MASTERING MEMORIZATION IN FLOW-BASED GENERATIVE MODELS VIA THE MOREAU ENVELOPE

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow-based models have shown unprecedented generation capabilities across a wide range of modalities, but they suffer from the not fully understood phenomenon of memorization - the verbatim reproduction of data samples despite valid model convergence. This degrades their creativity and applicability in low data regimes as well as their reliability in secure settings. On the other hand, memorization could be the key to a new form of controlled generation, by memorizing features corresponding to specific states to observe, but this has not yet been studied for this purpose. We thus introduce a novel framework to achieve controlled memorization, with a dial allowing model optimization to decide to what extent it should be mitigated or enforced. This is made possible by turning a memorization metric into a Kantorovich potential via the Moreau Envelope, leveraging the proximal of the envelope to build an attractive or repulsive memorization field. We further provide an Optimal Transport interpretation of the proposed method, and derive a proximal gradient descent algorithm for the computation of the proximal during model inference. Moreover, we propose a number of illustrative use cases across various modalities; these range from preventing memorization in low data regime image generation and leverage the perturbation in text domain to steer flow-based language models token generation towards using specific words, showcasing the versatility of the proposed method.

1 INTRODUCTION

Flow-Matching and Diffusion models are trained to transport noise onto the empirical data distribution, a finite sum of Dirac's deltas centered on the training samples. Reproducing those samples exactly is therefore what their objective encourages, and it is observed in practice: text-conditional Diffusion models have been shown to produce verbatim copies of their training images (Somepalli et al., 2023; Carlini et al., 2023). A deep dive into this insight has been carried out theoretically for both classes of models, revealing that the loss objective itself necessarily leads the generative vector field or the score to reproduce training data samples, as shown by Gao & Li (2026) for Flow-Matching and Buchanan et al. (2025) for Diffusion. Importantly, the verbatim reproduction of a subset from the training dataset is undesirable and can be harmful when the latter is made of copyrighted images or biometrics data.

In practice, previous theoretical findings on full memorization via the score or velocity does not explain the generalization capabilities of Flow and Diffusion models, as they routinely produce samples absent from the training dataset, memorizing only partially. Further studies (Bonnaire et al., 2025; Buchanan et al., 2025) have thus leaned on the question for Diffusion models and revealed the existence of distinct memorization and generalization phases emerging during training, depending of model capacity and the amount of training data, supporting the previous hypothesis that generalization in such models arises from the incapacity of fitting the training objective perfectly (Yoon et al., 2023). On the same note, Chen (2026) points to the direction that neural networks naturally smooth the empirical score learned by Diffusion models, explaining their unexpected ability to generalize by a form of interpolation. Despite these considerations, Bonnaire et al. (2025) and Gao & Li (2026) still suggest that in low data regimes memorization is bound to happen. This is a critical limitation in applications where data is scarce, which is typically the case, e.g., for medical applications.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Given that the subtleties of memorization are still relatively underexplored in flow-based models, we aim to achieve a systematic way of preventing verbatim generation of training samples while keeping generative capabilities of the model at hand, even in low data regimes. Pushing this idea further, we may even wonder if partially enforcing memorization in a controllable fashion is achievable and beneficial in some applications flow-based generative models. A natural question could thus be:

How to balance memorization mitigation and enforcement while preserving the model's learned generative capabilities?

We propose to answer this question by perturbing the learned velocity field with a memorization-aware vector field. We then introduce a framework to mitigate or enforce memorization based on the Moreau envelope of the memorization metric proposed by Yoon et al. (2023), leveraging the notion of prox-regularity to bypass its non convexity. Specifically, we prove that the Moreau envelope of the memorization metric is a valid Kantorovich potential whose induced Brenier map is the proximal mapping itself, transporting the sample distribution onto the locally most memorized distribution. The functional this yields is the Wasserstein-2 Moreau envelope of expected memorization, whose minimizer is exactly one JKO step of the corresponding gradient flow, and our perturbation is exactly its Wasserstein gradient. Ascending or descending it therefore enforces or mitigates memorization through an explicit transport mechanism. Our contributions are summarized as follows:

- We propose an inference-time perturbation for trained flow models using the gradient of the Moreau envelope to control to what extent memorization should be enforced or mitigated;
- We derive an Optimal Transport interpretation of the framework and a proximal algorithm to compute the perturbation during inference without significant overhead;
- We demonstrate the versatility of our framework by showing its ability to mitigate memorization in low data regimes, as well as its ability to steer flow-based language models to enforce the generation of specific words.

1.1 Related Works

Memorization in Flow-Matching models Much effort has been made to understand and mitigate the phenomenon of memorization in diffusion models, both empirically and theoretically, while flow-based generative models are still relatively underexplored. A notable drawback of models that memorize the training data is that an attacker can perform Membership Inference (Carlini et al. (2022) to query for data samples; this was recently investigated for Flow-Matching models by Rao & Moyer (2026), and Sesmat et al. (2026), providing valuable insights in the vector field dynamics and further motivating research for the roots of memorization in such models. A recent attempt by Gegenfurtner et al. (2026) perturbs the learned vector field through a Riemannian Laplace approximation, but operates in parameter space and shifts memorized images by adding noise rather than altering them semantically. In our work, we focus on the inference-time mitigation of memorization in flow-based models, leveraging a perturbation to mitigate memorization while semantically altering generated samples to avoid memorization.

Wasserstein Gradient Flows for generative modeling Unifying generative models through the Wasserstein Gradient Flow (WGF) is a research direction gaining traction, as many instances of generative models can be understood through this paradigm as argued in Caucheteux et al. (2026). This class of models mainly focuses on the description of the density evolution describing the steepest descent of a specific functional, later leveraging its Wasserstein-2 gradient to describe the transport dynamics between a source and target distribution. A few models building on the recent Drifting paradigm by Deng et al. (2026) specifically design the gradient flow of known divergences such as the Sinkhorn Divergence in W-Flow models by Han et al. (2026), or the Maximum Mean Discrepancy (MMD) in Cao et al. (2026); achieving excellent speed and image generation quality. In our work, we follow this insightful trend focusing on the WGF to unify flow dynamics through the continuity equation, seeking an addition to the used functional that would theoretically yield lower or higher memorization; later deriving its Wasserstein gradient to manipulate memorization.

Flow Based Models Flow-matching models (Lipman et al., 2023; Liu et al., 2023) transport a source distribution to a target distribution via a time-dependent vector field describing the density evolution from source to target given by a continuity equation. A notable framework encapsulating and bridging formulations across Flow-Matching and Diffusion can be seen in the Stochastic Interpolants

framework (Albergo et al., 2025; Albergo & Vanden-Eijnden, 2023), allowing findings for both class of model to regroup under Gaussianity of the sources for instance. Other approaches in generalizing flow models to a wider variety of sources and target distributions were presented by Tong et al. (2024) and Tong et al. (2023), notably providing dynamical Optimal Transport approximations of the generative vector field akin to the OT-fluid dynamic formulation of Benamou & Brenier (2000) through their OT-CFM models. Interestingly, the flexibility of the interpolant framework formulation led to the recent advent of flow-based language models such as Embedded Language Flows (ELFs) (Hu et al., 2026) and Discrete Flow Maps (Potaptchik et al., 2026) challenging successful diffusion alternatives in both speed and performance. Similar contributions include that of Lee et al. (2026). In this work, the flexibility of Flow-based formulations across modalities allows us to interpret them in a principled manner without the need to specialize our method to specific data modalities such as text or images.

2 Preliminaries & Mathematical background

2.1 Wasserstein Gradient Flows

Flow-based generative models (Han et al. (2026), Cao et al. (2026)) can be expressed under the general setting of the Wasserstein Gradient Flow (WGF), which describes the steepest descent of an energy functional in the manifold of probability measures under the Wasserstein-2 metric (Ambrosio et al., 2008). In the flow-matching landscape, a particular instance of that is the OT-CFM model (Tong et al., 2024), which by approximating the dynamical OT formulation from Benamou & Brenier (2000) naturally describes the WGF of the Wasserstein-2 distance.

With $\mathcal{F}[\rho]$ being a functional in the space of positive measures with finite second moments, the change of density leading to the steepest descent of $\mathcal{F}$ is expressed as follows:

$$\partial_t \rho_t = \nabla \cdot \left(\rho_t \nabla_x \frac{\delta \mathcal{F}[\rho]}{\delta \rho}(\rho_t) \right) \tag{1}$$

Where $\frac{\delta \mathcal{F}[\rho]}{\delta \rho}$ is the first variation of $\mathcal{F}$. The spatial gradient $\nabla_x \frac{\delta \mathcal{F}[\rho]}{\delta \rho}(x)$ is called the *Wasserstein Gradient* of $\mathcal{F}$ at ρ_t, and constitutes a vector field.

One can notably recover known continuity equations for Diffusion considering $\mathcal{F}$ to be the Kullback-Leibler divergence between ρ and a Gaussian source distribution. A recent class of generative models explicitly leverage the WGF formulation using functionals such as the MMD distance (Cao et al., 2026) or the Sinkhorn Divergence (Han et al., 2026) in place of $\mathcal{F}$.

$$\partial_t \rho_t = \nabla \cdot (\rho_t v_t) \quad \text{with} \quad v_t(x) = -\nabla_x \frac{\delta \mathcal{F}[\rho]}{\delta \rho}(\rho_t)(x) \tag{2}$$

2.2 Optimal Transport

The Wasserstein-2 metric introduced in the last section is a central object from the theory of Optimal Transport (OT), a field concerned with the cost-constrained transport of probability mass between measures (Villani, 2009). We recall here only what the construction of Section 3 requires, following the notation of Friesecke (2024); a fuller account of the Kantorovich problem and its duality is given in Appendix B.

Let $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$, $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$, and let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous cost. The Kantorovich formulation seeks the transport plan with μ and ν as marginals minimizing the expected cost, over the candidate space $\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \mid \pi_{1\#}\gamma = \mu,\ \pi_{2\#}\gamma = \nu\}$, which defines the Wasserstein distance for the cost c:

$$\mathcal{W}^c(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \mathrm{d}\gamma(x, y) \tag{3}$$

Through Fenchel-Rockafellar duality, this problem admits a dual formulation whose variables, the *Kantorovich potentials*, give valuable economic and geometric insights on the primal problem. A pair of potentials is admissible when it belongs to

$$\Phi(c) = \{(\varphi, \psi) \in \mathcal{C}_b(\mathcal{X}) \times \mathcal{C}_b(\mathcal{Y}) \mid \forall (x, y) \in \mathcal{X} \times \mathcal{Y},\ \varphi(x) + \psi(y) \leq c(x, y)\} \tag{4}$$

and Kantorovich duality states that $\mathcal{W}^c(\mu,\nu) = \sup_{(\varphi,\psi)\in\Phi(c)} \left(\int_{\mathcal{X}} \varphi \,\mathrm{d}\mu + \int_{\mathcal{Y}} \psi \,\mathrm{d}\nu\right)$.

For the halved squared Euclidean cost, Brenier's theorem expresses the primal transport plan through the first Kantorovich potential alone, gives unicity guarantees, and recovers the original formulation of Optimal Transport due to Monge. It extends to any cost satisfying the *twist condition*, which is the form we use.

Definition 1 (Twist Condition)**.** A differentiable cost $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ satisfies the twist condition iff the map $y \mapsto \nabla_x c(x, y)$ is injective on $\mathbb{R}^d$.

