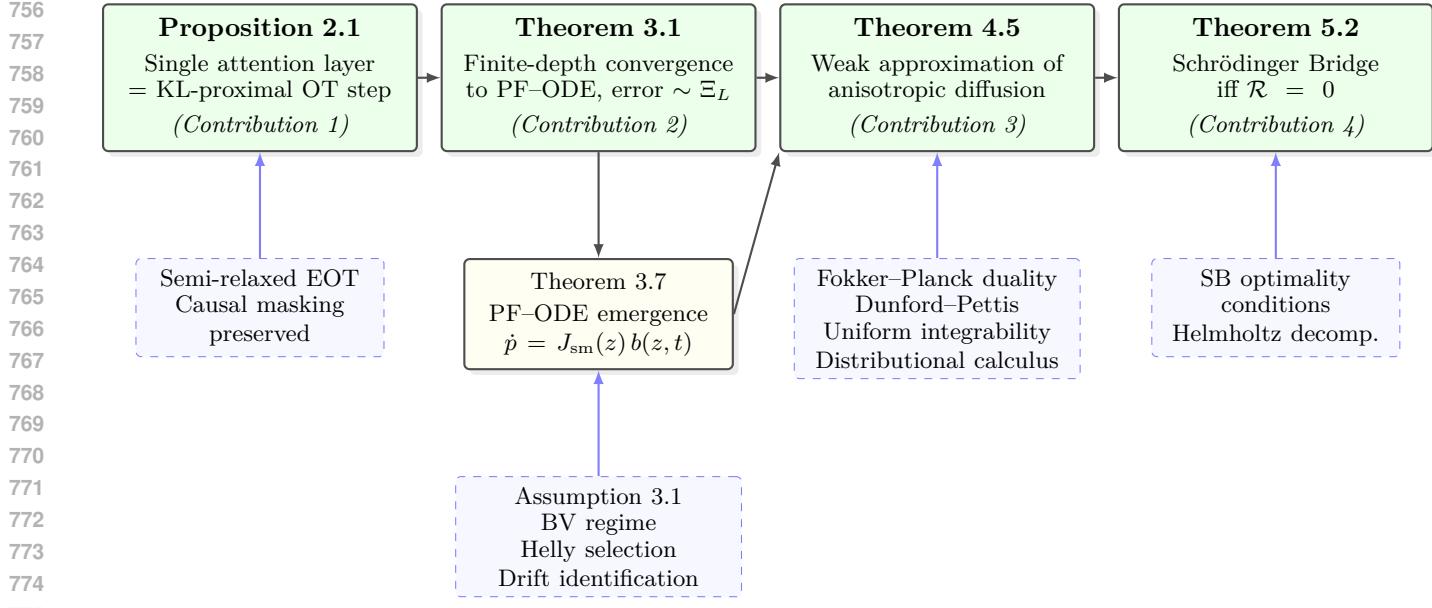


Figure 5: Logical structure and dependencies of main theoretical results. The framework establishes four contributions enumerated in Section 1: (1) Proposition 2.1 shows that a single attention layer implements a KL-proximal optimal transport step in the sense of Jordan-Kinderlehrer-Otto schemes, establishing the foundational connection between neural architecture and optimal transport geometry. (2) Theorem 3.1 proves that stacked attention layers converge to continuous probability flows on the simplex with explicit finite-depth error bounds controlled by the bounded variation budget Ξ_L , using Helly selection and architectural consistency to identify the limiting drift. (3) Theorem 4.5 demonstrates that the limiting probability flow weakly approximates time-inhomogeneous anisotropic reverse diffusions, unifying attention and diffusion through Fokker-Planck duality. (4) Theorem 5.2 provides a rotational energy characterization showing that vanishing \mathcal{R} is necessary and sufficient for exact Schrödinger Bridge alignment. The intermediate result Theorem 3.7 establishes well-posedness of the probability flow ODE with the softmax Jacobian as mobility tensor. Blue dashed boxes indicate foundational assumptions and mathematical tools; solid boxes with shadows indicate proven results. Green highlighting emphasizes the four main contributions; yellow indicates supporting infrastructure.

with temperature $\tau > 0$ and $\text{KL}(p\|u) = \sum_j p_j \log(p_j/u_j)$.

Step 1: Lagrangian formulation. Imposing $\sum_j p_j = 1$ via Lagrange multiplier λ gives

$$\mathcal{L}(p, \lambda) = \sum_{j=1}^V p_j c_j + \tau \sum_{j=1}^V p_j \log \frac{p_j}{u_j} - \lambda \left(\sum_{j=1}^V p_j - 1 \right).$$

Step 2: First-order optimality conditions. Setting $\partial \mathcal{L} / \partial p_j = 0$ yields

$$c_j + \tau \left(\log \frac{p_j}{u_j} + 1 \right) - \lambda = 0.$$

Solving for p_j gives

$$p_j = u_j \exp \left(\frac{\lambda - c_j - \tau}{\tau} \right) = C u_j \exp \left(-\frac{c_j}{\tau} \right),$$

where $C = \exp((\lambda - \tau)/\tau)$ is a normalization constant.

Step 3: Row normalization yields softmax. Enforcing $\sum_j p_j = 1$ determines C :

$$C = \left[\sum_{j=1}^V u_j \exp(-c_j/\tau) \right]^{-1}.$$

810 For uniform $u_j = 1/V$, we obtain
 811

$$812 \quad p_j = \frac{\exp(-c_j/\tau)}{\sum_{k=1}^V \exp(-c_k/\tau)} = \frac{\exp(q \cdot k_j/\tau)}{\sum_{k=1}^V \exp(q \cdot k_k/\tau)},$$

814 which is exactly the standard attention weight $p_j = \text{softmax}(qK^\top/\tau)_j$.
 815

816 Thus each attention row solves a semi-relaxed entropic OT problem where the row-stochastic
 817 constraint is enforced but column marginals are unconstrained, preserving autoregressive
 818 structure. Causal masking is implemented by assigning infinite cost $c_j = +\infty$ to masked
 819 positions; the resulting row-normalized solution coincides with the attention distribution
 820 induced by logits, with existence and uniqueness guaranteed by the strict convexity of the
 821 KL divergence.
 822

823 **Proof of Proposition 2.1.** The mirror-descent Euler step in KL geometry with objective
 824 $\langle c, p \rangle$ and step τ yields the variational form in Proposition 2.1. The unique minimizer has
 825 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. \square
 826

827 C EXTENDED RELATED WORK AND POSITIONING (FULL VERSION)

829 C.1 PROBABILITY FLOWS AND SCHRÖDINGER BRIDGES

830 Score-based diffusion established that reverse-time SDEs admit a *probability-flow* ODE with
 831 identical marginals (Song et al., 2021), while flow matching proposed simulation-free training
 832 of vector fields that realize desired probability paths (including OT geodesics) (Lip-
 833 man et al., 2022). The Schrödinger Bridge (SB) program casts diffusion as *entropic* OT
 834 on path space and provides scalable IPF-style solvers (De Bortoli et al., 2021; Shi et al.,
 835 2023). We leverage this geometry *inside* transformers: depth induces a PF-ODE on the sim-
 836 plex, weak/anisotropic FP theory gives a deterministic/stochastic duality for hidden-state
 837 evolution, and an *if&only-if* potential-plus-reference drift condition characterizes when a
 838 transformer’s probability path is exactly an SB.
 839

840 C.2 ATTENTION AS ENTROPIC OPTIMAL TRANSPORT

841 Balanced OT views of attention enforce doubly-stochastic constraints via Sinkhorn itera-
 842 tions (Sander et al., 2022; Tay et al., 2020), and OT-based co-attention improves multimodal
 843 learning (Xu et al., 2023). A complementary line shows transformers can be *programmed*
 844 to *solve* entropic OT with accuracy improving in depth (Daneshmand, 2024). In contrast,
 845 we work in the *causal* regime and prove that *standard row-softmax attention* is precisely
 846 the optimizer of a *semi-relaxed* entropic OT (row constraints only), which preserves autore-
 847 gressive masking and does not require imposing OT constraints at training time. From this
 848 equality we derive a BV depth→PF-ODE limit and the SB characterization in the causal
 849 setting; balanced OT results do not cover this regime and are fundamentally incompatible
 850 with the autoregressive structure essential to language modeling.
 851

852 C.3 CONTINUOUS-TIME VIEWS OF TRANSFORMERS

853 Continuous-depth interpretations of transformers address irregular time environments and
 854 ODE couplings (Zhang et al., 2021; Chen et al., 2023); OT-Transformer introduces OT as
 855 a *regularizer* in a continuous-time backbone (Kan et al., 2025). These works, however, do
 856 not *explicitly* endow the dynamics with an entropic-OT geometry that explains empirical
 857 phenomena. Our framework fills this gap: the softmax Jacobian acts as a *mobility tensor*
 858 on Δ^{V-1} , depth induces a PF-ODE with simplex invariance and well-posedness, and SB
 859 equivalence provides a variational certificate for transport optimality.
 860

861 C.4 AUTOREGRESSIVE–DIFFUSION HYBRIDS

862 Bridging autoregressive and diffusion/flow paradigms has shown strong empirical results
 863 (Hoogeboom et al., 2022; Ma et al., 2025). Our theory explains *why*: AR transformers and
 864

864 diffusion models are two discretizations (discrete in depth vs. continuous in time) of the same
 865 entropy-regularized transport principle. The PF-ODE/FP duality and SB tools provide
 866 quantitative diagnostics (e.g., rotational energy) for assessing alignment with entropic OT.
 867

868 C.5 ARCHITECTURAL UNIFICATION VIA DIFFUSION TRANSFORMERS 869

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

875 C.6 POSITIONING OF OUR CONTRIBUTIONS 876

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

880 **(ii) Depth → PF-ODE on the simplex.** Under bounded-variation scaling, stacking
 881 attention layers induces a PF-ODE for probe-induced probabilities, with simplex invariance
 882 and well-posedness.

883 **(iii) Weak FP duality with anisotropy.** Allowing time-inhomogeneous, anisotropic
 884 (and possibly ill-conditioned) diffusion, we establish deterministic/stochastic equivalence of
 885 marginals via Fokker–Planck in the renormalized/weak sense.

886 **(iv) SB equivalence (iff) & diagnostics.** The depth path is an SB iff its velocity is
 887 potential-plus-reference drift; deviations are quantified by a rotational-energy gap.

888 **(v) Mechanisms and predictions.** Identifying J_{sm} as mobility explains entropy collapse
 889 and representation locking; output-logit temperature scaling predicts mobility reductions
 890 that move locking earlier.

891 These theoretical advances translate directly into *actionable diagnostics and design principles*
 892 (e.g., mobility/locking metrics, SB alignment, anisotropy-aware regularization) for
 893 improving both transformer and diffusion architectures.

894 D SUPPLEMENTARY DETAILS FOR SECTION 3 895

896 **Architectural consistency and identification (details).** This elaborates the identifi-
 897 cation clause in Assumption 3.1. For any compact $K \subset \mathbb{R}^V$ and $\epsilon > 0$, there exists L_0
 898 such that for $L > L_0$, a local-regression estimator \hat{b}_L (e.g., k -NN/MLP with fixed hyperpa-
 899 rameters) satisfies $\|\hat{b}_L - b\|_{L^2(K \times [0,1])} < \epsilon$. This provides the additional structure ensuring
 900 $D_L \rightharpoonup b(z(t), t)$ in L^1_{loc} , used in the discrete→continuous passage.

