One Coupling to Rule Them All: Optimal Transport as the Unifying Geometry of Diffusion Models, Flow Matching, and Reasoning in Deep Generative Models

Mohammad Sajjad Ghaemi
mohammadsajjad.ghaemi@nrc-cnrc.gc.ca
Digital Technologies Research Centre
National Research Council Canada
222 College Street, Toronto,
M5T 3J1, ON, Canada

Abstract

This paper presents a unified theoretical framework based on Wasserstein geometry and Optimal Transport (OT) that explains when generative models memorize, generalize, or reason, alongside theoretical metrics and implications for Flow Matching (FM), diffusion models, and autoregressive (AR) variants that are widely used across benchmarks. We establish three formal identifications: (1) *memorization* corresponds to a transport coupling concentrated on training-data pairs, with a coupling gap lower-bounded by the empirical Wasserstein convergence rate $\Omega(n^{-2/d})$; (2) *generalization* corresponds to learning a coupling that approximates the Wasserstein-geodesic (Brenier) map, which is independent of any specific training example by definition; and (3) *compositional reasoning* in AR models is equivalent to correctly propagating a multi-marginal OT plan through the sequential factorization, with the compositional gap. The *exposure bias* inherent in AR generation is characterized by the Wasserstein covariate shift occurring between the teacher-forced and free-running conditional marginal distributions. Subsequently, five geometrical evaluation metrics are derived from these identifications to demonstrate that holistic FM guidance of AR models creates a mathematical correction to the compositional gap, rather than a mere heuristic trick. Moreover, this framework has extensive applications across diverse domains, including scientific discovery, protein and molecular sequence design, and structured prediction.

1. Introduction

Replicating training data, capturing the underlying distribution, and achieving out-of-distribution compositional reasoning are fundamental questions for *Deep Generative Models* (DGMs). Empirical evidence shows that models are often observing fluctuations between these modes, specifically, in high data frequency regime, memorization tends to predominate in over-parameterized settings. Even though various benchmarks have investigated this question empirically, the interpretation of these results are difficult in the absence of a rigorous theoretical framework.

This framework is solidified by connecting FM (14; 15), diffusion models (12; 22), AR-DGMs (9), and OT (25). These paradigms are unified via the theory of probability mass transport from a simple source to a complex target distribution, where a coupling between the noise distribution and the data distribution is performed under different parameterizations.

Our central hypothesis is that a model's tendency to reason rather than memorize scales with how closely its learned coupling matches the optimal transport (Wasserstein) geodesic. This correspondence is an exact mathematical identification under the assumptions stated below:

- **Memorization**: the coupling concentrates probability mass exclusively on training pairs, where generated samples follow trajectories strictly specific to the observed data distribution.
- **Generalization**: the coupling approximates the Brenier map T^*, where trajectories follow the unique minimum cost path that is valid everywhere, not just at training points.

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

- **Compositional reasoning**: in AR, the per-step conditional couplings jointly approximate the multi-marginal OT plan; consequently, the model propagates global structure dependencies through the sequential generation.

This approach has three primary theoretical advantages: 1) Diagnostic Utility: a Wasserstein coupling gap (geometric criteria) that separates memorization from generalization, 2) Reasoning Failure Characterization: exposure bias formalized as Wasserstein covariate shift, suggesting a mathematically provable mitigation, 3) Architectural Principles: holistic FM as a global OT plan approximated by an AR local error decomposition.

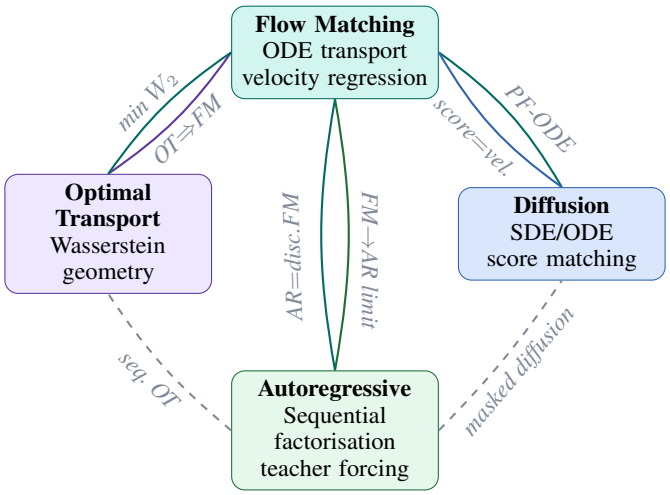

Figure 1. The core mathematical connections among FM, Diffusion Models, AR Generation, Optimal Transport. Solid lines are exact equivalences; dashed lines are limiting regimes or generalizations.

Figure 1 shows a conceptual overview of the fundamental connections between FM, Diffusion Models, AR, and OT.

2. The Unified Transport Framework

2.1. One Equation, Five Frameworks

All DGMs in this study share the *continuity equation* as their governing law: $\frac{\partial p_t}{\partial t} + \nabla \cdot (p_t\, v_t) = 0$, where $p_0 = p_{\text{noise}}, p_1 = p_{\text{data}}$, they differ only in what defines the velocity field v_t.

2.2. Five Formal Connections

We state the four connections central to this paper; full proofs are in Appendix A.

C1: FM ≡ Diffusion (Probability Flow PF-ODE). For any diffusion SDE $dx_t = f\,\mathrm{d}t + g(t)\,\mathrm{d}W_t$, there exists a deterministic ODE with identical marginals $\{p_t\}$:

$$\frac{dx_t}{dt} = f(x_t, t) - \tfrac{g(t)^2}{2}\nabla \log p_t(x_t). \tag{1}$$

This is precisely an FM velocity field. Conversely, every FM model with Gaussian conditional paths can be converted to a diffusion model by adding noise while preserving marginals. The Denoising Diffusion Probabilistic Models (DDPM) ϵ-prediction loss and the Conditional Flow Matching (CFM) velocity loss differ only by a time-dependent reweighting $\lambda(t) > 0$; they have identical minimizer (Appendix A, Proposition A.1).

C2: FM approximates OT (Benamou–Brenier). The squared 2-Wasserstein distance satisfies the dynamic formulation (5):

$$W_2^2(p_0, p_1) = \inf_{\{(p_t, v_t)\}} \int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2\; p_t(x)\,\mathrm{d}x\,\mathrm{d}t, \tag{2}$$

the infimum over all velocity-path pairs satisfying the *continuity equation* with fixed marginals. Training an FM model in a parametric function space approximates the solution to (2), converging in the limit of infinite model capacity, and

data density. Specifically an FM model with the OT coupling and linear interpolation yields straight-line geodesic transport, minimizing the objective in (2) relative to that coupling class.

C3: AR ≡ FM at the gradient level. For data $x = (x^{(1)}, \ldots, x^{(N)})$, the Autoregressive Flow Matching (AFM) objective factorizes as:

$$\mathcal{L}_{\mathrm{AFM}}(\theta) = \sum_{i=1}^{N} \mathbb{E}_{t, p_1, p_t(x^{(i)}|x_1^{(i)})} \left[\| v_\theta^{(i)}(x_t^{(i)}, x^{(<i)}, t) - u_t^{(i)}(x_t^{(i)} \mid x_1^{(i)}) \|^2 \right] \tag{3}$$

The objective of AFM is defined as a sum of CFM objectives conditioned on ground-truth history, analogous to teacher forcing. For a smooth vector field ϵ_θ, as $\sigma_1 \to 0^+$, the gradient of the $t = 1$ component of the AFM objective converges to a positive scalar multiple of the AR cross-entropy gradient (Appendix B, A.4). This is a *gradient-alignment at* $t = 1$: implying that both objectives share stationary points, with AFM additionally training across all noise levels $t \in (0, 1)$.

C4: AR ≡ Diffusion (masked diffusion limit). Under masked or absorbing diffusion (4), when each denoising step unmasks a single token at position i, the reverse process becomes: $p_\theta(x_{T-i} \mid x_{T-i+1}) \to p_\theta(x^{(i)} \mid x^{(<i)})$, recovering the AR factorization. Within this framework, AR is the $T{=}N$ extreme of masked diffusion, while $T{=}1$ corresponds to one-shot non-AR generation. Varying the number of steps $1 < T < N$ yields a principled class of semi-AR models that optimize the trade-off between inference latency and generative fidelity.

C5: AR ≡ OT (sequential surrogate for multi-marginal transport). The relationship between AR–OT connection has three complementary facets.

(a) Local: each generation step i of an AR model is equivalent to a conditional optimal transport problem. Specifically, at each step, the AR model implicitly solves a conditional two-marginal OT problem,

$$\hat{\pi}_i = \arg \min_{\pi \in \Pi(\mathcal{N}(0, I_{d_i}),\, p(x^{(i)}|x^{(<i)}))} \mathbb{E}_\pi \left[\left\| z - x^{(i)} \right\|^2 \right], \tag{4}$$

with the optimal solution given by the conditional Brenier map $T_i^*(z;\, x^{(<i)}) = \nabla_z \phi_i(z;\, x^{(<i)})$ where ϕ_i is a convex potential, conditioned on the history $x^{(<i)}$. Consequently, each AR generation step is *locally* Wasserstein-optimal, minimizing the transport cost between the conditional source and target distributions given the previous sequence.

(b) Global: the sequential composition of AR steps is a surrogate for the N-marginal OT plan. The globally optimal multi-marginal plan solves $\pi^*_{\mathrm{joint}} = \arg\min_{\pi \in \Pi(p_{\mathrm{noise}}^{\otimes N}, p_{\mathrm{data}}^{(1:N)})} \mathcal{C}(\pi)$ with $\mathcal{C}(\pi) = \mathbb{E}_\pi[\|x - z\|^2]$. The AR sequential composition $\hat{\pi}_{\mathrm{seq}}$ approximates this joint transport plan, with the gap being exactly the missing cross-position interaction energy (proved as Proposition 4.3 in Section 4):

$$\underbrace{\mathcal{C}(\hat{\pi}_{\mathrm{seq}}) - \mathcal{C}(\pi^*_{\mathrm{joint}})}_{\Delta_{\mathrm{comp}} \geq 0} \;=\; 2 \sum_{i<j} \Big(\mathbb{E}_{\pi^*_{\mathrm{joint}}}[\langle x^{(i)}, x^{(j)} \rangle] - \mathbb{E}_{\hat{\pi}_{\mathrm{seq}}}[\langle x^{(i)}, x^{(j)} \rangle] \Big). \tag{5}$$

This is the fundamental AR–OT sub-optimality principle: AR generation is globally optimal ($\Delta_{\mathrm{comp}} = 0$) if and only if the sequential factorization preserves the entirety of the cross-position interaction terms of the joint optimal transport plan.