Theorem 1 (Brenier, twisted form)**.** *Let $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$ be compact, $\mu \in \mathcal{P}(\mathcal{X})$ absolutely continuous, $\nu \in \mathcal{P}(\mathcal{Y})$, and let c satisfy Definition 1. Then the optimal transport plan $\gamma \in \Pi(\mu,\nu)$ exists uniquely and is of Monge form $\gamma = (\mathrm{id}, T)_{\#}\mu$, with T the unique measurable map pushing μ forward to ν, given in terms of the optimal first Kantorovich potential φ by*

$$T(x) = (\nabla_x c(x,\cdot))^{-1}(\nabla_x \varphi(x)) \tag{5}$$

The potentials φ and ψ are c-conjugate, and for $c(x,y) = \frac{1}{2}\|x-y\|^2$ Equation 5 reduces to the classical Brenier map $T(x) = x - \nabla\varphi(x)$.

3 METHODOLOGY

To prevent or enforce memorization in flow-based generative models, we derive a geometric intervention acting on the sampling trajectories: particles still reach the data manifold, but whether they collapse onto a training point becomes something we control. To this end, the generative vector field responsible for pushing the particles forward is combined with a "memorization" vector field that achieves the desired attractive or repulsive effect.

We posit that the design of such a geometric intervention at the vector field level can be made by finding an appropriate functional to calculate a Wasserstein gradient, thus yielding a convenient vector field. Considering the dual formulation of the optimal transport problem, we analyze the memorization metric proposed by Yoon et al. (2023) and work to turn it into an appropriate Kantorovich potential.

Let $\mathcal{D} = \{y_i\}_{i=1}^n$ be a dataset and $\hat{\nu} = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ the corresponding empirical data distribution. The memorization metric we consider is:

$$S(x) = \frac{\left\|x - \mathrm{nn}_1^{\mathcal{D}}(x)\right\|^2}{\left\|x - \mathrm{nn}_2^{\mathcal{D}}(x)\right\|^2} \tag{6}$$

where $\mathrm{nn}_i^{\mathcal{D}}(x)$ corresponds to the ith nearest neighbor of x in the dataset $\mathcal{D}$, and a sample is defined as memorized when $S(x) < \tau$ for a chosen threshold τ.

The function S provides a direct way of assessing how a sample may qualify as memorized. Its value is close to one when the generated sample is nearly as far from its first nearest neighbor as it is from its second, signaling a sample that does not resemble any particular element of the dataset according to the metric above. Conversely, a value close to zero signals that the sample is much closer to its first neighbor than to its second, indicating a verbatim copy of that neighbor.

3.1 VORONOI BOUNDARIES

A useful way of partitioning the space around the datapoints is to consider the Voronoi tesselation that they define. In the context of memorization, they directly provide a notion of neighbors and interpolation. We define the Voronoi sets as follows:

$$\mathcal{V}_i = \left\{ x \in \mathbb{R}^d \;\middle|\; \|x - y_i\|^2 \le \|x - y_j\|^2 ,\ \forall j \neq i \right\} \tag{7}$$

With this partitioning of the input space, we can refine the function S to make it valid per Voronoi space, thus fixing the value of the first nearest neighbor. This simple consideration is powerful in practice as it allows our function to unlock Moreau's envelope (Moreau, 1965), a useful regularization that has been initially developed for convex functions, but that can be extended to non-convex

functions that verify prox-regularity (Poliquin & Rockafellar, 1996). We thus write the per-Voronoi memorization potential in a form that allows us to access prox-regularity by strong amenability everywhere.

We consider $S_i : \mathcal{V}_i \to [0, 1]$ and fix $p_i = y_i$ as follows:

$$S_i(x) = \max_{y \in \mathcal{D} \setminus \{p_i\}} \left[\frac{\|x - p_i\|^2}{\|x - y\|^2} \right] \tag{8}$$

Proposition 1. *The functions S_i (for $i \in \{1, \dots, n\}$) are strongly amenable everywhere in their respective Voronoi cells. As such, the latter functions are prox-regular.*

Prox-regularity is what makes the proximal mapping single-valued and the envelope continuously differentiable in a neighborhood of each anchor, for λ small enough. The proof is deferred to Appendix A.

3.2 The Moreau envelope

With prox-regularity we can rigorously define the Moreau envelope of the memorization potential. Its role here is twofold. First, the envelope shares its minima with the function itself while being $\mathcal{C}^{1+}$, which lets us optimize the envelope in place of S without altering its interpretation. Secondly, under Optimal Transport lenses, the Moreau envelope is the c-transform of minus the function at hand, which will let us recover a second Kantorovich potential once we have a guess for the first one.

Let f be a suitable function. The Moreau envelope of f and its associated proximal mapping are defined as:

$$\mathcal{M}_\lambda f(x) = \inf_y \left\{ f(y) + \frac{1}{2\lambda} \|x - y\|^2 \right\}, \qquad \operatorname{prox}_{\lambda f}(x) = \arg\min_y \left\{ f(y) + \frac{1}{2\lambda} \|x - y\|^2 \right\} \tag{9}$$

Intuitively, the additional distance term acts as a regularizer controlling the closeness of the candidate points of the infimum to the evaluated point. In the context of memorization, it returns the point closest to x that is the most memorized, within a radius controlled by λ. When the $\arg\min$ is a singleton, the envelope is differentiable at x and its gradient reads:

$$\nabla_x \mathcal{M}_\lambda f(x) = \frac{x - \operatorname{prox}_{\lambda f}(x)}{\lambda} \tag{10}$$

Geometrically distancing from the proximal point in a consistent way would therefore allow a sample to avoid memorization, and conversely moving towards it would enforce memorization.

3.3 Inverse optimal transport problem

We now show that $\mathcal{M}_\lambda S$ is a legitimate first Kantorovich potential, identify the transport problem it solves, and derive from it the functional whose Wasserstein gradient is exactly the perturbation we seek.

The cost $c(x, y) = \frac{1}{2\lambda} \|x - y\|^2$ satisfies the twist condition, with $(\nabla_x c(x, \cdot))^{-1}(z) = x - \lambda z$. Taking $\varphi = \mathcal{M}_\lambda S$ as a candidate first potential and applying Equation 10 in the generalized Brenier map of Equation 5:

$$T(x) = (\nabla_x c(x, \cdot))^{-1}(\nabla_x \mathcal{M}_\lambda S(x)) = x - \lambda \left(\frac{x - \operatorname{prox}_{\lambda S}(x)}{\lambda} \right) = \operatorname{prox}_{\lambda S}(x) \tag{11}$$

The Brenier map induced by our candidate potential is therefore the proximal mapping itself. Since the transport plan is $\gamma = (\mathrm{id}, T)_{\#}\mu$ with $\gamma \in \Pi(\mu, \nu)$, the target measure is necessarily $\nu = \operatorname{prox}_{\lambda S \#}\mu$: the potential $\mathcal{M}_\lambda S$ solves the transport problem from μ to the pushforward of μ by the proximal mapping, that is, towards the distribution locally closest to μ that yields the most memorization. This inverse reading of the dual problem is made exact by the following proposition.

Proposition 2 (Transport interpretation). *$\mathcal{M}_\lambda S$ is c-concave for $c(x,y) = \frac{1}{2\lambda}\|x-y\|^2$, with c-transform $(\mathcal{M}_\lambda S)^c = -S$ on the range of $\mathrm{prox}_{\lambda S}$. Consequently $\mathrm{prox}_{\lambda S}$ is an optimal map from μ to $\mathrm{prox}_{\lambda S\#}\,\mu$, and*

$$\mathcal{W}_c\big(\mu, \mathrm{prox}_{\lambda S\#}\,\mu\big) \;+\; \int_{\mathcal{X}} S \circ \mathrm{prox}_{\lambda S}\;\mathrm{d}\mu \;=\; \int_{\mathcal{X}} \mathcal{M}_\lambda S\,\mathrm{d}\mu. \tag{12}$$

Equation 12 decomposes the duality gap into two interpretable pieces. The first term is the distributional distance between the source and its proximal pushforward; minimizing it contributes to keeping the flow close to its own trajectory. The second is the expected memorization of the proximal distribution, which is non-negative since $S \geq 0$. The sum of the two is the expected Moreau-memorization, and because the second term is non-negative we immediately obtain

$$\mathcal{W}_c\big(\mu, \mathrm{prox}_{\lambda S\#}\,\mu\big) \;\leq\; \int_{\mathcal{X}} \mathcal{M}_\lambda S\,\mathrm{d}\mu \;=:\; \mathcal{E}_\lambda[\mu]. \tag{13}$$

We therefore take $\mathcal{E}_\lambda$ as the functional driving the flow: acting on it controls the transport cost towards the memorized distribution through the bound, while being considerably better behaved than $\mathcal{W}_c(\mu, \mathrm{prox}_{\lambda S\#}\,\mu)$ as argued in Appendix A.4. The next proposition identifies $\mathcal{E}_\lambda$ within the Wasserstein gradient flow picture of Section 2.1 and gives its Wasserstein gradient.

Proposition 3 (JKO step and Wasserstein gradient). *Let $\mathcal{V}[\mu] = \int S\,\mathrm{d}\mu$ denote expected memorization. Then $\mathcal{E}_\lambda$ is the $\mathcal{W}_2$-Moreau envelope of $\mathcal{V}$ at scale λ,*

$$\mathcal{E}_\lambda[\mu] \;=\; \inf_\nu \left\{ \mathcal{V}[\nu] + \frac{1}{2\lambda}\mathcal{W}_2^2(\mu,\nu) \right\}, \tag{14}$$

whose unique minimizer is $\nu^\star = \mathrm{prox}_{\lambda S\#}\,\mu$. Moreover $\mathcal{E}_\lambda$ is linear in μ, so its first variation is $\mathcal{M}_\lambda S$ and its Wasserstein gradient is

$$\nabla_x \frac{\delta \mathcal{E}_\lambda}{\delta \mu}(\mu)(x) = \nabla_x \mathcal{M}_\lambda S(x) = \frac{x - \mathrm{prox}_{\lambda S}(x)}{\lambda}. \tag{15}$$

Equation 14 is precisely one JKO step for the functional $\mathcal{V}$, with λ playing the role of the step size. The proximal mapping is thus twice the same object: the transport map of Proposition 2, and the map performing a single step of the Wasserstein gradient flow of expected memorization. Descending $\mathcal{E}_\lambda$ drives the distribution away from its locally most memorized form, and ascending it drives the distribution towards it. Since the perturbation is available in closed form once the proximal is known, and is locally Lipschitz by prox-regularity, we can add it to the velocity field of a trained model and reroute generation during ODE integration with a signed strength γ:

$$v_t(x) = v_\theta(x,t) + \gamma \nabla_x \mathcal{M}_\lambda S(x) \tag{16}$$

where $\gamma > 0$ mitigates memorization and $\gamma < 0$ enforces it, the magnitude acting as the dial announced in the introduction. We further let λ depend on the integration time.

3.4 Algorithmic considerations

Computing the perturbation of Equation 16 amounts to computing the proximal mapping, which is out of reach in closed form. We therefore approximate it with a fixed-point iteration obtained by reasoning on the subdifferential of the proximal objective, as done in the convex case by Parikh & Boyd (2014); prox-regularity (Proposition 1) replaces the convexity requirement, with the limiting subgradients of S_i reducing to $\{\nabla S_i\}$ on the relevant neighborhood.