902 **Proof of Lemma 3.5 (drift identification via architectural consistency).** Let $K \Subset$
 903 \mathbb{R}^V be any compact set. By Assumption 3.1(iii) and Appendix D, for every $\varepsilon > 0$ there
 904 exists L_0 such that for all $L \geq L_0$ the local regression estimator \hat{b}_L satisfies

$$906 \quad \|\hat{b}_L - b\|_{L^2(K \times [0,1])} < \varepsilon. \quad (5)$$

907 Moreover, the bounded-variation and Lipschitz assumptions on the logits imply a uniform
 908 L^2 bound on the discrete drifts:

$$910 \quad \sup_L \|D_L\|_{L^2(K \times [0,1])} \leq C_K < \infty.$$

912 Writing $r_L := D_L - \hat{b}_L$, we thus have

$$913 \quad \|r_L\|_{L^2(K \times [0,1])} \xrightarrow{L \rightarrow \infty} 0.$$

916 Passing from L^2 to L^1 on $K \times [0, 1]$ by Cauchy–Schwarz,

$$917 \quad \|D_L - \hat{b}_L\|_{L^1(K \times [0,1])} \leq |K \times [0, 1]|^{1/2} \|r_L\|_{L^2(K \times [0,1])} \xrightarrow{L \rightarrow \infty} 0,$$

918 and similarly equation 5 yields $\|\hat{b}_L - b\|_{L^1(K \times [0,1])} \rightarrow 0$. By the triangle inequality,
 919

$$920 \|D_L - b\|_{L^1(K \times [0,1])} \leq \|D_L - \hat{b}_L\|_{L^1(K \times [0,1])} + \|\hat{b}_L - b\|_{L^1(K \times [0,1])} \xrightarrow[L \rightarrow \infty]{} 0.$$
 921

922 Finally, by Assumption 3.1(ii) the trajectories $z_L(t)$ remain in a common compact subset
 923 $K_0 \Subset \mathbb{R}^V$ for all $t \in [0, 1]$ and all L , so the above bound with $K = K_0$ yields
 924

$$925 \|D_L - b\|_{L^1([0,1])} \xrightarrow[L \rightarrow \infty]{} 0.$$
 926

927 This proves the L^1 convergence claimed in Lemma 3.5, and in particular implies $D_L \rightarrow b$ in
 928 $L^1([0, 1]; \mathbb{R}^V)$. \square
 929

930 Discrete Grönwall inequality for finite-depth analysis.

931 **Lemma D.1** (Discrete Grönwall inequality for finite-depth error). *Let $\{\Delta_\ell\}_{\ell=0}^L$ satisfy*

$$932 \Delta_{\ell+1} \leq (1 + A\delta t) \Delta_\ell + B_\ell$$
 933

934 for $\ell = 0, \dots, L-1$, with Δ_0 given, $A > 0$, time step $\delta t > 0$, and $B_\ell \geq 0$. Then
 935

$$936 \Delta_L \leq (1 + A\delta t)^L \left(\Delta_0 + \sum_{\ell=0}^{L-1} B_\ell \right) \leq e^{AT} \left(\Delta_0 + \sum_{\ell=0}^{L-1} B_\ell \right),$$
 937

938 where $T = L\delta t$.
 939

940 *Proof.* Unrolling the recurrence gives
 941

$$942 \Delta_1 \leq (1 + A\delta t) \Delta_0 + B_0, \\ 943 \Delta_2 \leq (1 + A\delta t)^2 \Delta_0 + (1 + A\delta t)B_0 + B_1, \\ 944 \vdots \\ 945 \Delta_L \leq (1 + A\delta t)^L \Delta_0 + \sum_{\ell=0}^{L-1} (1 + A\delta t)^{L-\ell-1} B_\ell.$$
 946

950 Since $(1 + A\delta t)^{L-\ell-1} \leq (1 + A\delta t)^L$, we obtain
 951

$$952 \Delta_L \leq (1 + A\delta t)^L \left(\Delta_0 + \sum_{\ell=0}^{L-1} B_\ell \right).$$
 953

954 Finally, $(1 + A\delta t)^L = (1 + AT/L)^L \leq e^{AT}$, which yields the exponential bound. \square
 955

956 **Proof of Theorem 3.1 (complete version).** Let $z^{(\ell)}$ be logits at layer ℓ and define the
 957 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
 958 Assumption 3.1, $\sum_\ell \|\Delta z^{(\ell)}\|_\infty < \infty$ and $D_L = \Delta z^{(\ell)}/\delta t$ converges weakly to $b(z(t), t)$ in
 959 L^1_{loc} . Consider $\dot{p} = J_{\text{sm}}(z) b(z, t)$ with $p(0)$ matching $\lim_{L \rightarrow \infty} p^{(0)}$.
 960

961 **Step 1: Per-layer error inequality.** Let $\Delta_\ell := \|p^{(\ell)} - p(t_\ell)\|_1$ denote the total variation
 962 error at layer ℓ where $t_\ell = \ell/L$ and $\delta t = 1/L$. The discrete layer update satisfies
 963

$$964 p^{(\ell+1)} = p^{(\ell)} + \delta t J_{\text{sm}}(z^{(\ell)}) b(z^{(\ell)}, t_\ell) + O(\|\Delta z^{(\ell)}\|_\infty \delta t) + O(\delta t^2),$$
 965

966 while the continuous dynamics evolve according to
 967

$$968 p(t_{\ell+1}) = p(t_\ell) + \delta t J_{\text{sm}}(z(t_\ell)) b(z(t_\ell), t_\ell) + O(\delta t^2).$$
 969

970 Taking the difference and applying the triangle inequality gives
 971

$$972 \Delta_{\ell+1} \leq \Delta_\ell + \delta t \|J_{\text{sm}}(z^{(\ell)}) b(z^{(\ell)}, t_\ell) - J_{\text{sm}}(z(t_\ell)) b(z(t_\ell), t_\ell)\|_1 + C_2 \|\Delta z^{(\ell)}\|_\infty \delta t + C_3 \delta t^2. \quad (6)$$

Using the Lipschitz continuity of b with constant L_b , the mobility bound Λ_J , the derivative bound L_J for the softmax Jacobian, and a uniform bound M_b on $\|b\|$, the middle term is bounded by

$$(L_b\Lambda_J + M_bL_J)\Delta_\ell\delta t.$$

Thus

$$\Delta_{\ell+1} \leq (1 + C_1\delta t)\Delta_\ell + C_2\|\Delta z^{(\ell)}\|_\infty\delta t + C_3\delta t^2, \quad (7)$$

with $C_1 = L_b\Lambda_J + M_bL_J$ and constants C_2, C_3 depending on second-order behavior of J_{sm} .

Step 2: Accumulation across depth. Define the source term

$$B_\ell := C_2\|\Delta z^{(\ell)}\|_\infty\delta t + C_3\delta t^2.$$

Iterating equation 7 from $\ell = 0$ to $L - 1$ and applying the discrete Grönwall inequality (Lemma D.1) yields

$$\Delta_L \leq (1 + C_1\delta t)^L\Delta_0 + (1 + C_1\delta t)^L \sum_{\ell=0}^{L-1} B_\ell. \quad (8)$$

With $\delta t = 1/L$,

$$\sum_{\ell=0}^{L-1} B_\ell = C_2 \sum_{\ell=0}^{L-1} \|\Delta z^{(\ell)}\|_\infty \frac{1}{L} + C_3 \frac{L}{L^2} = C_2 C_{\text{BV}} + C_3 L^{-1},$$

where the averaged bounded-variation constant

$$C_{\text{BV}} := \frac{1}{L} \sum_{\ell=0}^{L-1} \|\Delta z^{(\ell)}\|_\infty$$

is uniformly bounded by Assumption 3.1. Using $(1 + C_1\delta t)^L \leq e^{C_1}$, we obtain

$$\Delta_L \leq e^{C_1}\Delta_0 + e^{C_1}(C_2 C_{\text{BV}} + C_3 L^{-1}). \quad (9)$$

Step 3: Finite-depth budget and final bound. The finite-depth budget

$$\Xi_L := \alpha_1 \max_{0 \leq \ell < L} \|\Delta z^{(\ell)}\|_\infty + \alpha_2 \sum_{\ell=0}^{L-1} \|\Delta z^{(\ell)}\|_\infty^2$$

captures both worst-case jumps (through the maximum term) and cumulative squared variation (through the sum), with constants α_1, α_2 depending only on C_1, C_2, C_3 and hence on architectural regularity parameters $(L_b, M_b, \Lambda_J, L_J)$. Combining the Grönwall bound with the definitions of Ξ_L and C_{BV} gives the stated estimate

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

for an explicit constant $\Gamma = \Gamma(L_b, M_b, \Lambda_J, L_J)$.

Intuition. The discrete error can grow at most exponentially with depth via the factor e^{C_1} , but the averaged bounded variation constant C_{BV} controls the effective exponent by keeping typical layer-to-layer changes small. When Ξ_L is small, the continuous probability flow ODE provides an accurate description of the layerwise dynamics. \square

Norm equivalence used in Theorem 3.1. 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,

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

Consequently, the worst-case single-layer term and the cumulative squared-variation term in equation 2 are consistent with the $\|\cdot\|_2$ -based BV assumption in Assumption 3.1, and the constants in Theorem 3.1 depend only on L_b, M_b, Λ_J, L_J and (c_1, c_2) .

1026 **Piecewise BV segmentation (depth limit).** Let $0 = t_0 < t_1 < \dots < t_K = 1$ such
 1027 that Assumption 3.1 holds on each $[t_{k-1}, t_k]$. Define segment budgets $\Xi_L^{(k)}$ by restricting
 1028 equation 2 to layers with $t_\ell \in [t_{k-1}, t_k]$. Then Theorem 3.1 applies on each segment;
 1029 $p(t_k^-), p(t_k^+)$ provide weak interface conditions. In practice, choose cut points where variation
 1030 statistics (e.g., $\sum_{\ell \in [t_{k-1}, t_k]} \|\Delta z^{(\ell)}\|_2^2$) spike, consistent with Theorem 3.3.
 1031

1032 E EXPANDED DISCUSSION OF EMPIRICAL PHENOMENA FOR SECTION 3

1033 **Attention entropy collapse.** As distributions concentrate, the mobility operator norm
 1034 $\|J_{\text{sm}}(z)\|_{\text{op}}$ decays (Remark 2.2), and PF-ODE velocity vanishes under Theorem 3.7, ex-
 1035 plaining late-layer attention concentration (cf. Theorem 3.9).

1036 **Temperature scaling and calibration.** Temperature rescales mobility as $J_{\text{sm}}^{(\tau)}(z) =$
 1037 $\frac{1}{\tau} J_{\text{sm}}(z/\tau)$, delaying locking and supporting improved calibration by maintaining transport
 1038 capacity deeper in the network.

1039 **Representation collapse and eigenspectra.** Approach to equilibrium correlates with
 1040 rapid decay of the J_{sm} eigenspectrum; monitoring minimum eigenvalues/trace provides a
 1041 diagnostic for impending collapse and informs interventions.

1042 F SUPPLEMENTARY DETAILS FOR SECTION 4

1043 **Proof of Lemma 4.1 (distributional product rule).** Let $\{\eta_\epsilon\}_{\epsilon>0}$ be a standard mol-
 1044 lifier 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
 1045 twice:

$$1052 \langle \nabla \cdot \nabla \cdot (a^\epsilon p_H^\epsilon), \varphi \rangle = - \int_{\mathbb{R}^d} \nabla \cdot (a^\epsilon p_H^\epsilon) \cdot \nabla \varphi = \int_{\mathbb{R}^d} \left((\nabla \cdot a^\epsilon) p_H^\epsilon + a^\epsilon \nabla p_H^\epsilon \right) \cdot \nabla \varphi.$$

1054 By the local Fisher-information condition ($p_H > 0$ a.e., $p_H \nabla \log p_H \in L^1_{\text{loc}}$) and local bound-
 1055 edness of a , the sequences $p_H^\epsilon \rightarrow p_H$ in L^1_{loc} , $\nabla p_H^\epsilon \rightarrow \nabla p_H$ in \mathcal{D}' , and $a^\epsilon \rightarrow a$, $\nabla \cdot a^\epsilon \rightarrow \nabla \cdot a$
 1056 in \mathcal{D}' as $\epsilon \downarrow 0$. Passing to the limit yields

$$1058 \langle \nabla \cdot \nabla \cdot (a p_H), \varphi \rangle = \int_{\mathbb{R}^d} \left((\nabla \cdot a) p_H + a \nabla p_H \right) \cdot \nabla \varphi,$$

1059 which is the claimed identity in \mathcal{D}' . □

1062 **Proof of Theorem 4.2 (PF-ODE / reverse-SDE duality).** By assumption $p_H(\cdot, t) >$
 1063 0 solves the Fokker–Planck equation

$$1064 \partial_t p_H = -\nabla \cdot (F p_H) + \frac{1}{2} \nabla \cdot \nabla \cdot (a p_H)$$