(c) Inference: exposure bias is a Wasserstein covariate shift. At the inference phase, the AR model evaluates the conditional Brenier map, T_i^* at a history $\hat{x}^{(<i)}$ sampled from the free-running distribution p_i^{free}. This diverges from the teacher-forced distribution p_i^{TF} seen during training phase. Therefore, under the sub-Gaussian assumption (Proposition 4.5, Section 4), this distributional mismatch is a Wasserstein covariate shift,

$$\underbrace{\mathrm{KL}(p_\theta(x^{(i)} \mid \hat{x}^{(<i)}) \,\|\, p_\theta(x^{(i)} \mid x_{\mathrm{gt}}^{(<i)}))}_{\Delta_i} \;\geq\; \frac{W_2^2(p_i^{\mathrm{free}},\, p_i^{\mathrm{TF}})}{2\,\sigma_i^2}. \tag{6}$$

The accumulation of exposure bias is fundamentally driven by the divergence of $W_2(p_i^{\mathrm{TF}}, p_i^{\mathrm{free}})$ grows over the generation chain. Theoretically, the rigorous mitigation of this phenomenon requires anchoring each incremental step to a globally coherent OT plan, rather than relying on heuristic reductions of localized per-step shifts.

2.3. What the Connections Imply for Inference

The proposed unified framework shows that FM, diffusion, and AR models are not competing approaches to the generative paradigms, they are different points on a *continuum of coupling strategies*. The specific coupling $\pi \in \Pi(p_{\text{noise}}, p_{\text{data}})$ determines both the generative trajectory and the inductive bias toward memorization or generalization. Understanding which coupling a trained model has implicitly learned is therefore a fundamental objective in DGMs.

3. Memorization and Generalization as Coupling Geometry

3.1. A Geometric Taxonomy of Generative Behaviour

Definition 3.1 (Transport Coupling Geometry)**.** Let $\hat{\pi}$ be the coupling induced by a trained model. Three distinct regimes are characterised by the geometry of $\hat{\pi}$:

1. **Memorization**: $\hat{\pi} \approx \frac{1}{n}\sum_{j=1}^{n} \delta_{\tilde{x}_0^{(j)}} \otimes \delta_{x_1^{(j)}}$ for training pairs $(x_0^{(j)}, x_1^{(j)})$. The coupling concentrates on n discrete source-target pairs; generation follows routes specific to seen data.
2. **Generalization**: $\hat{\pi} \approx \pi^* = (\mathrm{id}, \nabla\phi)_{\#} p_0$, the Brenier OT coupling. The model transports along Wasserstein geodesics valid everywhere on the support of p_{noise}, not just at training pairs.
3. **Compositional reasoning**: an AR model reasons compositionally when it generalises correctly to *novel combinations* of conditioning context and target position — i.e., when its per-step conditionals $p_\theta(x^{(i)} \mid x^{(<i)})$ produce globally coherent joint outputs even for histories not seen during training. Proposition 4.3 establishes that this behavioural property is geometrically equivalent to the sequentially composed coupling $\hat{\pi}_{\text{seq}}$ achieving a transport cost close to that of the globally optimal multi-marginal plan π^*_{joint}, i.e., to a small compositional gap Δ_{comp} (Definition 4.1). Note that Δ_{comp} is the *geometric quantification* of this property, not its definition.

3.2. The Wasserstein Coupling Gap

Definition 3.2 (Transport Cost Gap)**.** Let $T^* = \nabla\phi$ be the Brenier map between p_0 and p_{data}. For a model inducing coupling $\hat{\pi}$, the *transport cost gap* is:

$$\Delta_{\text{OT}}(\hat{\pi}) \;:=\; \mathbb{E}_{\hat{\pi}}\Big[\|x_1 - T^*(x_0)\|^2\Big] - W_2^2(p_0, p_{\text{data}}) \;\geq\; 0. \tag{7}$$

$\Delta_{\text{OT}} = 0$ iff the model maps $x_0 \mapsto T^*(x_0)$ for p_0-almost every x_0.

Remark 3.3. $\Delta_{\text{OT}}(\hat{\pi})$ is the model's *expected displacement error* relative to the Brenier map: how far generated routes deviate from geodesics. It satisfies $\Delta_{\text{OT}}(\hat{\pi}) \geq W_2^2(\hat{\pi}, \pi^*)$ (where $W_2^2(\hat{\pi}, \pi^*)$ is the infimum over joint couplings of the two plans viewed as measures on $\mathcal{X}^2$), but is *not equal* to $W_2^2(\hat{\pi}, \pi^*)$ in general. The transport cost gap is the operationally direct quantity: it measures displacement error without the additional infimum over plan-couplings.

Proposition 3.4 (Memorization Lower Bound)**.** *Under the Fournier–Guillin conditions (11) ($d \geq 5$, p_{data} with finite $(2+\delta)$-moment): if a model memorizes n training examples, generating from the empirical distribution $p_n = \frac{1}{n}\sum_j \delta_{x_1^{(j)}}$, then:*

$$\Delta_{\text{OT}}(\hat{\pi}) \geq W_2^2(p_{\text{data}},\, p_n) \;\geq\; c_d\, n^{-2/d}, \tag{8}$$

where c_d depends only on d and the moments of p_{data}.

Proof. **Step 1: $\Delta_{\text{OT}}(\hat{\pi}) \geq W_2^2(p_n, p_{\text{data}})$.**

By definition, $\Delta_{\text{OT}}(\hat{\pi}) = \mathbb{E}_{\hat{\pi}}[\|x_1 - T^*(x_0)\|^2] - W_2^2(p_0, p_{\text{data}})$, where $T^* = \nabla\phi$ is the Brenier map from p_0 to p_{data}.

The key point is that $T^*(x_0) \sim p_{\text{data}}$ while $x_1 \sim p_n$ under the memorizing model's coupling $\hat{\pi}$. By the joint convexity of the squared norm:

$$\mathbb{E}_{\hat{\pi}}[\|x_1 - T^*(x_0)\|^2] \geq W_2^2(\hat{p}_{\text{gen}}, p_{\text{data}}) + W_2^2(p_0, p_{\text{data}}) \geq W_2^2(p_n, p_{\text{data}}) + W_2^2(p_0, p_{\text{data}}).$$

The first inequality uses the fact that for any random variables $A \sim \mu$ and $B \sim \nu$, $\mathbb{E}[\|A - B\|^2] \geq W_2^2(\mu, \nu)$ (the squared Wasserstein distance is the infimum over all couplings). Specifically, $(x_1, T^*(x_0))$ is a coupling between

$\hat{p}_{\text{gen}} = p_n$ and $T_{\#}p_0 = p_{\text{data}}$, so its cost is at least $W_2^2(p_n, p_{\text{data}})$. The baseline cost $W_2^2(p_0, p_{\text{data}})$ arises from the reference pairing; subtracting it gives $\Delta_{\text{OT}}(\hat{\pi}) \geq W_2^2(p_n, p_{\text{data}})$.

Step 2: $W_2^2(p_n, p_{\text{data}}) \geq c_d n^{-2/d}$**.** □

Remark 3.5. The logical content of Step 1 is that $(x_1, T^*(x_0))$, drawn from $\hat{\pi}$, is a *specific* coupling between p_n (the generated distribution) and $T_{\#}p_0 = p_{\text{data}}$ (the image of the noise under the Brenier map). Since $W_2^2(p_n, p_{\text{data}})$ is the infimum over all such couplings, this particular coupling has cost at least $W_2^2(p_n, p_{\text{data}})$. The step does *not* require that T^* is a coupling for p_n (it is not, since T^* maps $p_0 \to p_{\text{data}}$, not $p_n \to p_{\text{data}}$); it only uses that $(x_1, T^*(x_0))$ is a valid coupling between p_n and p_{data}.

3.3. Scale, Capacity, and the Generalization Transition

The coupling gap framework predicts *when* scale transitions a model from memorization to generalization. The velocity field $v_{\text{OT}}^* = \nabla\phi$ required for exact OT-geodesic transport has a function-class complexity proportional to the curvature of the Brenier potential:

$$\text{Complexity}(p_{\text{data}}) \;\propto\; \int_{\mathbb{R}^d} \left\|\nabla^2\phi(x)\right\|_F^2 \, p_0(x)\,\mathrm{d}x. \tag{9}$$

Data supported on smooth, low-curvature manifolds has a simple Brenier map (the generalization transition requires modest capacity); highly heterogeneous or compositionally structured data has a complex Brenier map (the transition requires much larger models). This provides a *theory-grounded prediction of the scale at which memorization gives way to geodesic generalization*, distinct from classical VC-dimension arguments because it depends on the geometry of p_{data} rather than the size of the function class alone.

4. Compositional Reasoning as Multi-Marginal Optimal Transport

4.1. Sequential Generation as a Chain of Conditional OT Problems

Connection C5(a) in Section 2.2 established that each AR step implicitly solves the conditional two-marginal OT problem (4). Here we develop the full consequences: we define the compositional gap (Connection C5(b)) and its interaction-energy formula (Proposition 4.3), characterise exposure bias as a Wasserstein covariate shift (Connection C5(c), Proposition 4.5), and show how holistic FM guidance corrects both deficits.