Algorithm 1 APGM: Accelerated Proximal Gradient Method

Input: y, λ, $\bar{\lambda}$
Return: $\mathrm{prox}_{\lambda S}(y)$

Initialize $x^0 \leftarrow y,\ x^1 \leftarrow y$
for $k = 1 \dots N$ **do**
  $z^{k+1} = x^k + \omega^k(x^k - x^{k-1})$
  $x^{k+1} = \frac{1}{\lambda+\bar{\lambda}}\left[\bar{\lambda}y + \lambda(z^{k+1} - \bar{\lambda}\nabla S(z^{k+1}))\right]$
end for
return x^N

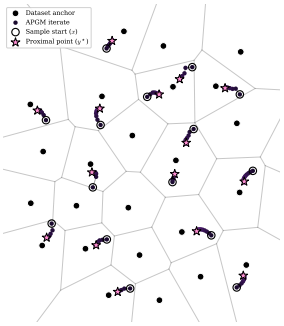


Figure 1: APGM iteration traces

Finding the proximal at a point y is the composite problem $\min_x\{S(x) + g(x)\}$ with $g(x) = \frac{1}{2\lambda}\|x - y\|^2$. A minimizer $x^\star$ is a fixed point of the forward-backward operator $\mathrm{prox}_{\bar{\lambda}g}(\mathrm{id} - \bar{\lambda}\nabla S)$, and since g is a scaled squared distance its proximal is available in closed form, $\mathrm{prox}_{\bar{\lambda}g}(z) = \frac{\bar{\lambda}y+\lambda z}{\lambda+\bar{\lambda}}$. This yields the fixed-point condition

$$x^\star = \frac{1}{\lambda+\bar{\lambda}}\left[\bar{\lambda}y + \lambda\left(x^\star - \bar{\lambda}\nabla S(x^\star)\right)\right] \tag{17}$$

which we solve with the accelerated scheme of Algorithm 1, using $\omega^k = \frac{k}{k+3}$. The two scales play distinct roles and we keep them separate in practice: λ sets the radius of the proximal search, and therefore the locality of the perturbation, while $\bar{\lambda}$ is the step size of the inner solver and is constrained by the local Lipschitz modulus of ∇S. The derivation of Equation 17, the role of $\bar{\lambda}$, and the convergence behaviour of the scheme are detailed in Appendix D.

4 Experiments

We evaluate the proximal perturbation across three modalities, testing both signs of the dial of Equation 16. Our main study is the mitigation regime on a memorizing latent flow model trained on medical images (Section 4.1), where we isolate the contribution of the envelope against natural baselines and characterize the interpolation behavior it induces. We then test the enforcement regime on a flow-based language model, steering unconditional generation towards a set of anchor words (Section 4.2). A two-dimensional study, which exhibits both regimes on the same model and serves as a sanity check of the geometry described in Section 3, is reported in Appendix E.

4.1 Image domain: latent flow matching

Setup. We train a latent flow matching model with minibatch OT-CFM (Tong et al., 2024) on the full RetinaMNIST dataset (Yang et al., 2023), 1000 images at 64×64, encoded with the Stable Diffusion variational autoencoder (Rombach et al., 2022) into 256-dimensional latents. The dataset is deliberately small: it reflects the data scarcity inherent to medical imaging, and places the model in the regime where memorization is expected to be unavoidable (Bonnaire et al., 2025; Gao & Li, 2026). The resulting model indeed memorizes, with 91.8% of its samples falling below $\tau = 1/3$ in latent space. Training and sampling details are given in Appendix F.

Measurement. We report the memorization rate $S < \tau$ both in latent space and, independently, in raw pixel space: the perturbation acts in a projection of the former, so the latter is a genuinely different measurement domain and agreement between the two rules out the method merely optimizing the quantity it is scored on. Fidelity is measured by the Fréchet distance in DINOv2 feature space (Oquab et al., 2024) against a *held-out* split, since a model reproducing its training data would score best against the training set by construction. We additionally report the fraction of samples having a near-duplicate elsewhere in the generated set, a diversity diagnostic involving no other distribution. Appendix F.2 discusses the calibration of the perceptual metrics.

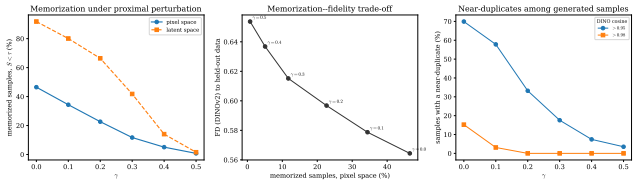


Figure 2: Memorization control on RetinaMNIST. **Left:** the memorization rate falls monotonically with γ and agrees between the latent space, where the perturbation acts, and pixel space. **Center:** the induced trade-off against held-out fidelity **Right:** near-duplicates among generated samples.

Memorization is controlled monotonically. Figure 2 shows memorization falling smoothly with γ in both measurement spaces, with the near-duplicate rate tracking it. We note that held-out fidelity degrades gradually rather than collapsing. Figure 3 shows the corresponding distributions of S: the perturbation redistributes mass from the memorized region across the whole range.

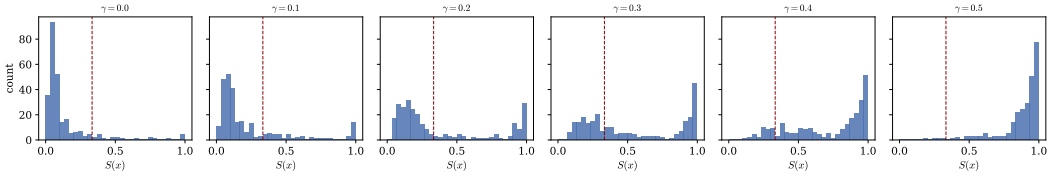


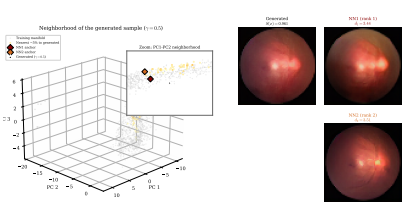
Figure 3: Distribution of the memorization metric S in latent space across γ. The dashed line marks the threshold $\tau = 1/3$.

Effect of the Moreau envelope A natural question is asking if any field pushing samples away from their nearest training point would achieve the same effect. Table 1 compares the Moreau field against the raw gradient ∇S and against unit repulsion from the nearest anchor, all at $\gamma = 0.3$ with identical initial noise and integration budget. At their natural scale, both leave memorization at the unperturbed level. This is the envelope's scale selection at work, since λ sets the magnitude of $\nabla \mathcal{M}_\lambda S$ adaptively. Norm-matching each baseline to the median perturbation norm of the Moreau field at every integration time makes both effective, but neither best the envelope (Table 1). The excess integration cost of ∇S is thus attributable to its irregularity rather than its strength.

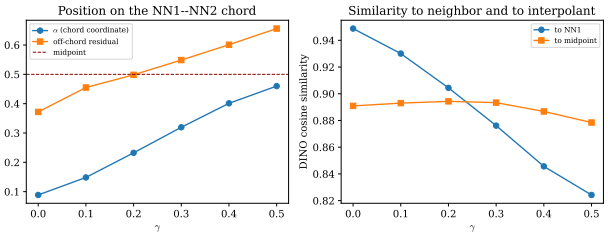
Table 1: Perturbation field ablation at $\gamma = 0.3$, with identical initial noise and a fixed-step integration budget. NFE is the number of function evaluations of an adaptive `dopri5` solver on the same problem. Matched rows are rescaled to the median perturbation norm of the Moreau field.

Field	$S < \tau$ latent (%)	$S < \tau$ pixel (%)	FD (held-out) ↓	NFE ↓
Unperturbed	90.6	50.8	0.577	602
∇S (raw)	89.8	50.0	0.578	602
NN repulsion	91.4	49.2	0.583	602
∇S (norm-matched)	50.0	16.4	0.614	872
NN repulsion (norm-matched)	53.1	17.2	0.659	602
$\nabla \mathcal{M}_\lambda S$ (ours)	**46.1**	**11.7**	**0.625**	**602**

Samples move towards interpolants of their neighbors. Writing z_1, z_2 for the two nearest training latents of a generated sample x, we measure the coordinate $\alpha = \langle x - z_1, z_2 - z_1 \rangle / \|z_2 - z_1\|^2$ along the chord between them. Figure 4 shows α moving from near zero towards the midpoint as γ increases, with essentially every sample ending up between its two neighbors, while perceptual similarity to the nearest neighbor falls below similarity to the decoded midpoint of the two. The generated images acquire features of both neighbors accordingly. The full measurements, including the off-chord residual, are reported in Appendix F.3.



(a) A generated sample and its two nearest neighbors

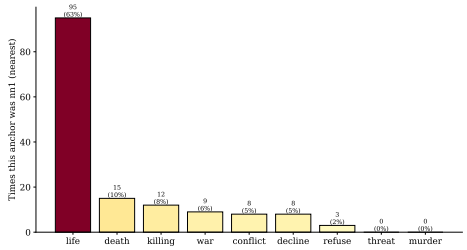


(b) Position relative to the neighbors, across γ

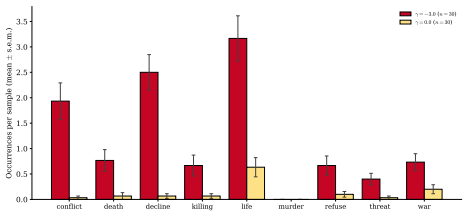
Figure 4: Neighbor interpolation. As γ grows the sample leaves its nearest training image and moves towards the midpoint of its two neighbors, sitting near the boundary where S approaches one.

4.2 Embedded Language Flows steering

We now showcase the attractive regime ($\gamma < 0$) on a modality where the anchors are not training data to demonstrate the generality of the construction, perturbing the ELF-B flow-based language model (Hu et al., 2026). The experimental setup of this experiment is explained in Appendix G.



(a) Anchor Voronoi cell occupancy



(b) Anchor word occurrences, steered against baseline

Figure 5: Steering ELF-B towards 9 anchor words. **Left**: how often each anchor is the nearest neighbor of an ordinary generated token position. **Right**: occurrences per sample (mean $\pm$ s.e.m.) across 30 paired samples at $\gamma = -3$ against $\gamma = 0$.

Results. The attractive field increases the usage of eight of the nine anchor words (Figure 5b), every one of them by more than 2.8 standard errors of the paired difference. The ninth word, *murder*, is never produced in either condition, so the field has no occurrence to act on rather than failing to steer one. Figure 5a explains why the effect is uneven across words. Since S is defined against the nearest anchor, a position is steered towards whichever anchor cell it occupies, and the anchor cells are strongly unequal: *life* is the nearest anchor for 63% of ordinary generated positions while *threat* and *murder* are nearest for none of the 150 sampled. Because the field acts at every token position, the whole sequence is displaced relative to its unperturbed counterpart at the same initial noise: paired generations differ throughout, carrying different subject matter and different surrounding vocabulary (shown in Appendix G.6). What distinguishes those positions is that their latent lies close enough to an anchor for the attractive field to collapse them onto it, while the remainder drift without reaching a cell centre. Lastly, we note that the perturbation does not collapse entropy and perplexity; further details are deferred to Appendix G.5.

