# BEYOND EXPLICIT MIRROR MAPS: CONSTRAINED GENERATIVE MODELING VIA BRENIER MAPS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Mirror-lifting paradigms for constrained generation are concise and efficient, but they suffer from fundamental limitations. Analytical mirror functions are only valid for specific convex constraint domains, lacking a unified construction method and failing to handle non-convex constraints. Neural approximate mirror functions broaden applicable scenarios yet brings uncontrollable geometric distortion and heavy training overhead. To address these issues, we propose a unified framework for constructing constrained generative models based on Brenier maps. Theoretically, we prove that mirror functions on convex domains constitute a special case of dual Brenier functions. Leveraging singular set theory, we break the convexity constraint of mirror maps and enable unified adaptation to arbitrary convex and non-convex domains. Algorithmically, we develop a universal multi-stage solving strategy for Brenier maps. Anchor sampling and geometric variational principles are adopted to capture precise geometric structures, and a controllable smoothing mechanism is introduced to preserve critical geometric properties and improve generalization. Experimental results show that on three classical convex constraint domains, our method achieves better fitting accuracy for mirror functions than neural approximation baselines. For the watermark image generation task and five other complex real-world constrained generation tasks involving both convex and non-convex scenarios, our method surpasses existing SOTA baselines in generation quality and constraint satisfaction with lower computational cost, substantially improving the accuracy and efficiency of constrained generation.

## 1 INTRODUCTION

Practical tasks including physical simulation, astronomical imaging, and material design demand generative models that strictly comply with geometric, partial differential equations (PDEs), and counting priors (Feng et al., 2025). Generative outputs inherently satisfying constraints drastically cut post-correction overhead and improve the reliability of inverse problem solving.

The mirror-lifting paradigm has emerged as a promising framework for constrained generative modeling (Tae, 2023; Liu et al., 2024; Ha et al., 2025; Guan et al., 2026; Li et al., 2026): a mirror map $\nabla\phi$ transforms data confined within a constraint set $\mathcal{C} \subset \mathbb{R}^d$ to an unconstrained Euclidean space $\mathbb{R}^d$ for standard generative training, then maps samples back to the original domain via its inverse operation $(\nabla\phi)^{-1}$ to guarantee constraint compliance by construction. However, analytical mirror maps admit closed-form expressions only for a small set of convex constraints (Bubeck, 2015; Tan et al., 2023; Tae, 2023; Liu et al., 2024). For general convex constraints, as well as non-convex, piecewise and discrete constraints, there exists no universal construction strategy, and the potential function $\phi$ has to be manually derived case-by-case, which greatly raises the practical barrier of deployment.

To bypass this barrier, Feng et al. (2025) proposed Neural Approximate Mirror Maps (NAMMs), which employ Input Convex Neural Networks (ICNNs) to fit bidirectional mirror maps without handcrafted analytical potentials and achieve state-of-the-art (SOTA) performance across multiple non-convex generative tasks. Still, this purely neural framework bears two critical limitations: (1) it requires joint optimization over two networks with a multi-objective loss composed of cyclic consistency, constraint, and regularization terms, incurring substantial computational overhead; (2) lacking rigorous geometric foundations, learned maps easily induce local distortions and distribution support shifts, leading to drastically degraded constraint satisfaction on highly non-convex constraints. Al-



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.

Table 1: Comparison of Our Proposed Method with Mirror Maps and Their Neural Approximations

| Methods | Admissible Constraints | Construction | Geometric Fidelity |
|---|---|---|---|
| Analytical Mirror Maps (Tae, 2023; Liu et al., 2024) (Ha et al., 2025; Guan et al., 2026) | Convex Constraint | Manual Derivation Required, with No Universal Automated Construction Pipeline | Distortion-Free on Convex Constraints Yet Fails Outright on Non-Convex Constraints |
| Neural Approximate Mirror Maps (Feng et al., 2025) | Arbitrary Constraints | Training ICNNs with Cycle-Consistency Loss at Substantial Optimization Cost | Satisfactory for Convex Constraints; Prone to Local Geometric Distortion Under Non-Convex Constraints. |
| Brenier Maps (Ours) | Arbitrary Constraints | General Pipeline with Controllable Geometric Distortion | Negligible Geometric Distortion for Arbitrary Constraints |

ternative methods such as equivariant generative models (Song et al., 2023; Yim et al., 2023; Midgley et al., 2023; Battiloro et al., 2025), manifold diffusion (De Bortoli et al., 2022; Huang et al., 2022; Fu et al., 2024), Metropolis-Hastings sampling (Fishman et al., 2023b), and soft constraint penalties (Giannone et al., 2023b; Daras et al., 2023; Upadhyay et al., 2023) each suffer narrow applicability, rigid manifold assumptions, high rejection rates, or expensive iterative simulation costs—none simultaneously deliver universal applicability, computational efficiency, and geometric fidelity.

Rooted in optimal transport theory, this work establishes a critical theoretical insight: all valid mirror maps defined on convex constraints constitute special cases of Brenier maps under quadratic transport costs. Equipped with singular set theory for optimal transport (Figalli, 2010; Luo et al., 2022), Brenier maps impose no convexity restrictions and admit numerically computable bidirectional transforms over arbitrary feasible domains. Drawing on this finding, we propose a unified constrained generative framework that fully replaces both analytical mirror maps and their ICNN-based neural approximations with Brenier maps, fundamentally resolving the core limitations of existing mirror paradigms (Tab. 1). Our main contributions are summarized as follows:

• We propose a unified construction scheme for constrained generative models based on Brenier map. Compared with conventional mirror lifting paradigm, the proposed method is theoretically adaptable to arbitrary constraint domains, and is equipped with a general solution pipeline with controllable geometric distortion, which significantly enhances its practical applicability.

• We demonstrate that mirror functions over convex constraint domains are essentially a subclass of dual Brenier functions. Meanwhile, we reveal that the convexity-preserving property of mirror maps makes them inherently inapplicable to non-convex constraint domains. In contrast, Brenier functions, supported by singular set theory, can accommodate arbitrary non-convex scenarios, fundamentally breaking through the theoretical limitations of mirror-based methods.

• We proposed a general procedure for solving Brenier functions. Specifically, we first sample anchor points within the constraint domain. Then we identify critical positions of geometric distortion, namely singular sets, via the based geometric variational principle, and solve for a piecewise-linear Brenier function to lock the core geometric structure. Finally, we introduce a controllable smoothing technique to preserve critical geometric properties while improving generalization capability.

• Experiments show that our method outperforms neural approximation baselines in mirror function fitting across three classical convex constraint domains. For the watermark image and five complex real-world constrained generation tasks spanning both convex and non-convex scenarios, it surpasses existing SOTA baseline in both generation quality and constraint satisfaction at a reduced computational cost, delivering substantial gains in the accuracy and efficiency of constrained generation.

## 2 Preliminaries and Related Works

**Analytical or Neural Approximate Mirror Maps.** Within the lifting paradigm for constrained generative modeling, mirror maps serve as the dominant domain transformation tool, whose theoretical foundation hinges on convexity assumptions. Specifically, a strongly smooth convex mirror potential $\phi$ (Bubeck, 2015; Tan et al., 2023) and its gradient mirror map $\nabla\phi$ (Tae, 2023; Liu et al., 2024)—an invertible mapping from the constrained set $\mathcal{C} \subset \mathbb{R}^d$ to Euclidean space-are well-defined only if $\mathcal{C}$ is convex. Every valid mirror map admits an inverse $(\nabla\phi)^{-1} : \mathbb{R}^d \to \mathcal{C}$, yet this inverse does not necessarily correspond to the gradient of another strongly convex function (Zhou, 2018; Tan et al., 2023). Originating from constrained optimization (Beck & Teboulle, 2003; Borovykh

et al., 2021; Sun & Richtárik, 2022; Tzen et al., 2023; Nguyen & Sakurai, 2023; Yu et al., 2025; Bonet et al., 2024) and mirrored Langevin sampling (Hsieh et al., 2018; Zhang et al., 2020; Ahn & Chewi, 2021; Jiang, 2021; Shi et al., 2021; Lau et al., 2022; Li et al., 2022; Fishman et al., 2023b; Sharrock et al., 2023; Karimi Jaghargh et al., 2023; Zhu, 2023; Liu et al., 2024; Capdeville et al., 2026), mirror maps have recently been extended to constrained generative tasks (Tae, 2023; Liu et al., 2024; Ha et al., 2025; Guan et al., 2026; Li et al., 2026). Nevertheless, closed-form analytical mirror maps exclusively exist for convex domains, inherently restricting their applicability to non-convex constraints. Prior works alleviate this limitation via neural network approximations of mirror maps (Tan et al., 2023; Alfano et al., 2024; Feng et al., 2025), yet such purely neural parameterizations disregard the intrinsic geometric structure of transport mappings, leading to prohibitive training overhead, severe mapping distortion, and shifted distribution supports. To address these drawbacks, we reconstruct the constrained generative pipeline from the geometric perspective of optimal transport, replacing naive neural mirror approximations with exact Brenier maps.

**Brenier's Theorem and Quadratic Optimal Transport.** This work only consider the quadratic transport cost function $c(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$, and there exists strong correlation between the Brenier theory and optimal transport under this premise. Let $\mu$ and $\nu$ be the probability measures defined on $\mathbb{R}^d$ and $\mathcal{C}$ respectively, transport map $M : \mathbb{R}^d \to \mathcal{C}$ is called to be measure-preserving $M_{\#}\mu = \nu$ if, for any Borel set $B \subset \mathcal{C}$, there is $\int_B d\nu(\boldsymbol{y}) = \int_{M^{-1}(B)} d\mu(\boldsymbol{x})$. The famous Monge problem (Monge, 1781) aims to find the optimal transport map $M_{ot}$ with the lowest total transport cost

$$M_{ot} = \arg \inf_{M_{\#}\mu=\nu} \frac{1}{2} \int_{\mathbb{R}^d} \|\boldsymbol{x} - M(\boldsymbol{x})\|_2^2 d\mu(\boldsymbol{x}), \tag{1}$$

in all measure-preserving transport maps. Specially, there exists a convex Brenier potential function $u(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$ (Brenier, 1991) that satisfies that $\nabla u = M_{ot}$ is the unique solution to equation 1 when quadratic cost. Practical tasks frequently involve non-convex target supports $\mathcal{C}$. Figalli (2010) proves that zero-Lebesgue-measure singular set $\Sigma$ of $\mathbb{R}^d$ can render the Brenier map $\nabla u$ discontinuous; this theoretical observation is further elaborated by (Luo et al., 2022).

# 3 Constrained Generative Modeling via Brenier Maps

## 3.1 Unified Characterization of Mirror Maps via Brenier Maps

We first clarify the notation and direction of the mirror lifting paradigm. Given a convex constrained feasible set $\mathcal{C} \subset \mathbb{R}^d$, the mirror potential $\phi$ is a strongly convex function defined on $\mathcal{C}$. The associated mirror map takes the gradient form (Tae, 2023; Liu et al., 2024; Ha et al., 2025; Guan et al., 2026): $\nabla\phi : \mathcal{C} \to \mathbb{R}^d$. This map transforms data inside $\mathcal{C}$ into the unconstrained Euclidean space $\mathbb{R}^d$ for training standard generative models (GMs). During sampling, the inverse operator of the mirror map plays the central role: $S = (\nabla\phi)^{-1} : \mathbb{R}^d \to \mathcal{C}$, which maps samples drawn from the unconstrained space back to $\mathcal{C}$, ensuring all samples satisfy constraints by construction.

We then connect mirror maps to Brenier maps under the quadratic cost $c(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$. Let $\mu$ denote the source measure supported on $\mathbb{R}^d$, and $\nu$ denote the target measure supported on $\mathcal{C}$. By Brenier's theorem (Brenier, 1991), if $\mu$ is absolutely continuous with respect to the Lebesgue measure, there exists a unique lower-semicontinuous convex potential $u : \mathbb{R}^d \to \mathbb{R}$ such that the Brenier (optimal transport) map can be written as the gradient of this potential: $M_{ot} = \nabla u : \mathbb{R}^d \to \mathcal{C}$. For any valid mirror potential $\phi$, its inverse $S = (\nabla\phi)^{-1}$ satisfies all regularity requirements for a Brenier map. Using Fenchel conjugacy, we define the conjugate potential of $\phi$: $u(\boldsymbol{x}) = \sup_{\boldsymbol{y}\in\mathcal{C}}\{\langle \boldsymbol{x}, \boldsymbol{y}\rangle - \phi(\boldsymbol{y})\}$. Since the valid mirror potential $\phi$ is strongly convex and continuously differentiable on $\mathcal{C}$, its Fenchel conjugate $u$ is a convex function over $\mathbb{R}^d$, and the gradient inversion identity holds: $S(\boldsymbol{x}) = (\nabla\phi)^{-1}(\boldsymbol{x}) = \nabla u(\boldsymbol{x})$. Detailed derivations are provided in Appendix B.1.

In summary, the core sampling operator of the mirror lifting paradigm (the inverse mirror map) forms a proper subset of all Brenier maps under quadratic transport cost. The conventional mirror map framework is therefore a restricted special case of the general Brenier optimal transport framework, limited to scenarios where analytical mirror potentials exist on convex constrained domains.

Furthermore, Theorem 1 in Appendix B.3 demonstrates that any mirror map is theoretically inapplicable to non-convex constrained domains due to its intrinsic convexity constraint. Nevertheless, the singular set theory of Brenier maps endows them with the capability to adapt to arbitrary constraint

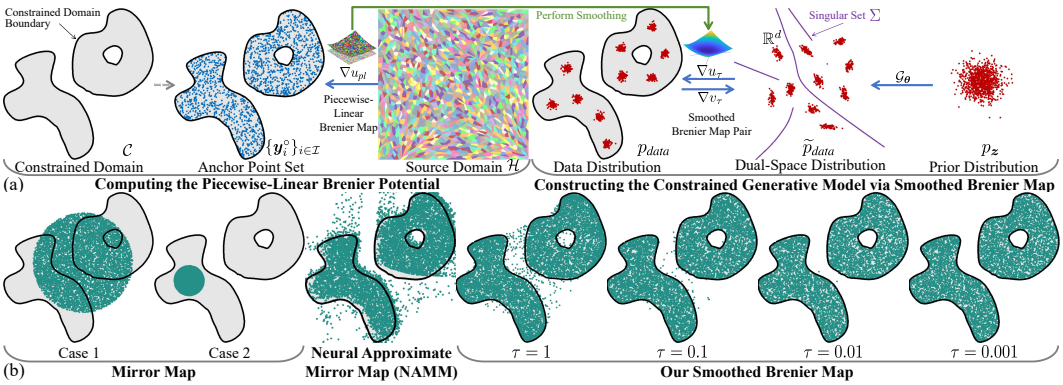



Fig 1: (a) The Overall Framework of Our Method. *Computing the Brenier Potential (Left)*: Given a constrained domain $\mathcal{C} \subset \mathbb{R}^d$, we first uniformly sample a set of anchor points $\{\boldsymbol{y}_i^\circ\}_{i\in\mathcal{I}} \subset \mathcal{C}$ to cover its main region. Using the geometric variational principle (Gu et al., 2013; An et al., 2020), we solve the piecewise-linear Brenier function $u_{pl}$ from a sufficiently large origin-centered hypercube $\mathcal{H}$ to $\mathcal{C}$. The controllably smoothed output $u_\tau$ serves as the foundation for constructing our constrained GMs. Notably, $u_\tau$ induces a singular set $\Sigma$ in $\mathbb{R}^d$, which corresponds to the region outside $\mathcal{C}$ under the Brenier map $\nabla u_\tau$. As $|\mathcal{I}| \to \infty$ and $\tau \to 0^+$, the Lebesgue measure of $\Sigma$ converges to zero, which theoretically enables distortion-free transformation between an arbitrary $\mathcal{C}$ and $\mathbb{R}^d$. *Constructing the Constrained GMs (Right)*: During training, we employ the dual map $\nabla v_\tau$ of $\nabla u_\tau$ to project real data distribution $p_{data}$ supported on $\mathcal{C}$ onto an unconstrained distribution $\widetilde{p}_{data}$ defined on the dual space $\mathbb{R}^d$. A GM $\mathcal{G}_{\boldsymbol{\theta}}$ that maps the prior distribution $p_{\boldsymbol{z}}$ to $\widetilde{p}_{data}$ is trained on the dual space $\mathbb{R}^d$. At the generation stage, we sample from $p_{\boldsymbol{z}}$, and the outputs of $\mathcal{G}_{\boldsymbol{\theta}}$ are mapped back to the constrained domain $\mathcal{C}$ via $\nabla u_\tau$. (b) Visual Comparison of Geometric Distortion Across Different Maps. We project 10000 random points in $\mathbb{R}^d$ onto $\mathcal{C}$ and analyze geometric distortion via point-distribution patterns. Conventional mirror maps cannot handle non-convex domains due to convex-preserving limitations (Theorem 1 in Appendix B.3). Although NAMMs (Feng et al., 2025) mitigate this issue, they still exhibit severe uncontrolled geometric distortion. Our method yields Brenier maps with controllable geometric distortion.

domains $\mathcal{C}$. Specifically, when the constraint domain $\mathcal{C}$ is non-convex, there exists a zero-measure singular set $\Sigma \subset \mathbb{R}^d$, on which the convex Brenier potential $u$ is non-differentiable, while all other theoretical conclusions still hold. The zero-measure property of $\Sigma$ does not affect the training of GMs constructed in the unconstrained Euclidean space. Motivated by this superior property, we propose to replace mirror maps with Brenier maps.

### 3.2 Efficient Computation of Brenier Maps and Their Inverse Operators

Direct computation of the continuous Brenier potential from $\mathbb{R}^d$ to an arbitrary constrained domain $\mathcal{C}$ is analytically intractable. We present a general geometric numerical approximation procedure: (1) we first sample anchor points uniformly inside $\mathcal{C}$, and employ a large origin-centered hypercube $\mathcal{H}$ (rather than unbounded $\mathbb{R}^d$) as the source domain, motivated by the compact support of training data; (2) we then solve the piecewise-linear Brenier potential from $\mathcal{H}$ to the set of sampled anchors via the Monte Carlo (MC)-based geometric variational principle (Gu et al., 2013; An et al., 2020), which enables exact identification of the singular set of the Brenier map (Luo et al., 2022); (3) the controlled convex smoothing procedure (Mazumder et al., 2019) is further applied to obtain a smoothed numerical approximation of the Brenier potential. Our Brenier-potential solver circumvents the difficult analytic-construction barrier of mirror-map constrained generation. Empirically, our method consumes roughly $\frac{2}{3}$ of the training cost of NAMMs, and outperforms them on both approximation quality and constraint satisfaction for constrained generation (Figs. 2, 5, 6. Tabs. 2, 3).