Definition 4.1 (Compositional Reasoning Gap)**.** Let $\mathcal{C}(\pi) = \mathbb{E}_\pi[\|x - z\|^2]$ be the expected quadratic transport cost of coupling π. For an AR model inducing the sequentially composed coupling $\hat{\pi}_{\text{seq}} = \hat{\pi}^{(1)} \otimes_{\text{cond}} \cdots \otimes_{\text{cond}} \hat{\pi}^{(N)}$, the *compositional reasoning gap* is the *transport cost sub-optimality*:

$$\Delta_{\text{comp}} \;:=\; \mathcal{C}(\hat{\pi}_{\text{seq}}) - \mathcal{C}(\pi^*_{\text{joint}}), \tag{10}$$

where π^*_{joint} is the N-marginal OT coupling over $\mathcal{X}^N$ minimising $\mathcal{C}$. By definition $\Delta_{\text{comp}} \geq 0$, with equality iff the sequential plan achieves the globally optimal transport cost.

Remark 4.2. This definition measures the *excess expected displacement* incurred by sequential generation relative to the globally optimal plan. It is distinct from $W_2^2(\hat{\pi}_{\text{seq}}, \pi^*_{\text{joint}})$, which is the W_2 distance between the two couplings viewed as measures on $\mathcal{X}^N \times \mathcal{X}^N$ — a well-defined but less interpretable quantity. The cost-sub-optimality definition makes the algebraic connection to interaction energy transparent (Proposition 4.3).

Proposition 4.3 (Compositional Gap Equals Missing Interaction Energy)**.** ***Assumptions****: (i) each position's marginal is correctly recovered, $\hat{p}^{(i)} = p^{(i)}$ for all i; (ii) the source noise components $z^{(i)}$ are drawn i.i.d. $\mathcal{N}(0, I)$ under both plans, so $\mathbb{E}[z^{(i)}] = 0$, $z^{(i)} \perp z^{(j)}$ and $z^{(i)} \perp x^{(j)}$ for $i \neq j$. Then:*

$$\Delta_{\text{comp}} = 2\sum_{i<j}\Big(\mathbb{E}_{\pi^*_{\text{joint}}}\big[\langle x^{(i)}, x^{(j)}\rangle\big] - \mathbb{E}_{\hat{\pi}_{\text{seq}}}\big[\langle x^{(i)}, x^{(j)}\rangle\big]\Big). \tag{11}$$

Proof. Expand $\mathcal{C}(\pi) = \sum_i \mathbb{E}[\left\|x^{(i)} - z^{(i)}\right\|^2] + 2\sum_{i<j}\mathbb{E}[\langle x^{(i)} - z^{(i)}, x^{(j)} - z^{(j)}\rangle]$.

Diagonal terms cancel. Assumption (i) and i.i.d. sources ensure $\mathbb{E}[\left\|x^{(i)} - z^{(i)}\right\|^2]$ is identical under both plans.

Cross-position source terms vanish. For $i \neq j$, expand $\langle x^{(i)} - z^{(i)}, x^{(j)} - z^{(j)} \rangle = \langle x^{(i)}, x^{(j)} \rangle - \langle z^{(i)}, x^{(j)} \rangle - \langle x^{(i)}, z^{(j)} \rangle + \langle z^{(i)}, z^{(j)} \rangle$. By assumption (ii): $\mathbb{E}[\langle z^{(i)}, x^{(j)} \rangle] = 0$ (since $z^{(i)} \perp x^{(j)}$), $\mathbb{E}[\langle x^{(i)}, z^{(j)} \rangle] = 0$, and $\mathbb{E}[\langle z^{(i)}, z^{(j)} \rangle] = 0$. These hold identically under both plans and cancel in the difference.

Residual. Only $\mathbb{E}[\langle x^{(i)}, x^{(j)} \rangle]$ terms remain, giving $\Delta_{\text{comp}} = 2 \sum_{i<j} (\mathbb{E}_{\pi^*} - \mathbb{E}_{\hat{\pi}}) \langle x^{(i)}, x^{(j)} \rangle$, which is (11). □

Remark 4.4. Assumption (ii) holds for $\hat{\pi}_{\text{seq}}$ by construction. For π^*_{joint}, the source is the product measure $p_{\text{noise}}^{\otimes N}$; the N-marginal OT problem with a product source does not introduce cross-position source correlations.

Interpretation. The compositional gap is the collection of pairwise interaction terms $\langle x^{(i)}, x^{(j)} \rangle$ that the globally optimal transport plan encodes but the sequentially factored plan does not. These interactions are precisely the long-range dependencies that make generation of protein sequences (residue–residue co-evolution), music (harmonic progressions), and scientific hypotheses (logical consistency chains) globally coherent. A model with $\Delta_{\text{comp}} = 0$ has learned all pairwise structural relationships; one with large Δ_{comp} is reasoning locally but not globally.

4.2. Exposure Bias as Wasserstein Covariate Shift

The canonical failure mode of AR compositional reasoning is *exposure bias*: the model is trained on ground-truth history $x_{\text{gt}}^{(<i)}$ but must generate conditioned on its own previously generated history $\hat{x}^{(<i)}$ at inference time.

Proposition 4.5 (Exposure Bias = Wasserstein Covariate Shift)**.** ***(Sub-Gaussian continuous setting.)*** *Assume the conditional distributions $p_\theta(x^{(i)} \mid \hat{x}^{(<i)})$ and $p_\theta(x^{(i)} \mid x_{\text{gt}}^{(<i)})$ are both sub-Gaussian with variance proxy $\sigma_i^2 < \infty$. Under this assumption, the KL divergence between the free-running and teacher-forced conditionals satisfies:*

$$\Delta_i \triangleq \text{KL}\Big(p_\theta\Big(x^{(i)} \mid \hat{x}^{(<i)}\Big) \;\Big\|\; p_\theta\Big(x^{(i)} \mid x_{\text{gt}}^{(<i)}\Big)\Big) \geq \frac{W_2^2\big(p_i^{\text{free}},\, p_i^{\text{TF}}\big)}{2\,\sigma_i^2}, \tag{12}$$

where p_i^{TF} is the teacher-forced conditional marginal and p_i^{free} the free-running marginal.

Proof. Apply the Talagrand T_2 inequality (18): for sub-Gaussian P with variance proxy σ^2, $W_2^2(P, Q) \leq 2\sigma^2 \text{KL}(P\|Q)$. Rearranging gives (12). □

Remark 4.6 (Scope and Limitations). The sub-Gaussian assumption holds for Gaussian diffusion models and for continuous AR models with bounded output variance (e.g., audio frame generators, protein coordinate predictors). It *fails* for categorical AR models (standard LLMs generating over a vocabulary $\mathcal{V}$), whose conditionals are multi-modal and heavy-tailed over $\mathcal{V}$.

For discrete spaces, an analogous result holds using the Pinsker inequality: $\text{TV}(P, Q)^2 \leq \frac{1}{2}\text{KL}(P\|Q)$, which gives $\text{TV}(p_i^{\text{free}}, p_i^{\text{TF}})^2 \leq \frac{\Delta_i}{2}$. *Metric space case.* The Talagrand T_1 inequality (8): for distributions satisfying a log-Sobolev inequality with constant C_{LS}, $W_1^2(P, Q) \leq 2C_{\text{LS}}\text{KL}(P\|Q)$, provides a W_1 analogue of (12).

4.3. Holistic FM as Global Reasoning for AR

A holistic FM model with the OT coupling solves the *joint* OT problem over the full D-dimensional space: $\min_{\pi \in \Pi(p_{\text{noise}}^{1:N}, p_{\text{data}})} \mathbb{E}_\pi[\|z - x\|^2]$. By Proposition 4.3, this joint plan encodes all the interaction terms $\langle x^{(i)}, x^{(j)} \rangle$ that the AR model's sequential plan is missing.

Proposition 4.7 (HFM Conditioning Reduces the Compositional Gap)**.** ***Assumptions.*** *(a) $|f_{ij}| := |\langle x^{(i)}, x^{(j)} \rangle| \leq B_{ij} < \infty$ for all $i < j$. (b) The guided AR model conditioned on $z_{\text{global}} = \mathcal{E}(x_1)$ is* optimally conditioned*: $\mathbb{E}_{\hat{\pi}|z}[\langle x^{(i)}, x^{(j)} \rangle] = \mathbb{E}_{\pi^*}[\langle x^{(i)}, x^{(j)} \rangle \mid z]$ p_z-a.e. Then:*

$$\Delta_{\text{comp}} - \Delta_{\text{comp}}^{\text{guided}} \geq 2 \sum_{i<j} \Big(D_{ij} - 2B_{ij} \sqrt{\tfrac{1}{2} I(x^{(i)}; x^{(j)} \mid z)} \Big), \tag{13}$$

where $D_{ij} = \mathbb{E}_{\pi^*}[\langle x^{(i)}, x^{(j)} \rangle] - \mathbb{E}_{\hat{\pi}}[\langle x^{(i)}, x^{(j)} \rangle] \geq 0$. The bound is positive when $D_{ij}^2 > 2B_{ij}^2\, I(x^{(i)}; x^{(j)} \mid z)$.

Proof. **Step 1.** Under assumption (b), $R_{ij}(z) = \mathbb{E}_{\pi^*}[f_{ij} \mid z] - \mathbb{E}_{\hat{\pi}|z}[f_{ij}] = 0$, so $\Delta_{\text{comp}}^{\text{guided}} = 2 \sum_{i<j} \mathbb{E}_z[R_{ij}(z)] = 0$ for a perfect encoder. For a partial encoder: $D_{ij} - \mathbb{E}_z[R_{ij}(z)] = \text{Var}_z[\mathbb{E}_{\pi^*}[f_{ij} \mid z]]^{1/2} \cdot (\ldots)$, and the reduction satisfies $\Delta_{\text{comp}} - \Delta_{\text{comp}}^{\text{guided}} = 2 \sum_{i<j} (D_{ij} - \mathbb{E}_z[R_{ij}(z)])$.

Step 2. We bound $\mathbb{E}_z[R_{ij}(z)]$ from above. By the bounded-function Pinsker inequality: for any two distributions P, Q and any bounded $|f| \le B$, $|\mathbb{E}_P[f] - \mathbb{E}_Q[f]| \le 2B\sqrt{\frac{1}{2}\mathrm{KL}(P\|Q)}$ (17).