5 Conclusion

We introduced an inference-time perturbation that mitigates or enforces memorization in flow-based generative models through a term obtained by turning a standard memorization metric into a Kantorovich potential via its Moreau envelope. The resulting field admits an exact Optimal Transport reading: its proximal mapping is simultaneously the optimal map towards the locally most memorized distribution and the map realizing one JKO step of the Wasserstein gradient flow of expected memorization, which makes the perturbation the Wasserstein gradient of an explicit functional rather than a heuristic correction.

AI USE STATEMENT

We used large language model assistants during the preparation of this work, and describe their role here in the interest of transparency.

The research contribution is our own. The central idea of turning a memorization metric into a Kantorovich potential through its Moreau envelope, the optimal transport reading of the resulting field, the research questions, the methodology, the narrative of the paper, and the design of all three experimental settings are entirely ours. We note that LLM assistants were initially unpersuaded by several of the claims we ultimately prove here, and that the arguments which resolved those disagreements were ours.

LLM assistants were used as collaborative tools in three ways. First, for writing: polishing prose, tightening exposition, and reorganizing material between the main text and the appendices. Second, for implementation: writing and refining experiment, evaluation and plotting code in dialogue with us, substantially beyond autocompletion, but always against an experimental design and a set of measurements that we specified. Third, for the mathematics: the proofs in Appendix A were restructured and expanded for clarity with LLM assistance, and the discussion that produced them helped us identify and correct an error in an earlier version of our transport derivation. Every resulting statement was checked by us.

The authors take full responsibility for the entire content of this paper, including any errors introduced or left uncorrected during this process.

REPRODUCIBILITY STATEMENT

All theoretical claims are stated as propositions in the main text with complete proofs in Appendix A; Appendix B collects the optimal transport background they rely on, and Appendix A.4 records the alternative derivation we did not adopt, together with the reason. The proximal solver is given as Algorithm 1, its fixed-point derivation and the distinct roles of its two scale parameters in Appendix D.

Every experiment uses public data and public checkpoints. The two-dimensional study (Appendix E), the latent flow matching study on RetinaMNIST (Appendix F) and the ELF-B steering study (Appendix G) each report the training or checkpoint provenance, the sampler and step count, the envelope and solver parameters and the sample counts behind every reported number. Appendix F.2 documents the calibration of the perceptual metrics and states which comparisons are made against raw and which against reconstructed training images, since that choice materially affects the numbers.

ETHICS STATEMENT

This work is dual-use by construction: the same field that steers a model away from reproducing its training data steers it towards doing so. Mitigation addresses the concrete harm of verbatim reproduction of copyrighted or sensitive data, acute in the low-data settings where our image experiments sit. Enforcement carries two risks we state plainly. It can be used to extract training content from a model, and, more subtly, the degree to which a model can be forced towards a candidate sample is itself a membership signal, so the construction suggests a privacy attack on models we did not set out to build. We judge disclosure the better course, since the mechanism follows from published components and the same signal is a privacy-auditing tool in the hands of a data owner.

Our text anchors are words denoting violence, chosen solely because they are frequent enough in unconditional generation to be harvested and counted; we neither endorse nor promote that content, and any lexical set with the same statistics would have served. We report the experiment plainly because it doubles as a warning: steering a public language model towards an arbitrary concept set required no training and no privileged access, only a modification of the ODE integration at inference. Vector-field intervention is an unguarded surface on open-weight flow-based models, and we would rather it be characterized in the open, with the diagnostics of Appendix G.5 that let its strength be measured. Finally, our method reduces measured memorization but does not eliminate it, and is not a privacy guarantee. All datasets are public benchmarks with no personally identifying information.

REFERENCES

Michael Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025. URL http://jmlr.org/papers/v26/23-1605.html.

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2 edition, 2008. ISBN 978-3-7643-8721-1. doi: 10.1007/978-3-7643-8722-8.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, January 2000. ISSN 0945-3245. doi: 10.1007/s002110050002. URL http://dx.doi.org/10.1007/s002110050002.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mezard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 141266–141286. Curran Associates, Inc., 2025. doi: 10.52202/085713-4720. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/ceb7f3cc876a6dcb15130a645b5a4507-Paper-Conference.pdf.

Sam Buchanan, Druv Pai, Yi Ma, and Valentin De Bortoli. On the edge of memorization in diffusion models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 96113–96157. Curran Associates, Inc., 2025. doi: 10.52202/085713-3213. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/8a6a3f289e3e2654509ccfdd631e592e-Paper-Conference.pdf.

Jiarui Cao, Zixuan Wei, and Yuxin Liu. Gradient flow drifting: Generative modeling via wasserstein gradient flows of kde-approximated divergences, 2026. URL https://arxiv.org/abs/2603.10592.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2022. doi: 10.1109/SP46214.2022.9833649.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, Anaheim, CA, August 2023. USENIX Association. ISBN 978-1-939133-37-3. URL https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

Paul Caucheteux, Clément Bonet, and Anna Korba. A unifying view of variational generative wasserstein flows. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=sJ7ngz2eQx.

Zhengdao Chen. On the interpolation effect of score smoothing in diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=O33LAUliUF.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting, 2026. URL https://arxiv.org/abs/2602.04770.

Gero Friesecke. *Optimal Transport: A Comprehensive Introduction to Modeling, Analysis, Simulation, Applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2024. doi: 10.1137/1.9781611978094. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611978094.

Weiguo Gao and Ming Li. How do flow matching models memorize and generalize in sample data subspaces. *npj Artificial Intelligence*, August 2026. ISSN 3005-1460. doi: 10.1038/s44387-026-00149-4. URL http://dx.doi.org/10.1038/s44387-026-00149-4.

Johanna Marie Gegenfurtner, Albert Kjøller Jacobsen, Naima Elosegui Borras, Alejandro Valverde Mahou, and Georgios Arvanitidis. Reducing memorisation in generative models via riemannian bayesian inference, 2026. URL https://arxiv.org/abs/2602.00199.

Jiaqi Han, Puheng Li, Qiushan Guo, Renyuan Xu, Stefano Ermon, and Emmanuel J. Candès. One-step generative modeling via wasserstein gradient flows, 2026. URL https://arxiv.org/abs/2605.11755.

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

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998. doi: 10.1137/S0036141096303359. URL https://doi.org/10.1137/S0036141096303359.

Chanhyuk Lee, Jaehoon Yoo, Manan Agarwal, Sheel Shah, Jerry Huang, Aditi Raghunathan, Seunghoon Hong, Nicholas M. Boffi, and Jinwoo Kim. Flow map language models: One-step language modeling via continuous denoising, 2026. URL https://arxiv.org/abs/2602.16813.

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

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

J.J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965. doi: 10.24033/bsmf.1625. URL https://www.numdam.org/articles/10.24033/bsmf.1625/.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 01 2014. ISSN 2167-3888. doi: 10.1561/2400000003. URL https://doi.org/10.1561/2400000003.

R. A. Poliquin and R. T. Rockafellar. Prox-regular functions in variational analysis. *Transactions of the American Mathematical Society*, 348(5):1805–1838, 1996. ISSN 00029947, 10886850. URL http://www.jstor.org/stable/2155081.

Peter Potaptchik, Jason Yim, Adhi Saravanan, Peter Holderrieth, Eric Vanden-Eijnden, and Michael S. Albergo. Discrete flow maps, 2026. URL https://arxiv.org/abs/2604.09784.

Mingxing Rao and Daniel Moyer. Generalization and memorization in rectified flow. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, pp. 537–554, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-37383-0.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

Thomas Sesmat, Gabriel Meseguer-Brocal, and Geoffroy Peeters. Where flow matching leaks: Characterising the membership signals along the interpolation path. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Ty5X41WbJw.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6048–6058, 2023. doi: 10.1109/CVPR52729.2023.00586.

Alexander Tong, Nikolay Malkin, Kilian Fatras, Lazar Atanackovic, Yanlei Zhang, Guillaume Huguet, Guy Wolf, and Yoshua Bengio. Simulation-free schrödinger bridges via score and flow matching. *arXiv preprint 2307.03672*, 2023.

Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW. Expert Certification.

Cédric Villani. *Optimal Transport*. Springer Berlin Heidelberg, 2009. ISBN 9783540710509. doi: 10.1007/978-3-540-71050-9. URL http://dx.doi.org/10.1007/978-3-540-71050-9.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1), January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01721-8. URL http://dx.doi.org/10.1038/s41597-022-01721-8.

TaeHo Yoon, Joo Young Choi, Sehyun Kwon, and Ernest K. Ryu. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023. URL https://openreview.net/forum?id=shciCbSk9h.

Appendix

This appendix is organized as follows. Appendix A proves the three propositions of Section 3 and records the alternative derivation we did not adopt. Appendix B collects the optimal transport background these proofs rely on. Appendix C gives the expressions for operating in a latent or projected space, together with the dimension normalization used throughout. Appendix D details the proximal solver, the roles of its two scale parameters, and the projection-based variant used in two dimensions. Appendices E, F and G report the full experimental details and additional results for the two-dimensional, image and text experiments respectively.

A Proofs

Throughout this appendix we write

$$c(x, y) = \frac{1}{2\lambda} \|x - y\|^2$$

for the cost of Section 3.3, and we recall that for a function ψ the c-transform is

$$\psi^c(x) = \inf_y \{c(x, y) - \psi(y)\}.$$

A function is said to be c-concave when it is the c-transform of some function.

A.1 Proof of Proposition 1

We restate the result in the form in which it is proved. Recall from Poliquin & Rockafellar (1996) that f is *strongly amenable* at $\bar{x}$ when it can be written $f = g \circ F$ on a neighborhood of $\bar{x}$, with F of class $\mathcal{C}^2$, g proper, lower semicontinuous and convex, and the constraint qualification

$$\partial^\infty g\big(F(\bar{x})\big) \cap \ker\left(\nabla F(\bar{x})^\top\right) = \{0\}$$

holding. Strong amenability implies prox-regularity.

Proof. Fix i and let $\mathcal{D} \setminus \{p_i\} = \{q_1, \dots, q_m\}$. For each $j \in \{1, \dots, m\}$ define

$$h_j(x) = \frac{\|x - p_i\|^2}{\|x - q_j\|^2},$$

so that on the cell $\mathcal{V}_i$, where the first nearest neighbor is fixed to p_i, the metric of Equation 6 reads

$$S_i(x) = \max_{1 \le j \le m} h_j(x).$$

Step 1: each h_j is smooth on $\mathcal{V}_i$. The only singularity of h_j is at $x = q_j$. We claim $q_j \notin \mathcal{V}_i$. Indeed, by definition of the Voronoi cell,

$$x \in \mathcal{V}_i \quad \Longrightarrow \quad \|x - p_i\| \le \|x - q_j\|,$$

so $x = q_j$ would force $\|q_j - p_i\| \le 0$, hence $p_i = q_j$, contradicting $q_j \in \mathcal{D} \setminus \{p_i\}$. Therefore $h_j \in \mathcal{C}^\infty(\mathcal{V}_i)$, and on any compact subset of $\mathcal{V}_i$ its denominator is bounded away from zero.

Step 2: the amenable decomposition. Set

$$F = (h_1, \dots, h_m) : \mathcal{V}_i \to \mathbb{R}^m, \qquad g = \max : \mathbb{R}^m \to \mathbb{R},$$

so that $S_i = g \circ F$. By Step 1, F is of class $\mathcal{C}^\infty$, and in particular $\mathcal{C}^2$. The function g is finite everywhere, convex and continuous.