Let $\{\boldsymbol{y}_i^\circ\}_{i\in\mathcal{I}}$ denote the discrete anchor points sampled uniformly within the constrained domain $\mathcal{C}$. The target Dirac probability density is $p^\circ = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \delta(\boldsymbol{y} - \boldsymbol{y}_i^\circ)$. A large origin-centered hypercube $\mathcal{H}$ is chosen as the source domain equipped with measure $\mu$. Gu et al. (2013) construct a cluster of optimal hyperplanes $\{\pi_i(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{y}_i^\circ\rangle + h_i\}_{i\in\mathcal{I}}$ and denote their upper envelope as $env(\{\pi_i\}_{i\in\mathcal{I}})$, which corresponds to the graph of the convex piecewise-linear Brenier function

$$u_{pl}(\boldsymbol{x}) = \max_{i\in\mathcal{I}}\{\langle \boldsymbol{x}, \boldsymbol{y}_i^\circ\rangle + h_i\}, \boldsymbol{x} \in \mathcal{H}, \tag{2}$$

which converges to continuous case as $|\mathcal{I}| \to \infty$. Further, Gu et al. (2013) design an energy functional $E(\boldsymbol{h}) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} h_i - \int_0^{\boldsymbol{h}} \sum_{i\in\mathcal{I}} w_i(\boldsymbol{\eta}) d\eta_i$ to obtain the optimal $\boldsymbol{h}$, where $w_i(\boldsymbol{\eta}) = \mu(W_i)$, $\boldsymbol{h} = (h_i)_{i\in\mathcal{I}}$ and $W_i = \{\boldsymbol{x} \in \mathcal{H} | \langle \boldsymbol{x}, \boldsymbol{y}_i^\circ - \boldsymbol{y}_j^\circ \rangle \geq h_j - h_i, \forall j \in \mathcal{I}\}$ such that $\mathcal{H} = \bigcup_{i\in\mathcal{I}} W_i$, within the admissible potential space $\{\boldsymbol{h} | w_i(\boldsymbol{h}) > 0, i \in \mathcal{I}\} \cap \{\boldsymbol{h} | \sum_{i\in\mathcal{I}} h_i = 1\}$. We utilize the MC estimator (An et al., 2020) to maximize $E(\boldsymbol{h})$ until convergence, and then substitute the optimal $\boldsymbol{h}$ into equation 2 to yield the approximate piecewise-linear Brenier potential $u_{pl}(\boldsymbol{x})$.

We smooth $u_{pl}(\boldsymbol{x})$ with a smoothing temperature $\tau > 0$ (Mazumder et al., 2019):

$$u_\tau(\boldsymbol{x}) = \tau \log \left( \sum_{i\in\mathcal{I}} \exp\left( \frac{\langle \boldsymbol{x}, \boldsymbol{y}_i^\circ \rangle + h_i}{\tau} \right) \right) - \tau \log |\mathcal{I}|. \tag{3}$$

The Fenchel conjugate dual potential of $u_\tau(\boldsymbol{x})$ is defined as $v_\tau(\boldsymbol{y}) = \sup_{\boldsymbol{x}\in\mathbb{R}^d} \{\langle \boldsymbol{x}, \boldsymbol{y}^\circ \rangle - u_\tau(\boldsymbol{x})\}$, which amounts to the following maximization problem over the probability simplex $\Delta = \{\boldsymbol{\lambda} \mid \lambda_i \geq 0, \sum_{i\in\mathcal{I}} \lambda_i = 1, \boldsymbol{\lambda} = (\lambda_i)_{i\in\mathcal{I}}\}$. Detailed derivations are provided in Appendix B.2.

$$v_\tau(\boldsymbol{y}) = \begin{cases} \tau \log |\mathcal{I}| + \sup\limits_{\boldsymbol{\lambda}\in\Delta,\, \sum_{i\in\mathcal{I}} \lambda_i \boldsymbol{y}_i^\circ = \boldsymbol{y}} \sum\limits_{i\in\mathcal{I}} \lambda_i \left(-h_i - \tau \log \lambda_i\right), & \boldsymbol{y} \in \mathbf{conv}\{\boldsymbol{y}_i^\circ\}, \\ +\infty, & \boldsymbol{y} \notin \mathbf{conv}\{\boldsymbol{y}_i^\circ\}. \end{cases} \tag{4}$$

By convex analysis, Fenchel conjugates are always convex, so $v_\tau(\boldsymbol{y})$ is convex. The objective $\mathcal{L}(\boldsymbol{\lambda}) = \sum_{i\in\mathcal{I}} \lambda_i(-h_i - \tau \log \lambda_i)$ is strictly concave over the simplex; thus for every $\boldsymbol{y} \in \mathbf{conv}\{\boldsymbol{y}_i^\circ\}$, there exists a unique optimal weight vector $\boldsymbol{\lambda}^*(\boldsymbol{y})$, ensuring well-posedness of the dual potential.

We define the Brenier map as the gradient of the smoothed potential $u_\tau(\boldsymbol{x})$:

$$\boldsymbol{T}_\tau(\boldsymbol{x}) \triangleq \nabla u_\tau(\boldsymbol{x}) = \sum_{i\in\mathcal{I}} \left( \frac{\exp\left(\frac{\langle \boldsymbol{x}, \boldsymbol{y}_i^\circ \rangle + h_i}{\tau}\right)}{\sum_{j\in\mathcal{I}} \exp\left(\frac{\langle \boldsymbol{x}, \boldsymbol{y}_j^\circ \rangle + h_j}{\tau}\right)} \right) \boldsymbol{y}_i^\circ, \tag{5}$$

which maps latent source points to the constrained target domain. The entire workflow is detailed in Algorithm 1.

Then the inverse Brenier map $\boldsymbol{S}_\tau(\boldsymbol{y}) \triangleq \nabla v_\tau(\boldsymbol{y})$ is defined to lift samples in $\mathcal{C}$ back to $\mathbb{R}^d$. From differentiable Fenchel conjugate properties, the Brenier and inverse maps satisfy a bijective reciprocal relation on smooth regions of the potentials: $\boldsymbol{y} = \boldsymbol{T}_\tau(\boldsymbol{x}) \iff \boldsymbol{x} = \boldsymbol{S}_\tau(\boldsymbol{y})$.

The inverse map $\boldsymbol{S}_\tau$ admits no elementary closed-form expression. For a valid $\boldsymbol{y} \in \mathbf{conv}\{\boldsymbol{y}_i^\circ\}$, $\boldsymbol{S}_\tau(\boldsymbol{y})$ equals the unique solution $\boldsymbol{x}^*$ to the nonlinear equation $\boldsymbol{y} = \boldsymbol{T}_\tau(\boldsymbol{x})$. To compute $\boldsymbol{S}_\tau(\boldsymbol{y})$ numerically, we solve least-squares minimization problem by inverting the explicit Brenier map: $\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \left\| \boldsymbol{T}_\tau(\boldsymbol{x}) - \boldsymbol{y} \right\|_2^2$. Within $\mathbf{conv}\{\boldsymbol{y}_i^\circ\}$, $\boldsymbol{T}_\tau$ is smooth and strictly monotonic, making the loss strictly convex with unique global minimum. We employ gradient-based iterative optimization to recover $\boldsymbol{x}^*$ and set $\boldsymbol{S}_\tau(\boldsymbol{y}) = \boldsymbol{x}^*$. This procedure is listed in Algorithm 2.

**Algorithm 1** The Computation Process of the Brenier Potential and the Brenier Map

**Input**: Constraint domain $\mathcal{C} \subset \mathbb{R}^d$, initial MC sample size $N_{mc}$, source domain $\mathcal{H} = [-R_\mathcal{H}, R_\mathcal{H}]^d$ with measure $\mu$, learning rate $lr$, maximum iterations $T$, integer $s > 0$, termination thresholds $\{\kappa_1, \kappa_2\}$, and smoothness coefficient $\tau > 0$.

1: $\{\boldsymbol{y}_i^\circ\}_{i\in\mathcal{I}} \overset{i.i.d}{\sim} \mathcal{C}$; $\boldsymbol{h}^{(0)} \leftarrow \mathbf{0}_{1\times|\mathcal{I}|}$.
2: **for** $t = 0 : T-1$ **do**
3: $\quad W_i(\boldsymbol{h}^{(t)}) \leftarrow \{\boldsymbol{x} \in \mathcal{H} \mid \langle \boldsymbol{x}, \boldsymbol{y}_i^\circ \rangle + h_i^{(t)} \geq \langle \boldsymbol{x}, \boldsymbol{y}_j^\circ \rangle + h_j^{(t)}, \forall j \neq i, j \in \mathcal{I}\}, i \in \mathcal{I}$.
4: $\quad \nabla_{h_i} \hat{E}(\boldsymbol{h}^{(t)}) \leftarrow \frac{1}{N_{mc}} \sum_{j=1}^{N_{mc}} \mathbb{I}\{\boldsymbol{x}_j \in W_i(\boldsymbol{h}^{(t)})\} - \frac{1}{|\mathcal{I}|}$, $i \in \mathcal{I}$, where $\{\boldsymbol{x}_j\}_{j=1}^{N_{mc}} \overset{i.i.d}{\sim} (\mathcal{H}, \mu)$.
5: $\quad \nabla\hat{E}(\boldsymbol{h}^{(t)}) \leftarrow \nabla\hat{E}(\boldsymbol{h}^{(t)}) - mean(\nabla\hat{E}(\boldsymbol{h}^{(t)}))$.
6: $\quad \boldsymbol{h}^{(t+1)} \leftarrow \boldsymbol{h}^{(t)} - lr \cdot \nabla\hat{E}(\boldsymbol{h}^{(t)})$; $\boldsymbol{h}^* \leftarrow \boldsymbol{h}^{(t+1)}$.
7: $\quad N_{mc} \leftarrow 2 \cdot N_{mc}$, $lr \leftarrow 0.9 \cdot lr$ if $\|\nabla\hat{E}(\boldsymbol{h}^{(t)})\|_2$ has not decreased for $s$ steps.
8: $\quad$ break if $\|\nabla\hat{E}(\boldsymbol{h}^{(t)})\|_2 < \kappa_1$ or $lr < \kappa_2$.
9: **end for**
10: $u_\tau(\cdot) \leftarrow \tau \log \left( \sum_{i\in\mathcal{I}} \exp\left( \frac{\langle \cdot, \boldsymbol{y}_i^\circ \rangle + h_i^*}{\tau} \right) \right) - \tau \log |\mathcal{I}|$.
11: $\boldsymbol{T}_\tau(\cdot) \leftarrow \sum_{i\in\mathcal{I}} \left( \frac{\exp\left(\frac{\langle \cdot, \boldsymbol{y}_i^\circ \rangle + h_i^*}{\tau}\right)}{\sum_{j\in\mathcal{I}} \exp\left(\frac{\langle \cdot, \boldsymbol{y}_j^\circ \rangle + h_j^*}{\tau}\right)} \right) \boldsymbol{y}_i^\circ$.

**Return**: Brenier potential $u_\tau(\cdot)$ and Brenier map $\boldsymbol{T}_\tau(\cdot)$.

### 3.3 Brenier-Map-Driven Generative Modeling

After numerically computing the Brenier map $\boldsymbol{T}_\tau : \mathbb{R}^d \to \mathcal{C}$ and its conjugate inverse map $\boldsymbol{S}_\tau : \mathcal{C} \to \mathbb{R}^d$, we construct a fully generic constrained-generation pipeline decoupled from the architecture of the underlying GM. We uniformly denote the arbitrary $\boldsymbol{\theta}$-parameterized GM as $\mathcal{G}_{\boldsymbol{\theta}}$.

**Projecting Constrained Training Data to Dual Space.** Let the training set be $\{\boldsymbol{y}_j\}_{j\in\mathcal{J}} \overset{i.i.d}{\sim} p_{data}$, where each sample satisfies $\boldsymbol{y}_j \in \mathcal{C}$. Using the inverse map $\boldsymbol{S}_\tau$, we push forward the constrained data distribution $p_{data}$ onto the unconstrained dual space and obtain the dual-space distribution $\widetilde{p}_{data} = (\boldsymbol{S}_\tau)_{\#}\, p_{data}(\boldsymbol{y})$, corresponding to the transformed dual-space training set $\{\boldsymbol{x}_j = \boldsymbol{S}_\tau(\boldsymbol{y}_j)\}_{j\in\mathcal{J}}$. This transformation is executed offline once before training and is not recomputed during model optimization, which substantially reduces the computational overhead in the training phase.

**Training of the Generic Generative Model $\mathcal{G}_{\boldsymbol{\theta}}$.** We train the GM $\mathcal{G}_{\boldsymbol{\theta}}$ purely on the unconstrained dual space $\mathbb{R}^d$ to learn $\widetilde{p}_{data}$ from the prior distribution $p_{\boldsymbol{z}}$. The model never directly observes constrained geometry, and no additional constraint-penalty terms are required. The training objective remains identical to that of standard unconstrained GMs.

**Mapping from Unconstrained Latents Back to Constrained Domain.** After training completes, constraint-compliant samples are produced in three stages: (1) Dual-Space Prior Sampling. Draw latent variable $\boldsymbol{z} \sim p_{\boldsymbol{z}}$ from the dual-space prior. (2) GM Forward Pass. Feed latent variable into the trained generator and obtain dual-space output $\hat{\boldsymbol{x}} = \mathcal{G}_{\boldsymbol{\theta}}(\boldsymbol{z})$. (3) Brenier Map Projection onto Constrained Domain. Pass $\hat{\boldsymbol{x}}$ through the pre-computed Brenier map to obtain final generated sample $\hat{\boldsymbol{y}} = \boldsymbol{T}_\tau\big(\mathcal{G}_{\boldsymbol{\theta}}(\boldsymbol{x})\big)$ residing inside the constrained domain $\mathcal{C}$.

# 4 Experiments

To fully verify the superiority of the proposed smooth Brenier map module, we select the analytical mirror map (Liu et al., 2024; Guan et al., 2026) and its neural approximation, i.e. NAMM (Feng et al., 2025), as baseline methods. The analytical mirror map has been demonstrated to adapt to various GMs $\mathcal{G}_{\boldsymbol{\theta}}$ including diffusion models (Liu et al., 2024) and flow matching (Guan et al., 2026), and achieves SOTA performance in watermarked image generation with convex polytope constraints. NAMM (Feng et al., 2025), to the best of our knowledge, is the only mirror-lifting paradigm applicable to non-convex constraints, and delivers SOTA generation performance under complex constraints. To further show the generalizability of our approach, we adopt three classical GMs, i.e., diffusion models, flow matching, and VAE, as the unconstrained $\mathcal{G}_{\boldsymbol{\theta}}$ in the dual space, , with their architectures and hyperparameters detailed in Appendices C.7, C.8.

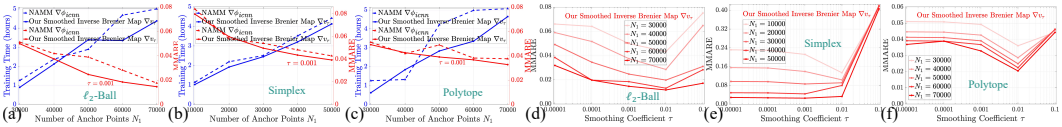
## 4.1 Map Approximation Error Evaluation across $\ell_2$-ball, Simplex and Polytope

**Experimental Setup.** We adopt the analytical mirror maps derived in (Liu et al., 2024) for three typical convex constrained domains, i.e., the arbitrary-dimensional $\ell_2$-ball, simplex and polytope, as the ground-truth references. Their analytical forms and detailed parameter settings are provided in Appendices C.1, C.4. For each convex constrained domain $\mathcal{C}$, let $\phi$ denote its corresponding analytical mirror function. Crucially, the convexity allows us to directly take $\mathcal{C}$ as source domain to solve for $v_\tau$. Specifically, we first uniformly sample discrete points $\{\boldsymbol{y}_i\}_{i=1}^{N_1}$ within $\mathcal{C}$, and use $\{\nabla\phi(\boldsymbol{y}_i)\}_{i=1}^{N_1}$ as anchor points. The hyperparameters are configured as: source domain $\mathcal{C}$ equipped with uniform measure $\mu$, initial MC sample size $N_{mc} = 10000$, maximum iteration $T = 1000000$, convergence thresholds $\{\kappa_1 = 10^{-4}, \kappa_2 = 10^{-12}\}$, $s = 100$, and learning rate $lr = 0.0001$. Following the core idea of Algorithm 1, we compute the piecewise-linear dual Brenier potential from $\mathcal{C}$ to the anchor set, yielding the smoothed dual Brenier function $v_\tau$ modulated by the smoothing coefficient $\tau$. For comparison, we feed $\{(\boldsymbol{y}_i, \nabla\phi(\boldsymbol{y}_i))\}_{i=1}^{N_1}$ into NAMM (structure and all hyperparameters are listed in Appendices C.2, C.4) and obtain the trained neural approximate mirror function denoted as $\phi_{icnn}$.

**Evaluation Metrics.** We employ the Mean Map Approximation Relative Error (MMARE) to quantify how well $v_\tau$ and $\phi_{icnn}$ approximate $\phi$. Specifically, we randomly sample $N_2 = 50000$ points $\{\boldsymbol{y}_i\}_{i=1}^{N_2}$ within the constrained domain, then MMARE $= \frac{1}{N_2}\sum_{i=1}^{N_2} \frac{\|\nabla\phi(\boldsymbol{y}_i) - \nabla\phi'(\boldsymbol{y}_i)\|_2}{\|\nabla\phi(\boldsymbol{y}_i)\|_2}$, where $\phi'$ stands for the candidate approximation, i.e., $v_\tau$ or $\phi_{icnn}$. Furthermore, the Training Time of $v_\tau$ and $\phi_{icnn}$ measured on single RTX 4090 GPU is reported to compare computational efficiency.

**Experimental Results.** Figs. 2, 5, 6 compare the approximation accuracy (referenced to analytical mirror maps) and computational efficiency of our smoothed Brenier map against NAMM across three classic convex constraint domains. Subplots (a)-(c) reveal that increasing the anchor number $N_1$ increases training overhead, yet yields lower MMARE under the same smoothing coefficient $\tau$ after training. Even so, our method still outperforms NAMM in both accuracy and computational

efficiency. Subplots (d)-(f) further show that, with $N_1$ fixed, either excessively large or small $\tau$ increases the MMARE, which is consistent with theoretical intuition.




(a) (b) (c) (d) (e) (f)

Fig 2: Effects of Anchor Number $N_1$ and Smoothness Coefficient $\tau$ on the Approximation Accuracy and Computational Efficiency of Different Map Pairs over Three Classical Convex Constraint Domains (Dimension $d = 5$). The results for $d = 2$ and 3 are shown in Figs. 5, 6.