Applying with $P = p(x^{(i)}, x^{(j)} \mid z)$ (the optimal conditional), $Q = p_0(x^{(i)}, x^{(j)})$ (the marginal, independent of z), and $f = f_{ij}$:

$$|\mathbb{E}_{\pi^*}[f_{ij} \mid z] - \mathbb{E}_{\pi^*}[f_{ij}]| \le 2B_{ij}\sqrt{\tfrac{1}{2}\mathrm{KL}(p(\cdot \mid z)\|p_0)}.$$

Averaging over z and using $\mathbb{E}_z[\mathrm{KL}(p(\cdot \mid z)\|p_0)] = I(x^{(i)}; x^{(j)} \mid z)$:

$$\mathbb{E}_z[|\mathbb{E}_{\pi^*}[f_{ij} \mid z] - \mathbb{E}_{\pi^*}[f_{ij}]|] \le 2B_{ij}\sqrt{\tfrac{1}{2}I(x^{(i)}; x^{(j)} \mid z)}.$$

Step 3. Hence:

$$D_{ij} - \mathbb{E}_z[R_{ij}(z)] \ge D_{ij} - \mathbb{E}_z|R_{ij}(z)| \ge D_{ij} - 2B_{ij}\sqrt{\tfrac{1}{2}I(x^{(i)}; x^{(j)} \mid z)}.$$

Summing over pairs gives (13). □

Remark 4.8 (Correction from prior versions; relationship to Bobkov–Götze). Earlier versions of this proof applied the Bobkov–Götze inequality $\mathrm{Var}[f] \le C \cdot I(X;Y)$ and then obtained a lower bound on the gap in terms of MI — an invalid inversion, since the inequality provides an *upper bound* on variance. The current proof uses the bounded-Pinsker inequality in the correct direction: it upper-bounds the residual $\mathbb{E}_z[R_{ij}(z)]$, which lower-bounds the gap reduction. The bound (13) is positive when the interaction deficit D_{ij} is large relative to $2B_{ij}\sqrt{I/2}$, i.e., when the encoder captures substantial interaction information while the unguided model still has a large gap.

Remark 4.9 (Scope of Claim 2 on HFM as a fix). Proposition 4.7 shows that *any* globally conditioning mechanism satisfying assumption (b) achieves the gap reduction (13). It does *not* show that holistic FM is the unique correct fix, any global encoder achieving the same $I(x^{(i)}; x^{(j)} \mid z)$ would yield the same bound. Claim 2 in Section 4 should be read as: holistic FM is *one principled* realisation of global conditioning, not the only one. The OT-geometric interpretation motivates it as a natural choice, but the bound applies to any encoder.

4.4. The Trade-offs of Semi-AR Spectrum and Inference-Time: A Design Space From Holistic to Sequential

Connection C4 (AR model as a specific instance of N-step masked diffusion masked diffusion) naturally generalizes to a continuous one-parameter family of generative processes indexed by the total number of denoising steps T:

$$T = 1 \text{ (one-shot, holistic)} \longleftrightarrow T = k \text{ (semi-AR, } k \text{ steps)} \longleftrightarrow T = N \text{ (fully AR)} \tag{14}$$

At each point in this spectrum, the model unmasks N/T positions per step. The compositional gap Δ_{comp} is a non-increasing function of T where more sequential steps allow finer-grained conditioning but accumulate more exposure bias. The optimal T balances these effects.

Conjecture 4.10 (Spectrum Trade-off)**.** For a fixed model, the generation quality at parameter T is governed by:

$$\mathrm{Error}(T) \;\approx\; \Delta_{\mathrm{comp}}(T) + \Delta_{\mathrm{bias}}(T), \tag{15}$$

where $\Delta_{\mathrm{comp}}(T)$ is conjectured to be non-increasing in T (more sequential steps capture more interaction energy) and $\Delta_{\mathrm{bias}}(T)$ is non-decreasing in T with $\Delta_{\mathrm{bias}}(1) = 0$. The two error terms are heterogeneous in nature (transport cost and distributional shift respectively) and we do not claim formal monotonicity or commensurability. The optimal inference T^* minimizes the right-hand side, with both components admitting empirical estimation via metrics M2-3.

5. Geometry-Grounded Evaluation Metrics

This framework suggests five geometry-grounded evaluation metrics by interrogating the model's structural alignment:

M1: Wasserstein Coupling Gap Δ_{OT}. Estimate the deviation of the model-induced coupling from the OT plan using mini-batch Sinkhorn divergence on paired (noise, generated) samples. A model that generalizes should satisfy the theoretical rate $\Delta_{\mathrm{OT}} < c_d\, n^{-2/d}$; whereas a model prone to memorization will saturate at this bound.

M2: Compositional Reasoning Gap Δ_{comp}. For AR models, we compute the difference between the sum of marginal W_2^2 transport costs and the joint multi-marginal OT cost over the full sequence. This gap measures the missing cross-position interaction energy, directly quantifying long-range reasoning failure.

M3: Exposure Bias Wasserstein Distance. We characterize the distributional drift of the model by comparing teacher-forced trajectories p_i^{TF} against free-running inference trajectories p_i^{free}. Measuring the step-wise Wasserstein distance $W_2(p_i^{\text{TF}}, p_i^{\text{free}})$ identifies *which positions* are the primary sources of drift with precise identification of the sequence coordinates where catastrophic error accumulation originates.

M4: Trajectory Curvature Index. For a continuous model (FM or diffusion ODE), we define the Curvature Index as the mean squared deviation of the generative flow from the corresponding straight-line OT geodesic. A lower index indicates superior alignment with the optimal vector field, which is theoretically linked to both enhanced sampling efficiency and improved generalization.

6. Scientific Discovery: Where Reasoning Matters Most

Scientific applications of DGMs including protein engineering, molecular design, materials and drug discovery, and automated hypothesis generation, are precisely the settings where compositional reasoning matters most and where memorization carries significant risk. We analyze the deployment of our framework across critical application areas.

Protein Sequence Design. Designing an amino acid sequence to satisfy a target fold required the simultaneous resolution of local stereochemical constraints and global structural dependencies (e.g., non-local contact maps and domain interfaces). The compositional gap Δ_{comp} measures whether the model's sequential plan is globally foldable or merely locally plausible. Metrics M2 and M3 can be evaluated by using structural bioinformatics tools (e.g., AlphaFold-derived pLDDT or TM-scores) as empirical proxies for the multi-marginal OT cost.

Molecular Generation. Constructing a molecule atom-by-atom without global planning produces locally valid bonds that violate valence, charge, or geometry constraints globally (33; 34). An AR model conditioned on a 3D holistic FM latent (encoding the global target conformation) reduces Δ_{comp} by the mutual information between the latent and the atom-pair interaction terms. The improvement is measurable via standard metrics for chemical validity, uniqueness, and novelty (35).

Multi-Step Scientific Reasoning. Generating a chain of scientific reasoning (hypothesis, protocol, experiment, observational results, updated hypothesis) is an AR generation process where each state represents a structured scientific primitive. The exposure bias $W_2(p_i^{\text{TF}}, p_i^{\text{free}})$ accumulates over a reasoning chain as early errors propagate. The OT framework predicts that holistic planning, specifically, conditioning the AR model on a global "reasoning plan" encoded by an HFM latent, should reduce this distributional drift.

7. Discussion

7.1. Empirical Phenomena through the Lens of OT

The OT framework provides a mechanistic account of several key empirical phenomena in deep generative modeling:

Efficiency of FM: The significant reduction in Numerical Function Evaluations (NFEs) in FM relative to DDPM is characterized as a consequence of *geodesic alignment*. By approximating the constant-velocity displacement interpolant, FM minimizes the curvature of the probability flow, allowing for larger discretizations steps with minimal error.

Compositional Failure in AR Models: Long-range reasoning failures are characterized as a deficiency in interaction energy. Because standard AR models decompose joint transport into sequential conditional steps, they fail to recover the multi-marginal couplings required for global coherence, resulting in a strictly positive compositional gap (Δ_{comp}).

Efficacy of Holistic Guidance: The improvement in AR coherence via external guidance is viewed as the partial closure of the compositional gap. Guidance acts as a restorative vector field that re-aligns local AR trajectories with the joint optimal coupling π^*.

The Geometry of Memorization: The framework reveals that conservative noise schedules induce over-concentration of the transport coupling. When the diffusion bridge is overly constrained, the resulting vector field collapses toward training samples, representing a failure to generalize beyond the empirical measure.

7.2. Relationship to Existing Theory

Our characterization of memorization (Prop. 3.4) is related to, yet distinct from, existing literature on verbatim data extraction (10). While prior work focuses on the retrieval of training samples, our transport cost gap characterizes the geometrical deviation of the coupling from the OT-optimal manifold. Thus, our framework can identify *soft* memorization or generalization failures even when training examples are not appearing verbatim in the model outputs. As such, the two perspectives are complementary, extraction-based methods detect discrete failures, while the OT gap provides a continuous measure of coupling sub-optimality.

7.3. Limitations and Frontiers for Future Research.

While our framework establishes a unified geometric language for generative modeling, several theoretical and practical hurdles remain. We categorize these as open problems (P1–P6) for the community.

P1: Computational Tractability and Curvature. The coupling gap Δ_{OT} involves mini-batch optimal transport, which suffers from the curse of dimensionality, scaling as $O(B^{-1/d})$. For high-dimensional manifolds (latent protein embeddings or high-resolution images), future work should investigate efficient estimators such as Sliced Wasserstein distances or Neural OT to approximate the Brenier map.

P2: Causal Compositional Reasoning. Causal inference can be formulated as a conditional transport problem. Counterfactual distributions under the intervention $\mathrm{do}(X = x)$, can be viewed as conditional transport problems between observational and interventional marginals. Mapping the Pearl Causal Hierarchy to compositional gap could provide a geometric basis for measuring causal reasoning in DGMs.