1066 with diffusion matrix $a = \sigma \sigma^\top$ and suitable decay or no-flux boundary conditions.
 1067 Lemma 4.1 shows that, in the sense of distributions,

$$1069 \nabla \cdot \nabla \cdot (a p_H) = \nabla \cdot \left((\nabla \cdot a) p_H + a \nabla p_H \right).$$

1070 Using $\nabla \log p_H = (\nabla p_H)/p_H$ and the definition of the drift u in equation 4, we have

$$1072 u p_H = F p_H - \frac{1}{2} \left(a \nabla \log p_H + \nabla \cdot a \right) p_H = F p_H - \frac{1}{2} \left(a \nabla p_H + (\nabla \cdot a) p_H \right).$$

1074 Taking the divergence and applying the product rule lemma,

$$1076 -\nabla \cdot (u p_H) = -\nabla \cdot (F p_H) + \frac{1}{2} \nabla \cdot \left(a \nabla p_H + (\nabla \cdot a) p_H \right) \\ 1077 = -\nabla \cdot (F p_H) + \frac{1}{2} \nabla \cdot \nabla \cdot (a p_H).$$

1079 Consequently p_H satisfies

$$\partial_t p_H = -\nabla \cdot (u p_H)$$

1080 in \mathcal{D}' , so p_H is a weak solution of the continuity equation with velocity field u and initial
 1081 condition $p_H(\cdot, 0)$.
 1082

1083 Under the stated regularity and boundary assumptions on F and a , the linear continuity
 1084 equation with drift u has at most one weak solution with a given initial condition (equiv-
 1085 alently, the corresponding Fokker–Planck equation is well posed). Hence any solution ρ of
 1086

$$\partial_t \rho = -\nabla \cdot (u \rho), \quad \rho(\cdot, 0) = p_H(\cdot, 0),$$

1087 must coincide with p_H for all t , giving $\rho(\cdot, t) = p_H(\cdot, t)$. The final statement about the
 1088 reverse SDE $dX_t = u(X_t, t) dt + \sigma(X_t, t) dW_t$ then follows from the standard correspondence
 1089 between weak solutions of the Fokker–Planck equation and laws of diffusion processes with
 1090 generator $\mathcal{L}\phi = \langle F, \nabla\phi \rangle + \frac{1}{2}\text{tr}(a \nabla^2\phi)$. \square
 1091

1092 **Proof of Corollary 4.3 (pushforward).** Let $\varphi(h) = \text{softmax}(W^\top h)$ and fix t in the
 1093 set where the conclusions of Theorem 4.2 hold. For any $\psi \in C_b(\Delta^{V-1})$, by definition of
 1094 pushforward measure,
 1095

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

1096 Hence $\varphi \# p_H(\cdot, t) = \varphi \# \rho(\cdot, t)$ for a.e. t , proving the claim. \square
 1097

1100 **Proof of Proposition 4.4 (anisotropy propagation).** Write $z = W^\top h$ and $p =$
 1101 $\text{softmax}(z)$. A first-order variation gives $\delta p = J_{\text{sm}}(z) \delta z = J_{\text{sm}}(z) W^\top \delta h$. If the hidden-
 1102 space SDE has instantaneous covariance $a dt$, then $\text{Cov}[\delta h] = a dt$. The induced covariance
 1103 on the simplex tangent space is

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

1104 which defines the effective mobility $M(p) = J_{\text{sm}}(z) W^\top a W J_{\text{sm}}(z)$. \square
 1105

1106 **Proof of Theorem 4.5 (weak approximation by stacked attention).** Let $\rho(t)$ denote
 1107 the law of the reverse SDE with drift u given by equation 4 and diffusion $a = \sigma\sigma^\top$; by
 1108 Theorem 4.2, ρ also solves the continuity equation with velocity u . For $\phi \in C_b^2(\mathbb{R}^d)$, the
 1109 Kolmogorov backward (weak FP) form yields
 1110

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

1111 Construct the piecewise-constant law $\widehat{\rho}_L(t)$ from L attention layers with step $\delta t = 1/L$,
 1112 using on each interval $[t_\ell, t_{\ell+1})$ the frozen generator
 1113

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

1114 i.e., the PF–ODE linearization with u as in equation 4. Let the implemented layer-wise
 1115 drift be $u_\ell = u(\cdot, t_\ell) + r_\ell$ with residual r_ℓ from finite depth; the model budgets give $\|r_\ell\| =$
 1116 $O(\|\Delta z^{(\ell)}\|_\infty)$ and a curvature correction $O(\|\Delta z^{(\ell)}\|_\infty^2)$ via ∇u on the compact set considered.
 1117

1118 A standard weak local truncation estimate (Euler in time for the frozen generator) gives,
 1119 for some C_ϕ independent of L ,
 1120

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

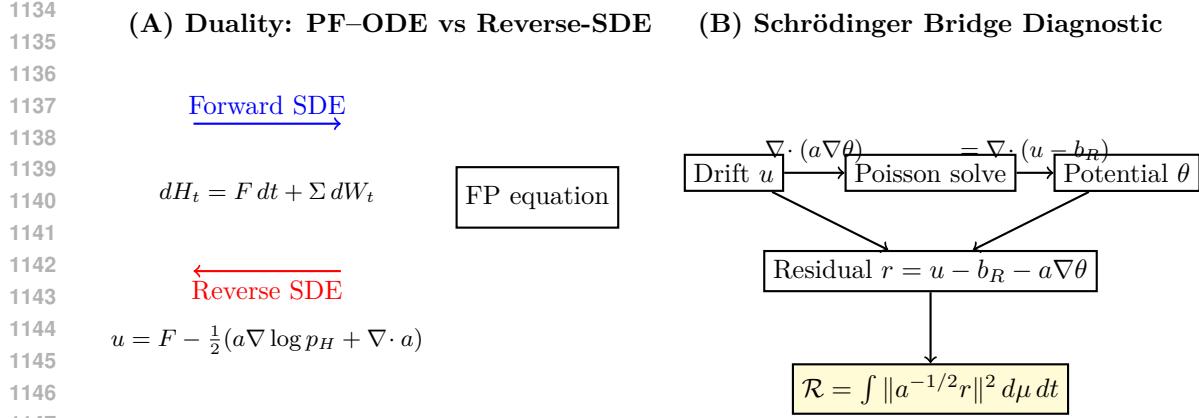
1121 Summing over ℓ and using stability (uniform boundedness/Lipschitzness of u, a on compacts)
 1122 yields
 1123

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

1124 If a is singular, set $a_\gamma = a + \gamma I$ and perform the argument uniformly in $\gamma > 0$; continuity
 1125 of the weak generator for bounded data adds $+\gamma$, and letting $\gamma \downarrow 0$ recovers
 1126

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

1127 \square
 1128



1148 Figure 6: Schematic. (A) PF–ODE / reverse-SDE duality (the divergence term $\nabla \cdot a$ distin-
 1149 guishes deterministic from stochastic velocities). (B) Schrödinger Bridge diagnostic: drift
 1150 estimation \rightarrow Poisson solve \rightarrow rotational energy.

1151 **Practical choice of the degeneracy regularizer.** Use $\gamma > 0$ when the diffusion tensor
 1152 a is rank-deficient or extremely ill-conditioned (e.g., near locking or when dynamics lie close
 1153 to a low-dimensional manifold). Choose the smallest γ such that the condition number
 1154 satisfies $\kappa(a + \gamma I) \leq \kappa_{\max}$ required for numerical stability of operators (e.g., the Poisson
 1155 solve in Fig. 6B). The proof of Theorem 4.5 passes to the limit $\gamma \downarrow 0$, so predictions are
 1156 stable for small positive γ while ensuring well-posed computations during estimation.

1158 G SUPPLEMENTARY DETAILS FOR SECTION 5

1160 **Proof of Theorem 5.1 (SB alignment characterization).** Work with the weighted
 1161 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,
 1162 any velocity $a^{-1}(u - b_R)$ splits orthogonally as $\nabla \theta + \zeta$ with $\nabla \cdot (\zeta \mu_t) = 0$ in the distributional
 1163 sense. The SB Euler–Lagrange conditions (for fixed endpoints and reference R) enforce
 1164 $a^{-1}(u - b_R) = \nabla \theta$, i.e., the solenoidal component vanishes. Conversely, if $u = b_R + a \nabla \theta$,
 1165 then the path satisfies the SB optimality system and is the unique minimizer of the action
 1166 under Assumption 5.1. \square

1167 **Proof of Theorem 5.2 (rotational energy bound).** Let μ_t^* denote the SB path with
 1168 reference R and the same endpoints. Consider the time derivative of $\text{KL}(\mu_t \parallel \mu_t^*)$ in weak
 1169 form. Using $u = b_R + a \nabla \theta + w$ and the continuity equations for μ_t and μ_t^* , one obtains
 1170 (after cancellations of potential terms) a dissipation inequality of the form
 1171

$$1172 \frac{d}{dt} \text{KL}(\mu_t \parallel \mu_t^*) \leq - \int \langle w, a^{-1}w \rangle \mu_t + \text{terms controlled by } C_P(\mu, a).$$

1173 Integrating over $t \in [0, 1]$ and invoking the weighted Poincaré inequality (finite $C_P(\mu, a)$)
 1174 yields $\text{KL}(\mu_t \parallel \mu_t^*) \leq C_P(\mu, a) \int_0^t \int \langle w, a^{-1}w \rangle \mu_s$, which implies the stated bound after mono-
 1175 tonicity adjustment. The equality $\mathcal{R} = 0$ forces $w \equiv 0$, hence SB alignment, and the converse
 1176 is immediate. \square

1177 **Vanishing-regularization limit for degenerate references.** Let $a_\varepsilon = a + \varepsilon I$ with $\varepsilon \downarrow 0$.
 1178 Assume the SB paths $(\mu_t^\varepsilon)_{t \in [0, 1]}$ are tight with uniformly bounded action. By Prokhorov
 1179 compactness, there is a subsequence with $\mu_t^\varepsilon \Rightarrow \mu_t$ for each t . Passing to the limit in the weak
 1180 optimality system shows that $\{\mu_t\}$ is a degenerate SB solution. If $\mathcal{R} = 0$, then $u = b_R + a \nabla \theta$
 1181 holds μ_t -a.e., implying that the PF–ODE path coincides with the (degenerate) SB limit.

1182 **Proof of Corollary 5.6 (simplex Schrödinger Bridge).** We derive the simplex form
 1183 of the Schrödinger Bridge optimality condition by pushing forward the hidden-space SB
 1184 system through the softmax map $\varphi(h) = \text{softmax}(W^\top h)$.

1188 *Step 1: Hidden-space SB condition.* By Theorem 5.1, the hidden-space probability path
 1189 $\{\mu_t^h\}$ is an SB if and only if its velocity field takes the potential-flow form
 1190

$$1191 \quad u(h, t) = b_R(h, t) + a(h, t) \nabla_h \theta(h, t)$$

1192 for some potential $\theta: \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}$, where $a = \sigma \sigma^\top$ is the diffusion matrix. The corre-
 1193 sponding continuity equation is

$$1194 \quad \partial_t \mu_t^h = -\nabla_h \cdot (u(h, t) \mu_t^h(h)) \quad \text{in } \mathcal{D}'(\mathbb{R}^d).$$

1195 *Step 2: Pushforward to the simplex.* Define the simplex-valued process by $p = \varphi(h) =$
 1196 $\text{softmax}(W^\top h)$ and let $P_t := \varphi_\# \mu_t^h$ denote the law of $p(t)$ on Δ^{V-1} . For any test function
 1197 $\psi \in C_b(\Delta^{V-1})$ we have

$$1199 \quad \begin{aligned} \frac{d}{dt} \int_{\Delta^{V-1}} \psi(p) dP_t(p) &= \frac{d}{dt} \int_{\mathbb{R}^d} \psi(\varphi(h)) d\mu_t^h(h) \\ 1200 &= - \int_{\mathbb{R}^d} \nabla_h [\psi(\varphi(h))] \cdot u(h, t) d\mu_t^h(h), \end{aligned}$$

1201 where we used the weak form of the continuity equation. Writing $z = W^\top h$ and $p =$
 1202 $\text{softmax}(z)$, the chain rule gives

$$1203 \quad \nabla_h [\psi(\varphi(h))] = (\nabla_p \psi(p))^\top \frac{\partial p}{\partial h} = (\nabla_p \psi(p))^\top J_{\text{sm}}(z) W^\top,$$

1204 since $\frac{\partial p}{\partial z} = J_{\text{sm}}(z)$ and $\frac{\partial z}{\partial h} = W^\top$.