### 4.2 Watermarked Image Generation

**Experimental Setup.** Following (Liu et al., 2024), we evaluate the proposed method on the task of watermarked $64 \times 64$ image generation using the FFHQ (Karras et al., 2019) and AFHQv2 (Choi et al., 2020) datasets. We first generate parameters $(\boldsymbol{a}_k, b_k, c_k)_{k\in\mathcal{K}}$ as private keys for distinct users, which define the constrained polytope $\mathcal{C} = \{\boldsymbol{y} \mid c_k < \langle \boldsymbol{a}_k, \boldsymbol{y}\rangle < b_k, k \in \mathcal{K}\}$ (Hyperparameters are set as described in Appendix C.5.). The corresponding analytical mirror function pair $\{\phi, \phi^*\}$ can be derived from equation 27. We select $N_3 = 50000$ real training samples within $\mathcal{C}$ as the anchor points $\{\boldsymbol{y}_i^\circ\}_{i=1}^{N_3}$ and solve for the smoothed forward-backward Brenier function pair $\{u_\tau, v_\tau\}$ via Algorithms 1, 2. The hyperparameters are set as: initial MC sample size $N_{mc} = 10000$, maximum iteration $T = 1000000$, source domain $\mathcal{H} = [-100, 100]^d$ with uniform measure $\mu$, convergence threshold $\{\kappa_1 = 10^{-4}, \kappa_2 = 10^{-12}\}$, $s = 100$, learning rate $lr = 0.1$, and smoothing coefficient $\tau = 0.5$. Then we adopt $\{\boldsymbol{y}_i^\circ\}_{i=1}^{N_3}$ to learn the neural approximate mirror function pair $\{\phi_{icnn}, \psi_{icnn}\}$ (Feng et al., 2025), as listed in Appendix C.6. For training, we first project raw FFHQ and AFHQv2 images onto $\mathcal{C}$ to embed watermarks and construct two watermarked datasets. Afterwards, the watermarked samples are transformed into the unconstrained dual space via $\nabla\phi$, $\nabla\phi_{icnn}$, or $\nabla v_\tau$ to train $\mathcal{G}_{\boldsymbol{\theta}}$. At inference time, images can be converted into vector representations for watermark verification: an image is confirmed to be generated by the corresponding model if it lies inside $\mathcal{C}$.

**Evaluation Metrics.** For each model, we generate 50000 images and evaluate the generation quality using Fréchet Inception Distance (FID) (Heusel et al., 2017) and CMMD (Jayasumana et al., 2024), where CMMD integrates CLIP embeddings with the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) metric. We measure the Constraint Satisfaction Ratio (CSR) of GMs by computing the proportion of these samples satisfying constraint $\mathcal{C}$. Furthermore, we report the total Training Time on single RTX 4090 GPU to assess the computational efficiency of each model.

**Experimental Results.** Tab. 2 summarizes the training time and performance of all methods on the watermark generation task. Notably, the reported NAMM training time in Tabs. 2, 3 covers both standard model training and reverse fine-tuning (As described in Appendix C.2) using the unconstrained GM in the dual space, which has been proven to boost performance in the original work (Feng et al., 2025). Experimental results demonstrate that mirror- and Brenier-lifting schemes outperform unconstrained GMs in both generation quality and constraint satisfaction. Our method achieves the best performance, with a CSR above 97%. In terms of computational efficiency, by transforming complex data distributions into regular, easy-to-learn dual distributions, these lifting methods require far less training time for the GMs than direct training on the original dataset. Compared with the analytical mirror map, the Brenier map needs additional training time but yields comparable or lower total training time, as its induced dual distribution is more regular whereas that of the analytical mirror map exhibits exponential divergence (Guan et al., 2026). Moreover, each module of NAMM takes longer to train than our method. Fig. 3, 7, 8 show the generation results of different models.

### 4.3 More Complex Scenarios

**Experimental Setup.** To fully verify the universal applicability of our smoothed Brenier map framework in more complex convex domains, non-convex geometries, PDE-governed physical systems, and scenarios with discrete counting constraints, we follow the setup in (Feng et al., 2025) and adopt five standard constrained generative benchmarks: fixed total brightness (Akiyama et al., 2019) (linear convex constraint), 1D Burgers' PDE (Bateman, 1915; Burgers, 1948) (nonlinear non-convex temporal constraint), 2D divergence-free fluid flow (Chandler & Kerswell, 2013; Boffetta & Ecke, 2012; Rozet & Louppe, 2023) (differential geometric constraint), periodic tiled geometry (Ogren

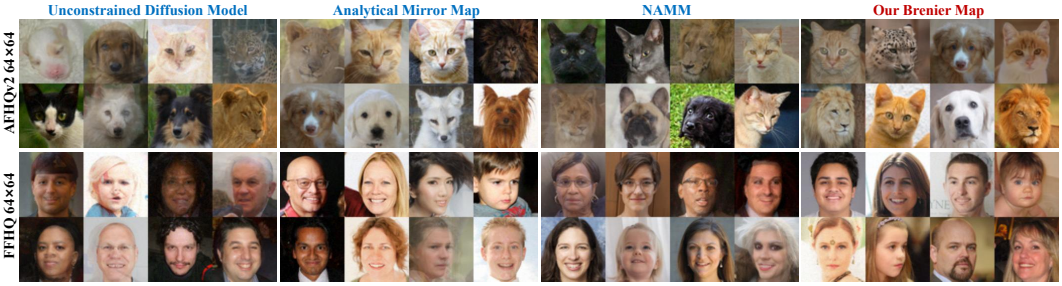



Fig 3: Comparison of Different Map Pairs Coupled with Diffusion Model via Generated Watermarked Image Visualization. More results using Flow Matching and VAE as $\mathcal{G}_{\boldsymbol{\theta}}$ are shown in Figs. 7,8.

Table 2: Performance Comparison of Different Methods in Watermarked Image Generation. h: hours.

| FFHQ 64 × 64 | +Diffusion Model | | | | | +Flow Matching | | | | | +VAE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | FID↓ | CMMD↓ | CSR↑ | $\phi'$ Training↓ | $\mathcal{G}_\theta$ | FID↓ | CMMD↓ | CSR↑ | $\phi'$ Training↓ | $\mathcal{G}_\theta$ | FID↓ | CMMD↓ | CSR↑ | $\phi'$ Training↓ | $\mathcal{G}_\theta$ |
| Unconstrained GM | 9.07 | 0.0753 | 70.3% | — | 37.2h | 9.63 | 0.0617 | 69.4% | — | 26.2h | 26.30 | 0.5661 | 62.7% | — | 8.95h |
| Analytical Mirror Map | 5.26 | 0.0338 | **100%** | — | 24.2h | 6.88 | 0.0421 | **100%** | — | 10.5h | 19.83 | 0.3206 | **100%** | — | 5.33h |
| NAMM | 6.72 | 0.0288 | 93.8% | 6.37h | 23.8h | 7.19 | 0.0483 | 94.7% | 6.37h | 10.1h | 22.74 | 0.2877 | 88.1% | 6.37h | 6.08h |
| Our Brenier Map | **4.94** | **0.0166** | 98.5% | **4.25h** | **19.3h** | **6.06** | **0.0374** | 98.9% | **4.25h** | **8.36h** | **17.13** | **0.2125** | 97.0% | **4.25h** | **3.89h** |
| AFHQv2 64 × 64 | +Diffusion Model | | | | | +Flow Matching | | | | | +VAE | | | | |
| Methods | FID↓ | CMMD↓ | CSR↑ | $\phi'$ Training↓ | $\mathcal{G}_\theta$ | FID↓ | CMMD↓ | CSR↑ | $\phi'$ Training↓ | $\mathcal{G}_\theta$ | FID↓ | CMMD↓ | CSR↑ | $\phi'$ Training↓ | $\mathcal{G}_\theta$ |
| Unconstrained GM | 8.95 | 0.4628 | 68.9% | — | 37.1h | 9.16 | 0.0747 | 65.6% | — | 28.8h | 25.51 | 0.5014 | 57.3% | — | 9.47h |
| Analytical Mirror Map | 4.10 | 0.1531 | **100%** | — | 24.8h | 4.59 | 0.0271 | **100%** | — | 12.4h | 16.49 | 0.2852 | **100%** | — | 6.88h |
| NAMM | 5.47 | 0.1917 | 91.4% | 7.54h | 22.6h | 4.91 | 0.0553 | 93.6% | 7.54h | 10.7h | 19.36 | 0.3255 | 89.5% | 7.54h | 4.74h |
| Our Brenier Map | **3.98** | **0.1338** | 97.7% | **5.08h** | **16.9h** | **4.36** | **0.0189** | 98.1% | **5.08h** | **7.76h** | **13.39** | **0.1861** | 97.4% | **5.08h** | **4.15h** |

et al., 2024), and fixed astronomical object counting (Paiss et al., 2023) (highly non-convex semantic constraint). Detailed descriptions of these scenarios and the corresponding datasets are provided in Appendix C.9. For the proposed Brenier map, we select all real training samples within each constrained domain $\mathcal{C}$ as discrete anchor points and employ the same hyperparameters as in Section 4.2 to compute the smooth forward-backward Brenier function pair $\{u_\tau, v_\tau\}$. $\tau = 0.1$ for Total Brightness and Count, and 0.5 for Burgers, Divergence-free and Periodic tasks. For the NAMM (Feng et al., 2025), we use identical training samples and strictly follow the network architecture and hyperparameter settings reported in the original paper (Also listed in Appendix C.6) to train the neural approximate mirror function pair $\{\phi_{icnn}, \psi_{icnn}\}$. Nevertheless, the results reproduced from its publicly released code deviate from the reported values in the reference paper. Therefore, all comparisons in this paper are based on our reproduced results to reflect the practical performance of this method.

**Evaluation Metrics.** We evaluate the performance of constrained GMs from four perspectives. Constraint Distance (CD) quantifies the degree of constraint violation for generated samples; a value closer to zero implies better constraint satisfaction. MMD (Gretton et al., 2012) and Kernel Inception Distance (KID) (Bińkowski et al., 2018) measure the similarity between the generated distribution and the ground-truth data distribution. Mean Map Reconstruction Error (MMRE) assesses the bidirectional invertibility of forward-backward Brenier map or NAMM pairs. Given $N_4 = 50000$ samples $\{\boldsymbol{y}_i\}_{i=1}^{N_4}$ within $\mathcal{C}$, MMRE $= \frac{1}{N_4}\sum_{i=1}^{N_4} \|\nabla\psi'(\nabla\phi'(\boldsymbol{y}_i)) - \boldsymbol{y}_i\|_2$, here $\{\phi', \psi'\}$ implies $\{u_\tau, v_\tau\}$ or $\{\phi_{icnn}, \psi_{icnn}\}$. Moreover, we record the total Training Time on single RTX 4090 GPU to compare the construction efficiency of different constrained GMs.

**Experimental Results.** Experimental results in Tab. 3 demonstrate that the smoothed Brenier map reduces the computational cost by an average of 41.44% and lowers the MMRE by over 100× compared with NAMM. This indicates that the smoothed Brenier map possesses superior reversibility, which lays a solid foundation for the promising performance of our proposed constrained GM.

Table 3: Evaluation of Invertibility and Computational Efficiency for Different Map Pairs in Complex Constrained-Generation Scenarios. h: hours.

| Tasks | Total Brightness | | 1D Burgers | | Divergence-Free | | Periodic | | Count | |
|---|---|---|---|---|---|---|---|---|---|---|
| Methods | MMRE↓ | Training↓ | MMRE↓ | Training↓ | MMRE↓ | Training↓ | MMRE↓ | Training↓ | MMRE↓ | Training↓ |
| NAMM | 0.0478 | 5.5053h | 0.1019 | 5.1194h | 3.6386 | 15.741h | 0.0342 | 9.7819h | 6.3397 | 9.7344h |
| Our Brenier Map | **0.000084** | **3.7217h** | **0.000876** | **2.5156h** | **0.000986** | **7.3650h** | **0.000223** | **6.4333h** | **0.000979** | **6.1806h** |

Furthermore, we transform the original training data into unconstrained dual-space datasets via the well-trained NAMM and smoothed Brenier map, and then adopt the unconstrained GM $\mathcal{G}_{\boldsymbol{\theta}}$ for training. Figs. 4, 9, 10 present a comprehensive comparison of generation visualization and sample-level constraint satisfaction distributions, verifying that our model achieves the optimal performance in both visual quality and constraint compliance. Tabs. 4, 5 report the training overhead and final perf-

ormance of different $\mathcal{G}_{\boldsymbol{\theta}}$. Compared with conventional unconstrained GMs, the introduction of NAMM or Brenier map effectively improves generation quality and constraint satisfaction across five complex constrained generation tasks, among which the Brenier map yields the best overall results.

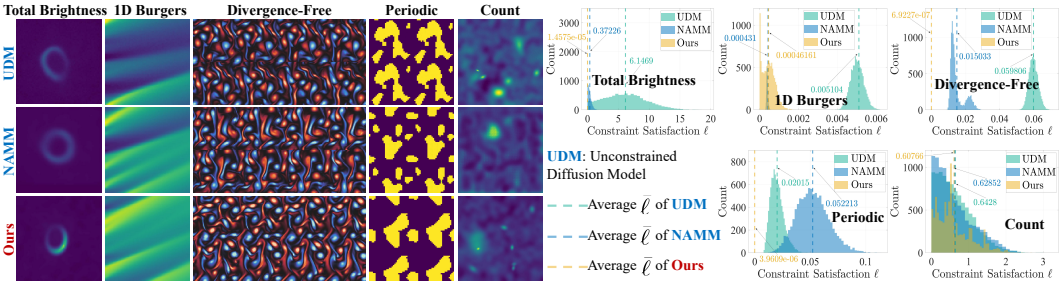



Fig 4: Comparison of Different Map Pairs Coupled with Diffusion Model via Sample Visualization, Per-Sample Distribution of the Data-Specific Constraint Satisfaction $\ell$ (Defined in Appendix C.9), and Its Average Value $\bar{\ell}$. Results using Flow Matching and VAE as $\mathcal{G}_{\boldsymbol{\theta}}$ are shown in Figs. 9, 10.

Table 4: Performance Comparison of Different Methods in Complex Constrained-Generation Scenarios. CD $= 100 \times \lambda_{constr} \times \ell$, where $\ell$ denotes data-specific constraint function defined in Appendix C.9. Experimental results using VAE as the GM $\mathcal{G}_{\boldsymbol{\theta}}$ are shown in Tab. 5. h: hours.

| Tasks | Total Brightness ($\lambda_{constr} = 0.01$) | | | | 1D Burgers ($\lambda_{constr} = 1$) | | | |
|---|---|---|---|---|---|---|---|---|
| Methods (+ Diffusion Model) | CD↓ | MMD↓ | KID↓ | Train $\mathcal{G}_{\boldsymbol{\theta}}$ ↓ | CD↓ | MMD↓ | KID↓ | Train $\mathcal{G}_{\boldsymbol{\theta}}$ ↓ |
| Unconstrained Diffusion Model | 6.1469 | 0.4860 | 0.1232 | 10.973h | 0.5104 | 0.1435 | 0.2462 | 9.7339h |
| NAMM (Tab. 3) | 0.3723 | 0.3198 | 0.0170 | 5.5053h | **0.0431** | 0.2690 | 0.0087 | 2.0161h |
| Our Brenier Map (Tab. 3) | **0.000015** | **0.1668** | **0.0120** | **3.7217h** | 0.0462 | **0.1148** | **0.0056** | **1.8864h** |

| Divergence-Free ($\lambda_{constr} = 1$) | | | | Periodic ($\lambda_{constr} = 1$) | | | | Count ($\lambda_{constr} = 0.01$) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CD↓ | MMD↓ | KID↓ | Train $\mathcal{G}_{\boldsymbol{\theta}}$ ↓ | CD↓ | MMD↓ | KID↓ | Train $\mathcal{G}_{\boldsymbol{\theta}}$ ↓ | CD↓ | MMD↓ | KID↓ | Train $\mathcal{G}_{\boldsymbol{\theta}}$ ↓ |
| 5.9806 | 0.1106 | 0.1799 | 98.5069h | 2.0150 | **0.0723** | **0.0045** | 29.586h | 0.6428 | **0.0534** | 0.2059 | 37.675h |
| 1.5033 | 0.0971 | 0.1778 | 35.2875h | 5.2213 | 0.1051 | 0.0179 | 6.0947h | 0.6285 | 0.4912 | 0.0544 | 8.1356h |
| **0.0001** | **0.0749** | **0.0012** | **22.9572h** | **0.0004** | 0.0952 | 0.0069 | **5.9708h** | **0.6077** | 0.3491 | **0.0166** | **7.6078h** |

| Tasks | Total Brightness ($\lambda_{constr} = 0.01$) | | | | 1D Burgers ($\lambda_{constr} = 1$) | | | |
|---|---|---|---|---|---|---|---|---|
| Methods (+ Flow Matching) | CD↓ | MMD↓ | KID↓ | Train $\mathcal{G}_{\boldsymbol{\theta}}$ ↓ | CD↓ | MMD↓ | KID↓ | Train $\mathcal{G}_{\boldsymbol{\theta}}$ ↓ |
| Unconstrained Flow Matching | 7.1104 | 0.4236 | 0.1847 | 4.0756h | 0.6254 | 0.0763 | 0.1817 | 2.0688h |
| NAMM (Tab. 3) | 2.3543 | 0.1856 | 0.0344 | 2.1067h | 0.0749 | 0.0708 | 0.0008 | 1.9999h |
| Our Brenier Map (Tab. 3) | **0.00002** | **0.1423** | **0.0304** | **1.8344h** | **0.0356** | **0.0699** | **0.0007** | **1.3517h** |

| Divergence-Free ($\lambda_{constr} = 1$) | | | | Periodic ($\lambda_{constr} = 1$) | | | | Count ($\lambda_{constr} = 0.01$) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CD↓ | MMD↓ | KID↓ | Train $\mathcal{G}_{\boldsymbol{\theta}}$ ↓ | CD↓ | MMD↓ | KID↓ | Train $\mathcal{G}_{\boldsymbol{\theta}}$ ↓ | CD↓ | MMD↓ | KID↓ | Train $\mathcal{G}_{\boldsymbol{\theta}}$ ↓ |
| 7.7833 | 0.0724 | 0.0038 | 10.967h | 5.1399 | 0.0769 | 0.0199 | 8.0084h | 1.0303 | 0.0959 | 0.0431 | 5.7087h |
| 4.9295 | 0.0445 | 0.0248 | 10.623h | 5.1281 | **0.0682** | **0.0017** | 6.4256h | 0.9386 | 0.0799 | 0.0422 | 5.6747h |
| **0.000068** | **0.0217** | **0.0027** | **10.043h** | **0.000071** | 0.0905 | 0.0024 | **3.8847h** | **0.7306** | **0.0648** | **0.0407** | **5.1617h** |

In particular, constrained GM built upon dual-space transformation achieves lower overall training costs than direct training on raw data, with the Brenier map exhibiting the most prominent efficiency advantage. The underlying reason is that the proposed paradigm converts complex constrained raw data into standardized dual data, enabling the GM to capture essential data features more efficiently.