Step 3: the constraint qualification. Since g is finite everywhere, it has no direction of recession in its domain, so its horizon subdifferential satisfies

$$\partial^\infty g(u) = \{0\} \qquad \text{for every } u \in \mathbb{R}^m.$$

The intersection with $\ker(\nabla F(\bar{x})^\top)$ is therefore $\{0\}$ for any $\bar{x}$, whatever the rank of ∇F, and the qualification holds trivially.

Hence S_i is strongly amenable at every point of $\mathcal{V}_i$, and therefore prox-regular there. □

Remark 1 (Consequences used in the main text). Prox-regularity of S_i at $\bar{x}$ with modulus r guarantees that for $\lambda < 1/r$ the proximal mapping $\mathrm{prox}_{\lambda S_i}$ is single-valued and Lipschitz on a neighborhood of $\bar{x}$, and that $\mathcal{M}_\lambda S_i$ is continuously differentiable there with the gradient of Equation 10.

A.2 Proof of Proposition 2

Proof. Step 1: c-concavity of the envelope. Directly from the definition in Equation 9,

$$\begin{aligned}\mathcal{M}_\lambda S(x) &= \inf_y \left\{ S(y) + \frac{1}{2\lambda}\|x-y\|^2 \right\} \\ &= \inf_y \left\{ c(x,y) - \big(-S\big)(y) \right\} \\ &= (-S)^c(x).\end{aligned}$$

So $\mathcal{M}_\lambda S$ is the c-transform of $-S$, and is therefore c-concave. We stress that this holds for *any* S, with no convexity or regularity requirement whatsoever. In particular the pair $\big(\mathcal{M}_\lambda S, (\mathcal{M}_\lambda S)^c\big)$ is admissible in the sense of Equation 4, since

$$\psi^c(x) + \psi(y) \le c(x,y)$$

holds by construction for any ψ and any pair (x,y).

Step 2: the c-transform on the range of the proximal. Let $x_0 \in \mathcal{X}$ and write $y^\star = \mathrm{prox}_{\lambda S}(x_0)$. We show $(\mathcal{M}_\lambda S)^c(y^\star) = -S(y^\star)$ by two inequalities.

For the first, note that for *every* x the point $y^\star$ is a candidate in the infimum defining $\mathcal{M}_\lambda S(x)$, so

$$\mathcal{M}_\lambda S(x) \le S(y^\star) + c(x,y^\star), \qquad \text{hence} \qquad c(x,y^\star) - \mathcal{M}_\lambda S(x) \ge -S(y^\star).$$

Taking the infimum over x on the left gives

$$(\mathcal{M}_\lambda S)^c(y^\star) \ge -S(y^\star).$$

For the reverse, evaluate the same infimum at the particular point $x = x_0$. Because $y^\star$ *attains* the infimum defining $\mathcal{M}_\lambda S(x_0)$, we have the equality $\mathcal{M}_\lambda S(x_0) = S(y^\star) + c(x_0,y^\star)$, and therefore

$$(\mathcal{M}_\lambda S)^c(y^\star) \;\le\; c(x_0,y^\star) - \mathcal{M}_\lambda S(x_0) \;=\; -S(y^\star).$$

The two inequalities give $(\mathcal{M}_\lambda S)^c = -S$ everywhere on the range of $\mathrm{prox}_{\lambda S}$.

Step 3: optimality and the splitting identity. Let

$$\nu = \mathrm{prox}_{\lambda S\#}\,\mu, \qquad \gamma = (\mathrm{id}, \mathrm{prox}_{\lambda S})_\#\mu \in \Pi(\mu,\nu).$$

Using Step 2 for the second integrand, and again the attainment of the infimum,

$$\begin{aligned}\int \mathcal{M}_\lambda S \,\mathrm{d}\mu + \int (\mathcal{M}_\lambda S)^c \,\mathrm{d}\nu &= \int \Big[\mathcal{M}_\lambda S(x) - S\big(\mathrm{prox}_{\lambda S}(x)\big)\Big]\,\mathrm{d}\mu(x) \\ &= \int c\big(x, \mathrm{prox}_{\lambda S}(x)\big)\,\mathrm{d}\mu(x) \\ &= \int c\,\mathrm{d}\gamma.\end{aligned}$$

By weak duality, the left-hand side, being the value of an admissible pair, is at most $\mathcal{W}_c(\mu,\nu)$; the right-hand side, being the cost of an admissible plan, is at least $\mathcal{W}_c(\mu,\nu)$. Since the two are equal, both coincide with $\mathcal{W}_c(\mu,\nu)$. Hence γ is an optimal plan and $\mathrm{prox}_{\lambda S}$ an optimal map, and rearranging the displayed equality gives

$$\mathcal{W}_c\big(\mu, \mathrm{prox}_{\lambda S\#}\,\mu\big) \;+\; \int S \circ \mathrm{prox}_{\lambda S}\;\mathrm{d}\mu \;=\; \int \mathcal{M}_\lambda S\,\mathrm{d}\mu,$$

which is Equation 12. When μ is absolutely continuous, uniqueness of the optimal plan follows from Theorem 1 through the twist condition of Definition 1, already verified for c in Section 3.3. □

Remark 2. The bound of Equation 13 follows at once, since $S \ge 0$ implies

$$\int S \circ \mathrm{prox}_{\lambda S}\;\mathrm{d}\mu \;\ge\; 0.$$

The gap between $\mathcal{E}_\lambda[\mu]$ and $\mathcal{W}_c(\mu, \mathrm{prox}_{\lambda S\#}\,\mu)$ is therefore exactly the expected memorization of the proximal distribution, and is small precisely when that distribution is strongly memorized, which is the regime the construction targets.

A.3 Proof of Proposition 3

Proof. Step 1: the envelope identity. We show

$$\int \mathcal{M}_\lambda S \,\mathrm{d}\mu \;=\; \inf_{\nu} \left\{ \int S \,\mathrm{d}\nu + \frac{1}{2\lambda} \mathcal{W}_2^2(\mu, \nu) \right\}$$

by proving the two inequalities separately.

Upper bound. Take the candidate $\nu = \mathrm{prox}_{\lambda S \#}\, \mu$ together with the coupling $\gamma = (\mathrm{id}, \mathrm{prox}_{\lambda S})_{\#}\mu$, which is admissible for (μ, ν) and therefore gives

$$\mathcal{W}_2^2(\mu, \nu) \;\le\; \int \|x - \mathrm{prox}_{\lambda S}(x)\|^2 \,\mathrm{d}\mu(x).$$

Substituting,

$$\begin{aligned} \int S \,\mathrm{d}\nu + \frac{1}{2\lambda} \mathcal{W}_2^2(\mu, \nu) &\le \int \left[S\big(\mathrm{prox}_{\lambda S}(x)\big) + \frac{1}{2\lambda} \|x - \mathrm{prox}_{\lambda S}(x)\|^2 \right] \mathrm{d}\mu(x) \\ &= \int \mathcal{M}_\lambda S \,\mathrm{d}\mu, \end{aligned}$$

the last equality being the definition of the envelope at each x, whose infimum is attained at $\mathrm{prox}_{\lambda S}(x)$.

Lower bound. Let ν be arbitrary and let $\pi \in \Pi(\mu, \nu)$ be any coupling. Bounding the integrand pointwise by the infimum that defines the envelope,

$$\begin{aligned} \int S \,\mathrm{d}\nu + \frac{1}{2\lambda} \int \|x - y\|^2 \,\mathrm{d}\pi &= \int \left[S(y) + \frac{1}{2\lambda} \|x - y\|^2 \right] \mathrm{d}\pi(x, y) \\ &\ge \int \mathcal{M}_\lambda S(x) \,\mathrm{d}\pi(x, y) \\ &= \int \mathcal{M}_\lambda S \,\mathrm{d}\mu, \end{aligned}$$

where the last equality uses only that the first marginal of π is μ. Taking the infimum over π, and then over ν, gives the reverse inequality.

The two bounds coincide, and the value is attained at $\nu^\star = \mathrm{prox}_{\lambda S \#}\, \mu$. It is the unique minimizer whenever $\mathrm{prox}_{\lambda S}$ is single-valued μ-almost everywhere, which Proposition 1 guarantees for λ below the prox-regularity threshold.

Step 2: the Wasserstein gradient. The functional

$$\mathcal{E}_\lambda[\mu] = \int \mathcal{M}_\lambda S \,\mathrm{d}\mu$$

is the integral of a *fixed* function against μ, hence linear in μ. Its first variation is therefore the integrand itself,

$$\frac{\delta \mathcal{E}_\lambda}{\delta \mu} = \mathcal{M}_\lambda S,$$

and by Equation 2 its Wasserstein gradient is the spatial gradient of that first variation,

$$\nabla_x \frac{\delta \mathcal{E}_\lambda}{\delta \mu}(x) = \nabla_x \mathcal{M}_\lambda S(x) = \frac{x - \mathrm{prox}_{\lambda S}(x)}{\lambda},$$

the last step being Equation 10. We emphasise that no approximation is involved: linearity in μ is what makes this identity exact. □

Remark 3 (Why this is a JKO step). Equation 14 is the variational problem defining one step of the JKO scheme (Jordan et al., 1998) for the functional $\mathcal{V}[\mu] = \int S \,\mathrm{d}\mu$, with λ in the role of the step size. The proximal mapping is thus simultaneously the optimal transport map of Proposition 2 and the map realizing a single implicit-Euler step of the Wasserstein gradient flow of expected memorization. Perturbing the velocity field by $\pm\gamma \nabla \mathcal{M}_\lambda S$ amounts to superposing, onto the learned dynamics, a flow that descends or ascends this functional.

A.4 THE DIRECT GRADIENT OF THE TRANSPORT COST

For completeness we record the cost of differentiating $\mathcal{W}_c(\mu, \operatorname{prox}_{\lambda S\#}\mu)$ itself rather than its upper bound, which justifies the choice made in Section 3.3.

Proposition 4. *With $P = \operatorname{prox}_{\lambda S}$ held fixed as a map, the functional $\mu \mapsto \mathcal{W}_c(\mu, P_{\#}\mu)$ is linear in μ, and at every point where $\mathcal{M}_\lambda S$ is twice differentiable its Wasserstein gradient is*

$$\big(\mathrm{id} - J_P^\top\big)\nabla_x \mathcal{M}_\lambda S(x) \;=\; \lambda\, \nabla_x^2 \mathcal{M}_\lambda S(x)\, \nabla_x \mathcal{M}_\lambda S(x) \;=\; \frac{\lambda}{2} \nabla_x \left\|\nabla_x \mathcal{M}_\lambda S(x)\right\|^2 . \tag{18}$$

Proof. By Proposition 2,

$$\mathcal{W}_c(\mu, P_{\#}\mu) = \int \big[\mathcal{M}_\lambda S - S \circ P\big]\, \mathrm{d}\mu,$$

which is linear in μ, so its first variation is the integrand and its Wasserstein gradient is the spatial gradient of that integrand,

$$\nabla_x \mathcal{M}_\lambda S(x) - J_P(x)^\top \nabla S\big(P(x)\big).$$

The optimality condition for the proximal point,

$$0 \in \nabla S\big(P(x)\big) + \frac{P(x) - x}{\lambda},$$

gives

$$\nabla S\big(P(x)\big) = \frac{x - P(x)}{\lambda} = \nabla_x \mathcal{M}_\lambda S(x),$$

so the expression factors as $\big(\mathrm{id} - J_P^\top\big)\nabla_x \mathcal{M}_\lambda S(x)$. Differentiating the identity $P(x) = x - \lambda \nabla_x \mathcal{M}_\lambda S(x)$ yields

$$J_P = \mathrm{id} - \lambda \nabla_x^2 \mathcal{M}_\lambda S, \qquad \text{whence} \qquad \mathrm{id} - J_P^\top = \lambda \big(\nabla_x^2 \mathcal{M}_\lambda S\big)^\top.$$

Since $\nabla \mathcal{M}_\lambda S$ is locally Lipschitz, $\nabla^2 \mathcal{M}_\lambda S$ exists almost everywhere by Rademacher's theorem and is symmetric almost everywhere, giving the stated form. □

Evaluating this field requires a Hessian-vector product of the envelope at every integration step, on top of the proximal solve already needed to obtain $\nabla \mathcal{M}_\lambda S$. The upper bound of Equation 13 avoids it entirely. We note in addition that for the small values of λ used in practice the two fields differ by $\mathcal{O}(\lambda)$, so the bound is also a close approximation of the direct gradient in the operating regime.