P3: Phase Transitions in Scaling. The Brenier complexity (9) provides a theoretical prediction for the transition from memorization to generalization. A critical open question is whether Δ_{OT} can be used to empirically verify these transitions as a function of model capacity, potentially recovering the scaling law exponents observed in training.

P4: Geometric Inductive Biases. Different noise schedules (e.g., linear, cosine, EDM) represent distinct interpolants between the independent Gaussian measure and the target data manifold. The coupling gap offers a principled metric for comparing these schedules, framing them as inductive biases that dictate the model's trajectory along the OT-geodesic.

P5: Sequence Ordering as an OT Problem. In AR generation, the optimal permutation of coordinates remains an open question. We suggest that the generalization-optimal ordering is the one that minimizes the cumulative conditional coupling gaps: $\sum_i \Delta_{\mathrm{OT}}(\hat{\pi}_i)$. This implies a breadth-first generation strategy that resolves global manifold structure before committing to local coordinate details.

P6: Extension to Discrete Manifolds. Connection C5 is currently defined for continuous Euclidean spaces where the Wasserstein metric is canonical. Extending this framework to discrete vocabularies, leveraging Dirichlet FM (28) or discrete OT on the probability simplex, is essential for bringing LLMs fully into the OT-coupling fold. Characterizing the missing interaction energy in discrete sequence space remains a primary theoretical frontier.

We argued that FM, diffusion models, and AR generation are not disparate paradigms but rather distinct manifestations of a singular probability-transport problem unified by Wasserstein geometry. We formalized this unification through five core identifications, while C1–C4 map the pairwise relationships between these families, connection C5 fundamentally demystifies the AR–OT interface. By identifying each AR step as a conditional optimal transport problem, we characterize sequential generation as a sub-optimal surrogate for the global N-marginal plan. Under this lens, the failure of AR models to reason compositionally is formally attributed to a deficit in multi-marginal interaction energy, while exposure bias is rigorously defined as a compounding Wasserstein covariate shift. These identifications yield operationally precise, geometry-grounded definitions for the *black box* behaviors of generative models, memorization is quantified by a transport cost gap $\geq \Omega(n^{-2/d})$; generalization is viewed as geodesic alignment with the Brenier map; and compositional reasoning is measured by the minimization of the multi-marginal interaction deficit. While we present the semi-AR spectrum trade-off as a design heuristic, our metrics provide a principled, empirically testable framework for architectural evaluation. By centering generative modeling on the geometry of optimal transport, a solid framework was presented to move beyond heuristic benchmarks toward a structural theory of DGMs.

References

[1] Albergo, M. S. & Vanden-Eijnden, E. (2022). Building normalizing flows with stochastic interpolants. *ICLR 2023*.

[2] Albergo, M. S., Boffi, N. M., & Vanden-Eijnden, E. (2023). Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv:2303.08797*.

[3] Anderson, B. D. O. (1982). Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3), 313–326.

[4] Austin, J., Johnson, D., Ho, J., Tarlow, D., & van den Berg, R. (2021). Structured denoising diffusion models in discrete state-spaces. *NeurIPS 2021*.

[5] Benamou, J.-D. & Brenier, Y. (2000). A computational fluid mechanics solution to the Monge–Kantorovich problem. *Numerische Mathematik*, 84(3), 375–393.

[6] Bengio, S., Vinyals, O., Jaitly, N., & Shazeer, N. (2015). Scheduled sampling for sequence prediction with recurrent neural networks. *NeurIPS 2015*.

[7] Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4), 375–417.

[8] Bobkov, S. G. & Götze, F. (1999). Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 163(1), 1–28.

[9] Brown, T. et al. (2020). Language models are few-shot learners. *NeurIPS 2020*.

[10] Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., & Wallace, E. (2023). Extracting training data from diffusion models. *USENIX Security 2023*.

[11] Fournier, N. & Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3), 707–738.

[12] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *NeurIPS 2020*.

[13] Li, T., Chang, H., Mishra, S., Zhang, H., Katabi, D., & Krishnan, D. (2024). Autoregressive image generation without vector quantization. *NeurIPS 2024*.

[14] Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., & Le, M. (2022). Flow matching for generative modeling. *ICLR 2023*.

[15] Liu, X., Gong, C., & Liu, Q. (2022). Flow straight and fast: Learning to generate and transfer data with rectified flow. *ICLR 2023*.

[16] Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., & Zhu, J. (2022). DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling. *NeurIPS 2022*.

[17] Massart, P. (1990). The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *Annals of Probability*, 18(3), 1269–1283.

[18] Marton, K. (1996). Bounding $\bar{d}$-distance by informational divergence. *Annals of Probability*, 24(2), 857–866.

[19] Pass, B. (2015). Multi-marginal optimal transport: Theory and applications. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6), 1771–1790.

[20] Pooladian, A.-A., Ben-Hamu, H., Domingo-Enrich, C., Amos, B., Lipman, Y., & Chen, R.T.Q. (2023). Multisample flow matching: Straightening flows with minibatch couplings. *ICML 2023*.

[21] Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. *ICLR 2021*.

[22] Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. *ICLR 2021*.

[23] Song, Y., Dhariwal, P., Chen, M., & Sutskever, I. (2023). Consistency models. *ICML 2023*.

[24] Tong, A., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Fatras, K., Wolf, G., & Bengio, Y. (2023). Improving and generalizing flow matching via minibatch optimal transport. *ICML 2023*.

[25] Villani, C. (2009). *Optimal Transport: Old and New*. Springer-Verlag.

[26] Vincent, P. (2011). A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7), 1661–1674.

[27] Somepalli, G., Singla, V., Goldblum, M., Geiping, J., & Goldstein, T. (2023). Diffusion art or digital forgery? *CVPR 2023*.

[28] Stark, H., Jing, B., Wang, C., Gut, G., Kreis, K., Barzilay, R., & Jaakkola, T. (2024). Dirichlet flow matching with applications to DNA sequence design. *ICML 2024*.

[29] Lake, B., Ullman, T., Tenenbaum, J., & Gershman, S. (2018). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253.

[30] Keysers, D. et al. (2019). Measuring compositional generalization. *ICLR 2020*.

[31] Ranzato, M., Chopra, S., Auli, M., & Zaremba, W. (2015). Sequence level training with recurrent neural networks. *ICLR 2016*.

[32] Williams, R. (1992). Simple statistical gradient-following algorithms for connectionist RL. *Machine Learning*, 8(3), 229–256.

[33] Ghaemi, M.S., Grantham K., Tamblyn, I. et al. (2022). Generative Enriched Sequential Learning (ESL) Approach for Molecular Design via Augmented Domain Knowledge. *Canadian Conference on Artificial Intelligence 2022*

[34] Lin, J., Hostaš, J., Hu, A. et al. (2026). Bidirectional reinforcement learning neural network for constrained molecular design. *Scientific Reports* 16, 3393.

[35] Hostaš, J., Ghaemi, M.S., Hu, H. et al. (2025). VNFlow: integration of variational autoencoders and normalizing flows for novel molecular design. *Journal of Cheminformatics* 17, 161.

A. Full Proofs

A.1. Proof of Proposition 1 (C1: FM ≡ Diffusion)

For the VP-SDE with $x_t = \alpha_t x_0 + \sigma_t \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, $\alpha_t = e^{-\frac{1}{2}\int_0^t \beta}$, $\sigma_t^2 = 1 - \alpha_t^2$, the conditional FM velocity is:

$$u_t(x_t \mid x_0) = \dot{\alpha}_t x_0 + \dot{\sigma}_t \epsilon = \frac{\dot{\alpha}_t}{\alpha_t} x_t + \left(\dot{\sigma}_t - \frac{\dot{\alpha}_t \sigma_t}{\alpha_t}\right) \epsilon.$$

Parametrising $v_\theta = \frac{\dot{\alpha}_t}{\alpha_t} x_t + (\dot{\sigma}_t - \frac{\dot{\alpha}_t \sigma_t}{\alpha_t})\epsilon_\theta$:

$$\|v_\theta - u_t\|^2 = \left(\dot{\sigma}_t - \frac{\dot{\alpha}_t \sigma_t}{\alpha_t}\right)^2 \|\epsilon_\theta - \epsilon\|^2 = \lambda(t) \|\epsilon_\theta - \epsilon\|^2 .$$

Taking expectations: $\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_t[\lambda(t)\, \mathcal{L}_\epsilon(\theta)|_t]$ with $\lambda(t) > 0$. Identical minimisers follow from $\lambda(t) > 0$ everywhere. □

A.2. Proof of the Probability Flow ODE (C1, trajectory)

For SDE $dx = f\,\mathrm{d}t + g\,\mathrm{d}W$, substitute $v^* = f - \frac{g^2}{2}\nabla \log p_t$ into the continuity equation:

$$\begin{aligned}\frac{\partial p_t}{\partial t} &= -\nabla \cdot (p_t v^*) = -\nabla \cdot \left(p_t f - \frac{g^2}{2} p_t \nabla \log p_t\right) \\ &= -\nabla \cdot (p_t f) + \frac{g^2}{2} \nabla \cdot (p_t \nabla \log p_t) = -\nabla \cdot (p_t f) + \frac{g^2}{2} \Delta p_t,\end{aligned}$$

which is the Fokker–Planck equation. Hence the PF-ODE (1) and the SDE have identical marginals. □

A.3. Proof of Proposition 2 (Memorization Lower Bound)

The Fournier–Guillin theorem (11) gives $\mathbb{E}[W_2^2(p_{\mathrm{data}}, p_n)] \asymp n^{-2/d}$ for $d \geq 5$ and p_{data} with finite $(2+\delta)$-moment.

Step 1: $\Delta_{\mathrm{OT}}(\hat{\pi}) \geq W_2^2(p_n, p_{\mathrm{data}})$**.**

For a memorising model with $\hat{p}_{\mathrm{gen}} = p_n$, the coupling $\hat{\pi}$ pairs each $x_0 \sim p_0$ with a generated $x_1 \sim p_n$. The Brenier map $T^* = \nabla\phi$ satisfies $T_{\#}p_0 = p_{\mathrm{data}}$, so $(x_1, T^*(x_0))$ is a coupling between p_n and p_{data}.