## 5 Conclusion

This paper proposes a unified framework for constructing constrained GMs based on Brenier maps. Theoretically, we prove that mirror functions on convex domains constitute a special case of dual Brenier functions. By leveraging singular set theory, our method removes the convexity limitation of mirror maps and enable unified adaptation to arbitrary convex and non-convex domains. Algorithmically, we develop a universal multi-stage solving strategy for Brenier map estimation: anchor sampling and geometric variational principles are adopted to capture precise geometric structures, and a controllable smoothing mechanism is introduced to preserve critical geometric properties and improve generalization. Experimental results show that our method achieves higher fitting accuracy of mirror functions than neural approximation baselines across three classical convex constraint domains. For the watermark image generation task and five additional complex real-world constrained generation tasks involving both convex and non-convex scenarios, our approach surpasses existing SOTA baselines in generation quality and constraint satisfaction with lower computational cost, substantially advancing the accuracy and efficiency of constrained generative modeling.

### AI USE STATEMENT

In this work, we used generative AI tools for polishing manuscript text and minor auxiliary code formatting. We have not used generative AI tools for formulating research ideas, designing experiments, analyzing experimental results, drawing conclusions, or writing the core content of the manuscript, and AI-assisted generation of research hypotheses, dataset construction, and full manuscript drafting are not applicable to this work. Additionally, we used generative AI tools for grammar refinement and minor code snippet adjustment. We have reviewed all AI-assisted work. All AI-polished text was manually revised and checked by the authors; AI-assisted code was fully examined, tested and modified by the authors to ensure correctness. No research ideas, interpretations or conclusions were generated by AI. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

All theoretical assumptions, derivations, and complete proofs are elaborated in the main text and appendix to ensure theoretical reproducibility. For empirical experiments, we adopt publicly available datasets, and all implementation details including data preprocessing pipelines, hyperparameters, and evaluation settings are fully documented in the appendix. The complete source code will be publicly released upon paper acceptance.

## REFERENCES

8bitmp3. Jax-flax-tutorial-image-classification-with-linen, 2023. URL `https://github.com/8bitmp3/JAX-Flax-Tutorial-Image-Classification-with-Linen`.

Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-langevin algorithm. *Advances in Neural Information Processing Systems*, 34:28405–28418, 2021.

Kazunori Akiyama, Antxon Alberdi, Walter Alef, Keiichi Asada, Rebecca Azulay, Anne-Kathrin Baczko, David Ball, Mislav Baloković, John Barrett, Dan Bintley, et al. First m87 event horizon telescope results. iv. imaging the central supermassive black hole. *The Astrophysical Journal Letters*, 875(1):L4, 2019.

Carlo Alfano, Sebastian Towers, Silvia Sapora, Chris Lu, and Patrick Rebeschini. Meta-learning the mirror map in policy mirror descent. *arXiv preprint arXiv:2402.05187*, 2024.

James Urquhart Allingham, Javier Antoran, Shreyas Padhy, Eric Nalisnick, and José Miguel Hernández-Lobato. Learning generative models with invariance to symmetries. In *NeurIPS 2022 Workshop on Symmetry and Geometry in Neural Representations*, 2022.

Dongsheng An, Yang Guo, Na Lei, Zhongxuan Luo, Shing-Tung Yau, and Xianfeng Gu. Ae-ot: A new generative model based on extended semi-discrete optimal transport. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

Harry Bateman. Some recent researches on the motion of fluids. *Monthly Weather Review*, 43(4): 163–170, 1915.

Claudio Battiloro, Ege Karaismailoglu, Mauricio Tec, George Dasoulas, Michelle Audirac, and Francesca Dominici. E (n) equivariant topological neural networks. In *International Conference on Learning Representations*, volume 2025, pp. 20267–20309, 2025.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *arXiv preprint arXiv:2101.03164*, 2021.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Guido Boffetta and Robert E Ecke. Two-dimensional turbulence. *Annual review of fluid mechanics*, 44(1):427–451, 2012.

Clément Bonet, Théo Uscidda, Adam David, Pierre-Cyril Aubin-Frankowski, and Anna Korba. Mirror and preconditioned gradient descent in wasserstein space. *Advances in Neural Information Processing Systems*, 37:25311–25374, 2024.

Anastasia Borovykh, Nikolas Kantas, Panos Parpas, and Grigorios A Pavliotis. On stochastic mirror descent with interacting particles: convergence properties and variance reduction. *Physica D: Nonlinear Phenomena*, 418:132844, 2021.

Denis Boyda, Gurtej Kanwar, Sébastien Racanière, Danilo Jimenez Rezende, Michael S Albergo, Kyle Cranmer, Daniel C Hackett, and Phiala E Shanahan. Sampling using su (n) gauge equivariant flows. *Physical Review D*, 103(7):074504, 2021.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44:375–417, 1991.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and trends in Machine Learning*, 8(3-4):231–357, 2015.

Johannes Martinus Burgers. A mathematical model illustrating the theory of turbulence. *Advances in applied mechanics*, 1:171–199, 1948.

Benjamin Capdeville, Young-Heon Kim, and Soumik Pal. Mirror langevin diffusions: Convergence rates and markov chain approximations. *arXiv preprint arXiv:2607.22892*, 2026.

Gary J Chandler and Rich R Kerswell. Invariant recurrent solutions embedded in a turbulent two-dimensional kolmogorov flow. *Journal of Fluid Mechanics*, 722:554–595, 2013.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. Guiding semi-supervision with constraint-driven learning. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 280–287, 2007.

Frédéric Chazal, David Cohen-Steiner, André Lieutier, Quentin Mérigot, and Boris Thibert. Inference of curvature using tubular neighborhoods. *Modern Approaches to Discrete Curvature*, pp. 133–158, 2017.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8185–8194. IEEE, 2020.

Jacob K Christopher, Stephen Baek, and Nando Fioretto. Constrained synthesis with projected diffusion models. *Advances in Neural Information Processing Systems*, 37:89307–89333, 2024.

Liam Connor, Katherine L Bouman, Vikram Ravi, and Gregg Hallinan. Deep radio-interferometric imaging with polish: Dsa-2000 and weak lensing. *Monthly Notices of the Royal Astronomical Society*, 514(2):2614–2626, 2022.

John Crank and Phyllis Nicolson. A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. In *Mathematical proceedings of the Cambridge philosophical society*, volume 43, pp. 50–67. Cambridge University Press, 1947.

Giannis Daras, Yuval Dagan, Alex Dimakis, and Constantinos Daskalakis. Consistent diffusion models: Mitigating sampling drift by learning to be consistent. *Advances in Neural Information Processing Systems*, 36:42038–42063, 2023.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. *Advances in Neural Information Processing Systems*, 35:2406–2422, 2022.

Neel Dey, Antong Chen, and Soheil Ghafurian. Group equivariant generative adversarial networks. *arXiv preprint arXiv:2005.01683*, 2020.

Berthy Feng, Ricardo Baptista, and Katherine Bouman. Neural approximate mirror maps for constrained diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 55155–55178, 2025.

Alessio Figalli. Regularity properties of optimal maps between nonconvex domains in the plane. *Communications in Partial Differential Equations*, 35:465–479, 2010.

Nic Fishman, Leo Klarner, Valentin De Bortoli, Emile Mathieu, and Michael Hutchinson. Diffusion models for constrained domains. *arXiv preprint arXiv:2304.05364*, 2023a.

Nic Fishman, Leo Klarner, Emile Mathieu, Michael Hutchinson, and Valentin De Bortoli. Metropolis sampling for constrained diffusion models. *Advances in Neural Information Processing Systems*, 36:62296–62331, 2023b.

Xingcheng Fu, Yisen Gao, Yuecen Wei, Qingyun Sun, Hao Peng, Jianxin Li, and Xianxian Li. Hyperbolic geometric latent diffusion model for graph generation, 2024. URL `https://arxiv.org/abs/2405.03188`.

Kuzman Ganchev, Joao Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 11:2001–2049, 2010.

G Giannone, L Regenwetter, A Srivastava, D Gutfreund, and F Ahmed. Learning from invalid data: On constraint satisfaction in generative models. arxiv 2023. *arXiv preprint arXiv:2306.15166*, 2023a.

Giorgio Giannone, Lyle Regenwetter, Akash Srivastava, Dan Gutfreund, and Faez Ahmed. Learning from invalid data: On constraint satisfaction in generative models. *arXiv preprint arXiv:2306.15166*, 2023b.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012.

Xianfeng Gu, Feng Luo, Jian Sun, and S-T Yau. Variational principles for minkowski type problems, discrete optimal transport, and discrete monge-ampere equations. *arXiv preprint arXiv:1302.5472*, 2013.

Yunrui Guan, Krishna Balasubramanian, and Shiqian Ma. Mirror flow matching with heavy-tailed priors for generative modeling on convex domains. In *International Conference on Learning Representations*, volume 2026, pp. 130098–130124, 2026.

Taeoh Ha, Hyunsoo Cha, and Daehyun Ji. Cdp: constrained diffusion policies with mirror diffusion model for safety-assured imitation learning. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9838–9845. IEEE, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6629–6640, 2017.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored langevin dynamics. *Advances in Neural Information Processing Systems*, 31, 2018.

Chin-Wei Huang, Milad Aghajohari, Joey Bose, Prakash Panangaden, and Aaron C Courville. Riemannian diffusion models. *Advances in Neural Information Processing Systems*, 35:2750–2761, 2022.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315. IEEE, 2024.

Qijia Jiang. Mirror langevin monte carlo: the case under isoperimetry. *Advances in Neural Information Processing Systems*, 34:715–725, 2021.

Mohammad Reza Karimi Jaghargh, Ya-Ping Hsieh, and Andreas Krause. A dynamical system view of langevin-based non-convex sampling. *Advances in Neural Information Processing Systems*, 36:41051–41075, 2023.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Patrick Kidger. On neural differential equations. *arXiv preprint arXiv:2202.02435*, 2022.

Leon Klein, Andreas Krämer, and Frank Noé. Equivariant flow matching. *Advances in Neural Information Processing Systems*, 36:59886–59910, 2023.

Dmitrii Kochkov, Jamie A Smith, Ayya Alieva, Qing Wang, Michael P Brenner, and Stephan Hoyer. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118(21):e2101784118, 2021.

J Köhler, L Klein, and F Noe. Equivariant flows: Exact likelihood generative learning for symmetric densities. in iii, hd and. In *Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research*, pp. 5361–5370. PMLR, 2020.

Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. *arXiv preprint arXiv:1909.13082*, 2019.

Tim Tsz-Kit Lau, Han Liu, and Thomas Pock. Non-log-concave and nonsmooth sampling via langevin monte carlo algorithms. In *INdAM Workshop: Advanced Techniques in Optimization for Machine learning and Imaging*, pp. 83–149. Springer, 2022.

Ruilin Li, Molei Tao, Santosh S Vempala, and Andre Wibisono. The mirror langevin algorithm converges with vanishing bias. In *International Conference on Algorithmic Learning Theory*, pp. 718–742. PMLR, 2022.

Xinpeng Li, Enming Liang, and Minghua Chen. Gauge flow matching: Efficient constrained generative modeling over general convex set and beyond. In *The Fourteenth International Conference on Learning Representations*, 2026.

Guan-Horng Liu, Tianrong Chen, Evangelos Theodorou, and Molei Tao. Mirror diffusion models for constrained and watermarked generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Aaron Lou and Stefano Ermon. Reflected diffusion models. In *International Conference on Machine Learning*, pp. 22675–22701. PMLR, 2023.

Zhongxuan Luo, Wei Chen, Na Lei, Yang Guo, Tong Zhao, Jiakun Liu, and Xianfeng Gu. The singularity set of optimal transportation maps. *Computational Mathematics and Mathematical Physics*, 62(8):1313–1330, 2022.

Gideon S Mann and Andrew McCallum. Simple, robust, scalable semi-supervised learning via expectation regularization. In *Proceedings of the 24th international conference on Machine learning*, pp. 593–600, 2007.

Rahul Mazumder, Arkopal Choudhury, Garud Iyengar, and Bodhisattva Sen. A computational framework for multivariate convex regression and its variants. *Journal of the American Statistical Association*, 114(525):318–331, 2019.

Quentin Mérigot, Alex Delalande, and Frederic Chazal. Quantitative stability of optimal transport maps and linearization of the 2-wasserstein space. In *International Conference on Artificial Intelligence and Statistics*, pp. 3186–3196. PMLR, 2020.

Laurence Midgley, Vincent Stimper, Javier Antorán, Emile Mathieu, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Se (3) equivariant augmented coupling flows. *Advances in Neural Information Processing Systems*, 36:79200–79225, 2023.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pp. 666–704, 1781.

Dai Hai Nguyen and Tetsuya Sakurai. Mirror variational transport: a particle-based algorithm for distributional optimization on constrained domains. *Machine Learning*, 112(8):2845–2869, 2023.

Chenhao Niu, Yang Song, Jiaming Song, Shengjia Zhao, Aditya Grover, and Stefano Ermon. Permutation invariant graph generation via score-based generative modeling. In *International conference on artificial intelligence and statistics*, pp. 4474–4484. PMLR, 2020.

Alexander C Ogren, Berthy T Feng, Katherine L Bouman, and Chiara Daraio. Gaussian process regression as a surrogate model for the computation of dispersion relations. *Computer Methods in Applied Mechanics and Engineering*, 420:116661, 2024.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3147–3157. IEEE, 2023.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Danilo Jimenez Rezende, Sébastien Racanière, Irina Higgins, and Peter Toth. Equivariant hamiltonian flows. *arXiv preprint arXiv:1909.13739*, 2019.

François Rozet and Gilles Louppe. Score-based data assimilation. *Advances in Neural Information Processing Systems*, 36:40521–40541, 2023.

Victor Garcia Satorras, Emiel Hoogeboom, Fabian Bernd Fuchs, Ingmar Posner, and Max Welling. E (n) equivariant normalizing flows. In *Advances in neural information processing systems*, 2021a.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021b.

Louis Sharrock, Lester Mackey, and Christopher Nemeth. Learning rate free sampling in constrained domains. *Advances in Neural Information Processing Systems*, 36:65380–65415, 2023.

Jiaxin Shi, Chang Liu, and Lester Mackey. Sampling with mirrored stein operators. *arXiv preprint arXiv:2106.12506*, 2021.

Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant flow matching with hybrid probability transport for 3d molecule generation. *Advances in Neural Information Processing Systems*, 36:549–568, 2023.

Lukang Sun and Peter Richtárik. A note on the convergence of mirrored stein variational gradient descent under $(l_0, l_1)-$ smoothness condition. *arXiv preprint arXiv:2206.09709*, 2022.

Jaesung Tae. Mirror diffusion models. *arXiv preprint arXiv:2308.06342*, 2023.

Hong Ye Tan, Subhadip Mukherjee, Junqi Tang, and Carola-Bibiane Schönlieb. Data-driven mirror descent with input-convex neural networks. *SIAM Journal on Mathematics of Data Science*, 5(2): 558–587, 2023.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

Belinda Tzen, Anant Raj, Maxim Raginsky, and Francis Bach. Variational principles for mirror descent and mirror langevin dynamics. *IEEE Control Systems Letters*, 7:1542–1547, 2023.

Rishi Upadhyay, Howard Zhang, Yunhao Ba, Ethan Yang, Blake Gella, Sicheng Jiang, Alex Wong, and Achuta Kadambi. Enhancing diffusion models with 3d perspective geometry constraints. *ACM Transactions on Graphics (TOG)*, 42(6):1–15, 2023.

George N Wong, Ben S Prather, Vedant Dhruv, Benjamin R Ryan, Monika Mościbrodzka, Chi-kwan Chan, Abhishek V Joshi, Ricardo Yarza, Angelo Ricarte, Hotaka Shiokawa, et al. Patoka: simulating electromagnetic observables of black hole accretion. *The Astrophysical Journal Supplement Series*, 259(2):64, 2022.

J Yim, B Trippe, V Bortoli, E Mathieu, A Doucet, R Barzilay, and T Jaakkola. Se (3) diffusion model with application to protein backbone generation. arxiv. doi: 10.48550. *arXiv preprint arXiv.2302.02277*, 2023.

Dingzhi Yu, Wei Jiang, Hongyi Tao, Yuanyu Wan, and Lijun Zhang. Mirror descent under generalized smoothness. *arXiv preprint arXiv:2502.00753*, 2025.

Kelvin Shuangjian Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. Wasserstein control of mirror langevin monte carlo. In *Conference on learning theory*, pp. 3814–3841. PMLR, 2020.

Xingyu Zhou. On the fenchel duality between strong convexity and lipschitz continuous gradient (2018). *arXiv preprint arXiv:1803.06573*, 2018.

Yuanzheng Zhu. Convergence rates for the constrained sampling via langevin monte carlo. *Entropy*, 25(8):1234, 2023.

# Appendix

## A Additional References

**Constrained Generative Models**. Explicitly embedding known prior constraints into generative models yields multiple merits, including improved data efficiency (Ganchev et al., 2010; Batzner et al., 2021), stronger generalization (Köhler et al., 2020), and inherently feasible generated samples (Giannone et al., 2023a). Numerous technical paradigms for constrained generation have been developed, yet each suffers from severe limitations in generalizability and practical deployment. The first line of work leverages equivariant neural networks to model symmetric constraints (Satorras et al., 2021a;b; Battiloro et al., 2025; Thomas et al., 2018; Allingham et al., 2022; Niu et al., 2020; Klein et al., 2023; Song et al., 2023; Boyda et al., 2021; Rezende et al., 2019; Köhler et al., 2020; Midgley et al., 2023; Dey et al., 2020; Hoogeboom et al., 2022; Yim et al., 2023). Nevertheless, such methods are highly task-specific: they cannot generalize to non-symmetric constraints, nor are they compatible with diverse generative architectures (Klein et al., 2023; Song et al., 2023; Boyda et al., 2021; Rezende et al., 2019; Satorras et al., 2021a; Köhler et al., 2020; Midgley et al., 2023; Niu et al., 2020; Hoogeboom et al., 2022; Yim et al., 2023; Dey et al., 2020; Allingham et al., 2022). Second, existing constrained diffusion algorithms (De Bortoli et al., 2022; Huang et al., 2022; Fishman et al., 2023a; Lou & Ermon, 2023; Tae, 2023; Liu et al., 2024; Fu et al., 2024) impose rigid mathematical prerequisites on target constraints. Some require the constraint space to form a Riemannian manifold (De Bortoli et al., 2022; Huang et al., 2022; Fu et al., 2024); others rely on well-defined reflection operators (Lou & Ermon, 2023) or projection operators (Christopher et al., 2024) for constraint projection correction. Many further restrict applicability to convex feasible sets (Fishman et al., 2023a; Tae, 2023; Liu et al., 2024), with ICNN-based approximation proposed as a makeshift workaround (Feng et al., 2025). Fishman et al. incorporate Metropolis-Hastings sampling to accommodate arbitrary constraints (Fishman et al., 2023b). However, this approach incurs prohibitively high sample rejection rates for hard-to-satisfy constraints such as PDEs, drastically degrading practical usability. Third, a widely adopted strategy introduces soft constraint penalties during training (Ganchev et al., 2010; Mann & McCallum, 2007; Chang et al., 2007; Giannone et al., 2023b; Daras et al., 2023; Upadhyay et al., 2023). Repeatedly computing constraint loss for all generated samples across training iterations induces prohibitive computational overhead. Although supplementing training with constraint-violating negative samples has been proposed (Giannone et al., 2023b), high-quality discriminative negative samples are typically difficult to acquire.