B OPTIMAL TRANSPORT BACKGROUND

We collect here the standard material condensed in Section 2.2. With the notation of that section, the Kantorovich problem of Equation 3 seeks a coupling of μ and ν minimizing the expected cost. Through Fenchel-Rockafellar duality it admits the dual formulation

$$\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, \mathrm{d}\gamma(x,y) = \sup_{(\varphi,\psi)\in\Phi(c)} \left(\int_{\mathcal{X}} \varphi(x)\, \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi(y)\, \mathrm{d}\nu(y) \right) \tag{19}$$

over the admissible pairs of Equation 4. The inequality $\varphi(x) + \psi(y) \le c(x,y)$ defining admissibility gives weak duality directly, and it is the direction used in the proof of Proposition 2: exhibiting an admissible pair and a coupling whose values coincide certifies optimality of both at once, which is how the splitting identity of Equation 12 is obtained without solving either problem.

In the classical case $c(x,y) = \frac{1}{2}\|x-y\|^2$, Brenier's theorem states that for μ absolutely continuous the optimal plan is unique, of Monge form $\gamma = (\mathrm{id}, T)_{\#}\mu$, and given by $T(x) = x - \nabla\varphi(x)$ with φ the optimal first potential, which is moreover locally Lipschitz. Theorem 1 is the extension of this statement to costs satisfying the twist condition of Definition 1, obtained by inverting $y \mapsto \nabla_x c(x,y)$ in place of the identity map; the assumptions on the measures are otherwise unchanged. Our cost $c(x,y) = \frac{1}{2\lambda}\|x-y\|^2$, or $\frac{1}{2\lambda D}\|x-y\|^2$ under the normalization of Appendix C, is a positive multiple of the squared Euclidean distance and therefore satisfies the condition with

$$\big(\nabla_x c(x,\cdot)\big)^{-1}(z) = x - \lambda z,$$

which is the inversion used in Equation 5.

C CHOICE OF SPACE AND DIMENSION NORMALIZATION

The construction is agnostic to the space in which the anchors live: nothing above requires $\mathcal{D}$ to be a dataset of images, only a finite point cloud in a Euclidean space. We may therefore operate directly in pixel space, or in the latent space of a pretrained model, whichever carries the semantics we wish to alter. For latent OT-CFM we take the latter, so that the perturbation produces semantic rather than pixel-level changes and the method applies to existing pretrained generative models without retraining.

Let $\{v_i\}_{i=1}^n \in \mathcal{X}$ be a dataset in pixel space and $\mathcal{Z} := \{z_i = f(v_i)\}_{i=1}^n$ its image under an encoder f. Latent memorization is

$$S(z) = \frac{\left\|z - \mathrm{nn}_1^{\mathcal{Z}}(z)\right\|^2}{\left\|z - \mathrm{nn}_2^{\mathcal{Z}}(z)\right\|^2},$$

with the same Voronoi partitioning as before. We may further restrict the perturbation to the leading principal directions of $\mathcal{Z}$, projecting a generated point by $x \mapsto z = V^T(x - \mu)$. The metric adjusts as

$$S(z) = \frac{\left\|V^T\big(x - \mathrm{nn}_1^{\mathcal{P}}(x)\big)\right\|^2}{\left\|V^T\big(x - \mathrm{nn}_2^{\mathcal{P}}(x)\big)\right\|^2},$$

and the chain rule lifts the envelope gradient back to the ambient space:

$$\nabla_x \mathcal{M}_\lambda S(z) = \left(\frac{\partial z}{\partial x}\right)^T \nabla_z \mathcal{M}_\lambda S(z) = V\left(\frac{z - \mathrm{prox}_{\lambda S}(z)}{\lambda}\right) \tag{20}$$

where $z = V^T(x - \mu)$.

DIMENSION NORMALIZATION

In dimension $D > 1$ we use the normalized squared distance $d(x, y) = \frac{1}{D}\|x - y\|^2$ throughout, so that the cost becomes $c(x, y) = \frac{1}{2\lambda D}\|x - y\|^2$. This is a constant rescaling and leaves every statement of Section 3.3 unchanged, amounting to the reparameterization $\lambda \mapsto \lambda D$ in the unnormalized problem. Its practical value is that it makes λ comparable across spaces of different dimension: a value chosen in a two-dimensional experiment transfers to a 200-dimensional latent subspace without rescaling, which matters here since the same construction is applied in spaces ranging over three orders of magnitude in dimension.

D ALGORITHMIC DETAILS

Derivation of the fixed point. Finding $\mathrm{prox}_{\lambda S}(y)$ is the composite minimization $\min_x\{S(x) + g(x)\}$ with $g(x) = \frac{1}{2\lambda}\|x - y\|^2$. A minimizer $x^\star$ satisfies

$$0 \in \bar{\lambda}\nabla S(x^\star) + \bar{\lambda}\,\partial g(x^\star)$$

for any $\bar{\lambda} > 0$, where ∇S denotes the limiting subgradient, which reduces to the classical gradient on the neighborhood where S_i is prox-regular and the active index in the maximum of Appendix A.1 is unique. Rearranging into a forward-backward form gives

$$x^\star = \big(\mathrm{id} + \bar{\lambda}\,\partial g\big)^{-1}\big(\mathrm{id} - \bar{\lambda}\nabla S\big)(x^\star) = \mathrm{prox}_{\bar{\lambda} g}\big(\mathrm{id} - \bar{\lambda}\nabla S\big)(x^\star).$$

Since g is a scaled squared distance, its proximal is available in closed form,

$$\mathrm{prox}_{\bar{\lambda} g}(z) = \arg\min_x \left\{\frac{1}{2\lambda}\|x - y\|^2 + \frac{1}{2\bar{\lambda}}\|x - z\|^2\right\} = \frac{\bar{\lambda} y + \lambda z}{\lambda + \bar{\lambda}}, \tag{21}$$

which, substituted into the fixed-point equation, gives Equation 17. Algorithm 1 iterates it with Nesterov extrapolation $\omega^k = k/(k + 3)$.

On the two scales. The parameters λ and $\bar{\lambda}$ play different roles and must not be identified, although an early version of our implementation did so. The envelope parameter λ is the tether: it defines the

radius within which the proximal point may be sought, and therefore the locality of the perturbation, and it appears in the denominator of the escape field. The inner step $\bar{\lambda}$ is a numerical parameter of the solver, constrained by the local Lipschitz modulus L of ∇S; the accelerated scheme converges at rate $\mathcal{O}(1/k^2)$ for $\bar{\lambda} \in (0, 1/L]$ in the convex case (Parikh & Boyd, 2014), with prox-regularity providing the local substitute here. Tying them means that any schedule imposed on λ silently rescales the solver's step, and in particular a large λ at small t forces large and poorly conditioned inner steps. In all reported experiments the two are separate quantities, with $\bar{\lambda}$ a small constant independent of t.

Dimension normalization. All experiments use the normalized distance $d(x, y) = \frac{1}{D} \|x - y\|^2$ of Appendix C. Since this amounts to $\lambda \mapsto \lambda D$ in the unnormalized problem, the values of λ_0 reported below are stated in normalized units unless the ambient form is given explicitly; for instance the RetinaMNIST setting uses $\lambda_0 = 0.05$ normalized with $D = 200$. Reported values of λ_0 are therefore directly comparable across the two-dimensional, latent-image and text experiments despite their dimensions spanning three orders of magnitude.

Restricting the proximal search. The infimum defining the envelope may be restricted to any nonempty closed subset $\mathcal{C} \subseteq \mathcal{V}_i$ without invalidating the construction, since the first nearest neighbor remains fixed throughout the cell. Writing $\iota_{\mathcal{C}}$ for the indicator of $\mathcal{C}$, the restricted envelope

$$\mathcal{M}_\lambda^{\mathcal{C}} S(x) = \inf_{y \in \mathcal{C}} \left\{ S(y) + \frac{1}{2\lambda} \|y - x\|^2 \right\} \tag{22}$$

is the same composite problem as before with $\iota_{\mathcal{C}}$ added to the non-smooth part. Since the two quadratics combine into a single one centred at the averaged point, the backward step of Appendix D acquires a projection and nothing else:

$$\operatorname{prox}_{\bar{\lambda} g}(z) = \arg\min_y \left\{ \frac{1}{2\lambda} \|y - x\|^2 + \frac{1}{2\bar{\lambda}} \|y - z\|^2 + \iota_{\mathcal{C}}(y) \right\} = \Pi_{\mathcal{C}} \left(\frac{\bar{\lambda} x + \lambda z}{\lambda + \bar{\lambda}} \right), \tag{23}$$

so Algorithm 1 carries over verbatim with $\Pi_{\mathcal{C}}$ applied to each iterate. This is the standard reduction of a constrained proximal step to a projected one (Parikh & Boyd, 2014).

A natural choice is the segment $\mathcal{C} = [x, \mathrm{nn}_1^{\mathcal{P}}(x)]$, which lies in $\mathcal{V}_i$ by convexity of the cell, and on which the projection is a clamp of the scalar coordinate $\alpha \in [0, 1]$ in $y(\alpha) = x + \alpha\big(\mathrm{nn}_1^{\mathcal{P}}(x) - x\big)$. The solve becomes one-dimensional, and the resulting field is

$$\nabla \mathcal{M}_\lambda^{\mathcal{C}} S(x) = \frac{x - y(\alpha^\star)}{\lambda} = \frac{\alpha^\star}{\lambda} \big(x - \mathrm{nn}_1^{\mathcal{P}}(x)\big). \tag{24}$$

This last identity is worth reading carefully, because it places the nearest-neighbor repulsion baseline of Table 1 inside our own framework rather than outside it. The line-restricted field points along exactly the direction that baseline uses, $x - \mathrm{nn}_1^{\mathcal{P}}(x)$; the two differ only in magnitude, which the baseline fixes by normalization while the envelope sets it through $\alpha^\star/\lambda$, the optimal position along the chord divided by the tether. The unconstrained envelope is thus the more expressive object, the line-restricted envelope recovers the baseline's direction with an adaptively chosen step, and the baseline itself is the further degenerate case in which that step is chosen externally. This is consistent with what the ablation reports: norm-matched nearest-neighbor repulsion reduces memorization but at a markedly worse fidelity cost than the unconstrained field.

We use the line-restricted variant in the two-dimensional experiment of Appendix E, where the one-dimensional solve is cheap and the restriction keeps the displacement visibly interpretable; the image and text experiments use the unconstrained solve throughout.