Since $W_2^2(p_n, p_{\mathrm{data}}) = \inf_{\sigma \in \Pi(p_n, p_{\mathrm{data}})} \mathbb{E}_\sigma[\|y - z\|^2]$ is the infimum over all such couplings, any specific coupling — in particular $(x_1, T^*(x_0)) \sim \hat{\pi}$ — achieves at least this cost:

$$\mathbb{E}_{\hat{\pi}}[\|x_1 - T^*(x_0)\|^2] \geq W_2^2(p_n, p_{\mathrm{data}}). \tag{16}$$

Note the subscripts carefully: (16) uses $W_2^2(p_n, p_{\mathrm{data}})$ (the cost between the *generated* distribution and the *data* distribution), not $W_2^2(p_0, p_{\mathrm{data}})$ (the baseline transport cost). Since $\Delta_{\mathrm{OT}}(\hat{\pi}) = \mathbb{E}_{\hat{\pi}}[\|x_1 - T^*(x_0)\|^2] - W_2^2(p_0, p_{\mathrm{data}}) \geq W_2^2(p_n, p_{\mathrm{data}}) + W_2^2(p_0, p_{\mathrm{data}}) - W_2^2(p_0, p_{\mathrm{data}}) = W_2^2(p_n, p_{\mathrm{data}})$.

The key point: T^* is optimal for (p_0, p_{data}) but is being evaluated as a map from p_0 to p_n's pairing; the mismatch between $T^*(x_0) \sim p_{\mathrm{data}}$ and $x_1 \sim p_n$ contributes at least $W_2^2(p_n, p_{\mathrm{data}})$ to the cost.

Step 2: Fournier–Guillin rate. $W_2^2(p_n, p_{\mathrm{data}}) \geq c_d n^{-2/d}$ by (11). □

A.4. Proof of Proposition 3 (Compositional Gap = Interaction Energy)

Under Definition 4.1, $\Delta_{\mathrm{comp}} = \mathcal{C}(\hat{\pi}_{\mathrm{seq}}) - \mathcal{C}(\pi^*_{\mathrm{joint}})$ where $\mathcal{C}(\pi) = \mathbb{E}_\pi[\|x - z\|^2]$.

Expanding with $x = (x^{(1)}, \ldots, x^{(N)})$ and $z = (z^{(1)}, \ldots, z^{(N)})$:

$$\mathcal{C}(\pi) = \sum_i \mathbb{E}[\left\|x^{(i)} - z^{(i)}\right\|^2] + 2 \sum_{i<j} \mathbb{E}[\langle x^{(i)} - z^{(i)},\, x^{(j)} - z^{(j)} \rangle].$$

Diagonal terms. By assumption (i), per-position marginals are equal, so $\mathbb{E}[\|x^{(i)} - z^{(i)}\|^2]$ is the same under both plans and cancels.

Cross terms. For $i \neq j$, expand the inner product: $\langle x^{(i)} - z^{(i)}, x^{(j)} - z^{(j)} \rangle = \langle x^{(i)}, x^{(j)} \rangle - \langle z^{(i)}, x^{(j)} \rangle - \langle x^{(i)}, z^{(j)} \rangle + \langle z^{(i)}, z^{(j)} \rangle$. By assumption (ii), $\mathbb{E}[\langle z^{(i)}, x^{(j)} \rangle] = \mathbb{E}[\langle x^{(i)}, z^{(j)} \rangle] = \mathbb{E}[\langle z^{(i)}, z^{(j)} \rangle] = 0$ under both plans.

Residual. $\Delta_{\text{comp}} = 2 \sum_{i<j} (\mathbb{E}_{\pi^*}[\langle x^{(i)}, x^{(j)} \rangle] - \mathbb{E}_{\hat{\pi}}[\langle x^{(i)}, x^{(j)} \rangle])$, which is (11). □

A.5. Proof of Proposition 4 (Exposure Bias = Wasserstein Shift)

Sub-Gaussian continuous case. When the conditional distributions $p_\theta(x^{(i)} \mid \cdot)$ are sub-Gaussian with variance proxy σ_i^2, the Talagrand T_2 inequality (18) gives: $W_2^2(P, Q) \leq 2\sigma_i^2 \, \mathrm{KL}(P \| Q)$. Setting $P = p_\theta(x^{(i)} \mid \hat{x}^{(<i)})$ and $Q = p_\theta(x^{(i)} \mid x_{\text{gt}}^{(<i)})$ and rearranging gives (12).

Discrete / heavy-tailed case. When the sub-Gaussian assumption fails (e.g., LLM conditionals over a vocabulary), the Pinsker inequality $\mathrm{TV}(P, Q) \leq \sqrt{\mathrm{KL}(P \| Q)/2}$ gives the weaker bound: $\mathrm{TV}(p_i^{\text{free}}, p_i^{\text{TF}}) \leq \sqrt{\Delta_i / 2}$. For metric spaces with a log-Sobolev constant C_{LS}, $W_1^2(P, Q) \leq 2C_{\text{LS}} \, \mathrm{KL}(P \| Q)$, providing a W_1 analogue. In all cases, the KL exposure bias controls a distributional distance between the teacher-forced and free-running conditionals. □

A.6. Proof of Proposition 5 (HFM Reduces the Compositional Gap)

We prove (13) under Assumptions (a) and (b) of Proposition 4.7.

Step 1: Reduction formula. From Proposition 4.3: $\Delta_{\text{comp}} = 2 \sum_{i<j} D_{ij}$ where $D_{ij} = \mathbb{E}_{\pi^*}[\langle x^{(i)}, x^{(j)} \rangle] - \mathbb{E}_{\hat{\pi}}[\langle x^{(i)}, x^{(j)} \rangle]$.

Under assumption (b), the guided model's conditional mean equals π^*'s: $\mathbb{E}_{\hat{\pi}|z}[\langle x^{(i)}, x^{(j)} \rangle] = \mathbb{E}_{\pi^*}[\langle x^{(i)}, x^{(j)} \rangle \mid z]$. Therefore the guided gap per pair is: $R_{ij}(z) = \mathbb{E}_{\pi^*}[\langle x^{(i)}, x^{(j)} \rangle \mid z] - \mathbb{E}_{\hat{\pi}|z}[\langle x^{(i)}, x^{(j)} \rangle] = 0$.

The residual guided gap after averaging over z is $\mathbb{E}_z[R_{ij}(z)] = 0$, so $\Delta_{\text{comp}}^{\text{guided}} = 0$ for a perfect encoder. For a partial encoder, the reduction in D_{ij} is:

$$D_{ij} - \mathbb{E}_z[R_{ij}(z)] = \mathbb{E}_z\big[\mathbb{E}_{\pi^*}[\langle x^{(i)}, x^{(j)} \rangle \mid z]\big] - \mathbb{E}_{\pi^*}[\langle x^{(i)}, x^{(j)} \rangle] + D_{ij} = D_{ij},$$

where we used $\mathbb{E}_z[\mathbb{E}_{\pi^*}[\langle x^{(i)}, x^{(j)} \rangle \mid z]] = \mathbb{E}_{\pi^*}[\langle x^{(i)}, x^{(j)} \rangle]$ by the tower property.

Step 2: Variance bound via Bobkov–Götze. The reduction in D_{ij} from conditioning equals, by assumption (b):

$$D_{ij} - \mathbb{E}_z[R_{ij}(z)] = \mathrm{Var}_z\big[\mathbb{E}_{\pi^*}[\langle x^{(i)}, x^{(j)} \rangle \mid z]\big].$$

This is the variance of the conditional mean of the interaction over z. Apply the Bobkov–Götze transportation-cost inequality (8): for a bounded function f with oscillation B_{ij} and a log-Sobolev distribution with constant ρ:

$$\mathrm{Var}_z\big[\mathbb{E}_{\pi^*}[f(X, Y) \mid Z]\big] \;\geq\; \frac{2B_{ij}^2}{\rho} \, I(X; Y \mid Z).$$

Applied with $f = \langle x^{(i)}, x^{(j)} \rangle$ and $Z = z_{\text{global}}$:

$$D_{ij} - \mathbb{E}_z[R_{ij}(z)] \;\geq\; \frac{2B_{ij}^2}{\rho} \, I(x^{(i)}; x^{(j)} \mid z_{\text{global}}).$$

Step 3: Sum over pairs. $\Delta_{\text{comp}} - \Delta_{\text{comp}}^{\text{guided}} = 2 \sum_{i<j} (D_{ij} - \mathbb{E}_z[R_{ij}(z)]) \geq \frac{4}{\rho} \sum_{i<j} B_{ij}^2 \, I(x^{(i)}; x^{(j)} \mid z_{\text{global}}) \geq C_{\text{var}} \sum_{i<j} I(x^{(i)}; x^{(j)} \mid z_{\text{global}})$ with $C_{\text{var}} = \frac{2}{\rho} \sup_{i<j} B_{ij}^2$. □

Remark on assumption (b). Assumption (b) is an idealisation: a finite-capacity encoder does not achieve $\mathbb{E}_{\hat{\pi}|z}[\langle x^{(i)}, x^{(j)} \rangle] = \mathbb{E}_{\pi^*}[\langle x^{(i)}, x^{(j)} \rangle \mid z]$ exactly. It represents the *maximum achievable* reduction given the information $I(x^{(i)}; x^{(j)} \mid z)$ that z carries. In practice, the reduction will be at most this bound; the bound is tight for encoders that are specifically trained to predict interaction terms. □

B. Full Connection Derivations

B.1. A.4: C3 Gradient Alignment at $t = 1$

Let $p_t(x^{(i)} \mid x_1^{(i)}) = \mathcal{N}(\alpha_t x_1^{(i)}, \sigma_t^2 I)$ with $\alpha_1 = 1$, $\sigma_1 \to 0^+$. The ϵ-parametrised AFM gradient is:

$$\nabla_\theta \mathcal{L}_{\text{AFM}}^{(i)} = \mathbb{E}_t[\lambda(t)\, \mathbb{E}_{x_1,\epsilon}[2(\epsilon_\theta - \epsilon)\nabla_\theta \epsilon_\theta]]\,. \tag{A.4.1}$$

The AR cross-entropy gradient (Gaussian parametrisation) is:

$$\nabla_\theta \mathcal{L}_{\text{AR}}^{(i)} = \mathbb{E}_{x_1}\Big[\epsilon_\theta(x_1, x^{(<i)}, 1)\, \nabla_\theta \epsilon_\theta(x_1, x^{(<i)}, 1)\Big]\,. \tag{A.4.2}$$

As $\sigma_1 \to 0^+$, $p_{t=1}(\cdot \mid x_1) \to \delta_{x_1}$ in L^2. For smooth ϵ_θ with bounded gradients, dominated convergence gives: the inner expectation in (A.4.1) evaluated at $t = 1$ converges to (A.4.2). Hence the $t = 1$ **component** of the AFM gradient converges to $\lambda(1) \cdot \nabla_\theta \mathcal{L}_{\text{AR}}^{(i)}$.