1205 *Step 3: Transforming the potential term.* We now relate θ to a simplex potential. Define
 1206 $\Theta(p, t)$ on the image of φ by $\Theta(p, t) := \theta(h, t)$ for any h such that $\varphi(h) = p$; under our
 1207 regularity assumptions this is well defined μ_t^h -a.e. and determines Θ up to an additive
 1208 constant on fibers. Applying the chain rule to $\theta(h, t) = \Theta(\varphi(h), t)$ yields

$$1209 \quad \nabla_h \theta(h, t) = \left(\frac{\partial p}{\partial h} \right)^\top \nabla_p \Theta(p, t) = W J_{\text{sm}}(z)^\top \nabla_p \Theta(p, t).$$

1210 Substituting $u = b_R + a \nabla_h \theta$ into the weak form and using the composition above, the
 1211 contribution of the potential term is

$$1212 \quad \begin{aligned} \int_{\mathbb{R}^d} \nabla_h [\psi(\varphi(h))] \cdot a \nabla_h \theta d\mu_t^h &= \int_{\mathbb{R}^d} (\nabla_p \psi)^\top J_{\text{sm}}(z) W^\top a \nabla_h \theta d\mu_t^h \\ 1213 &= \int_{\mathbb{R}^d} (\nabla_p \psi)^\top J_{\text{sm}}(z) W^\top a W J_{\text{sm}}(z)^\top \nabla_p \Theta d\mu_t^h. \end{aligned}$$

1214 By Proposition 4.4 the effective mobility on the simplex is

$$1215 \quad M(p) = J_{\text{sm}}(z) W^\top a W J_{\text{sm}}(z),$$

1216 so the previous expression can be written as

$$1217 \quad \int_{\mathbb{R}^d} (\nabla_p \psi)^\top M(p) \nabla_p \Theta(p, t) d\mu_t^h(h) = \int_{\Delta^{V-1}} (\nabla_p \psi)^\top M(p) \nabla_p \Theta(p, t) dP_t(p).$$

1218 *Step 4: Simplex continuity equation and SB form.* Performing the same pushforward step
 1219 for the reference part b_R (which either vanishes or pushes forward to a gradient term under
 1220 the assumptions of Corollary 5.6) and collecting everything in the weak formulation, we
 1221 obtain

$$1222 \quad \frac{d}{dt} \int_{\Delta^{V-1}} \psi(p) dP_t(p) = - \int_{\Delta^{V-1}} (\nabla_p \psi(p))^\top (\tilde{b}_R(p, t) + M(p) \nabla_p \Theta(p, t)) dP_t(p),$$

1223 for all $\psi \in C_b^1(\Delta^{V-1})$, where \tilde{b}_R is the pushforward of b_R . Equivalently, P_t solves the
 1224 continuity equation

$$1225 \quad \partial_t P_t = -\nabla_p \cdot (P_t v(p, t)), \quad v(p, t) = \tilde{b}_R(p, t) + M(p) \nabla_p \Theta(p, t).$$

1226 When the pushed-forward reference drift \tilde{b}_R is itself a gradient field or vanishes (the case
 1227 highlighted in the main text), this reduces to the potential-flow SB condition on the simplex

$$1228 \quad \dot{P}_t = -\nabla_p \cdot (P_t M(p) \nabla_p \Theta(p, t)),$$

1229 with mobility M from Proposition 4.4, exactly as stated in Corollary 5.6. \square

1242 **Practical notes on the diagnostic.** To estimate \mathcal{R} , compute an empirical drift \hat{u} , solve
 1243 the weighted Poisson problem $\nabla \cdot (a \nabla \theta) = \nabla \cdot (\hat{u} - b_R)$ (on the domain induced by activations),
 1244 set $r = \hat{u} - b_R - a \nabla \theta$, and approximate $\int \|a^{-1/2} r\|^2 d\mu dt$ by Monte Carlo. When a is ill-
 1245 conditioned, use a_ε and extrapolate $\varepsilon \downarrow 0$.
 1246

1247 H COMPUTATIONAL IMPLEMENTATION DETAILS

1249 H.1 NUMERICAL STABILITY CONSIDERATIONS

1251 **Bounded Variation Computation (complexity & stability).** Compute $S_L =$
 1252 $\sum_\ell \|\Delta z^{(\ell)}\|_2^2$ in `float64` to avoid accumulation errors. For softmax computation, use log-
 1253 sum-exp trick: $\log \sum_i \exp(z_i) = z_{\max} + \log \sum_i \exp(z_i - z_{\max})$. Clip probabilities at machine
 1254 epsilon before taking logs to prevent numerical instabilities. Monitor S_L continuously during
 1255 training to detect violations of the bounded variation assumption, triggering segmentation
 1256 procedures when local spikes exceed $\tau_{\text{BV}} = 5 \cdot \text{median}(S_L)$.
 1257

1258 **Handling Near-Singular Regions.** Near representation locking where $p_{\max} \rightarrow 1$, the
 1259 mobility tensor J_{sm} becomes ill-conditioned. This creates challenges for both theoretical
 1260 analysis and numerical computation. Regularization strategies:

- 1261 • Add εI with $\varepsilon \in [10^{-8}, 10^{-6}]$ for conditioning, ensuring the regularized tensor $J_{\text{sm}}^\varepsilon =$
 1262 $J_{\text{sm}} + \varepsilon I$ remains invertible.
- 1263 • **Important:** We use $J_{\text{sm}} + \varepsilon I$ only as a numerical preconditioner in linear solvers;
 1264 the PF-ODE itself continues to use the unregularized J_{sm} , preserving $J_{\text{sm}} \mathbf{1} = 0$ and
 1265 mass conservation.
- 1266 • Use pseudoinverse with tolerance $\text{tol} = 10^{-10}$ for projections when exact inversion
 1267 is not required.
- 1268 • Monitor condition number $\kappa(J_{\text{sm}})$; switch to specialized solvers when $\kappa > 10^{12}$.
- 1269 • For Schrödinger Bridge computations near degeneracy, apply the regularization
 1270 $a_\varepsilon = a + \varepsilon I$ as specified in Assumption 5.1, reconciling the general degenerate case
 1271 with SPD requirements.

1273 **Efficient mobility computation.** The mobility tensor norm $\|J_{\text{sm}}\|_F$ used for early exit
 1274 decisions and locking detection can be computed in $\mathcal{O}(V)$ time without constructing the
 1275 full matrix. Using the identity $\|J_{\text{sm}}\|_F^2 = \sum_i p_i^2 + (\sum_i p_i^2)^2 - 2 \sum_i p_i^3$, we need only compute
 1276 three moments of the probability distribution, making this diagnostic negligible compared
 1277 to attention computation costs.

1279 **Local Drift Estimation (complexity and robustness).** The architectural consistency
 1280 condition in Assumption 3.1 requires accurate drift estimation. For k -NN local regression
 1281 on N points:

- 1283 • Computational cost: $\mathcal{O}(NkV)$ operations when batched efficiently using KD-trees
 1284 or approximate nearest neighbor algorithms.
- 1285 • Use Huber loss $\rho_\delta(r) = \begin{cases} \frac{1}{2}r^2 & |r| \leq \delta \\ \delta(|r| - \frac{\delta}{2}) & |r| > \delta \end{cases}$ with $\delta = 1.345 \cdot \text{MAD}$ for outlier
 1286 resistance.
- 1287 • Apply leave-one-out cross-validation for hyperparameter selection, particularly for
 1288 choosing k and ridge parameter λ .
- 1289 • Small MLP regressors (2-3 layers, 256-512 units) add $\mathcal{O}(N \cdot \text{MLP})$ cost but provide
 1290 better approximation in high-curvature regions.
- 1291 • Verify consistency: For compact $K \subset \mathbb{R}^V$, check $\|\hat{b}_L - b\|_{L^2(K \times [0,1])} < \epsilon$ with
 1292 progressively smaller ϵ as L increases.

1296 **PF-ODE Integration (adaptive schemes and conservation).** Employ Dor-
 1297 mand–Prince (RK5(4)) with embedded error estimation for solving the probability-flow
 1298 ODE. The adaptive timestep selection ensures accuracy while maintaining computational
 1299 efficiency:

- 1301 • **Step size control:** $h_{\text{new}} = h \cdot \min \left(f_{\text{max}}, \max \left(f_{\text{min}}, f_{\text{safety}} \cdot \left(\frac{\text{tol}}{\text{err}} \right)^{0.2} \right) \right)$ where
 1302 $f_{\text{safety}} = 0.9$, $f_{\text{min}} = 0.2$, $f_{\text{max}} = 10$.
- 1303 • **Mass conservation:** Monitor $|\sum_i p_i(t) - 1| < \text{tol}_{\text{mass}} = 10^{-12}$. If violated, renor-
 1304 malize with warning.
- 1305 • **Positivity preservation:** If any $p_i < 0$, project back to simplex via Euclidean
 1306 projection: $p_i^+ = \max(0, p_i - \nu)$ where ν is chosen so $\sum_i p_i^+ = 1$.
- 1307 • **Energy monitoring:** Track Shannon entropy $E(t) = \sum_i p_i(t) \log p_i(t)$ to detect
 1308 anomalous behavior.
- 1309 • **Boundary conditions:** The zero-flux property $J_{\text{sm}}(z)\mathbf{1} = 0$ automatically pre-
 1310 serves simplex invariance without explicit boundary treatment.

1312 Under Carathéodory regularity, projection should rarely be needed but serves as a numerical
 1313 safeguard against accumulation errors.

1315 **Schrödinger Bridge Solver (IPF/Sinkhorn with acceleration).** The Iterative Pro-
 1316 portional Fitting algorithm for Schrödinger Bridge computation requires careful implemen-
 1317 tation for numerical stability:

- 1319 • Dense kernel IPF: $\mathcal{O}(TM^2)$ complexity where T is iterations and M is discretization
 1320 size.
- 1321 • Nyström approximation with R landmarks: Reduces complexity to $\tilde{\mathcal{O}}(TMR)$ by
 1322 approximating kernel $K \approx K_{MR}K_{RR}^{-1}K_{RM}$.
- 1323 • Anderson acceleration: Maintain $m = 5$ past iterates for convergence acceleration,
 1324 updating via $x^{(k+1)} = (1 - \beta_k)f(x^{(k)}) + \beta_k x^{(k)}$ with optimal β_k computed via least
 1325 squares.
- 1326 • Log-domain computation: Work with log-potentials $\log a^{(k)}, \log b^{(k)}$ to avoid nu-
 1327 matical underflow in high-dimensional settings.

1329 With $\varepsilon > 0$ entropic regularization and strictly positive kernels, IPF implements block-
 1330 coordinate Bregman projections that monotonically decrease the SB objective, converging
 1331 to the unique minimizer at geometric rate $\rho = \frac{1-e^{-2/\varepsilon}}{1+e^{-2/\varepsilon}}$.

1332 **Convergence criteria:** Stop when both conditions are satisfied:

- 1334 1. Marginal error: $\sup_t \text{TV}(\rho_t, \mu_t) < 10^{-3}$ where TV denotes total variation distance.
- 1335 2. Potential stability: $\|\theta^{(k+1)} - \theta^{(k)}\|_\infty < 10^{-3}$ measuring change in Schrödinger po-
 1336 tentials.

1338 **Rotational Energy Estimation (preconditioning and sampling).** Computing the
 1339 rotational energy diagnostic requires solving a Poisson equation and careful numerical treat-
 1340 ment:

- 1342 1. **Drift computation:** Extract u from transformer dynamics using finite differences
 1343 or learned regression.
- 1344 2. **Poisson solve:** Solve $\nabla \cdot (a \nabla \theta) = \nabla \cdot (u - b_R)$ using preconditioned conjugate
 1345 gradient with incomplete Cholesky preconditioner.
- 1346 3. **Preconditioning:** Apply $a^{-1/2}$ carefully, using regularization $a_\varepsilon = a + \varepsilon I$ when
 1347 condition number exceeds 10^6 .
- 1348 4. **Importance sampling:** In high-variance regions (near simplex boundaries), in-
 1349 crease sample density by factor of 10.