E TWO-DIMENSIONAL FLOW MATCHING

This experiment exhibits both regimes of the dial on a single model and in a setting where the geometry of Section 3 can be inspected directly.

Setup. We train an OT-CFM model (Tong et al., 2024) transporting an eight-mode Gaussian to the two-moons distribution, using an MLP velocity field with time conditioning. Training is deliberately configured to induce memorization: $N = 256$ training points, 30 000 iterations, a small conditional-path noise $\sigma = 0.02$, and a single static optimal transport plan computed once over the full dataset

rather than recoupled per minibatch. Removing the per-minibatch recoupling is the decisive choice, as it eliminates the regularization against fitting one specific assignment; the resulting model classifies 38.1% of its generated samples as memorized at $\tau = 1/3$. We generate 2000 samples for evaluation. The perturbation uses $\lambda_0 = 0.01$ with the schedule $\lambda(t) = \lambda_0/(t + \varepsilon)$ clamped at 50, a fixed inner step $\lambda = 0.05$, and five inner iterations.

Protocol. Rather than taking the whole training set as anchors, we select a datapoint and its 32 nearest neighbors among the training data, defining a localized zone of interest. The perturbation is computed against this subset alone, so that the rest of the model is left untouched. Applying the field with $\gamma = 0.6$ repels samples from the zone, and with $\gamma = -0.6$ attracts them into it.

Results. Figure 6 shows both regimes. Under repulsion, the fraction of in-zone samples classified as memorized drops from the baseline to 1%, while under attraction it rises to 69% (VS 4% and 80% in the line constrained case from 22), with samples generated very close to the protected training points. Samples away from the zone are essentially undisplaced in both cases, which is the expected consequence of the field being defined by a local proximal search: at a point far from every anchor, $S \approx 1$ and the proximal point does not move. This gives the construction two complementary readings, as a way to shield a sensitive region of the data distribution from reproduction, and as a way to impose conditions on generation by forcing samples into a chosen region, in both cases without retraining.

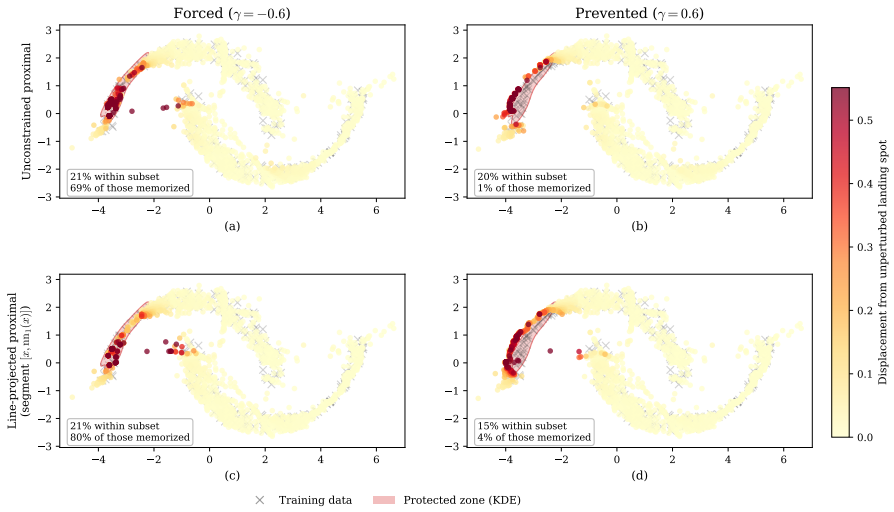


Figure 6: Attraction and repulsion from a localized zone of interest in two dimensions. Colour encodes displacement from the unperturbed sample at the same initial noise.

F Image domain: additional details and results

F.1 Training and sampling

The latent flow model is trained with minibatch OT-CFM on the 1000 RetinaMNIST training images at 64×64, encoded with the Stable Diffusion variational autoencoder into $4 \times 8 \times 8 = 256$-dimensional latents. The perturbation is computed in the subspace of the first 200 principal directions of the latent training set and lifted back by the chain rule of Equation 20. Sampling uses a fixed-step Euler integrator with 100 steps unless otherwise stated; the function-evaluation counts of Table 1 are obtained by running the same fields through an adaptive `dopri5` solver with tolerances 10^{-4}. All figures in Section 4.1 use 256 samples drawn from one fixed set of initial noises, so that curves across γ describe the same trajectories perturbed by different amounts. The envelope uses $\lambda_0 = 0.05$, the schedule $\lambda(t) = \lambda_0/t$ clamped at 50, a fixed inner step of 0.05, and 30 inner iterations.

F.2 Calibration of the perceptual metrics

Perceptual measurements in this setting require care, because generated images are decoder outputs whereas training images are not, and this asymmetry is large enough to dominate naive similarity scores. We quantify it directly. Encoding and decoding each training image through the variational autoencoder and measuring the DINOv2 cosine similarity between an image and its own round trip gives a mean of 0.661. The mean cosine similarity between two *unrelated* decoded training images is 0.725. Two different retinas, once decoded, therefore resemble each other more than a real retina resembles its own reconstruction: at this resolution DINOv2 cosine similarity encodes the decoder's signature at least as strongly as it encodes identity.

Two consequences follow, and we apply both. First, top-1 similarity of a generated sample against *raw* training images is uninformative here: it ranges over 0.737 to 0.710 across the sweep, entirely at or below the level of unrelated pairs, and a verbatim latent copy cannot score highly against a raw image it was never able to reproduce exactly. Measured instead against reconstructions, where both sides have passed through the decoder, the same statistic runs from 0.954 at $\gamma = 0$ to 0.886 at $\gamma = 0.5$, well above the chance level and behaving as a copy-detection score should. Second, all comparisons in the interpolation study of Appendix F.3 are made against reconstructed training images, since the reference interpolant is itself decoded and a decoded-against-raw comparison would not be meaningful.

We emphasize that this concerns the perceptual diagnostics only. The memorization rates reported in the main text are computed with the metric S of Equation 6 against the *raw* training data in pixel space, which is the conservative choice: the quantity of interest is whether the model reproduces a real training image, and scoring against reconstructions would grade the method against an easier target.

F.3 Interpolation measurements

Figure 7 reports the full interpolation geometry summarized in Section 4.1. In addition to the chord coordinate α, we report the off-chord residual, the distance from the sample to the line through its two nearest training latents, normalized by the distance between them. This second quantity guards against a degenerate reading of the metric S: a sample equidistant from both neighbors achieves $S \approx 1$ whether it lies between them or has drifted sideways into empty space, and only a small residual distinguishes the former.

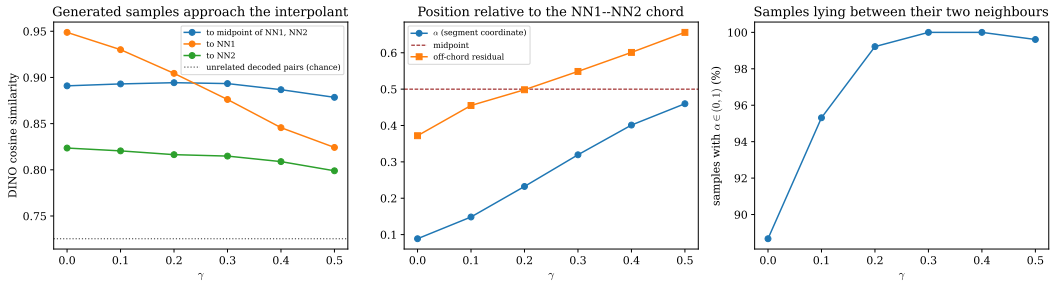


Figure 7: Interpolation geometry in latent space across γ. **Left:** perceptual similarity of a generated sample to its nearest training neighbor, to its second neighbor, and to the decoded midpoint of the two, measured in DINOv2 space against reconstructed training images. **Center:** the chord coordinate α, where 0 means coincident with the first neighbor and 0.5 the midpoint, together with the off-chord residual. **Right:** the fraction of samples lying strictly between their two neighbors.

The chord coordinate rises from 0.089 at $\gamma = 0$ to 0.460 at $\gamma = 0.5$, and the fraction of samples lying strictly between their two neighbors from 88.7% to essentially all. The residual also grows, by a factor of 1.76 over the sweep, but the chord coordinate grows by a factor of 5.2, so the ratio of off-chord to along-chord displacement falls from 4.2 to 1.4: the motion induced by the perturbation is predominantly along the chord. The residual does not vanish, and this is expected rather than a defect. A sample with zero residual would be exactly the latent midpoint of two training images, which is the naive interpolant; the perturbed samples approach that midpoint perceptually while

retaining an off-chord component, which is what keeps them on the manifold the velocity field has learned. We note for completeness that the naive midpoint, decoded, attains a slightly better held-out Fréchet distance than our samples at $\gamma = 0.5$ (0.576 against 0.654), which is unsurprising since it is a convex combination of real data encodings; it is however not a generative model, requiring the training set at inference and producing only pairwise midpoints.

F.4 Additional diagnostics

Figure 8 reports the measurements deferred from the main text: memorization in the PCA subspace where the perturbation acts, which tracks the full latent space closely and confirms that the projection is not distorting the effect; the copy-detection score against reconstructions discussed above; and the mean pairwise similarity within the generated set.

The last of these rises slightly with γ, from 0.7405 to 0.7626. This is a consequence of the interpolation behaviour rather than a collapse of diversity: averaging two samples from a distribution produces a point closer to the mean than either endpoint, so a set of near-midpoints is mechanically more concentrated than a set of endpoints. The alternative explanation, a collapse onto fewer modes, is ruled out by the anchor counts, which move in the opposite direction, the number of distinct training images serving as nearest neighbor rising from 84 to 96 over the sweep. The magnitude of the change is also small relative to the scale of the space, in which unrelated pairs sit at 0.725 and near-duplicates above 0.95.

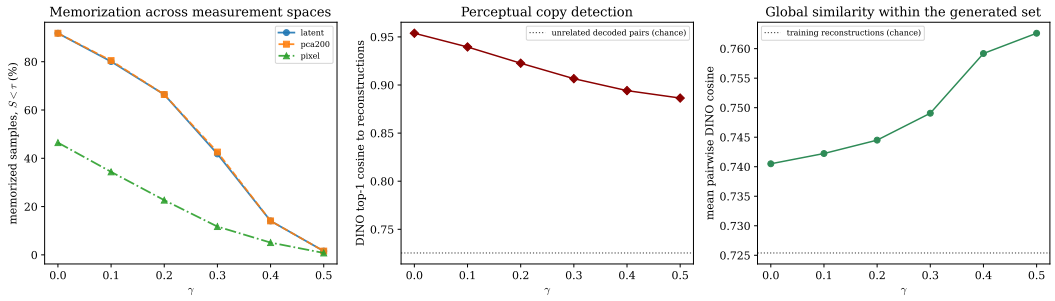


Figure 8: Additional measurements on RetinaMNIST. **Left:** memorization across the three measurement spaces. **Center:** perceptual copy detection against reconstructed training images, with the chance level of unrelated decoded pairs. **Right:** mean pairwise similarity within the generated set.