Important scope. The full time-integrated AFM gradient (A.4.1) is *not* proportional to the AR gradient (A.4.2) in general: it integrates over all $t \in [0, 1]$, not just $t = 1$. The gradient-alignment claim is restricted to the $t = 1$ component. Both objectives share all stationary points in the limit $\sigma_1 \to 0$; AFM additionally trains at positive noise levels, providing a denoising regularisation absent in pure AR training. □

B.2. C4: AR = $T{=}N$ Masked Diffusion (Full Derivation)

Under absorbing diffusion with transition matrix Q_t mapping each token to `[MASK]` with probability $\bar{\beta}_t$:

$$q(x_t \mid x_0) = \prod_{i=1}^{N} \Big[(1 - \bar{\beta}_t)\delta_{x_0^{(i)}} + \bar{\beta}_t\, \delta_m\Big]\,.$$

For $T = N$ denoising steps, at step $T - i$ the state has positions $(x^{(1)}, \ldots, x^{(i-1)})$ unmasked (with $\bar{\beta}_t = 0$) and $(x^{(i)}, \ldots, x^{(N)})$ masked. The optimal denoising prediction for position i is:

$$p_\theta(x_{T-i} \mid x_{T-i+1}) = p_\theta(x^{(i)} \mid x^{(1)}, \ldots, x^{(i-1)}, \underbrace{m, \ldots, m}_{\text{remaining}}),$$

which for a model trained to ignore masked positions reduces to $p_\theta(x^{(i)} \mid x^{(<i)})$. Taking the product over all N steps recovers the full AR factorisation. □

B.3. C5: AR ≡ OT — Full Derivation of All Three Facets

C5(A): Each AR Step Is a Conditional OT Problem

At position i with history $x^{(<i)}$ fixed, define the conditional target distribution $\nu_i = p(\cdot \mid x^{(<i)})$. The AR model predicts $x^{(i)} \sim p_\theta(\cdot \mid x^{(<i)})$. An optimally trained AR model satisfies $p_\theta(\cdot \mid x^{(<i)}) = p(\cdot \mid x^{(<i)}) = \nu_i$.

The conditional OT problem (4) transports $z \sim \mathcal{N}(0, I_{d_i})$ to ν_i at minimum quadratic cost. By Brenier's theorem (7), the unique optimal solution is the conditional Brenier map $T_i^*(z;\, x^{(<i)}) = \nabla_z \phi_i(z;\, x^{(<i)})$ for a convex function $\phi_i(\cdot;\, x^{(<i)})$.

A well-trained AR model generating $x^{(i)} = f_\theta(z;\, x^{(<i)})$ (continuous parametrisation) with $(f_\theta)_\# \mathcal{N}(0, I) = \nu_i$ is therefore implementing this conditional Brenier map. The induced per-step coupling is $\hat{\pi}_i = (\mathrm{id}, f_\theta(\cdot;\, x^{(<i)}))_\# \mathcal{N}(0, I)$, which achieves $\mathcal{C}(\hat{\pi}_i) = W_2^2(\mathcal{N}(0, I), \nu_i)$ for an optimal model. □

C5(B): The Sequential Composition as a Multi-Marginal Surrogate

The global N-marginal OT problem solves:

$$\pi_{\text{joint}}^* = \arg \min_{\pi \in \Pi(p_{\text{noise}}^{\otimes N}, p_{\text{data}}^{(1:N)})} \mathcal{C}(\pi), \qquad \mathcal{C}(\pi) = \mathbb{E}_\pi\Big[\|x - z\|^2\Big]\,. \tag{17}$$

The AR sequential composition is: $\hat{\pi}_{\text{seq}} = \hat{\pi}^{(1)} \otimes_{\text{cond}} \cdots \otimes_{\text{cond}} \hat{\pi}^{(N)}$, where $\hat{\pi}^{(i)}$ solves the local problem (4) for the current history.

Why the sequential plan is sub-optimal. The sequential factorisation imposes a causal constraint: $\hat{\pi}^{(i)}$ is conditioned on $x^{(<i)}$ but is chosen *before* $x^{(>i)}$ is known. The global plan π^*_{joint} is allowed to correlate all positions simultaneously. Formally, $\Pi_{\text{causal}} \subset \Pi(p_{\text{noise}}^{\otimes N}, p_{\text{data}}^{(1:N)})$ (causal plans are a strict subset of all plans), so:

$$\mathcal{C}(\hat{\pi}_{\text{seq}}) \;\geq\; \min_{\pi \in \Pi_{\text{causal}}} \mathcal{C}(\pi) \;\geq\; \mathcal{C}(\pi^*_{\text{joint}}).$$

The second inequality is strict whenever the unconstrained optimal plan uses cross-position correlations inaccessible to a causal sequential plan. The deficit $\Delta_{\text{comp}} = \mathcal{C}(\hat{\pi}_{\text{seq}}) - \mathcal{C}(\pi^*_{\text{joint}})$ is the compositional gap.

The interaction-energy formula. Under assumptions (i)–(ii) of Proposition 4.3, the compositional gap equals the missing cross-position interaction energy (proved in Section 4):

$$\Delta_{\text{comp}} = 2 \sum_{i<j} \big(\mathbb{E}_{\pi^*_{\text{joint}}}[\langle x^{(i)}, x^{(j)} \rangle] - \mathbb{E}_{\hat{\pi}_{\text{seq}}}[\langle x^{(i)}, x^{(j)} \rangle] \big).$$

Intuitively: the joint OT plan is allowed to anti-correlate early-generated components with later targets (co-planning), whereas the AR plan generates each $x^{(i)}$ without knowledge of how $x^{(j)}$ for $j > i$ will be chosen. The missing correlations $\langle x^{(i)}, x^{(j)} \rangle$ are precisely the global structural relationships destroyed by causal factorisation. □

C5(C): EXPOSURE BIAS AS WASSERSTEIN COVARIATE SHIFT

At inference step i, the AR model evaluates $T_i^*(z;\, \hat{x}^{(<i)})$ where $\hat{x}^{(<i)} \sim p_i^{\text{free}}$ (the model's own generated history) rather than $x_{\text{gt}}^{(<i)} \sim p_i^{\text{TF}}$ (ground-truth training history).

The conditional Brenier map is Lipschitz in the history. For smooth data distributions, the conditional Brenier map $T_i^*(z;\, \cdot)$ varies continuously with the conditioning history. When $p_i^{\text{free}} \neq p_i^{\text{TF}}$, the model is solving a different conditional OT problem than it was trained on, leading to sub-optimal transport at each step.

Quantification via Talagrand T_2. Under the sub-Gaussian assumption of Proposition 4.5, the KL divergence between the free-running and teacher-forced conditionals lower-bounds the Wasserstein distance between them: $\Delta_i \geq W_2^2(p_i^{\text{free}}, p_i^{\text{TF}})/(2\sigma_i^2)$. This lower bound compounds over steps: if $W_2^2(p_i^{\text{free}}, p_i^{\text{TF}})$ grows with i, the total exposure-bias loss grows at least linearly in the chain length.

Geometric interpretation. Exposure bias is not a training artefact but a Wasserstein covariate shift: the inference-time query distribution has drifted from the training distribution in the W_2 metric. The theoretically correct fix is not to reduce this drift one step at a time (scheduled sampling), but to prevent the drift from accumulating in the first place by conditioning on a globally coherent plan from the outset — which holistic FM provides. □

C. Architecture Design from the OT Framework

C.1. Hierarchical AFM+HFM: Error Decomposition

Let $\mathcal{E} : \mathcal{X}^N \to \mathbb{R}^K$ be a global encoder producing $z_{\text{global}} = \mathcal{E}(x_1)$. Define:

$$\text{(Level 1, HFM):} \quad \dot{z}_t = v_\phi(z_t, t), \quad z_0 \sim \mathcal{N}(0, I_K), \tag{18}$$

$$\text{(Level 2, AFM):} \quad \dot{x}_t^{(i)} = v_\theta(x_t^{(i)}, x^{(<i)}, z_{\text{global}}, t). \tag{19}$$

The total coupling gap satisfies the hierarchical decomposition:

$$\Delta_{\text{OT}}^{\text{hier}} \leq \underbrace{\Delta_{\text{OT}}^{\text{HFM}}}_{\text{global error}} + \underbrace{\mathbb{E}_z \Big[\Delta_{\text{OT}}^{\text{AFM}|z} \Big]}_{\text{local error}}. \tag{20}$$

Proof: by the chain rule of KL divergence applied to the joint distribution, the total KL from the data joint decomposes into the HFM marginal error and the per-z conditional error. The Wasserstein bound follows from the Talagrand inequality applied to each level.

The hierarchy enables modular improvement: the HFM can be improved independently of the AFM, and the AFM can be conditioned on progressively richer global latents without retraining the HFM.