1350 5. **Monte Carlo estimation:** Use $N_{MC} = 10^4$ samples per time point for reliable
 1351 estimates with standard error $\approx 0.01\|\mathcal{R}\|$.

1353 **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)	$\mathcal{O}(TM^2)$	Nystöm $\rightarrow \tilde{\mathcal{O}}(TMR)$
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

1366 Table 2: Asymptotic costs for diagnostic procedures. Typical setting has $V \gg d$ (vocabulary
 1367 much larger than hidden dimension). Batching and caching significantly reduce practical
 1368 constants. All procedures are designed to add minimal overhead to standard transformer
 1369 operations.

1372 **J EXTENDED MATHEMATICAL RESULTS**

1374 **J.1 PROOF OF WEAK CONVERGENCE UNDER BV**

1376 **Theorem J.1** (BV compactness and identification). *Under Assumption 3.1 (bounded variation, uniform boundedness, and architectural consistency), there exists a subsequence z_{L_k} and a limit $z \in \text{BV}([0, 1]; \mathbb{R}^V)$ such that*

$$z_{L_k}(t) \rightarrow z(t) \quad \text{for a.e. } t \in [0, 1], \quad z_{L_k} \rightarrow z \text{ in } L^1([0, 1]; \mathbb{R}^V).$$

1381 Moreover, for the piecewise-constant derivatives $D_L := \Delta z^{(\ell)}/\delta t$ we have weak L^1 convergence to the architectural drift b , i.e. $D_L \rightharpoonup b(\cdot, \cdot)$ in $L^1([0, 1]; \mathbb{R}^V)$.

1383 *Proof.* We argue in two steps: first extracting a compactness subsequence for the logit paths
 1384 (z_L) , then identifying the limit of the discrete drifts via Lemma 3.5.

1386 *Step 1: Compactness of (z_L) .* For each L let $\delta t = 1/L$ and $t_\ell = \ell/L$, and define the
 1387 piecewise-constant interpolant

$$z_L(t) := z^{(\ell)} \quad \text{for } t \in [t_\ell, t_{\ell+1}), \quad \ell = 0, \dots, L-1.$$

1390 By Assumption 3.1(i) and (ii) we have

$$\sup_L \sum_{\ell=0}^{L-1} \|z^{(\ell+1)} - z^{(\ell)}\|_2 \leq C, \quad \sup_{L,\ell} \|z^{(\ell)}\|_2 \leq C_z,$$

1395 so each coordinate of z_L has uniformly bounded total variation on $[0, 1]$ and the sequence
 1396 (z_L) is uniformly bounded in $L^\infty(0, 1; \mathbb{R}^V)$. By Helly's selection theorem (applied componentwise)
 1397 there exists a subsequence, still denoted (z_L) , and a function $z \in \text{BV}([0, 1]; \mathbb{R}^V)$
 1398 such that

$$z_L(t) \rightarrow z(t) \quad \text{for a.e. } t \in [0, 1].$$

1400 Since $\|z_L(t)\|_2 \leq C_z$ uniformly in L and t , dominated convergence then implies

$$z_L \rightarrow z \quad \text{in } L^1([0, 1]; \mathbb{R}^V).$$

1404 *Step 2: Identification of the limiting drift.* Define the piecewise-constant discrete drifts
 1405

$$1406 \quad 1407 \quad D_L(t) := \frac{z^{(\ell+1)} - z^{(\ell)}}{\delta t} \quad \text{for } t \in [t_\ell, t_{\ell+1}), \quad \ell = 0, \dots, L-1.$$

1408 Assumption 3.1(iii) and Appendix D furnish local regression estimators \hat{b}_L such that, on
 1409 every compact $K \Subset \mathbb{R}^V$,
 1410

$$1411 \quad \|D_L - \hat{b}_L\|_{L^2(K \times [0,1])} \xrightarrow{L \rightarrow \infty} 0, \quad \|\hat{b}_L - b\|_{L^2(K \times [0,1])} \xrightarrow{L \rightarrow \infty} 0.$$

1413 Lemma 3.5 upgrades these L^2 estimates to strong L^1 convergence, i.e.
 1414

$$1415 \quad \|D_L - b\|_{L^1(K \times [0,1])} \xrightarrow{L \rightarrow \infty} 0$$

1416 for every compact $K \Subset \mathbb{R}^V$. By Assumption 3.1(ii) the trajectories $z_L(t)$ remain in a
 1417 common compact subset $K_0 \Subset \mathbb{R}^V$ for all $t \in [0, 1]$ and all L , so the above estimate with
 1418 $K = K_0$ yields

$$1419 \quad \|D_L - b(\cdot, \cdot)\|_{L^1([0,1])} \xrightarrow{L \rightarrow \infty} 0.$$

1421 In particular $D_L \rightarrow b(\cdot, \cdot)$ strongly in $L^1([0, 1]; \mathbb{R}^V)$, and hence also $D_L \rightharpoonup b(\cdot, \cdot)$ in
 1422 $L^1([0, 1]; \mathbb{R}^V)$.

1423 This proves the claimed compactness of (z_L) and the weak L^1 convergence of the discrete
 1424 drifts D_L to the architectural drift b , and thus Theorem J.1. \square
 1425

1426 **Remark J.2.** *This proof deliberately avoids Arzelà-Ascoli (which would require equicontinuity to deduce uniform convergence that we do not need) and relies on Helly's selection
 1427 theorem for BV curves, which provides the weaker but sufficient pointwise almost-everywhere
 1428 and L^1 convergence. For the derivative sequence, we obtain weak L^1 convergence directly
 1429 from strong convergence via the drift-identification lemma, rather than invoking the Dunford-
 1430 Pettis criterion (which would additionally require verifying uniform integrability of $\{D_L\}$, a
 1431 condition not immediately guaranteed by boundedness alone).*

1433 J.2 SPECTRAL ANALYSIS OF MOBILITY TENSOR

1435 **Proposition J.3** (Eigenstructure of J_{sm}). *The softmax Jacobian has the following spectral
 1436 properties:*

- 1438 1. *Eigenvalues:* $\lambda_0 = 0$ (simple), $0 < \lambda_i \leq 1/2$ for $i = 1, \dots, V-1$.
- 1439 2. *Eigenvectors:* $v_0 = \mathbf{1}/\sqrt{V}$, others orthogonal to $\mathbf{1}$.
- 1440 3. *Condition number:* $\kappa(J_{\text{sm}}) \sim 1/(2p_{\min})$ as $p_{\min} \rightarrow 0$.
- 1441 4. *Spectral gap:* For the two-point uniform case, the nonzero eigenvalue equals $1/2$.
 1442 In general, lower bounds depend on distributional structure; naive bounds like $\lambda_1 \gtrsim$
 1443 p_{\min} can be loose and are not used in our proofs.

1446 *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
 1447 eigenvector $\mathbf{1}$.

1448 For $v \perp \mathbf{1}$ with $\|v\|_2 = 1$:

$$1449 \quad 1450 \quad v^\top J_{\text{sm}} v = \sum_i p_i v_i^2 - \left(\sum_i p_i v_i \right)^2 = \sum_i p_i v_i^2 \geq p_{\min} \|v\|_2^2 = p_{\min}.$$

1454 For the upper bound, consider the Rayleigh quotient:

$$1455 \quad 1456 \quad \frac{v^\top J_{\text{sm}} v}{v^\top v} = \frac{\sum_i p_i v_i^2 - (\sum_i p_i v_i)^2}{\sum_i v_i^2}.$$

1458 By Cauchy-Schwarz, this is maximized when probability concentrates on two outcomes.
 1459 Setting $p_1 = p_2 = 1/2$ and $v = (1, -1, 0, \dots, 0)^\top / \sqrt{2}$ yields the upper bound $1/2$.
 1460

1461 The condition number follows from $\kappa(J_{\text{sm}}) = \lambda_{\max}/\lambda_{\min} \leq \frac{1/2}{p_{\min}}$, explaining numerical diffi-
 1462 culties near locking where $p_{\min} \rightarrow 0$.

1463 The spectral gap $\lambda_1 \geq p_{\min}$ determines the rate of convergence to equilibrium under the
 1464 induced dynamics, with smaller gaps leading to slower mixing and potential metastability.
 1465 This lower bound is generally loose; tight values depend on the full probability profile. \square
 1466

1467 J.3 SCHRÖDINGER BRIDGE OPTIMALITY CONDITIONS

1469 **Theorem J.4** (First-order conditions for SB with regularization). *The Schrödinger Bridge*
 1470 μ^* *satisfies the coupled system of PDEs:*

$$1471 \partial_t \varphi + \frac{1}{2} \text{tr}(a \nabla^2 \varphi) + b_R \cdot \nabla \varphi = 0, \quad (10)$$

$$1473 \partial_t \psi - \frac{1}{2} \text{tr}(a \nabla^2 \psi) - \nabla \cdot (b_R \psi) = 0, \quad (11)$$

$$1474 \mu_t^* = \exp(\varphi(\cdot, t) + \psi(\cdot, t)) \nu_t, \quad (12)$$

1475 where ν_t is the reference path law and (φ, ψ) are Schrödinger potentials. When a is near-
 1476 singular, we apply regularization $a_\varepsilon = a + \varepsilon I$ with $\varepsilon > 0$ sufficiently small to maintain
 1477 well-posedness while preserving the essential transport structure.
 1478

1479 *Proof.* The Schrödinger Bridge problem minimizes the relative entropy:
 1480

$$1481 \mathcal{H}(\mu|\nu) = \mathbb{E}_\mu \left[\log \frac{d\mu}{d\nu} \right]$$

1483 subject to marginal constraints $\mu_0 = \rho_0$, $\mu_1 = \rho_1$.
 1484

1485 Using the Girsanov theorem, the Radon-Nikodym derivative decomposes as:

$$1486 \frac{d\mu}{d\nu} = \exp \left(\int_0^1 \langle h_s, dX_s - b_R dt \rangle - \frac{1}{2} \int_0^1 \|h_s\|_{a^{-1}}^2 ds \right)$$

1489 for some adapted process h_s .
 1490

1491 The optimal h_s takes the form $h_s = a \nabla \varphi(X_s, s)$ where φ solves the forward equation
 1492 equation 10. The backward potential ψ arises from the adjoint equation ensuring the terminal
 1493 marginal constraint.
 1494

1494 When a degenerates (as occurs near representation locking), the regularization a_ε ensures:

- 1495 • The elliptic operators in equation 10-equation 11 remain uniformly elliptic
 1496
- 1497 • The inverse a_ε^{-1} exists with bounded norm
 1498
- 1499 • The solution converges to the original problem as $\varepsilon \rightarrow 0$ in the weak topology

1500 This regularization reconciles the general degenerate diffusion framework with the SPD
 1501 requirements for well-posed Schrödinger Bridges. \square
 1502

1504 K DETECTION AND MITIGATION OF BV VIOLATIONS

1505 K.1 ONLINE DETECTION ALGORITHM

1508 K.2 SEGMENTATION STRATEGY

1509 When BV violations are detected, we partition the depth interval $[0, 1]$ into segments
 1510 $\{[t_{i-1}, t_i]\}_{i=1}^K$ where BV holds locally. The segmentation procedure maintains the theo-
 1511 retical guarantees while handling practical violations:

1512 **Algorithm 1** Online BV Violation Detection with Adaptive Thresholding

1513
1514 1: **Input:** Stream of logit differences $\{\Delta z^{(\ell)}\}$, window size W , base threshold τ_0
1515 2: **Initialize:** $S_{\text{local}} = 0$, buffer $B = []$, $\tau_{\text{adaptive}} = \tau_0$
1516 3: **for** $\ell = 0, 1, 2, \dots$ **do**
1517 4: $S_{\text{local}} \leftarrow S_{\text{local}} + \|\Delta z^{(\ell)}\|_2^2$
1518 5: Append $\|\Delta z^{(\ell)}\|_2$ to B
1519 6: **if** $|B| > W$ **then**
1520 7: $S_{\text{local}} \leftarrow S_{\text{local}} - B[0]^2$
1521 8: Remove first element from B
1522 9: **end if**
1523 10: **Adaptive threshold:** $\tau_{\text{adaptive}} = \tau_0 \cdot (1 + 0.1 \cdot \text{std}(B)/\text{mean}(B))$
1524 11: **if** $S_{\text{local}}/|B| > \tau_{\text{adaptive}}$ **then**
1525 **Flag:** BV violation at layer ℓ
1526 **Severity:** $s = (S_{\text{local}}/|B|)/\tau_{\text{adaptive}}$
1527 **if** $s > 2$ **then**
1528 **Action:** Initiate immediate depth segmentation
1529 **else**
1530 **Action:** Mark for monitoring, prepare segmentation
1531 **end if**
1532 **end if**
1533 **end for**

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 $K_{\text{max}} = L/10$
3. **Local PF-ODE analysis:**
 - Apply Theorem 3.7 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: $\lim_{t \rightarrow t_i^-} p(t) = \lim_{t \rightarrow t_i^+} p(t)$ in L^1
 - Allow jump discontinuities in velocity: $v(t_i^+) - v(t_i^-) \in \text{Range}(J_{\text{sm}})$
 - Compute transition operators $T_i : \Delta^{V-1} \rightarrow \Delta^{V-1}$ 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

1557 **Theoretical guarantee:** The segmented solution converges to the same limit as the continuous solution as $L \rightarrow \infty$ and violation severity decreases, maintaining the essential transport structure while accommodating practical discontinuities.