G Embedded Language Flows: additional details

G.1 Model, checkpoint and generation regime

We use the public ELF-B checkpoint `embedded-language-flows/ELF-B-owt-torch`, trained on OpenWebText, taken at training step $95\,085$ (epoch 5) and evaluated through its exponential moving average parameters. The tokenizer and text encoder are `t5-small`. The flow operates in the 512-dimensional raw output space of that encoder, which is the space the rectified-flow interpolant is trained in, and not the narrower internal bottleneck of the transformer's forward pass. Since unconditional generation never queries the encoder, the anisotropy of the encoder's embedding space is irrelevant to the mechanism; the interpolant being trained under plain mean squared error, the geometry is the Euclidean one our construction assumes. The encoder is moved off the accelerator for all runs reported here, which are unconditional.

Generation is unconditional throughout, with an Euler ODE sampler and 32 steps. All figures derive from a single invocation of the experiment driver in its `paper_figures` mode.

G.2 Anchor harvesting

The anchors define the set against which S is computed and are harvested from the model's own generations rather than produced by encoding words. An earlier version built anchors by passing

candidate words through the T5 encoder, optionally inside template sentences. A diagnostic rejected that approach: for each word, the distance between its standalone embedding and its mean in-context embedding was comparable to the distance between different words (ratio 0.95 to 1.03), and its spread across contexts was larger still (ratio 1.14 to 1.19). Context dominated word identity, and the resulting anchors moreover lived in a space the generative process never visits at inference.

The harvesting procedure instead runs unperturbed unconditional generation in batches of 16, up to five batches, decodes each sample, and locates occurrences of the target words by matching on the decoded string over token spans of at most three subtokens. Matching on decoded strings rather than raw token identifiers avoids tokenizer artifacts such as leading-space variants being distinct identifiers and words spanning several subtokens. At each match we take the model's own clean-embedding prediction at that position, mean-pooled across the matched subtokens, and average up to five occurrences into a single anchor vector. The harvested set is cached and invalidated automatically if the word list changes.

The reported runs use nine anchors, giving an anchor matrix of shape 9×512: *war, conflict, death, life, threat, killing, murder, refuse, decline*. These were chosen partly for harvestability, being frequent enough in unconditional generation for occurrences to be collected, and partly because S acts with respect to the nearest anchor, so a semantically coherent cluster makes the nearest-anchor assignment meaningful.

G.3 Perturbation and run parameters

The escape field is added to the predicted velocity inside the ODE step, and the existing Euler update then integrates the perturbed field. Three points are worth making explicit. First, we perturb the velocity rather than the model's x-prediction: the latter parameterization is a training device, whereas the derivation of Section 3 describes the evolution of the law of the particles themselves, so the field must act on the velocity. Recovering a velocity from a perturbed x-prediction would introduce an additional implicit time dependence on top of the explicit schedule and would no longer be the flow the theory describes. Second, the gradient is evaluated at the current latent, not at the clean prediction; the concern that nearest-anchor assignment is ill-posed when the latent is mostly noise is handled by the schedule, which makes the field vanish at small t. Third, the perturbation is masked at conditioning positions, so a source prefix is never perturbed.

Two steering modes are implemented. In `per_token`, used for all reported runs, each of the L positions is treated as its own point with its own proximal solve, which forces word choice. In `centroid`, a single gradient is computed from the sequence-mean embedding and broadcast to all positions, a pure translation of the point cloud that preserves pairwise token relations and is better suited to topic drift than to lexical substitution.

Table 2 lists the parameters of the reported run. The steered condition is $\gamma = -3$ throughout, the baseline $\gamma = 0$.

Table 2: Parameters of the reported ELF run. The envelope and solver parameters follow the conventions of Appendix D; λ_0 is stated in normalized units.

Sampler	Euler ODE, 32 steps
Latent dimension	512 (T5-small encoder output)
Steering mode	`per_token`
γ (steered condition)	-3
λ_0	0.25, schedule $\lambda(t) = \lambda_0/(t+\varepsilon)$, clamped at 50
Inner step $\bar{\lambda}$	0.05, constant in t
Inner iterations	30
Proximal search	unconstrained (segment restriction disabled)
Paired samples (usage, excerpts)	30 per condition
Samples for perplexity and entropy	30 per condition
Background positions (cell occupancy)	5 per baseline sequence, 150 total
Qualitative grid	one fixed initial noise, 336 tokens shown

The comparison is run on paired samples, which is essential and not merely convenient. The initial noise is drawn once per batch, the CPU and accelerator generator states are snapshotted, the unperturbed pass is run, the states are restored, and the steered pass is run from a clone of the same initial noise. An earlier version generating the two conditions as independent batches made every per-position comparison meaningless, and produced the misleading artifact of baseline excerpts containing no anchor words at all, purely because they were unrelated draws.

G.4 Per-word usage and cell occupancy

Table 3 gives the per-sample usage rates underlying Figure 5b, together with the cell occupancy of Figure 5a. We report rates rather than totals so that a frequent word's large absolute increase does not visually dominate a rare word's proportionally larger one, and we quote the standardized difference z, the paired difference in rate divided by the combined standard error, so that each word carries its own uncertainty.

Table 3: Anchor-word usage over 30 paired sequences per condition, and cell occupancy measured over 150 background positions from the unperturbed pass. Rates are occurrences per sample; z is the difference in rate divided by the combined standard error.

Word	occupancy (%)	$\gamma = 0$	$\gamma = -3$	fold	z
life	63.3	0.633	3.167	5.0×	5.2
decline	5.3	0.067	2.500	37.3×	6.9
conflict	5.3	0.033	1.933	58.6×	5.3
death	10.0	0.067	0.767	11.4×	3.1
war	6.0	0.200	0.733	3.7×	2.8
killing	8.0	0.067	0.667	10.0×	2.9
refuse	2.0	0.100	0.667	6.7×	2.9
threat	0.0	0.033	0.400	12.1×	3.1
murder	0.0	0.000	0.000	—	—

Three observations follow. First, every word that appears at all responds, and does so by at least 2.8 standard errors; the increases are not carried by one runaway sample, as the per-sample standard deviations behind these standard errors are comparable to the means. Second, occupancy predicts the magnitude of the effect across words but does not determine it: the rank correlation between occupancy and steered rate is $\rho = 0.69$ ($p = 0.04$), and the clearest departure is that *life*, which dominates the anchor geometry, attains the largest absolute rate but the smallest fold change, while *conflict* and *decline* occupy small cells and rise by factors of 59 and 37. Third, *threat* is never the nearest anchor for any of the 150 background positions sampled and yet responds clearly, so a cell too small to measure at this sample size is not a cell the field cannot reach; a zero in the occupancy column bounds the rate below the resolution of 150 draws rather than establishing that the cell is empty.

The remaining word, *murder*, is never produced in either condition. We do not read this as a failure to steer, since there is no occurrence to redirect, but neither can we distinguish it from one: with zero counts on both sides the experiment is uninformative about that anchor. Both its occupancy and its usage are consistent with the region of latent space decoding to that word being visited too rarely to observe at $n = 30$.

G.5 Fluency and diversity

Table 4 reports generative perplexity under GPT-2 and unigram entropy over 30 sequences per condition, generated in the same run and scored in a separate pass. Both quantities are read relatively rather than absolutely: ELF-B's unconditional generation is already partly incoherent, so the question is whether the perturbation degrades it, not whether the output is fluent in absolute terms.

Entropy is unchanged to three significant figures, which rules out the distributional collapse that strong classifier-free guidance produces in this model class and which motivates Hu et al. (2026) to report entropy alongside perplexity. The perplexity point estimate rises by roughly 8%, but the per-sample standard deviation is of order 45 in both conditions, so at $n = 30$ the difference lies well

Table 4: Fluency and diversity of unperturbed against steered generation, $n = 30$ sequences per condition with no empty generations. Perplexity is measured under GPT-2; the standard deviation is over per-sample values.

γ	Gen-PPL	median	s.d.	entropy
0	126.9	123.0	43.5	5.573
-3	137.2	139.7	50.5	5.580

inside the sampling spread and we do not read it as a measured degradation. We state this as an absence of detectable degradation rather than as an absence of degradation.

We note in this respect that a looped or spliced span under steering is not by itself evidence of damage. ELF-B produces subword splices and repeated spans in unperturbed generation, and the perturbation is a deterministic function of the latent, so identical tokens are displaced identically and any loop present in the unsteered generation is expected to persist. Figure 9 shows the two conditions exhibiting these artifacts at comparable rates.

G.6 Qualitative excerpts

Figure 9 reproduces paired generations at $\gamma = 0$ and $\gamma = -3$ from one fixed initial noise, with per-token background shading encoding displacement in the continuous space relative to the baseline and a distinct highlight where the decoded token is an anchor word. We reproduce long passages rather than isolated positions because the effect is not confined to the substituted tokens: the two passages share a syntactic skeleton but differ in subject matter and vocabulary throughout, while only the positions whose latent lies nearest an anchor collapse onto that anchor.

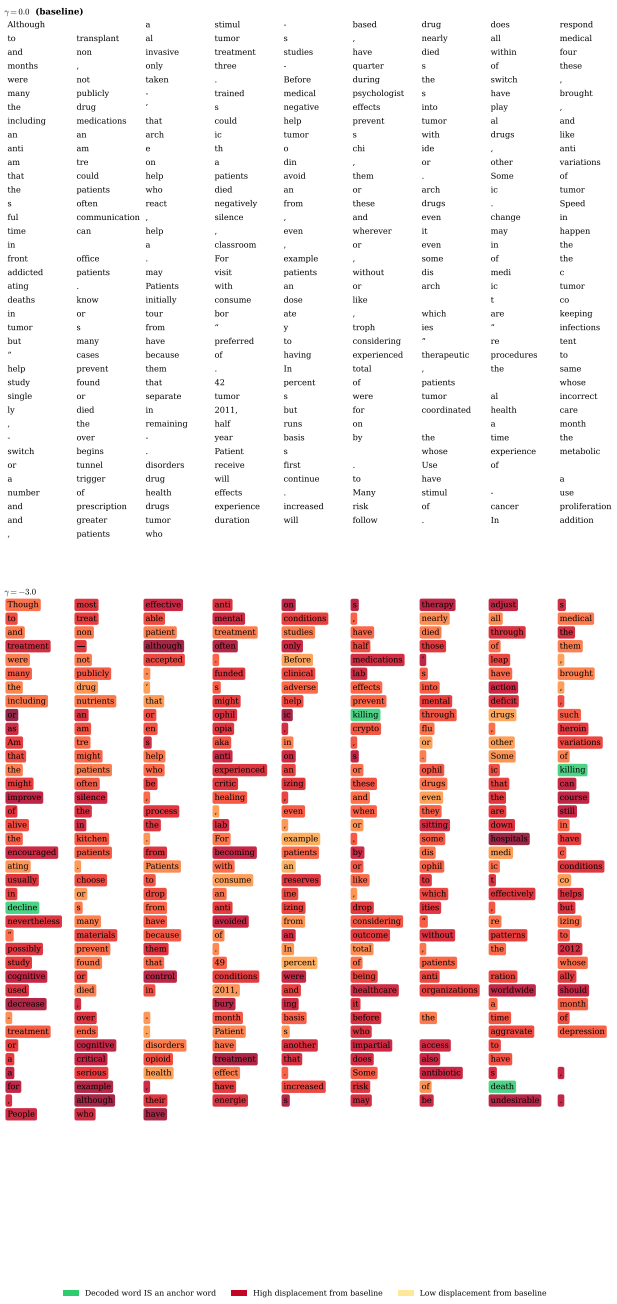


Figure 9: Paired generations from one fixed initial noise, unperturbed above and steered below. Background shading encodes per-token displacement from the unperturbed generation; highlighted tokens decode to anchor words.