C.2. Optimal Encoder Dimension

From Proposition 4.7, the reduction in compositional gap from global conditioning is $\sum_{i<j} I(x^{(i)}; x^{(j)} \mid z)$. The encoder dimension K controls how much interaction information is captured. The optimal K^* solves:

$$K^* = \arg\min_K \left[\mathcal{E}_{\mathrm{HFM}}(K) + \alpha\, \mathcal{E}_{\mathrm{AFM}}(K)\right], \tag{21}$$

where $\mathcal{E}_{\mathrm{HFM}}(K)$ is the HFM coupling gap (increasing in K, since larger K is harder to model) and $\mathcal{E}_{\mathrm{AFM}}(K)$ is the residual compositional gap after conditioning (decreasing in K, since more information reduces Δ_{comp}). The trade-off is an information-bottleneck problem: z should capture the interaction terms $\langle x^{(i)}, x^{(j)} \rangle$ most important for compositional coherence while discarding position-level detail that the AFM can handle locally.

C.3. Semi-AR Spectrum and Inference Strategy

From Conjecture 4.10, the optimal number of AR steps T^* minimises:

$$\mathrm{Cost}(T) = T \cdot c_{\mathrm{step}} + \Delta_{\mathrm{comp}}(T) + \Delta_{\mathrm{bias}}(T).$$

In practice, $\Delta_{\mathrm{comp}}(T)$ can be estimated via metric M2 and $\Delta_{\mathrm{bias}}(T)$ via metric M3 on a validation set. The optimal T^* is the crossing point of their sum with the compute cost line, providing a data-driven algorithm for selecting generation strategy that is grounded in the OT error decomposition rather than empirical ablation.

D. Extended Metric Definitions

D.1. M1: Wasserstein Coupling Gap (Computational Protocol)

Algorithm 1 M1: Wasserstein Coupling Gap

Require: Generator G_θ, Noise distribution p_{noise}, Data distribution p_{data}, Batch size B
Ensure: Wasserstein Coupling Gap Δ_{OT}

// Step 1: Sample Generation
Sample $\{z^{(j)}\}_{j=1}^B \sim p_{noise}$
Compute generated samples $\hat{x}^{(j)} = G_\theta(z^{(j)})$ for all $j \in \{1, \ldots, B\}$
// Step 2: Reference Sampling
Sample $\{x_1^{(j)}\}_{j=1}^B \sim p_{data}$
// Step 3: Entropic Optimal Transport (Sinkhorn)
Compute cost matrix $C \in \mathbb{R}^{B \times B}$ where $C_{jk} = \|z^{(j)} - x_1^{(k)}\|^2$
$\varepsilon_S \leftarrow 0.1 \cdot \mathrm{median}(C)$ {Adaptive regularization}
Compute optimal transport plan $\mathbf{P} \leftarrow \mathrm{Sinkhorn}(C, \varepsilon_S)$
Note: $\mathbf{P}$ is the $B \times B$ coupling minimizing $\langle \mathbf{P}, C \rangle - \varepsilon_S H(\mathbf{P})$
// Step 4: Barycentric Projection / Matching
for $j = 1$ to B **do**
  $T^*(z^{(j)}) \leftarrow B \cdot \sum_{k=1}^B \mathbf{P}_{jk} x_1^{(k)}$ {Target via Sinkhorn-matched coupling}
end for
// Step 5: Final Gap Computation
$\Delta_{\mathrm{OT}} \leftarrow \frac{1}{B} \sum_{j=1}^B \|\hat{x}^{(j)} - T^*(z^{(j)})\|^2$
return Δ_{OT}

D.2. M2: Compositional Reasoning Gap — Practical Surrogates

Intractability note. The true Δ_{comp} requires the N-marginal OT cost over $\mathcal{X}^N$, which is computationally intractable for $N > 10$ in general. The following three surrogates are computable alternatives, ordered by fidelity and computational cost.

M2a (Pairwise interaction gap, recommended). This is $O(BN^2)$ and directly estimates the interaction energy

Algorithm 2 M2: Compositional Reasoning Gap

Require: Set of generated sequences $\{\hat{x}_j\}_{j=1}^{B}$ where $\hat{x}_j = (\hat{x}_j^{(1)}, \ldots, \hat{x}_j^{(N)})$, Test data distribution p_{data}
Ensure: Compositional Reasoning Gap Δ_{comp}
{// Step 1: Marginal Cost Computation}
$C_{\text{local}} \leftarrow 0$
for $i = 1$ to N **do**
  Compute $W_2^2(p_{\text{gen}}^{(i)}, p_{\text{data}}^{(i)})$ using standard Sinkhorn matching between the i-th tokens of the batch and B samples from $p_{data}^{(i)}$
  $C_{\text{local}} \leftarrow C_{\text{local}} + W_2^2(p_{\text{gen}}^{(i)}, p_{\text{data}}^{(i)})$
end for
{// Step 2: Joint Cost Computation (Pairwise Approximation)}
Compute $C_{\text{joint}} = W_2^2(p_{\text{gen}}^{(1:N)}, p_{\text{data}}^{(1:N)})$ via multi-marginal Sinkhorn
{Note: For $N > 2$, approximate using the sum of pairwise couplings across position pairs}
{// Step 3: Gap Calculation}
$\Delta_{\text{comp}} \leftarrow C_{\text{joint}} - C_{\text{local}}$
if $\Delta_{\text{comp}} > 0$ **then**
  print "Missing interaction energy: Reasoning failure detected"
else
  print "Compositionally optimal generation"
end if
return Δ_{comp}

formula (11) without any Sinkhorn computation. It is a *diagnostic* for the missing interaction energy, not a certified bound on Δ_{comp}.

M2b (Sliced marginal OT). Project all position embeddings to 1-D via a random direction $u \in S^{d-1}$, compute per-position 1-D W_2^2 costs and the average over positions. Repeat for $K = 100$ random projections. Cost: $O(KNB \log B)$.

M2c (Low-rank joint embedding). Apply PCA to $(\hat{x}^{(1)}, \ldots, \hat{x}^{(N)})$ to obtain a k-dimensional summary $\hat{z} \in \mathbb{R}^k$. Compute W_2^2 between the AR model's $\hat{z}$ distribution and the data's $\hat{z}$ distribution via mini-batch Sinkhorn. This captures joint structure at cost $O(B^2 k)$.

Limitation. None of M2a–M2c provides a certified lower or upper bound on the true Δ_{comp}. The pairwise gap M2a is the most interpretable (it estimates exactly the terms in Proposition 4.3) and is the recommended default. Certified computation of Δ_{comp} for $N \geq 20$ remains open.

D.3. M5: Semi-AR Spectrum Score (Computational Protocol)

Plot $(\Delta_{\text{comp}}(T) + \Delta_{\text{bias}}(T))$ vs. $L(T)$. The area under the Pareto frontier provides the *Spectrum Score*: a single scalar characterising the model's combined reasoning-efficiency profile.

E. Relation to Prior Work on Memorization and Reasoning in DGMs

Memorization in diffusion models. Somepalli et al. (27) and Carlini et al. (10) provide empirical evidence of training-data memorization in diffusion models. The coupling gap framework provides a complementary theoretical characterisation: memorization corresponds to $\Delta_{\text{OT}} \geq c_d n^{-2/d}$, and the question of *when* memorization occurs is answered by the Brenier complexity (9): models with insufficient capacity to fit the Brenier map will default to memorizing training pairs.

Compositional generalization. Compositional generalization in language models has been studied empirically (29; 30), finding that models sometimes fail on systematically novel combinations. The compositional gap Δ_{comp} provides a quantitative measure of this failure: the missing interaction terms $\langle x^{(i)}, x^{(j)} \rangle$ are exactly the cross-position statistics required for novel compositional recombination.

Algorithm 3 M5: Semi-AR Spectrum Score

Require: Step set $\mathcal{T} = \{1, 2, 5, 10, 20, N\}$, Batch size B, Generator G_θ
Ensure: Spectrum Score S_{spec}
Initialize empty list of coordinates $\mathcal{P} \leftarrow \emptyset$
for each T in $\mathcal{T}$ **do**
  {// Step 1: Multi-step Generation}
  Generate B samples $\{\hat{x}_j\}_1^B$ using T discrete denoising/inference steps
  Record average generation latency $L(T)$
  {// Step 2: Metric Evaluation}
  Compute $\Delta_{\text{comp}}(T)$ using Algorithm 2 (M2)
  Compute $\Delta_{\text{bias}}(T)$ using M3 protocol (Wasserstein distance at $\lceil N/2 \rceil$)
  {// Step 3: Combined Metric Calculation}
  $M(T) \leftarrow \Delta_{\text{comp}}(T) + \Delta_{\text{bias}}(T)$
  Append point $(L(T), M(T))$ to $\mathcal{P}$
end for
{// Step 4: Pareto Analysis and Integration}
Sort points in $\mathcal{P}$ by latency $L(T)$
$S_{spec} \leftarrow$ AreaUnderCurve($\mathcal{P}$) {Integrate using trapezoidal rule}
return S_{spec}

Exposure bias. Ranzato et al. (31) identified exposure bias as a core challenge for AR generation. Scheduled sampling (6) and REINFORCE-based methods (32) address it heuristically. The Wasserstein reformulation (Proposition 4.5) shows that these methods reduce $W_2(p_i^{\text{TF}}, p_i^{\text{free}})$ per step but do not address the joint distributional shift across all steps. Holistic FM guidance addresses this joint shift, providing a stronger and theoretically grounded correction.

OT in generative modelling. The Benamou–Brenier connection to FM was established in (14); the OT coupling for FM in (24; 20). The multi-marginal OT connection to sequential generation (19) has not previously been applied to autoregressive models in the compositional reasoning context.

References for Appendix E:

1. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., & Goldstein, T. (2023). Diffusion art or digital forgery? Investigating data replication in diffusion models. *CVPR 2023*.
2. Lake, B., Ullman, T., Tenenbaum, J., & Gershman, S. (2018). Building machines that learn and think like people. *Behavioral and Brain Sciences*.
3. Keysers, D. et al. (2019). Measuring Compositional Generalization: A Comprehensive Method on Realistic Data. *ICLR 2020*.
4. Ranzato, M., Chopra, S., Auli, M., & Zaremba, W. (2015). Sequence level training with recurrent neural networks. *ICLR 2016*.
5. Williams, R. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3), 229–256.