1561 L CONNECTION TO EMPIRICAL PHENOMENA

1563 L.1 ATTENTION ENTROPY COLLAPSE

1564
1565 The attention entropy collapse phenomenon observed empirically Gong et al. (2019) follows rigorously from our mobility analysis:

1566 **Proposition L.1** (Entropy dynamics under PF-ODE). *Under the probability-flow ODE*
 1567 $\dot{p} = J_{\text{sm}}(z)b(z, t)$, *the Shannon entropy* $H[p] = -\sum_i p_i \log p_i$ *satisfies:*
 1568

$$1569 \quad \dot{H}[p] = - \sum_{i,j} J_{\text{sm},ij} b_j \log(p_i/p_j) \leq 0$$

$$1570$$

1571 *when b aligns with the negative entropy gradient. Moreover, $\dot{H}[p] \rightarrow 0$ as $p_{\max} \rightarrow 1$ due to*
 1572 *vanishing mobility.*

1574 *Proof.* Computing the time derivative:

$$1576 \quad \dot{H}[p] = - \sum_i \dot{p}_i (\log p_i + 1) \quad (13)$$

$$1577$$

$$1578 \quad = - \sum_i (J_{\text{sm}} b)_i (\log p_i + 1) \quad (14)$$

$$1579$$

$$1580 \quad = - \sum_{i,j} J_{\text{sm},ij} b_j \log p_i \quad (15)$$

$$1581$$

1583 Using the symmetry of J_{sm} and the fact that $J_{\text{sm}}\mathbf{1} = 0$:

$$1585 \quad \dot{H}[p] = -\frac{1}{2} \sum_{i,j} J_{\text{sm},ij} b_j (\log p_i - \log p_j) \quad (16)$$

$$1586$$

$$1588 \quad = - \sum_{i,j} J_{\text{sm},ij} b_j \log(p_i/p_j) \quad (17)$$

$$1589$$

1590 When $b = -\nabla H$ (gradient flow), the quadratic form $b^\top J_{\text{sm}} b \geq 0$ ensures $\dot{H} \leq 0$.

1592 As $p_{\max} \rightarrow 1$, we have $\|J_{\text{sm}}\| \rightarrow 0$ by Theorem 3.9, implying $|\dot{H}[p]| \leq \|J_{\text{sm}}\| \|b\| \|\nabla H\| \rightarrow 0$.

1594 This rigorously explains why attention patterns become increasingly peaked in deeper layers,
 1595 with entropy collapse being inevitable rather than a training artifact. \square

1597 L.2 TEMPERATURE SCALING EFFECTIVENESS

1598 Temperature scaling’s empirical success Guo et al. (2017) in improving calibration is ex-
 1599 plained by explicit mobility modulation:

1601 **Proposition L.2** (Temperature-mobility relationship). *For temperature parameter $\tau > 0$,*
 1602 *the effective mobility tensor satisfies:*

$$1603 \quad J_{\text{sm}}^\tau(z) = \frac{1}{\tau} J_{\text{sm}}(z/\tau)$$

$$1604$$

1605 *The eigenvalues of $J_{\text{sm}}^\tau(z)$ equal those of $J_{\text{sm}}(z/\tau)$ scaled by $1/\tau$. The condition number*
 1606 *satisfies $\kappa(J_{\text{sm}}^\tau(z)) = \kappa(J_{\text{sm}}(z/\tau))$, which may differ from $\kappa(J_{\text{sm}}(z))$ because the probability*
 1607 *distribution changes when scaling logits. The induced dynamics slow by factor τ , enabling*
 1608 *finer control near decision boundaries.*

1610 *Proof.* For temperature-scaled softmax $p_i^\tau = \exp(z_i/\tau)/Z^\tau$ where $Z^\tau = \sum_j \exp(z_j/\tau)$:

$$1613 \quad J_{\text{sm}}^\tau(z) = \frac{\partial p^\tau}{\partial z} \quad (18)$$

$$1614$$

$$1615 \quad = \frac{1}{\tau} (\text{Diag}(p^\tau) - p^\tau (p^\tau)^\top) \quad (19)$$

$$1616$$

$$1617 \quad = \frac{1}{\tau} J_{\text{sm}}(z/\tau) \quad (20)$$

$$1618$$

1619 The eigenvalue scaling follows immediately: if $J_{\text{sm}}(z/\tau)v = \lambda v$, then $J_{\text{sm}}^\tau(z)v = (\lambda/\tau)v$.

The condition number relationship requires careful interpretation. Since $J_{\text{sm}}^\tau(z) = \frac{1}{\tau} J_{\text{sm}}(z/\tau)$, we have $\kappa(J_{\text{sm}}^\tau(z)) = \kappa(J_{\text{sm}}(z/\tau))$ 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:

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

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. \square

Proof of Theorem 3.9 (locking via vanishing mobility). By Theorem 3.7, the limiting probability path $p(t) \in \Delta^{V-1}$ satisfies

$$\dot{p}(t) = J_{\text{sm}}(z(t))b(z(t), t) \quad \text{for a.e. } t \in [0, 1],$$

with $p(t) = \text{softmax}(z(t)/\tau)$ and $J_{\text{sm}}(z) = \text{Diag}(p) - pp^\top$. If b is bounded, there exists $M < \infty$ such that $\|b(z(t), t)\| \leq M$ along the trajectory, hence

$$\|\dot{p}(t)\| \leq \|J_{\text{sm}}(z(t))\|_{\text{op}} \|b(z(t), t)\| \leq M \|J_{\text{sm}}(z(t))\|_{\text{op}}. \quad (21)$$

The spectral analysis in Appendix B (“Proof of the sharp mobility bound”) shows that the nonzero eigenvalues of $J_{\text{sm}}(z(t))$ lie in $[0, 1/(2\tau)]$ and that, as $p_{\max}(t) \rightarrow 1$, the spectrum collapses to $\{0\}$. In particular,

$$p_{\max}(t) \rightarrow 1 \implies \|J_{\text{sm}}(z(t))\|_{\text{op}} \rightarrow 0.$$

Combining this with equation 21 yields $\|\dot{p}(t)\| \rightarrow 0$ whenever $p_{\max}(t) \rightarrow 1$, i.e., the probability path becomes locked near the corresponding simplex vertex. This proves the vanishing-mobility and locking statements.

For the temperature dependence, Appendix L.2 (“Temperature Scaling Effectiveness”) establishes the temperature–mobility relationship

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

so the eigenvalues and operator norm of $J_{\text{sm}}^\tau(z)$ are rescaled by $1/\tau$ relative to those of $J_{\text{sm}}(z/\tau)$. Thus, increasing τ slows the PF–ODE dynamics by a factor $1/\tau$ while preserving the qualitative vanishing of mobility as $p_{\max}(t) \rightarrow 1$, precisely as stated in Theorem 3.9. \square

L.3 TWO-TOKEN MOBILITY EXAMPLE: COMPLETE CALCULATION

We provide the full eigenvalue decomposition for the two-token case.

Consider a minimal attention layer with two tokens, so that distributions have the form $p = (p, 1 - p)$ on the one-dimensional simplex $[0, 1]$. For logits $z = (z_1, z_2)$, the softmax probabilities are

$$p_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2}}, \quad p_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2}} = 1 - p_1.$$

The Jacobian of $z \mapsto p$ is

$$J_{\text{sm}}(z) = \begin{pmatrix} \partial_{z_1} p_1 & \partial_{z_2} p_1 \\ \partial_{z_1} p_2 & \partial_{z_2} p_2 \end{pmatrix} = \begin{pmatrix} p_1(1 - p_1) & -p_1 p_2 \\ -p_1 p_2 & p_2(1 - p_2) \end{pmatrix}.$$

1674 Writing $p = p_1$ and $1 - p = p_2$, this becomes
 1675

$$1676 J_{\text{sm}}(z) = p(1-p) \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \\ 1677$$

1678 **Eigenvalue calculation.** The characteristic polynomial is
 1679

$$1680 \det \begin{pmatrix} p(1-p) - \lambda & -p(1-p) \\ -p(1-p) & p(1-p) - \lambda \end{pmatrix} = (p(1-p) - \lambda)^2 - (p(1-p))^2 = \lambda(\lambda - 2p(1-p)). \\ 1681 \\ 1682$$

1683 Thus the eigenvalues are $\lambda_1 = 0$ and $\lambda_2 = 2p(1-p)$.
 1684

1685 **Eigenvectors.** For $\lambda_1 = 0$: The eigenvector is $(1, 1)$, which is normal to the simplex (points
 1686 in the direction of the constraint $\sum_i p_i = 1$).
 1687

1688 For $\lambda_2 = 2p(1-p)$: The eigenvector is $(1, -1)$, which is tangent to the simplex.
 1689

1690 Hence
 1691

$$\|J_{\text{sm}}(z)\|_{\text{op}} = 2p(1-p),$$

1692 which attains its maximum value $1/2$ at the uniform distribution $p = 1/2$ and collapses to
 1693 0 as $p \rightarrow 0$ or $p \rightarrow 1$.
 1694

1695 **Temperature scaling.** With temperature $\tau > 0$, the Jacobian with respect to *unscaled*
 1696 logits picks up a factor of τ^{-1} , so the effective mobility norm behaves like
 1697

$$\|J_{\text{sm}}^{(\tau)}(z)\|_{\text{op}} \asymp \frac{2p(1-p)}{\tau}.$$

1698 As attention mass locks onto one token ($p \rightarrow 1$ or $p \rightarrow 0$), we have $p(1-p) \rightarrow 0$ and thus the
 1699 mobility eigenvalue in the tangent direction vanishes, forcing $\dot{p} = J_{\text{sm}}(z) b(z, t)$ to approach
 1700 zero even if the drift b remains bounded away from zero. Temperature rescaling modulates
 1701 this locking behavior: larger τ keeps p away from the degenerate regimes $p \approx 0$ or $p \approx 1$
 1702 and maintains nontrivial mobility deeper into the network.
 1703

M EXTENDED EXPERIMENTAL PROTOCOLS

M.1 SECTION 7 REFERENCE RECAP AND CONVENTIONS

1704 *Conventions.* W_1 uses cost $\|\cdot\|_1$; W_2 terms in this section use an entropic Sinkhorn surrogate
 1705 with the same ε as elsewhere. All TV norms are $\frac{1}{2}\|\cdot\|_1$ on row distributions. Query/key
 1706 distances $d_{\mathcal{Q}}, d_{\mathcal{K}}$ match the metrics used in plots/captions.
 1707