## B Our Main Theoretical Contribution

### B.1 Regularity Analysis for Inverse Mirror Maps

This appendix proves that the inverse mirror map associated with a valid mirror potential satisfies all regularity prerequisites for Brenier maps under quadratic transport cost.

Let $\mathcal{C} \subset \mathbb{R}^d$ be a closed convex feasible set. A valid mirror potential $\phi : \mathcal{C} \to \mathbb{R}$ satisfies (Liu et al., 2024): (1) $\phi$ is strongly convex and $C^1$ over $\mathcal{C}$; (2) $\nabla\phi : \mathcal{C} \to \mathbb{R}^d$ is a bijective mapping which is strongly monotone and Lipschitz continuous. Since the gradient of a strongly convex differentiable function is bijective, its inverse is well-defined:

$$S = (\nabla\phi)^{-1} : \mathbb{R}^d \to \mathcal{C}. \tag{6}$$

Let source measure $\mu$ on $\mathbb{R}^d$ and target measure $\nu$ on $\mathcal{C}$ be absolutely continuous w.r.t. the Lebesgue measure. A measurable map $M : \mathbb{R}^d \to \mathcal{C}$ is a Brenier (optimal transport) map for cost $c(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$ if (Brenier, 1991): (1) There exists a convex potential $u : \mathbb{R}^d \to \mathbb{R}$ such that $M = \nabla u$; (2) $M$ is single-valued almost everywhere and strictly monotone. Under these conditions $M$ pushes forward $\mu$ to $\nu = M_{\#}\mu$.

Define the Fenchel conjugate of mirror potential $\phi$:

$$u(\boldsymbol{x}) = \sup_{\boldsymbol{y} \in \mathcal{C}} \big\{ \langle \boldsymbol{x}, \boldsymbol{y} \rangle - \phi(\boldsymbol{y}) \big\}. \tag{7}$$

Because $\mathcal{C}$ is closed convex and $\phi$ is strongly convex and $C^1$ on $\mathcal{C}$, convex analysis gives: (1) The conjugate $u$ is convex and $C^1$ over the whole space $\mathbb{R}^d$; (2) Gradient inversion identity holds:

$$\nabla u(\boldsymbol{x}) = (\nabla\phi)^{-1}(\boldsymbol{x}) = S(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathbb{R}^d. \tag{8}$$

Hence the inverse mirror map admits the gradient form $S = \nabla u$ with convex potential $u$ defined on source domain $\mathbb{R}^d$, which verifies condition (1) of Brenier's theorem (Brenier, 1991).

Since $\nabla\phi$ is bijective, its inverse map $S$ is globally uniquely defined over $\mathbb{R}^d$. The inverse map of a strongly monotone bijective map inherits strong monotonicity. Therefore $S$ is strongly monotone, satisfying condition (2) of Brenier's theorem (Brenier, 1991). The Lipschitz property of $\nabla\phi$ transfers to $S$. Lipschitz maps are measurable; the push-forward measure $\nu = S_{\#}\mu$ is well-defined and supported within $\mathcal{C}$.

For any valid mirror potential $\phi$ on closed convex domain $\mathcal{C}$, its inverse mirror map $S = (\nabla\phi)^{-1}$ fulfills all regularity requirements of Brenier maps under quadratic cost. Every inverse mirror lifting sampling operator is therefore a special Brenier transport map. The strict inclusion holds:

$$\big\{S = (\nabla\phi)^{-1} \,\big|\, \text{Valid Mirror Potential } \phi\big\} \subsetneq \big\{M_{ot} = \nabla u \,\big|\, \text{Brenier Potential } u\big\}. \tag{9}$$

General Brenier maps do not require convex target domains or analytic mirror potentials, so the inclusion is proper.

**Remark 1.** *We clarify a potentially confusing existing result from (Zhou, 2018; Tan et al., 2023): the inverse mirror map $S = (\nabla\phi)^{-1}$ is not guaranteed to be the gradient of a strongly convex function. This is consistent with our analysis. Fenchel conjugacy only guarantees convexity of the conjugate potential $u$, yet strong convexity of the original mirror potential $\phi$ does not transfer to $u$. Importantly, Brenier's theorem imposes only convexity rather than strong convexity on the transport potential. Hence, $S$ remains a valid Brenier map even when the associated potential $u$ lacks strong convexity.*

### B.2 Fenchel Conjugate of the Log-Sum-Exp Smoothed Brenier Potential

This section derives the Fenchel-Legendre dual potential of the Log-Sum-Exp (LSE) smoothed convex Brenier potential, with all intermediate algebraic steps included. Let $\mathcal{I}$ denote the index set of discrete anchor points, $|\mathcal{I}|$ the total number of anchors, $\{\boldsymbol{y}_i^\circ\}_{i\in\mathcal{I}}$ the discrete anchor points within the constrained domain, $h_i$ the offset constants for Power-cell partitioning, and $\tau > 0$ the smoothing temperature parameter.

The LSE-regularized smoothed Brenier potential is defined as

$$u_\tau(\boldsymbol{x}) = \tau \log\left(\sum_{i\in\mathcal{I}} \exp\left(\frac{\langle \boldsymbol{x}, \boldsymbol{y}_i^\circ\rangle + h_i}{\tau}\right)\right) - \tau \log|\mathcal{I}|. \tag{10}$$

We introduce the shorthand LSE function

$$F(\boldsymbol{x}) \triangleq \tau \log \sum_{i\in\mathcal{I}} \exp\left(\frac{\langle \boldsymbol{x}, \boldsymbol{y}_i^\circ\rangle + h_i}{\tau}\right), \tag{11}$$

so that $u_\tau(\boldsymbol{x}) = F(\boldsymbol{x}) - \tau\log|\mathcal{I}|$. The term $-\tau\log|\mathcal{I}|$ is a constant normalization offset and does not affect convexity or gradient computations.

Following the standard definition of Fenchel-Legendre conjugation in convex analysis, the dual conjugate potential of $u_\tau$ reads

$$v_\tau(\boldsymbol{y}) = \sup_{\boldsymbol{x}\in\mathbb{R}^d} \Big\{\langle \boldsymbol{x}, \boldsymbol{y}\rangle - u_\tau(\boldsymbol{x})\Big\}. \tag{12}$$

In optimal transport, the forward and dual Brenier maps correspond to gradients of the primal and dual potentials: $\boldsymbol{T}_\tau(\boldsymbol{x}) = \nabla u_\tau(\boldsymbol{x})$ and $\boldsymbol{S}_\tau(\boldsymbol{y}) = \nabla v_\tau(\boldsymbol{y})$.

Substitute $u_\tau(\boldsymbol{x}) = F(\boldsymbol{x}) - \tau\log|\mathcal{I}|$ into equation 12. Since $\tau\log|\mathcal{I}|$ is independent of $\boldsymbol{x}$, it can be pulled outside the supremum:

$$v_\tau(\boldsymbol{y}) = \sup_{\boldsymbol{x}\in\mathbb{R}^d} \Big\{\langle \boldsymbol{x}, \boldsymbol{y}\rangle - \big(F(\boldsymbol{x}) - \tau\log|\mathcal{I}|\big)\Big\} = \tau\log|\mathcal{I}| + \sup_{\boldsymbol{x}\in\mathbb{R}^d} \Big\{\langle \boldsymbol{x}, \boldsymbol{y}\rangle - F(\boldsymbol{x})\Big\}. \tag{13}$$

Define the objective inside the supremum as $g_{\boldsymbol{y}}(\boldsymbol{x}) \triangleq \langle \boldsymbol{x}, \boldsymbol{y} \rangle - F(\boldsymbol{x})$. Because $F(\boldsymbol{x})$ is strictly convex and differentiable, $g_{\boldsymbol{y}}(\boldsymbol{x})$ is strictly concave; its unique global maximum satisfies the first-order optimality condition $\nabla_{\boldsymbol{x}} g_{\boldsymbol{y}}(\boldsymbol{x}) = \mathbf{0}$:

$$\nabla_{\boldsymbol{x}} g_{\boldsymbol{y}}(\boldsymbol{x}) = \boldsymbol{y} - \nabla F(\boldsymbol{x}) = \mathbf{0} \implies \boldsymbol{y} = \nabla F(\boldsymbol{x}). \tag{14}$$

Compute the gradient of the LSE function $F(\boldsymbol{x})$ via chain-rule differentiation:

$$\nabla F(\boldsymbol{x}) = \tau \cdot \frac{\sum_{i\in\mathcal{I}} \exp\left(\frac{\langle \boldsymbol{x}, \boldsymbol{y}_i^\circ \rangle + h_i}{\tau}\right) \cdot \frac{\boldsymbol{y}_i^\circ}{\tau}}{\sum_{j\in\mathcal{I}} \exp\left(\frac{\langle \boldsymbol{x}, \boldsymbol{y}_j^\circ \rangle + h_j}{\tau}\right)} = \sum_{i\in\mathcal{I}} \lambda_i(\boldsymbol{x}) \boldsymbol{y}_i^\circ,\ \lambda_i(\boldsymbol{x}) = \frac{\exp\left(\frac{\langle \boldsymbol{x}, \boldsymbol{y}_i^\circ \rangle + h_i}{\tau}\right)}{\sum_{j\in\mathcal{I}} \exp\left(\frac{\langle \boldsymbol{x}, \boldsymbol{y}_j^\circ \rangle + h_j}{\tau}\right)}. \tag{15}$$

The weights satisfy $\lambda_i(\boldsymbol{x}) \geq 0$ and $\sum_{i\in\mathcal{I}} \lambda_i(\boldsymbol{x}) = 1$. From equation 14, $\boldsymbol{y} = \sum_{i\in\mathcal{I}} \lambda_i(\boldsymbol{x}) \boldsymbol{y}_i^\circ$, which implies $\boldsymbol{y}$ must lie in the convex hull $\mathbf{conv}\{\boldsymbol{y}_i^\circ\}$ of anchor points.

We next eliminate the primal variable $\boldsymbol{x}$ and rewrite the supremum in terms of probability weights $\lambda_i$. Taking logarithm of $\lambda_i(\boldsymbol{x})$ and multiplying by $\tau$ yields

$$\tau \log \lambda_i(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{y}_i^\circ \rangle + h_i - \tau \log \sum_{j\in\mathcal{I}} \exp\left(\frac{\langle \boldsymbol{x}, \boldsymbol{y}_j^\circ \rangle + h_j}{\tau}\right). \tag{16}$$

Using the definition of $F(\boldsymbol{x})$, we obtain the key identity

$$\langle \boldsymbol{x}, \boldsymbol{y}_i^\circ \rangle = \tau \log \lambda_i(\boldsymbol{x}) - h_i + F(\boldsymbol{x}). \tag{17}$$

Let $\boldsymbol{x}^*$ denote the maximizer and set $\lambda_i^* = \lambda_i(\boldsymbol{x}^*)$. At optimality $\boldsymbol{y} = \sum_{i\in\mathcal{I}} \lambda_i^* \boldsymbol{y}_i^\circ$. Expand the inner product:

$$\begin{aligned} \langle \boldsymbol{x}^*, \boldsymbol{y} \rangle &= \langle \boldsymbol{x}^*, \sum_{i\in\mathcal{I}} \lambda_i^* \boldsymbol{y}_i^\circ \rangle = \sum_{i\in\mathcal{I}} \lambda_i^* \langle \boldsymbol{x}^*, \boldsymbol{y}_i^\circ \rangle \\ &= \sum_{i\in\mathcal{I}} \lambda_i^* [\tau \log \lambda_i^* - h_i + F(\boldsymbol{x}^*)] \\ &= \sum_{i\in\mathcal{I}} \lambda_i^* (-h_i) + \tau \sum_{i\in\mathcal{I}} \lambda_i^* \log \lambda_i^* + F(\boldsymbol{x}^*). \end{aligned} \tag{18}$$

Subtract $F(\boldsymbol{x}^*)$ to evaluate the objective inside the supremum:

$$\langle \boldsymbol{x}^*, \boldsymbol{y} \rangle - F(\boldsymbol{x}^*) = \sum_{i\in\mathcal{I}} \lambda_i^* (-h_i - \tau \log \lambda_i^*). \tag{19}$$

The variable $\boldsymbol{x}^*$ is fully eliminated, converting maximization over $\boldsymbol{x} \in \mathbb{R}^d$ into maximization over probability weights $\{\lambda_i\}$.

We split into two cases. If $\boldsymbol{y} \in \mathbf{conv}\{\boldsymbol{y}_i^\circ\}$, the optimality condition is attainable and the supremum runs over the probability simplex $\Delta = \{\{\lambda_i\} \mid \lambda_i \geq 0, \sum_{i\in\mathcal{I}} \lambda_i = 1,\ \sum_{i\in\mathcal{I}} \lambda_i \boldsymbol{y}_i^\circ = \boldsymbol{y}\}$. If $\boldsymbol{y} \notin \mathbf{conv}\{\boldsymbol{y}_i^\circ\}$, no $\boldsymbol{x}$ satisfies equation 14 and the objective is unbounded above. Substituting back into equation 13, the dual potential becomes

$$v_\tau(\boldsymbol{y}) = \begin{cases} \tau \log |\mathcal{I}| + \sup\limits_{\lambda\in\Delta,\, \sum_{i\in\mathcal{I}} \lambda_i \boldsymbol{y}_i^\circ = \boldsymbol{y}} \sum\limits_{i\in\mathcal{I}} \lambda_i \big( - h_i - \tau \log \lambda_i \big), & \boldsymbol{y} \in \mathbf{conv}\{\boldsymbol{y}_i^\circ\}, \\ +\infty, & \boldsymbol{y} \notin \mathbf{conv}\{\boldsymbol{y}_i^\circ\}. \end{cases} \tag{20}$$

### B.3 Convexity Constraint of Mirror Maps

Theorem 1 theoretically exposes inherent drawbacks of the conventional mirror map paradigm for non-convex constraint domains $\mathcal{C}$. Both analytical mirror maps $S$ induced by globally defined convex dual mirror potentials on $\mathbb{R}^d$ and their ICNN-based neural approximations fail to fit a non-convex $\mathcal{C}$ perfectly, falling into two exclusive failure cases: (1) portions of $\mathcal{C}$ lie outside the image of $S$ and cannot be covered; (2) there exists a source-sample set with non-zero Lebesgue measure mapped outside $\mathcal{C}$. Constrained by the convexity of $S$'s image set, such maps cannot discard spurious interior regions and must approximate target boundaries via global metric stretching and deformation, which inherently produces geometric distortion.

It is important to clarified that the smoothed Brenier potential $u_\tau$ proposed in this work is still convex, so it cannot fundamentally overcome the inherent drawbacks of convex-potential gradient maps on non-convex domains. Even so, our approach supports explicit control over out of domain geometric distortion, which marks a critical difference from conventional mirror map paradigms.

The unsmoothed semi-discrete Brenier map is piecewise linear and exhibits gradient jumps across singular set $\Sigma$ at power-diagram cell boundaries. These discontinuous singularities enable precise alignment with non-convex constraints, yet discontinuity degrades model generalization. Our controllable smoothing does not eliminate such singular structures; it only replaces sharp gradient jumps with smooth transition bands whose width is governed by the hyperparameter $\tau$. Larger jumps on $\Sigma$ yield smaller relative distortion after smoothing. Moreover, as $\tau \to 0^+$, the smoothed map converges pointwise to its semi-discrete counterpart, nearly preserving the original geometric configuration.

**Theorem 1.** *Assuming that $\mathcal{C}$ is a non-convex constraint domain, then for any differentiable strictly convex function $u$ defined on $\mathbb{R}^d$, there must be a set $\Gamma \subset \mathbb{R}^d$ whose Lebesgue measure is non-zero such that $\nabla u(\Gamma) \cap \mathcal{C} = \emptyset$ holds or $\Lambda \subset \mathcal{C}$ such that $\Lambda \cap \nabla u\left(\mathbb{R}^d\right) = \emptyset$ is true.*

*Proof.* Let $T = \nabla u \colon \mathbb{R}^d \to \mathbb{R}^d$ and denote its image set $G = T(\mathbb{R}^d)$. From Fenchel-conjugacy theory for convex functions, since $u$ is differentiable and strictly convex over the whole $\mathbb{R}^d$, the image $G = \nabla u(\mathbb{R}^d)$ must be a convex set. We briefly justify this fact. Define the Fenchel conjugate:

$$u^*(\boldsymbol{y}) = \sup_{\boldsymbol{x} \in \mathbb{R}^d} \left\{ \langle \boldsymbol{x}, \boldsymbol{y} \rangle - u(\boldsymbol{x}) \right\}. \tag{21}$$

The effective domain $\mathbf{dom}(u^*) = \{\boldsymbol{y} \mid u^*(\boldsymbol{y}) < +\infty\}$ of a convex function $u^*$ is always convex. For everywhere-differentiable convex $u \colon \mathbb{R}^d \to \mathbb{R}$, we have the identity

$$\nabla u(\mathbb{R}^d) = \mathbf{dom}(u^*). \tag{22}$$

First, for any $\boldsymbol{y} = \nabla u(\boldsymbol{x})$, convexity gives $u(\boldsymbol{z}) \geq u(\boldsymbol{x}) + \langle \boldsymbol{y}, \boldsymbol{z} - \boldsymbol{x} \rangle$ for all $\boldsymbol{z} \in \mathbb{R}^d$, which yields $u^*(\boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle - u(\boldsymbol{x}) < +\infty$, hence $\boldsymbol{y} \in \mathbf{dom}(u^*)$. Conversely, take $\boldsymbol{y} \in \mathbf{dom}(u^*)$. Since $u$ is differentiable convex over $\mathbb{R}^d$, the strictly concave function $\langle \boldsymbol{z}, \boldsymbol{y} \rangle - u(\boldsymbol{z})$ attains its maximum at some unique $\boldsymbol{x}^*$. The first-order optimality condition gives $\boldsymbol{y} = \nabla u(\boldsymbol{x}^*)$, so $\boldsymbol{y} \in \nabla u(\mathbb{R}^d)$. As a result, $G = \nabla u(\mathbb{R}^d) = \mathbf{dom}(u^*)$ is convex.

Then we consider two mutually exclusive cases:

**Case 1:** $\mathcal{C} \not\subset G$**.** Define $\Lambda = \mathcal{C} \setminus G$. Then $\Lambda \subset \mathcal{C}$ and $\Lambda \cap \nabla u(\mathbb{R}^d) = \emptyset$, which satisfies the second condition of the theorem.

**Case 2:** $\mathcal{C} \subset G$**.** Recall that $\mathcal{C}$ is non-convex while $G$ is convex. There exist two points $\boldsymbol{c}_1, \boldsymbol{c}_2 \in \mathcal{C}$ such that some point $\boldsymbol{c}_0$ on the segment $[\boldsymbol{c}_1, \boldsymbol{c}_2]$ satisfies $\boldsymbol{c}_0 \notin \mathcal{C}$. Since $G$ is convex and contains $\boldsymbol{c}_1, \boldsymbol{c}_2$, the full segment satisfies $[\boldsymbol{c}_1, \boldsymbol{c}_2] \subset G$. Because $u$ is strictly convex and differentiable on $\mathbb{R}^d$, its gradient $T = \nabla u$ defines a global diffeomorphism between $\mathbb{R}^d$ and $\mathrm{int}(G)$. The pre-image $\Gamma = T^{-1}\big([\boldsymbol{c}_1, \boldsymbol{c}_2]\big) \subset \mathbb{R}^d$ has strictly positive Lebesgue measure. Since $\boldsymbol{c}_0 \in T(\Gamma)$ and $\boldsymbol{c}_0 \notin \mathcal{C}$, there exists a subset $\Gamma' \subset \Gamma$ of non-zero measure satisfying $\nabla u(\Gamma') \cap \mathcal{C} = \emptyset$. This fulfills the first condition.

In both cases the statement holds. □

# C Experiment

## C.1 Analytical Mirror Function on $\ell_2$-ball, Simplex and Polytope

**$\ell_2$-Ball.** Let $\mathcal{C} := \{\boldsymbol{y} \in \mathbb{R}^d : \|\boldsymbol{y}\|_2^2 < R^2\}$ denote the $\ell_2$-ball of radius $R$ in $\mathbb{R}^d$. The mirror function is explicitly defined as:

$$\phi_{ball}(\boldsymbol{y}) = -\gamma \log\left(R^2 - \|\boldsymbol{y}\|_2^2\right), \tag{23}$$

where $\gamma \in \mathbb{R}^+$. The mirror map and its inverse can be computed analytically by

$$\nabla \phi_{ball}(\boldsymbol{y}) = \frac{2\gamma \boldsymbol{y}}{R^2 - \|\boldsymbol{y}\|_2^2}, \; \nabla \phi^*_{ball}(\boldsymbol{x}) = \frac{R^2 \boldsymbol{x}}{\sqrt{R^2 \|\boldsymbol{x}\|_2^2 + \gamma^2} + \gamma}. \tag{24}$$

where $\phi^*_{ball}$ is the Fenchel conjugate of $\phi_{ball}$.

**Simplex.** Given a $(d+1)$-dimensional simplex $\Delta_{d+1} = \{\boldsymbol{y} \in \mathbb{R}^{d+1} \mid \sum_{i=1}^{d+1} y_i = 1, y_i \geq 0\}$, we constructing $\mathcal{C} = \{\boldsymbol{y} \in \mathbb{R}^d \mid \sum_{i=1}^{d} y_i \leq 1, y_i \geq 0\}$ and define $y_{d+1} = 1 - \sum_{i=1}^{d} y_i$. The mirror function is explicitly defined as:

$$\phi_{\textit{simplex}}(\boldsymbol{y}) = \sum_{i=1}^{d} y_i \log(y_i) + \left(1 - \sum_{i=1}^{d} y_i\right) \log\left(1 - \sum_{i=1}^{d} y_i\right). \tag{25}$$

The mirror map and its inverse can be computed analytically by

$$\left[\nabla \phi_{\textit{simplex}}(\boldsymbol{y})\right]_i = \log y_i - \log\left(1 - \sum_{i=1}^{d} y_i\right), \quad \left[\nabla \phi^*_{\textit{simplex}}(\boldsymbol{x})\right]_i = \frac{e^{x_i}}{1 + \sum_{i=1}^{d} e^{x_i}}, \tag{26}$$

where $\phi^*_{\textit{simplex}}$ is the Fenchel conjugate of $\phi_{\textit{simplex}}$.

**Polytope.** Let $\mathcal{C} = \{\boldsymbol{y} \in \mathbb{R}^d \mid c_i < \langle \boldsymbol{a}_i, \boldsymbol{y} \rangle < b_i, \forall i \in \{1, \cdots, m\}\}$, where $c_i, b_i \in \mathbb{R}$ and $\boldsymbol{a}_i \in \mathbb{R}^d$ are linearly independent to each other. The mirror function is explicitly defined as:

$$\phi_{\textit{polytope}}(\boldsymbol{y}) := -\sum_{i=1}^{m} \big(\log(\langle \boldsymbol{a}_i, \boldsymbol{y} \rangle - c_i) + \log(b_i - \langle \boldsymbol{a}_i, \boldsymbol{y} \rangle)\big) + \sum_{j=m+1}^{d} \frac{1}{2} \langle \boldsymbol{a}_j, \boldsymbol{y} \rangle^2. \tag{27}$$

The mirror map and its inverse can be computed analytically by

$$\begin{cases} \nabla \phi_{\textit{polytope}}(\boldsymbol{y}) = \sum_{i=1}^{m} s_i(\langle \boldsymbol{a}_i, \boldsymbol{y} \rangle) \boldsymbol{a}_i + \sum_{j=m+1}^{d} \langle \boldsymbol{a}_j, \boldsymbol{y} \rangle \boldsymbol{a}_j, \\ \nabla \phi^*_{\textit{polytope}}(\boldsymbol{x}) = \sum_{i=1}^{m} s_i^{-1}(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle) a_i + \sum_{j=m+1}^{d} \langle \boldsymbol{a}_j, \boldsymbol{x} \rangle \boldsymbol{a}_j, \end{cases} \tag{28}$$

where $s_i = \frac{-1}{\langle \boldsymbol{a}_i, \boldsymbol{y} \rangle - c_i} + \frac{-1}{\langle \boldsymbol{a}_i, \boldsymbol{y} \rangle - b_i}$ and $s_i^{-1}$ is its inverse, $\phi^*_{\textit{polytope}}$ is the Fenchel conjugate of $\phi_{\textit{polytope}}$.

### C.2 Neural Approximate Mirror Maps

Neural Approximate Mirror Maps (NAMM) (Feng et al., 2025) use neural networks to fit the forward mirror map $g_\phi : \mathcal{C} \to \mathbb{R}^d$ and inverse mirror map $f_\psi : \mathbb{R}^d \to \mathcal{C}$. This framework circumvents the limitation of analytical mirror maps that only support convex constraints, and can be applied to arbitrary differentiable constraint sets $\mathcal{C}$. The overall optimization objective is formulated as

$$\min_{\phi, \psi} \mathcal{L}_{\textit{cycle}} + \lambda_{\textit{constr}} \mathcal{L}_{\textit{constr}} + \lambda_{\textit{reg}} \mathcal{R}(g_\phi). \tag{29}$$

**Noise-Augmented Distribution for Robust Training.** To accommodate noisy perturbations from GM sampling, we consider noise-corrupted pushforward distributions over a range of noise scales:

$$(g_\phi)_{\#} p_{\textit{data}} * \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}), \quad \sigma \in [0, \sigma_{\textit{max}}]. \tag{30}$$

**Loss Functions.** The cycle-consistency loss enforces bijective invertibility for both clean samples and noisy samples across all noise magnitudes:

$$\mathcal{L}_{\textit{cycle}} = \mathbb{E}_{\boldsymbol{y} \sim p_{\textit{data}}} \Big[ \|\boldsymbol{y} - f_\psi(g_\phi(\boldsymbol{y}))\|_1 + \int_0^{\sigma_{\textit{max}}} \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})} \big\| g_\phi(\boldsymbol{y}) + \sigma \boldsymbol{z} - g_\phi(f_\psi(g_\phi(\boldsymbol{y}) + \sigma \boldsymbol{z})) \big\|_1 d\sigma \Big]. \tag{31}$$

The constraint loss leverages a differentiable constraint distance function $\ell_{\textit{constr}}$ to force outputs of the inverse map to lie within the constraint manifold:

$$\mathcal{L}_{\textit{constr}} = \mathbb{E}_{\boldsymbol{y} \sim p_{\textit{data}}} \Big[ \int_0^{\sigma_{\textit{max}}} \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})} \, \ell_{\textit{constr}}\big(f_\psi(g_\phi(\boldsymbol{y}) + \sigma \boldsymbol{z})\big) d\sigma \Big]. \tag{32}$$

The regularization term penalizes deviation of the forward map from the identity function to guarantee solution uniqueness and geometric stability:

$$\mathcal{R}(g_\phi) = \mathbb{E}_{\boldsymbol{y} \sim p_{\textit{data}}} \|\boldsymbol{y} - g_\phi(\boldsymbol{y})\|_1. \tag{33}$$

**Network Architectures.** The forward map $g_\phi$ is parameterized as the gradient of an ICNN, satisfying the theoretical requirement for convex mirror maps. The inverse map $f_\psi$ is implemented with residual networks (MLP for point clouds) to reconstruct points on the constraint domain.

**Fine-Tuning of Inverse Map.** To mitigate distribution shift between pre-trained maps and samples from the GM, the authors fine-tune $f_\psi$ using generated samples $\hat{\boldsymbol{x}} \sim \widetilde{p}_{\boldsymbol{\theta}}$ in the dual space:

$$\begin{aligned}&\min_{\psi} \mathbb{E}_{\boldsymbol{y}\sim p_{data}} \|\boldsymbol{y} - f_\psi(g_\phi(\boldsymbol{y}))\|_1 \\ &+\mathbb{E}_{\hat{\boldsymbol{x}}\sim\widetilde{p}_{\boldsymbol{\theta}}} \int_0^{\sigma_{max}} \mathbb{E}_{\boldsymbol{z}}\Big[\big\|\hat{\boldsymbol{x}} + \sigma\boldsymbol{z} - g_\phi(f_\psi(\hat{\boldsymbol{x}} + \sigma\boldsymbol{z}))\big\|_1 + \lambda_{constr}\ell_{constr}(f_\psi(\hat{\boldsymbol{x}} + \sigma\boldsymbol{z}))\Big] d\sigma.\end{aligned} \tag{34}$$

Ultimately, the learned approximate mirror map and its inverse can be expressed as the gradients of convex potential functions, i.e., $g_\phi = \nabla\phi_{icnn}$, $f_\psi = \nabla\psi_{icnn}$.

### C.3 The Computation Process of the Inverse Brenier Map

**Algorithm 2** The Computation Process of the Inverse Brenier Map

**Input**: $\boldsymbol{y}$, Brenier map $\boldsymbol{T}_\tau(\cdot)$, learning rate $lr$, maximum iterations $T$, integer $s > 0$ and termination thresholds $\{\kappa_1, \kappa_2\}$, smoothness coefficient $\tau > 0$.

1: $\boldsymbol{x}^{(0)} \leftarrow R_{\mathcal{H}} \cdot \boldsymbol{y}$.
2: **for** $t = 0 : T-1$ **do**
3: $\quad\mathcal{L}(\boldsymbol{x}^{(t)}) \leftarrow \|\boldsymbol{T}_\tau(\boldsymbol{x}^{(t)}) - \boldsymbol{y}\|_2^2$.
4: $\quad\boldsymbol{x}^{(t+1)} \leftarrow \boldsymbol{x}^{(t)} - lr \cdot \nabla\mathcal{L}(\boldsymbol{x}^{(t)})$.
5: $\quad lr \leftarrow 0.9 \cdot lr$ if $\|\nabla\hat{E}(\boldsymbol{h}^{(t)})\|_2$ has not decreased for $s$ steps.
6: $\quad$break if $\|\nabla\mathcal{L}(\boldsymbol{x}^{(t)})\|_2 < \kappa_1$ or $lr < \kappa_2$.
7: **end for**

**Return**: Dual Brenier map $\boldsymbol{S}_\tau(\cdot)$.

### C.4 Experimental Parameter Settings for Map Approximation Error Evaluation

For the three classical convex domains adopted in the experiments of Section 4.1, the parameters are specified as follows. For the $\ell_2$-ball in $d$-dimensional space, we set $R \equiv 1.0$ and $\gamma \equiv 0.5$. For the $d$-dimensional polytope, we set $m = d$, $c_i \equiv -0.8$, $b_i \equiv 0.8$, where $\boldsymbol{a}_i = (0, \cdots, \underbrace{1}_{i\text{-th}}, \cdots, 0)^\top$ denotes the standard basis vector. Moreover, we set the point cloud dimension $d$ to 2, 3 or 5 on a single RTX 4090. The hidden dimension of the forward map network $g_\phi$ (ICNN) and inverse map network $f_\psi$ (residual MLP) is 128. The maximum noise scale $\sigma_{max}$ is set to 1.2, the constraint loss weight $\lambda_{constr} = 8.0$, and the regularization weight $\lambda_{reg} = 1.5$. We use a batch size of 256 and the Adam optimizer with an initial learning rate of $1 \times 10^{-3}$, accompanied by a cosine annealing scheduler. For anchor count $N_1 = 50000$, we train for 3000000 iterations in total; for $N_1 = 40000$, we use 2000000 iterations; all other settings use 1000000 iterations. To evaluate the approximation quality of NAMM with respect to the analytical mirror map $\nabla\phi$, we replace the original regularization term $\mathcal{R}(g_\phi)$ with $\mathbb{E}_{\boldsymbol{y}\sim p_{data}}\|g_\phi(\boldsymbol{y}) - \nabla\phi(\boldsymbol{y})\|_1$.

### C.5 Experimental Parameter Settings for Watermarked Image Generation

For the experiments presented in Section 4.2, a convex polytope $\mathcal{C} = \{\boldsymbol{y} \mid c_k < \langle \boldsymbol{a}_k, \boldsymbol{y}\rangle < b_k, k \in \mathcal{K}\}$ is adopted as the constraint domain for watermarked image generation, with parameters configured as follows. The total number of constraints is uniformly set to $|\mathcal{K}| = 100$. For the FFHQ dataset, the lower bound $c_k$ and upper bound $b_k$ of each constraint are fixed at $-1.05$ and $1.05$, respectively; for the AFHQ dataset, the corresponding bounds are set to $-0.9$ and $0.9$. The watermark direction vectors $\boldsymbol{a}_k$ are mutually orthogonal. Specifically, we first generate a standard Gaussian random matrix, and then derive 100 unit orthogonal directions via modified Gram-Schmidt orthogonalization. The identical direction matrix is shared by both datasets. Once generated, the direction matrix is frozen and applied consistently across target construction, generation mapping, and watermark detection, rather than being resampled for each individual image.

### C.6 Hyperparameter of NAMM

The input dimensions of NAMM are $3 \times 64 \times 64$ for watermarked AFHQ and FFHQ, $1 \times 64 \times 64$ for Periodic, Burgers, and Total Brightness, $1 \times 128 \times 128$ for Galaxies, and $2 \times 128 \times 256$ for Divergence-

free. The base channel counts for the forward and inverse mappings are set to 32 and 64 across all datasets, respectively. Watermarked AFHQ and FFHQ, Periodic, Burgers, and Total Brightness use 3-layer mapping networks, while Galaxies and Divergence-free use 2-layer architectures. We set $\lambda_{constr} = 1.0$ for watermarked AFHQ and FFHQ, Periodic, Burgers, and Divergence-free, and $\lambda_{constr} = 0.01$ for Total Brightness and Galaxies. Bounded random perturbations are applied to mirror inputs during training: the maximum noise scale is $\sigma_{max} = 0.1$ for Periodic, Burgers, Total Brightness, and Galaxies, and $\sigma_{max} = 0.5$ for Divergence-free, watermarked AFHQ and FFHQ. All NAMM models are trained from random initialization, with joint optimization of all parameters in both forward and inverse mappings. The learning rate is $2 \times 10^{-4}$ for watermarked AFHQ and FFHQ, Periodic, Burgers, Total Brightness, and Galaxies, and $1 \times 10^{-3}$ for Divergence-free; the batch size is 16 for all datasets. The Sparse ICNN has 3 layers for watermarked AFHQ and FFHQ, Periodic, Burgers, and Total Brightness, and 2 layers for Galaxies and Divergence-free. After base NAMM training converges, the inverse mapping is further fine-tuned using samples generated by the mirror-space GM to adapt to the generated mirror distribution, while preserving cycle reconstruction and constraint losses.

### C.7 The architecture of the generative model and the training hyperparameters for Watermark image generation

**Diffusion Model.** Experiments on both the FFHQ $64 \times 64$ and AFHQv2 $64 \times 64$ datasets employ a unified class-unconditional EDM GM built on a DDPM++-compatible SongUNet backbone. The model takes $3 \times 64 \times 64$ inverse-point tensors as input and outputs corresponding denoising estimates. It has a base channel count of 128 with channel multipliers $[1, 2, 2, 2]$, and features resolution stages at $64^2/128$, $32^2/256$, $16^2/256$, and $8^2/256$, with a symmetric upsampling structure on the decoder side. The encoder consists of 4 main residual blocks and 1 downsampling block per resolution stage; the decoder comprises 5 main residual blocks, 1 upsampling block, and 2 bottleneck inlet blocks per stage, with encoder skip connections integrated. Single-head attention is deployed across the 4 blocks of the $16 \times 16$ encoder, the final block of the $16 \times 16$ decoder, and the bottleneck inlet block at the $8 \times 8$ resolution. The network follows a standard encoder-decoder design: noise is encoded via 128-dimensional positional embeddings, which are projected to 512-dimensional conditioning embeddings through two fully connected layers; the data augmentation conditions are 9-dimensional and independent of class labels. SiLU activations and GroupNorm ($\varepsilon = 10^{-6}$) are used throughout, with $[1, 1]$ resampling filters and a residual scaling factor of $\sqrt{0.5}$. Both models have approximately 61.806 million parameters. Training is performed in full FP32 precision without AMP/FP16 or TF32 acceleration, and the EDM preconditioning parameter $\sigma_{data}$ is set to 0.5. All training runs are initialized with official MDM pre-trained weights for the respective dataset, followed by full fine-tuning of all parameters. Input data are reshaped from the fixed inverse-point repository into $3 \times 64 \times 64$ tensors, with no source-point scaling calibration applied. Sampling is implemented using a reproducible equivalent of InfiniteSampler, with a random seed of 0 and a shuffle window ratio of 0.5. The Adam optimizer is adopted with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, and weight_decay $= 0$. The effective batch size is 256, with a per-GPU microbatch size of 4; parameter updates are executed after gradient accumulation over 64 microbatches. The peak learning rate is $2 \times 10^{-4}$, with a 10000 kimg linear warmup phase and no additional cosine or stepwise decay schedule. The EMA base half-life is 500 kimg, with a ramp-up ratio of 0.05. Data augmentation is performed via AugmentPipe with a probability strength $p = 0.15$, applying horizontal flip, vertical flip, scaling, rotation, anisotropic scaling, and translation to the inverse-point tensors, corresponding to 9 augmentation dimensions and excluding color augmentation. The noise distribution follows $\log \sigma \sim \mathcal{N}(-1.2, 1.2^2)$, with $\sigma_{data}$ fixed at 0.5.