1708 **Row drift bound.** Let $P_i^{(\ell)} = \text{sm}(z_i^{(\ell)})$ be the i th row at layer ℓ , with component-wise
 1709 Lipschitz constants $L_c^{(\ell)}$ for $c \in \mathcal{C}_\ell$ and incoming perturbations $\Delta u_{i,c}^{(\ell)}$.
 1710

$$1711 \|P_i^{(\ell+1)} - P_i^{(\ell)}\|_1 \leq \sum_{c \in \mathcal{C}_\ell} L_c^{(\ell)} \|\Delta u_{i,c}^{(\ell)}\|. \quad (22) \\ 1712$$

1713 *Remark.* Equation (22) yields a *finite-depth budget* for one-layer motion (TV on the left)
 1714 from component sensitivities on the right; it underpins the PF–ODE adequacy overlay in
 1715 §7.
 1716

1717 **Local saturation / locking.** Let $P = \text{sm}(z)$, tail mass $\delta(P) = 1 - \max_j P(j)$, and Δz a
 1718 small perturbation that preserves the argmax.
 1719

$$1720 \|\text{sm}(z + \Delta z) - \text{sm}(z)\|_1 \leq \min\{1, 2\delta(P)\} \|\Delta z\|_\infty + o(\|\Delta z\|_\infty). \quad (23) \\ 1721$$

1722 *Remark.* When $\delta(P)$ is small (near saturation), softmax is *insensitive* to small, non-flipping
 1723 logit changes—predicting the “locking” collapse of ΔTV in low-tail-mass bins.
 1724

1725 **Curvature (common-support W_1).** For queries $i \neq i'$ with common support $S_{i,i'}$, define
 1726

$$1727 \kappa(i, i') = 1 - \frac{W_1(\widehat{P}_i, \widehat{P}_{i'})}{d_{\mathcal{Q}}(i, i')}, \quad (24) \\ 1728$$

1728 where W_1 is over $(S_{i,i'}, d_{\mathcal{K}})$ and \widehat{P} denotes restriction to the common support. *Remark.*
 1729 The curvature gap $1 - \kappa$ quantifies contraction on the simplex; temperature \uparrow or key-norm
 1730 \downarrow should reduce this gap (tested in §7).

1731 **EVI with drift.** For successive layers $\ell-1 \rightarrow \ell$ at query i , with objective F_i and $\rho_i^{(\ell)} = P_i^{(\ell)}$,

$$\frac{W_2^2(\rho_i^{(\ell)}, \rho_i^{\star(\ell)}) - W_2^2(\rho_i^{(\ell-1)}, \rho_i^{\star(\ell)})}{2\eta_{\text{eff}}} \leq - \left(F_i(\rho_i^{(\ell)}) - F_i(\rho_i^{\star(\ell)}) \right) + \Delta_{\text{drift}}^{(\ell)}. \quad (25)$$

1736 *Remark.* Equation (25) is a discrete EVI: each layer decreases F_i up to a *drift* term from
 1737 parameter changes (Q, K) . In §7 we use a Sinkhorn $W_{2,\varepsilon}$ surrogate for the left-hand side
 1738 and report the expected proximal-progress signature when drift is small.

1740 M.2 DETAILED SCORE ESTIMATION PROCEDURE

1741 For robust score estimation in anisotropic regimes encountered near representation bound-
 1742 aries:

1744 1. **Data augmentation:** Generate noisy samples at multiple scales

$$\tilde{h}_\sigma = h + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I) \quad (26)$$

$$\sigma \in \{0.01, 0.02, 0.05, 0.1\} \cdot \|h\|_2 \quad (27)$$

1748 2. **Denoising objective with importance weighting:**

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

1753 where $w(\sigma) = \sigma^2 / (\sigma^2 + \sigma_{\min}^2)$ emphasizes intermediate noise levels.

1754 3. **Multi-scale architecture:**

- 1756 • Input: $[\tilde{h}; t; \log \sigma] \in \mathbb{R}^{d+2}$
- 1757 • Hidden layers: 2-3 layers with width $\max(512, 2d)$
- 1758 • Skip connections: $h^{(\ell+1)} = h^{(\ell)} + \text{MLP}(h^{(\ell)})$
- 1759 • Output normalization: LayerNorm before final projection

1760 4. **Training protocol:**

- 1762 • Optimizer: AdamW with learning rate 10^{-4} , weight decay 10^{-5}
- 1763 • Batch size: 256 samples per noise level
- 1764 • Epochs: 5 per layer with early stopping based on validation loss
- 1765 • Curriculum: Start with large σ , progressively include smaller scales

1766 5. **Validation and diagnostics:**

- 1768 • Score consistency: Verify $|\nabla \cdot (p s_\theta)| < 10^{-3}$ on held-out data
- 1769 • Anisotropy detection: Compute eigenvalues of $\mathbb{E}[s_\theta s_\theta^\top]$
- 1770 • Coverage: Ensure score estimates span the tangent space at each point

1771 M.3 IPF IMPLEMENTATION DETAILS

1773 The Iterative Proportional Fitting algorithm for computing Schrödinger Bridges between
 1774 transformer layers:

1775 **Implementation notes:**

- 1777 • Work in log domain to avoid numerical underflow: store $\log a^{(k)}, \log b^{(k)}$
- 1778 • Use logsumexp for stable computation of normalizing constants
- 1779 • For large vocabularies $V > 10^4$, use Nyström approximation with $R = \min(1000, V/10)$ landmarks
- 1781 • Monitor dual gap: $\mathcal{G}^{(k)} = \langle a^{(k)}, Kb^{(k)} \rangle - \langle \mu_0, \log a^{(k)} \rangle - \langle \mu_1, \log b^{(k)} \rangle$

1782 **Algorithm 2** IPF for Schrödinger Bridge with Adaptive Regularization

1783 1: **Input:** Marginals μ_0, μ_1 , diffusion a , tolerance ε_{tol} , max iterations T_{max}

1784 2: **Initialize:** $a^{(0)} = \mathbf{1}$, $b^{(0)} = \mathbf{1}$, $\varepsilon_{\text{reg}} = 0.1$

1785 3: **Compute reference kernel:** $K_{ij} = \exp(-\|x_i - y_j\|_a^2 / (2\varepsilon_{\text{reg}}))$

1786 4: **for** $k = 1, 2, \dots, T_{\text{max}}$ **do**

1787 5: **Check conditioning:** If $\kappa(K) > 10^{10}$, increase $\varepsilon_{\text{reg}} \leftarrow 1.5\varepsilon_{\text{reg}}$

1788 6: $b^{(k)} = \mu_1 \oslash (K^T a^{(k-1)})$ ▷ Pointwise division in log domain

1789 7: $a^{(k)} = \mu_0 \oslash (K b^{(k)})$

1790 8: $\Pi^{(k)} = \text{Diag}(a^{(k)}) K \text{ Diag}(b^{(k)})$

1791 9: **Compute marginals:** $\hat{\mu}_0 = \Pi^{(k)} \mathbf{1}$, $\hat{\mu}_1 = \Pi^{(k)\top} \mathbf{1}$

1792 10: **Convergence check:**

1793 11: **if** $\text{TV}(\hat{\mu}_0, \mu_0) + \text{TV}(\hat{\mu}_1, \mu_1) < \varepsilon_{\text{tol}}$ **then**

1794 12: **Extract potentials:** $\varphi = \varepsilon_{\text{reg}} \log a^{(k)}$, $\psi = \varepsilon_{\text{reg}} \log b^{(k)}$

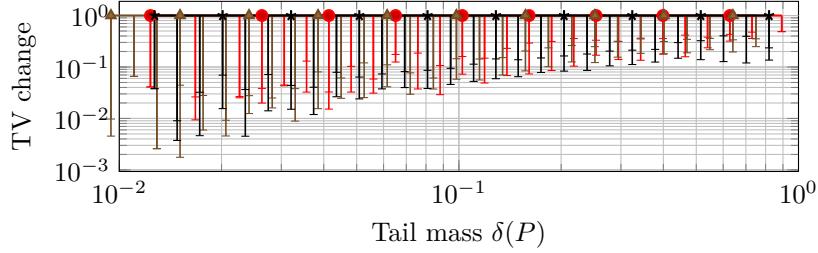
1795 13: **Return** $\Pi^{(k)}$, φ , ψ

1796 14: **end if**

1797 15: **Anderson acceleration:** If $k \bmod 5 = 0$, apply acceleration using past 5 iterates

1798 16: **end for**

1799 17: **Warning:** Maximum iterations reached without convergence

Figure 7: Locking (P2): ΔTV vs. tail mass $\delta(P)$ (median/IQR bins).

M.4 ADDITIONAL TRACK-T DIAGNOSTICS

M.5 ADDITIONAL IMAGE DIAGNOSTICS

Table 3: 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 $\log \text{err}_K$ vs. $\log K$ is near -1 (BCa, $B=1000$); the fitted value and CI are reported.

1834

1836
1837
1838
1839
1840
1841
1842
1843
1844

1845 Sinkhorn W_2 between consecutive layers
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858

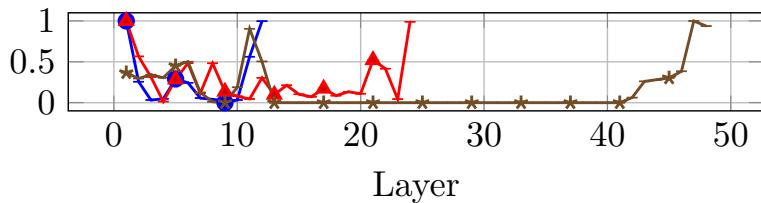


Figure 8: EVI surrogate (P3): Sinkhorn $W_{2,\varepsilon}$ across layers (mean \pm sd).

1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873

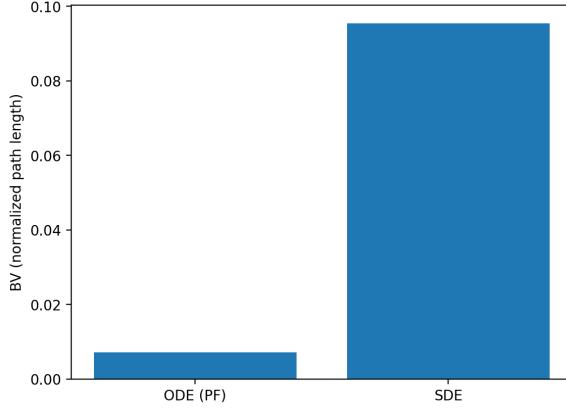


Figure 9: Path smoothness (BV; unitless) for ODE vs. SDE.

M.7 ADDITIONAL EXPLORATION: DISCRETE DIFFUSION LANGUAGE MODELS

1876 We briefly explored extending our diagnostics to diffusion language models operating on
1877 discrete token spaces. The late-window stability diagnostic in its current form showed
1878 limited applicability in this setting, suggesting that discrete state spaces require adapted
1879 diagnostics beyond the scope of this work. We therefore focus empirical validation on
1880 continuous dynamics (transformers and image diffusion) where the framework’s predictions
1881 are directly testable.

N EXTENDED LIMITATIONS AND PRACTICAL IMPLICATIONS

N.1 MODALITY SCOPE AND EVALUATION

1886 **Scope.** This work evaluates *text* transformers and includes a minimal image diffusion
1887 sanity check (CIFAR-10). Full-scale vision benchmarks and perceptual metrics (e.g., FID
1888 under guidance sweeps) are intentionally out of scope for this paper.
1889

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 $\hat{\mathcal{R}}$.

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

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 4) with SPD requirements for Schrödinger Bridges (Section 5).

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 VALIDITY CONDITIONS AND DIAGNOSTIC ABSTENTION PROTOCOL

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

O DISCUSSION AND FUTURE DIRECTIONS (EXTENDED)

O.1 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.