**Flow Matching.** Experiments on both the FFHQ $64 \times 64$ and AFHQv2 $64 \times 64$ datasets employ a unified flow matching model built on a DDPM++-compatible SongUNet backbone for velocity field estimation. The model takes $3 \times 64 \times 64$ interpolated sample tensors and time step embeddings as input, and outputs the predicted continuous velocity field. It shares an identical architectural configuration with the aforementioned diffusion model, and is trained from scratch using a linear interpolation path between clean data samples and standard Gaussian noise. The training objective is the mean squared error between the predicted velocity field and the analytical ground-truth velocity. Time steps are uniformly sampled from $[0, 1]$ during training.

**VAE.** Experiments on both datasets use a residual convolutional VAE with a symmetric encoder-decoder architecture. The encoder takes $3 \times 64 \times 64$ RGB images as input, and features 4 downsampling stages: each stage contains 2 residual blocks followed by a strided-2 convolution for spatial downsampling, with channel dimensions increasing from 64 to 128, 256, and 512 sequentially. The resulting $8 \times 8 \times 512$ feature map is projected to the mean $\mu$ and log-variance $\log \sigma^2$ of the latent Gaussian distribution via two fully connected layers, with a latent dimension $d_z = 256$. The decoder mirrors the encoder with 4 upsampling stages: each stage consists of 2 residual blocks followed by bilinear upsampling and a convolution layer, progressively reconstructing the $3 \times 64 \times 64$ output normalized to $[-1, 1]$ via tanh activation. SiLU activations and GroupNorm (32 channels per group, $\varepsilon = 10^{-6}$) are used throughout, with a residual scaling factor of $\sqrt{0.5}$. The total parameter count is approximately 58.2 million, and training is performed in full FP32 precision. The VAE model is trained from scratch with He normal initialization. The training objective combines a per-pixel MSE reconstruction loss and a KL divergence regularization term with a weight coefficient $\beta = 0.001$ to balance reconstruction fidelity and latent space smoothness. The Adam optimizer is used with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, and weight_decay $= 10^{-5}$. The effective batch size is 256, with a per-GPU microbatch size of 8 and gradient accumulation over 32 microbatches per update. The peak learning rate is $5 \times 10^{-4}$, with a 5000 kimg linear warmup phase followed by a constant learning rate. No EMA is applied to model weights. Data augmentation only includes random horizontal flipping with probability 0.5 to avoid compromising reconstruction accuracy.

### C.8 The architecture of the generative model and the training hyperparameters for Five Complex Scenarios

**Diffusion Model.** We employ a unified class-unconditional continuous-time score diffusion model as the GM across Periodic, Burgers, Total Brightness, Galaxies, and Divergence-free datasets. Built on the NCSN++ backbone, the model is trained under the continuous-time variance-preserving stochastic differential equation (VP-SDE). It takes tensors matching each dataset's spatial dimensions as input and outputs corresponding noise-conditioned score estimates. The model uses 64 base channels with a specified channel multiplier. Input dimensions and the four-stage spatial size/channel progression follow dataset-specific configurations. The decoder adopts an upsampling architecture symmetric to the encoder. Each encoder resolution stage contains 4 main residual blocks and 1 downsampling block, while each decoder stage has 5 main residual blocks and 1 upsampling block, with skip connections from the encoder fused at corresponding levels. DDPM-style self-attention is applied at the resolution stage where the feature map has a vertical dimension of 16, mapping to specific stages for square and Divergence-free inputs respectively. The network follows a standard encoder-decoder design: continuous diffusion time is first encoded into 64-dimensional positional embeddings, then projected into 256-dimensional conditioning embeddings via two fully connected layers. Swish activation and GroupNorm are used throughout, with a preset GroupNorm epsilon, fixed convolution kernel size, and 0.1 dropout rate. Residual blocks follow the BigGAN-style structure. FIR resampling modules with learnable convolutions are inserted between adjacent scales with a specified filter kernel size, and a scaling factor is applied to the residual branch. The network uses residual progressive input without progressive output, and the output layer is initialized with a scale of 0. All five datasets share the same backbone architecture except for input and output channel counts. All training starts from random initialization, optimizing all network parameters. Input data are reshaped into tensors per dataset spatial dimensions, with no class labels or additional data augmentation. The diffusion process follows the continuous-time VP-SDE with specified linear noise schedule parameters. Training adopts continuous-time modeling with uniformly sampled diffusion times, and a predefined number of time steps for SDE numerical discretization and reverse sampling. Training is conducted in full FP32 precision without explicit AMP/FP16 mixed precision. The Adam optimizer is used with specified hyperparameters, zero weight decay, and a gradient clipping threshold of 1.0. EMA with a decay rate of 0.9999 is applied to model parameters, and EMA weights are uniformly used for final generation. Sampling follows a predictor-corrector framework with Euler-Maruyama as the predictor and no corrector step. The reverse process comprises 1000 time steps, integrated from the VP-SDE's Gaussian prior to the final step with a concluding denoising update. Sampling is configured with 'probability_flow=False' and 'noise_removal=True'. The final checkpoint is selected based on both validation loss and generation quality.

**Flow Matching.** We adopt a unified class-unconditional continuous-time flow matching model over Periodic, Burgers, Total Brightness, Galaxies, and Divergence-free. Built on an NCSN++ backbone,

the model takes time-conditioned state tensors as input and outputs corresponding instantaneous velocity estimates. The input size is $1 \times 64 \times 64$ for the Periodic, Burgers, and Total Brightness datasets, $1 \times 128 \times 128$ for Galaxies, and $2 \times 128 \times 256$ for Divergence-free. The network has a base channel count of 64 with channel multipliers $[1, 2, 2, 2]$. For the first three datasets, the spatial dimension / channel count across the four resolution stages are $64 \times 64/64$, $32 \times 32/128$, $16 \times 16/128$, and $8 \times 8/128$, respectively; for Galaxies, they are $128 \times 128/64$, $64 \times 64/128$, $32 \times 32/128$, and $16 \times 16/128$; for Divergence-free, they are $128 \times 256/64$, $64 \times 128/128$, $32 \times 64/128$, and $16 \times 32/128$. The encoder contains 4 main residual blocks per stage, with downsampling modules between consecutive stages; the decoder comprises 5 main residual blocks per stage and recovers spatial resolution via upsampling modules and skip connections. DDPM-style self-attention is applied to feature layers with a longitudinal spatial dimension of 16. The time variable is encoded via 64-dimensional positional embeddings, which are projected to 256-dimensional conditioning embeddings through two fully connected layers. The network employs Swish activations, GroupNorm, $3 \times 3$ convolutions, and a dropout rate of 0.1. Residual blocks follow the BigGAN-style architecture; resampling modules with learnable convolutions and $[1, 3, 3, 1]$ FIR filter kernels are used for scaling, and residual branches are scaled by $1/\sqrt{2}$. The model adopts residual progressive input without progressive output, and the output layer is initialized with a scale of 0. All models are trained from random initialization with full parameter optimization. Let $\boldsymbol{x}_1$ denote the data endpoint and $\boldsymbol{x}_0 \sim \mathcal{N}(0, I)$ the Gaussian endpoint. During training, time steps are uniformly sampled from $t \sim \mathcal{U}(0, 1)$, and a linear interpolation path is constructed as $\boldsymbol{x}_t = (1 - t)\boldsymbol{x}_0 + t\boldsymbol{x}_1$. The network is trained to predict the target velocity $\boldsymbol{x}_1 - \boldsymbol{x}_0$ by minimizing the element-wise mean squared error between predicted and target velocities, averaged first over non-batch dimensions and then over the batch dimension. Training is conducted in FP32 precision without explicit AMP/FP16 mixed precision. The Adam optimizer is used with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, and weight decay = 0. The gradient clipping threshold is 1.0, and the learning rate is $5 \times 10^{-5}$. The effective batch size is 64 for the Periodic, Burgers, and Total Brightness datasets, and 16 for Galaxies and Divergence-free. An EMA of model parameters is maintained with a decay rate of 0.9999. Both evaluation and generation use the EMA parameters.

**VAE.** We employ a unified convolutional residual VAE across five datasets: Periodic, Burgers, Total Brightness, Galaxies, and Divergence-free. The model takes as input tensors matching the spatial dimensions of each dataset. The encoder predicts the mean and log-variance of the posterior distribution over latent variables, which are sampled via the reparameterization trick. A symmetric decoder then outputs reconstructions with the same dimensionality as the input. The input size is $1 \times 64 \times 64$ for Periodic, Burgers, and Total Brightness, $1 \times 128 \times 128$ for Galaxies, and $2 \times 128 \times 256$ for Divergence-free. The model has a base channel count of 64 with channel multipliers $[1, 2, 4, 4]$. For the first three datasets, the four stages have spatial size / channel count pairs of $64{\times}64/64$, $32{\times}32/128$, $16{\times}16/256$, and $8{\times}8/256$, respectively. For Galaxies, the corresponding pairs are $128{\times}128/64$, $64{\times}64/128$, $32{\times}32/256$, and $16{\times}16/256$; for Divergence-free, they are $128{\times}256/64$, $64{\times}128/128$, $32{\times}64/256$, and $16{\times}32/256$. Both the encoder and decoder contain 2 residual blocks per stage. Downsampling between consecutive encoder stages is performed via $3{\times}3$ convolutions with stride 2. Upsampling in the decoder uses nearest-neighbor interpolation followed by a $3{\times}3$ convolution. Each residual block employs SiLU activations, GroupNorm with $\varepsilon = 10^{-6}$, and two $3{\times}3$ convolutions, with dropout set to 0. A $1{\times}1$ convolution shortcut is used when channel dimensions change, and residual outputs are scaled by $1/\sqrt{2}$. At the end of the encoder, a $1{\times}1$ convolution produces 16 moment channels, split equally into 8 channels for the mean and 8 for the log-variance. The spatial latent dimensions for the three input types are $8{\times}8{\times}8$, $8{\times}16{\times}16$, and $8{\times}16{\times}32$, respectively. The decoder output layer is an unbounded linear $3{\times}3$ convolution initialized with zeros. The VAE uses a diagonal Gaussian approximate posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\mu(\boldsymbol{x}), \mathrm{diag}\left(\exp\left(\log \sigma^2(\boldsymbol{x})\right)\right)\right)$ and a standard Gaussian prior $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Spatial latent variables are sampled via reparameterization, with log-variance clamped to $[-20, 10]$. The training objective is $\mathcal{L}_{VAE} = \mathcal{L}_{rec} + \beta\mathcal{L}_{KL}$, where the reconstruction term is the element-wise mean squared error averaged over channel and spatial dimensions, and the KL term is first averaged over non-batch dimensions and then over the batch dimension. The KL weight $\beta$ increases linearly from 0 to $\beta_{max} = 10^{-3}$ over the first 20 epochs to mitigate posterior collapse in early training. All VAEs are trained from random initialization with joint optimization of all encoder and decoder parameters. Training is conducted in full FP32 precision without explicit AMP/FP16 mixed precision. The Adam optimizer is used with a learning rate of $2{\times}10^{-4}$, hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$,

$\varepsilon = 10^{-8}$, and weight decay = 0, and a gradient norm clipping threshold of 1.0. The batch size is 16 for Periodic, Burgers, and Total Brightness, 8 for Galaxies, and 2 for Divergence-free.

### C.9 Detailed Descriptions of Complex Scenarios and Corresponding Datasets

**Total Brightness.** In astronomical imaging, the total brightness (or total flux) of an observed source can be tightly constrained (Akiyama et al., 2019) without prior knowledge of the source morphological information. For an image $\boldsymbol{y} \in \mathbb{R}^d$ composed of discrete pixels, we define the total brightness as the summation of all pixel values $V(\boldsymbol{y}) = \sum_{i=1}^{d} y_i$. Correspondingly, the constraint distance function for total brightness is formulated as $\ell_{\textit{flux}}(\boldsymbol{y}) = |V(\boldsymbol{y}) - \bar{V}|$, where $\bar{V} \in \mathbb{R}_{\geq 0}$ denotes the predefined target total brightness of the image. We validate this constraint on general relativistic magnetohydrodynamic simulation snapshots of Sgr A* from (Wong et al., 2022). The original $400 \times 400$ simulation images are downsampled to $64 \times 64$ resolution and intensity-rescaled to ensure a unified total flux of 120 for all samples. The dataset consists of 10000 training and 100 validation images.

**1D Burgers'.** We study the one-dimensional viscous Burgers' equation (Bateman, 1915; Burgers, 1948), a fundamental nonlinear partial differential equation in fluid mechanics. The governing equation for the viscous fluid field $u(t, y)$ is given by $\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial y} = \nu \frac{\partial^2 u}{\partial y^2}$, where $u(0, y)$ denotes the initial spatial field and $\nu \in \mathbb{R}_{\geq 0}$ denotes the viscosity coefficient. We numerically solve this equation using the Crank-Nicolson discretization scheme (Crank & Nicolson, 1947) implemented via Diffrax (Kidger, 2022). The resulting spatio-temporal solutions are arranged as $n_y \times n_t$ images $\boldsymbol{y}$, where $n_y$ and $n_t$ correspond to the spatial and temporal grid resolutions, respectively. To measure the physical consistency of generated trajectories, we define the constraint metric $\ell_{\textit{burgers}}(\boldsymbol{y})$ to quantify the deviation between each predicted time step and the solver output conditioned on the previous state $\ell_{\textit{burgers}}(\boldsymbol{x}) = \frac{1}{n_t - 1} \sum_{t=0}^{n_t - 2} \|\boldsymbol{y}[:, t+1] - f_{\textit{CN}}(\boldsymbol{y}[:, t])\|_1$, where $f_{\textit{CN}} : \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ represents the Crank-Nicolson forward integration function, and Python-style array indexing is used for compact representation. While finite-difference losses commonly adopted in physics-informed neural networks (Raissi et al., 2019) provide alternative physics constraints, they yield systematic errors for our setting, since Crank-Nicolson trajectories do not rigorously satisfy low-order finite-difference approximations. For dataset generation, we fix the viscosity as $\nu = 0.5$ and initialize solutions from Gaussian processes with a Matérn kernel (smoothness 1.5, length scale 1.0). The spatio-temporal domain is uniformly discretized into a $64 \times 64$ grid over $y \in [0, 10]$ and $t \in [0, 8]$. We adopt a timestep of $\Delta t = 0.025$ and store snapshots every five integration steps to form $64 \times 64$ temporal sequences. The resulting dataset contains 10000 training and 1000 validation samples.

**Divergence-Free.** Incompressible fluid dynamics is governed by the divergence-free constraint, which requires the divergence of the time-varying two-dimensional velocity field $\boldsymbol{u}(t, y_1, y_2)$ to vanish everywhere: $\nabla \cdot \boldsymbol{u} = 0$. For a trajectory-encoded image $\boldsymbol{y}$ that stacks consecutive flow states, we compute the point-wise divergence field $\text{div}(\boldsymbol{y})$ over the entire spatial domain and define the constraint distance as $\ell_{\textit{div}}(\boldsymbol{y}) = \|\text{div}(\boldsymbol{y})\|_1$. This metric quantitatively penalizes deviations from strict incompressibility. We evaluate this constraint on synthetic Kolmogorov flow datasets generated by solving the incompressible Navier-Stokes equations $\frac{\partial \boldsymbol{u}}{\partial t} = -(\boldsymbol{u} \cdot \nabla)\boldsymbol{u} + \frac{1}{\textit{Re}} \nabla^2 \boldsymbol{u} - \frac{1}{\rho} \nabla p + \boldsymbol{f}, \nabla \cdot \boldsymbol{u} = 0$, where $\textit{Re}$ is the Reynolds number, $\rho$ is the fluid density, $p$ represents pressure, and $\boldsymbol{f}$ encodes external forcing (Chandler & Kerswell, 2013; Boffetta & Ecke, 2012; Rozet & Louppe, 2023). Following standard simulation setups (Kochkov et al., 2021; Rozet & Louppe, 2023), we fix $\textit{Re} = 1000$ and $\rho = 1$, and adopt linearly damped Kolmogorov forcing for $\boldsymbol{f}$. The spatial domain $[0, 2\pi]^2$ is discretized on a uniform $64 \times 64$ grid with periodic boundary conditions. We utilize the `jax-cfd` library (Kochkov et al., 2021) to initialize spectrally filtered, divergence-free velocity fields and integrate the Navier-Stokes system via forward Euler updates with a timestep of $\Delta t = 0.01$. Flow snapshots are sampled every 20 steps within the time window $t \in [3, 4.6]$, yielding 8 sequential states per trajectory. These snapshots are horizontally concatenated into a single two-channel $128 \times 256$ image $\boldsymbol{x}$, where the two channels correspond to the horizontal and vertical velocity components, respectively. The final dataset consists of 10000 training and 1000 validation samples for constraint evaluation.

**Periodic.** We consider images $\boldsymbol{y}$ formed by periodic tiling of a fundamental unit cell, a common symmetry encountered in materials science such as the design of metamaterials constructed from repeating unit cells (Ogren et al., 2024). We assume each $\boldsymbol{y}$ is assembled from $n_{\textit{tiles}}$ identical unit-cell tiles, and define the constraint distance $\ell_{\textit{periodic}}(\boldsymbol{y}) = \sum_{i=1}^{n_{\textit{tiles}}} \frac{1}{n_{\textit{tiles}}} \sum_{j=1}^{n_{\textit{tiles}}} \|\boldsymbol{t}_i(\boldsymbol{y}) - \boldsymbol{t}_j(\boldsymbol{y})\|_1$, where $\boldsymbol{t}_i(\boldsymbol{y})$ denotes the $i$-th tile extracted from $\boldsymbol{y}$. For our experiments, each $64 \times 64$ image is formed by

arranging $32 \times 32$ unit cells in a $2 \times 2$ tiling layout. Leveraging the unit-cell generation code from (Ogren et al., 2024), we construct a dataset containing 30000 training and 300 validation samples.

**Count.** Generative models often produce images with an incorrect number of target objects (Paiss et al., 2023). We introduce a differentiable counting constraint that employs a convolutional neural network (CNN) to estimate object counts (Feng et al., 2025) from an image $\boldsymbol{y}$. A key caveat is that this neural-network-based constraint is non-analytic and highly non-convex. Let $f_{CNN} : \mathbb{R}^d \to \mathbb{R}$ denote the trained counting CNN and $\bar{c}$ denote the target object count. We define the constraint distance function as $\ell_{count}(\boldsymbol{y}) = |f_{CNN}(\boldsymbol{y}) - \bar{c}|$. We evaluate this constraint on synthetic astronomical data of radio galaxies (Connor et al., 2022). The dataset consists of $128 \times 128$ simulated images contaminated with background noise, each containing exactly $\bar{c} = 8$ radio galaxies with SNR $\geq 15$ dB. We split the dataset into 10000 training images and 1000 validation images. To train the counting CNN, we construct an auxiliary mixed dataset covering object counts of 6, 7, 8, 9 and 10 galaxies, with 10000 training and 1000 validation samples per count class. The network architecture is adapted from a simple MNIST classification backbone (8bitmp3, 2023): it contains two convolutional layers followed by two fully-connected layers with ReLU activations. We train the CNN by minimizing the mean squared error between the continuously predicted object count and the ground truth.

### C.10 More Experimental Results

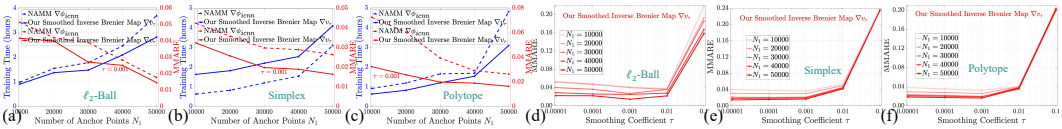



Fig 5: Effects of Anchor Number $N_1$ and Smoothness Coefficient $\tau$ on the Approximation Accuracy and Computational Efficiency of Different Map Pairs Relative to the Analytical Mirror Map over Three Classical Convex Constraint Domains. Dimension $d = 2$.

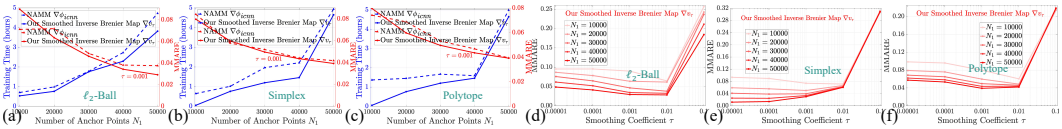



Fig 6: Effects of Anchor Number $N_1$ and Smoothness Coefficient $\tau$ on the Approximation Accuracy and Computational Efficiency of Different Map Pairs Relative to the Analytical Mirror Map over Three Classical Convex Constraint Domains. Dimension $d = 3$.

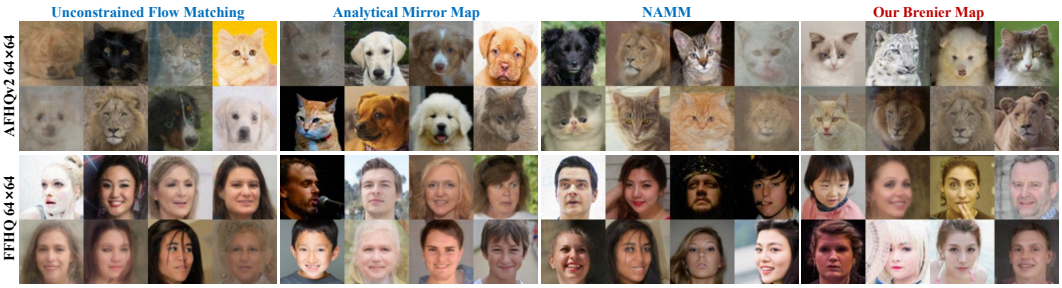



Fig 7: Comparison of Different Map Pairs Coupled with Flow Matching via Generated Watermarked Image Visualization.

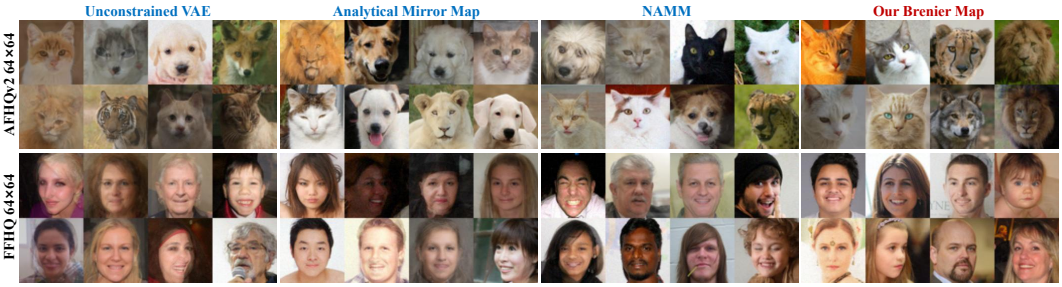



Fig 8: Comparison of Different Map Pairs Coupled with VAE via Generated Watermarked Image Visualization.

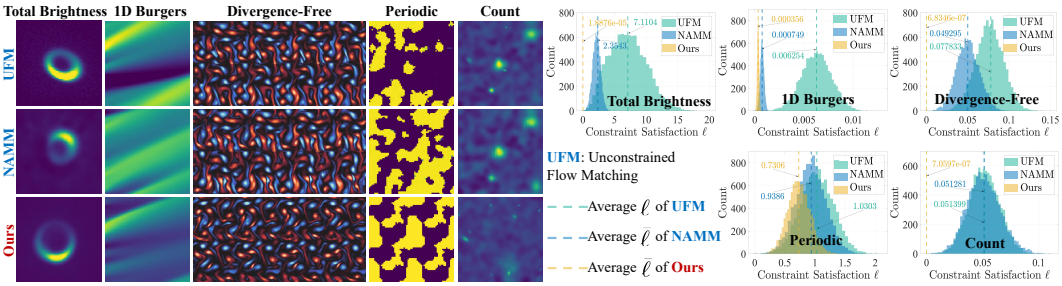



Fig 9: Comparison of Different Map Pairs Coupled with Flow Matching via Sample Visualization, the Per-Sample Distribution of the Data-Specific Constraint Satisfaction $\ell$ (Defined in Appendix C.9), and Its Average Value $\bar{\ell}$.

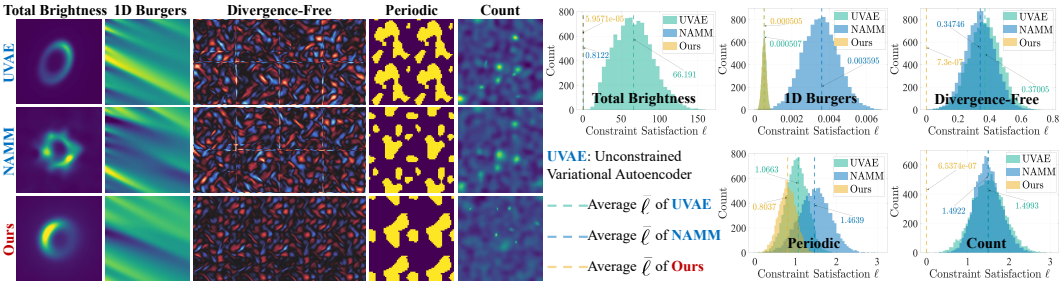



Fig 10: Comparison of Different Map Pairs Coupled with VAE via Sample Visualization, the Per-Sample Distribution of the Data-Specific Constraint Satisfaction $\ell$ (Defined in Appendix C.9), and Its Average Value $\bar{\ell}$.

Table 5: Performance Comparison of Different Methods in Complex Constrained-Generation Scenarios. CD $= 100 \times \lambda_{constr} \times \ell$, where $\ell$ denotes data-specific constraint function defined in Appendix C.9.

| Tasks | Total Brightness ($\lambda_{constr} = 0.01$) | | | | 1D Burgers ($\lambda_{constr} = 1$) | | | |
|---|---|---|---|---|---|---|---|---|
| Methods (+ VAE) | CD↓ | MMD↓ | KID↓ | Train $\mathcal{G}_{\boldsymbol{\theta}}$ ↓ | CD↓ | MMD↓ | KID↓ | Train $\mathcal{G}_{\boldsymbol{\theta}}$ ↓ |
| Unconstrained VAE | 66.191 | 0.6432 | 0.2109 | 0.5742h | 0.0507 | 0.2993 | 0.1638 | 0.4319h |
| NAMM (Tab. 3) | 0.8122 | 0.5974 | 0.0707 | 0.3681h | 0.3595 | 0.1983 | 0.0832 | 3.0328h |
| Our Brenier Map (Tab. 3) | **0.00006** | **0.2016** | **0.0151** | **0.3381h** | **0.0505** | **0.1058** | **0.0269** | **0.3361h** |
| Divergence-Free ($\lambda_{constr} = 1$) | | | | | Periodic ($\lambda_{constr} = 1$) | | | |
| CD↓ | MMD↓ | KID↓ | Train $\mathcal{G}_{\boldsymbol{\theta}}$ ↓ | CD↓ | MMD↓ | KID↓ | Train $\mathcal{G}_{\boldsymbol{\theta}}$ ↓ | |
| 37.005 | 0.2133 | 0.1899 | 3.2984h | 149.93 | 0.0907 | **0.0745** | 1.2339h | |
| 34.746 | 0.1954 | 0.2143 | 3.0328h | 149.22 | **0.0905** | 0.1202 | 1.0517h | |
| **0.00007** | **0.0945** | **0.0443** | **2.7647h** | **0.000064** | 0.1171 | 0.1154 | **1.0083h** | |
| Count ($\lambda_{constr} = 0.01$) | | | | | | | | |
| CD↓ | MMD↓ | KID↓ | Train $\mathcal{G}_{\boldsymbol{\theta}}$ ↓ | | | | | |
| 1.0663 | 0.3659 | 0.0708 | 1.2878h | | | | | |
| 1.4639 | 0.2129 | 0.2349 | 1.2175h | | | | | |
| **0.8037** | **0.1634** | **0.0591** | **1.1083h** | | | | | |

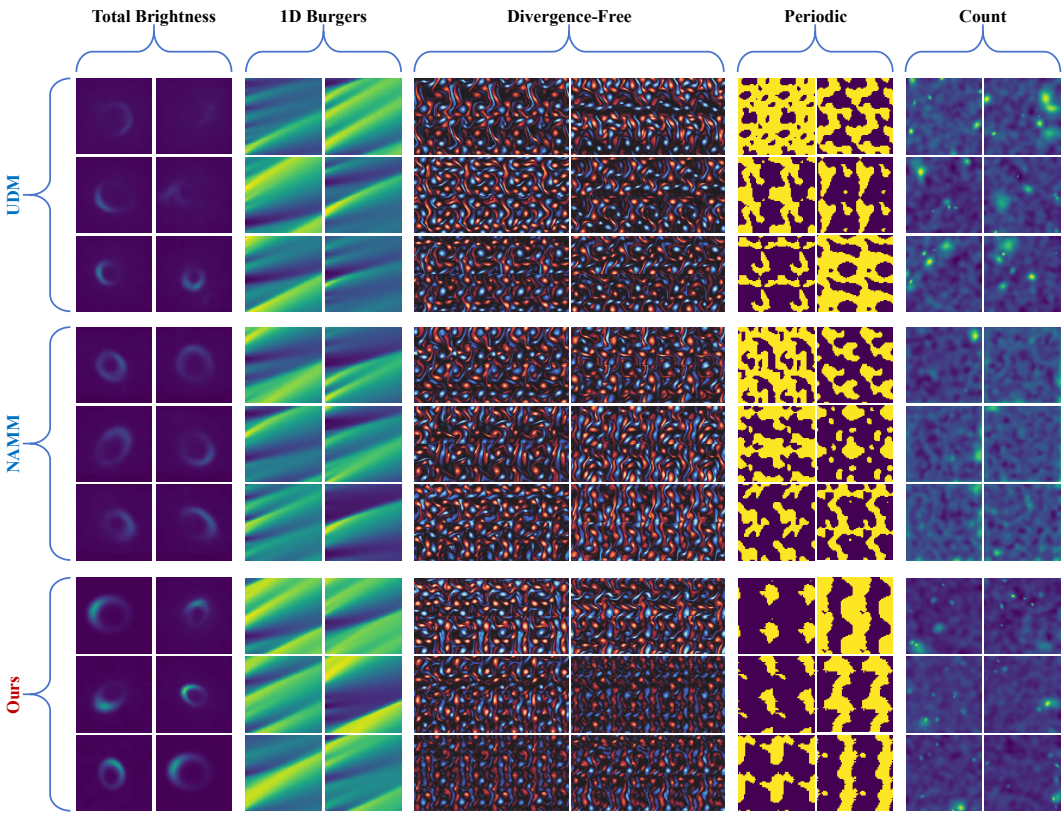



Fig 11: More Comparison of Different Map Pairs Coupled with Diffusion Model via Sample Visualization. UDM: Unconstrained Diffusion Model.

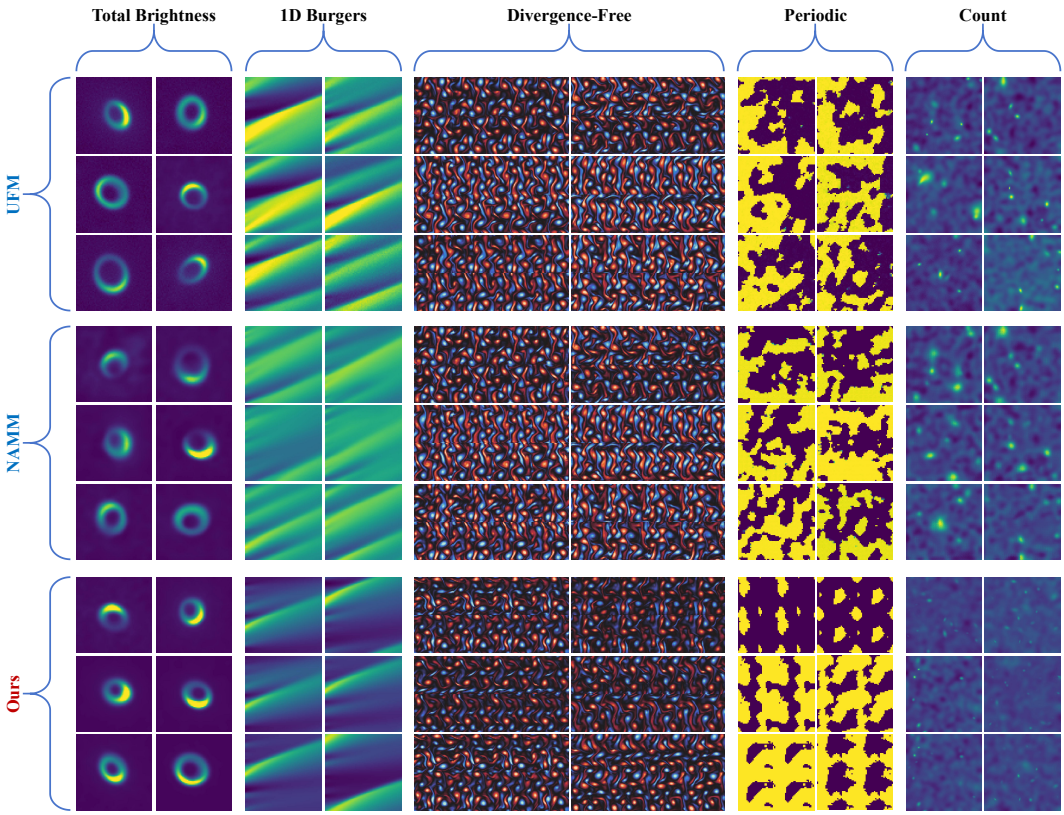



Fig 12: More Comparison of Different Map Pairs Coupled with Flow Matching via Sample Visualization. UFM: Unconstrained Flow Matching.

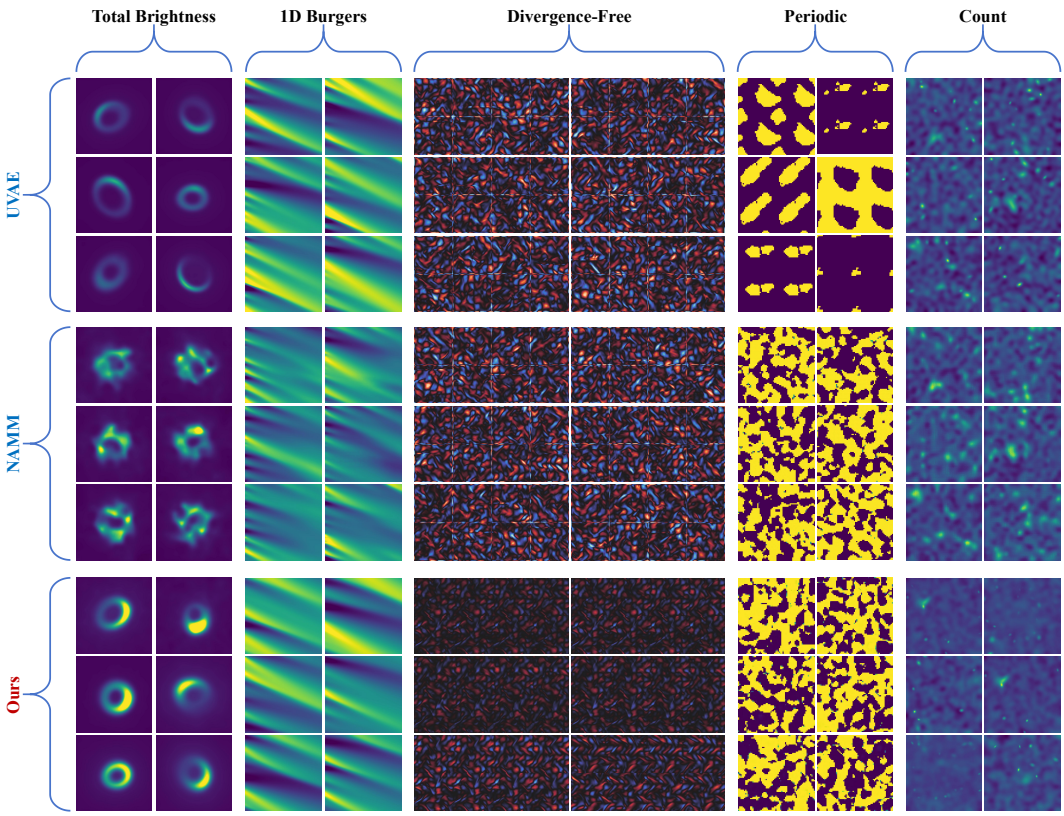



Fig 13: More Comparison of Different Map Pairs Coupled with VAE via Sample Visualization. UVAE: Unconstrained Variational Autoencoder.