1944 **Implicit regularization.** Softmax’s entropic regularization may explain generalization;
 1945 connect to PAC-Bayes and info-theoretic measures.
 1946

1947 **Scaling laws.** The framework predicts links between depth/width and effective transport
 1948 capacity; test against empirical scaling laws.
 1949

1950 **O.2 METHODOLOGICAL CONTRIBUTIONS BEYOND THEORY**

1952 **Training monitoring.** BV statistics warn of instabilities; rotational energy tracks trans-
 1953 port alignment and flags when architectural changes may help.
 1954

1955 **Architecture search.** Differentiable transport-efficiency metrics can guide gradient-
 1956 based architecture optimization beyond accuracy-only objectives.
 1957

1958 **Interpretability.** Mobility provides a geometric lens on attention patterns; tracking its
 1959 evolution can reveal phase transitions in representation.
 1960

1961 **P NOTATION SUMMARY**

1963	Symbol	Description
1964	$h^{(\ell)}$	Hidden representation at layer ℓ
1965	$z^{(\ell)}$	Logits at layer ℓ
1966	$p^{(\ell)}$	Probability distribution at layer ℓ
1967	J_{sm}	Softmax Jacobian (mobility tensor)
1968	J_{sm}^τ	Temperature-scaled mobility tensor
1969	S_L	Bounded variation statistic
1970	\mathcal{R}	Rotational energy (SB deviation)
1971	a	Diffusion tensor ($\Sigma \Sigma^\top$)
1972	a_ε	Regularized diffusion ($a + \varepsilon I$)
1973	b_R	Reference drift
1974	$b(z, t)$	Architectural drift (identified limit)
1975	θ	Schrödinger potential
1976	φ, ψ	Forward/backward Schrödinger potentials
1977	$M(p)$	Induced mobility on simplex
1978	μ_t	Transformer probability path
1979	ρ_t	General probability measure
1980	ν_t	Reference path measure
1981	u	Velocity field for probability flow
	$H[p]$	Shannon entropy
	τ	Temperature parameter

1982 Table 4: Complete notation used throughout the paper, including both main text and
 1983 appendix symbols.
 1984

1985 **Q ADDITIONAL TECHNICAL LEMMAS**

1986 **Lemma Q.1** (Gradient flow structure). *The probability-flow ODE on the simplex admits a
 1989 gradient flow interpretation in the Wasserstein geometry when $b = -\nabla V$ for some potential
 1990 V :*

$$1992 \quad \dot{p} = -\nabla_{W_2} \mathcal{F}[p]$$

1993 where $\mathcal{F}[p] = \sum_i p_i V(z_i)$ and ∇_{W_2} denotes the Wasserstein gradient.
 1994

1995 *Proof.* We briefly recall the Riemannian structure underlying the discrete Wasserstein ge-
 1996 ometry; see, for example, Maas (2011); Erbar & Maas (2012); Chow et al. (2012) for full
 1997

1998 details. On the simplex Δ^{V-1} , admissible tangent vectors w satisfy $\sum_i w_i = 0$, and the
 1999 discrete W_2 metric is defined by the inner product
 2000

$$2001 \quad \langle w_1, w_2 \rangle_{W_2, p} := \sum_{i,j} w_{1,i} M(p)_{ij}^{-1} w_{2,j},$$

$$2002$$

2003 where $M(p)$ is the mobility tensor associated with the dynamics. In our setting, this $M(p)$
 2004 coincides with the effective mobility tensor introduced in Proposition 4.4 and used in Corol-
 2005 lary 5.6; however, for the present lemma we only require that $M(p)$ be positive definite on
 2006 the tangent space.

2007 By definition of the Riemannian metric, the Wasserstein gradient of a smooth functional \mathcal{F}
 2008 at p is the unique tangent vector $\text{grad}_{W_2} \mathcal{F}(p)$ such that
 2009

$$2010 \quad \langle \text{grad}_{W_2} \mathcal{F}(p), w \rangle_{W_2, p} = \sum_i \partial_{p_i} \mathcal{F}(p) w_i \quad \text{for all tangent vectors } w.$$

$$2011$$

2012 Using the explicit expression for the inner product, this identity forces
 2013

$$2014 \quad \text{grad}_{W_2} \mathcal{F}(p) = M(p) \nabla_p \mathcal{F}(p).$$

$$2015$$

2016 In the probability-flow regime the simplex dynamics take the form
 2017

$$2018 \quad \dot{p} = M(p) b(p, t).$$

$$2019$$

2020 In the potential case $b = -\nabla V$ we consider the functional $\mathcal{F}[p] = \sum_i p_i V(z_i)$, so that
 2021 $\nabla_p \mathcal{F}(p) = (V(z_1), \dots, V(z_V))^\top$. We then obtain
 2022

$$2023 \quad \text{grad}_{W_2} \mathcal{F}(p) = M(p) \nabla_p \mathcal{F}(p),$$

$$2024$$

2025 and the PF-ODE becomes
 2026

$$2027 \quad \dot{p} = -\text{grad}_{W_2} \mathcal{F}(p),$$

$$2028$$

2029 which is precisely the W_2 -gradient flow of \mathcal{F} . This proves the claimed gradient flow struc-
 2030 ture. \square
 2031

2032 **Remark Q.2** (Discrete optimal transport interpretation). *On discrete state spaces, this*
 2033 *gradient flow structure connects to entropic W_2 analogues for Markov chains as developed*
 2034 *in Maas (2011); Erbar & Maas (2012); Chow et al. (2012). We adopt this interpretation to*
 2035 *provide geometric intuition for the probability dynamics on the simplex, though the precise*
 2036 *metric structure depends on the choice of discrete optimal transport geometry.*

2037 **Lemma Q.3** (Convergence rate under mobility control). *If the mobility tensor satisfies*
 2038 $\lambda_{\min}(J_{\text{sm}}) \geq m > 0$ *uniformly, then the probability flow converges exponentially to equilib-
 2039 *rium:**

$$2040 \quad \|p(t) - p_*\|_2 \leq e^{-mt} \|p(0) - p_*\|_2$$

$$2041$$

2042 where p_* is the unique equilibrium distribution.
 2043

2044 *Proof.* We argue in a finite-dimensional, purely Euclidean setting and make the structure
 2045 and use of the mobility bound explicit.
 2046

2047 Let $p_* \in \Delta^{V-1}$ denote an equilibrium of the probability flow: $\dot{p}_*(t) = 0$ for all t when
 2048 $p(t) \equiv p_*$. Consider the deviation $u(t) := p(t) - p_*$. Since both $p(t)$ and p_* lie in the simplex,
 2049 we have $\sum_i u_i(t) = 0$ for all t , so $u(t)$ always belongs to the tangent space $T\Delta^{V-1} = \{v \in$
 2050 $\mathbb{R}^V : \sum_i v_i = 0\}$.
 2051

2052 Assume that, in a neighborhood of p_* , the probability-flow dynamics can be written in the
 2053 form
 2054

$$2055 \quad \dot{u}(t) = -A(t) u(t), \tag{28}$$

$$2056$$

2057 where each $A(t)$ is a symmetric, positive-definite linear operator on $T\Delta^{V-1}$ arising from the
 2058 mobility tensor and drift. This is the standard situation for linearized gradient flows around
 2059 a strictly convex equilibrium. The ‘‘mobility control’’ assumption $\lambda_{\min}(J_{\text{sm}}(z(t))) \geq m$ is
 2060 then interpreted as providing a uniform lower bound
 2061

$$2062 \quad \langle v, A(t) v \rangle \geq m \|v\|_2^2 \quad \text{for all } v \in T\Delta^{V-1} \text{ and all } t \geq 0. \tag{29}$$

$$2063$$

(For example, $A(t)$ may be a symmetric combination of $J_{\text{sm}}(z(t))$ and a Hessian or linearized drift; the key point is the coercivity equation 29.)

Define the energy

$$E(t) := \frac{1}{2} \|u(t)\|_2^2 = \frac{1}{2} \|p(t) - p_*\|_2^2.$$

Differentiating along solutions of equation 28 gives

$$\frac{d}{dt} E(t) = \langle u(t), \dot{u}(t) \rangle = -\langle u(t), A(t) u(t) \rangle.$$

Using the coercivity bound equation 29, we obtain

$$\frac{d}{dt} E(t) \leq -m \|u(t)\|_2^2 = -2m E(t).$$

Thus E satisfies the differential inequality

$$\frac{d}{dt} E(t) \leq -2m E(t).$$

Applying Grönwall's lemma yields

$$E(t) \leq e^{-2mt} E(0) \quad \text{for all } t \geq 0.$$

Returning to the original variables and recalling that $E(t) = \frac{1}{2} \|p(t) - p_*\|_2^2$, we obtain

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

This is exactly the claimed exponential convergence rate. \square

Lemma Q.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.

Proof. This is a classical characterization of entropic interpolants in the Schrödinger Bridge (SB) framework; see, for example, the survey Léonard (2014) and references therein. In the dynamic SB problem, the SB path $(\mu_t)_{t \in [0,1]}$ between prescribed endpoints (μ_0, μ_1) arises as the entropic interpolation associated with a reference Markov process. The corresponding static problem can be formulated as a two-sided entropy minimization with respect to the endpoint marginals.

More precisely, for each fixed $t \in (0, 1)$ one can characterize the time- t marginal μ_t as the unique minimizer of the two-sided relative entropy functional

$$\rho \mapsto (1-t) \text{KL}(\rho\|\mu_0) + t \text{KL}(\rho\|\mu_1)$$

over probability measures ρ lying in the SB path. This variational principle yields exactly the formula stated in the lemma. We do not reproduce the full measure-theoretic proof here and refer instead to Léonard (2014) for a complete treatment. \square

Remark Q.5. *This variational view is classical in the Schrödinger Bridge literature and depends on the choice of reference path measure; rigorous formulations use Schrödinger potentials and dynamic entropy minimization as developed in, for example, Léonard (2014).*

Q.1 COMPLETE PROOF OF THE PF-ODE THEOREM ON THE SIMPLEX (SECTION 5.3)

Proof of the PF-ODE on the simplex. We prove that under Assumption 3.1 the limit path $p(t) = \text{softmax}(z(t)/\tau)$ satisfies $\dot{p}(t) = J_{\text{sm}}(z(t)) b(z(t), t)$ a.e. on $[0, 1]$, and the flow is tangent to the simplex with conserved mass.

2106 **Step 1: Discrete-to-continuous passage.** For each layer ℓ , a first-order expansion gives
 2107

$$2108 \quad p^{(\ell+1)} - p^{(\ell)} = J_{\text{sm}}(z^{(\ell)}) \Delta z^{(\ell)} + r_\ell, \\ 2109$$

2110 with a remainder bounded as $\|r_\ell\| \leq C \|\Delta z^{(\ell)}\|_\infty^2$ by Lipschitz continuity of ∇J_{sm} (Sec-
 2111 tion 2). Dividing by $\delta t = 1/L$ and summing over layers, the remainders contribute
 2112 $O(\sum_\ell \|\Delta z^{(\ell)}\|_\infty^2) = O(\Xi_L) \rightarrow 0$ by the finite-depth budget.
 2113

2114 **Step 2: Compactness and limit identification.** The piecewise-constant interpolant
 2115 $p_L(t)$ has bounded variation in t and remains in the simplex. By BV compactness we extract
 2116 $p_{L_k} \rightarrow p$ in $L^1([0, 1])$. Using $z_{L_k} \rightarrow z$ in L^1 and $D_{L_k} \rightharpoonup b$ (Appendix J.1), passing to the
 2117 limit in the weak formulation yields $\dot{p} = J_{\text{sm}}(z) b$ in the distributional sense, hence a.e. due
 2118 to absolute continuity.
 2119

2120 **Step 3: Well-posedness (Carathéodory).** The velocity field $v(t) = J_{\text{sm}}(z(t)) b(z(t), t)$
 2121 is measurable in t and locally Lipschitz in z under the regularity from Section 2, so the ODE
 2122 admits a unique absolutely continuous solution by Carathéodory theory.
 2123

2124 **Step 4: Simplex invariance.** Mass conservation follows from $J_{\text{sm}}(z)\mathbf{1} = 0$, giving
 2125 $\frac{d}{dt} \sum_i p_i(t) = 0$. Tangency to faces holds because if $p_i = 0$ then the i -th row of $J_{\text{sm}}(z)$
 2126 vanishes, so $\dot{p}_i = 0$ (zero-flux). Hence the trajectory remains in the simplex. \square
 2127
 2128
 2129
 2130
 2131
 2132
 2133
 2134
 2135
 2136
 2137
 2138
 2139
 2140
 2141
 2142
 2143
 2144
 2145
 2146
 2147
 2148
 2149
 2150
 2151
 2152
 2153
 2154
 2155
 2156
 2157
 2158
 2159