A Gradient Flow Approach to Solving Inverse Problems with Latent Diffusion Models

Tim Y. J. Wang
Department of Mathematics
Imperial College London
tw1320@ic.ac.uk

O. Deniz Akyildiz
Department of Mathematics
Imperial College London
deniz.akyildiz@imperial.ac.uk

Abstract

Solving ill-posed inverse problems requires powerful and flexible priors. We propose leveraging pretrained latent diffusion models for this task through a new training-free approach, termed Diffusion-regularized Wasserstein Gradient Flow (DWGF). Specifically, we formulate the posterior sampling problem as a regularized Wasserstein gradient flow of the Kullback-Leibler divergence in the latent space. We demonstrate the performance of our method on standard benchmarks using StableDiffusion (Rombach et al., 2022) as the prior.

1 Introduction

Inverse problems are ubiquitous in science and engineering. They involve finding the underlying true signal x_0 from the corrupted observation y. In this work, we are primarily concerned with inverse problems of the form:

$$y = \mathcal{A}(x_0) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma_y^2 I), \tag{1}$$

where $\mathcal{A} : \mathsf{X} \to \mathsf{Y}$ is a known forward corruption operator and ϵ is the additive Gaussian noise. Through the lens of Bayesian inference (Stuart, 2010), the solution to the problem in (1) can be elegantly viewed as sampling from the posterior $p(x_0|y) \propto p(y|x_0)p_{data}(x_0)$ by placing a prior on x_0. While traditional hand-crafted priors are often chosen for mathematical convenience, they struggle with high-dimensional, ill-posed inverse problems.

In recent years, alternative priors based on diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) have become the state-of-the-art for solving imaging inverse problems. A common strategy is to modify the sampling process of a *pixel-space* diffusion model to guide the sampler towards the posterior $p(x_0|y)$ through various techniques, for instance, proximal (Zhu et al., 2023; Wu et al., 2024), gradient-based (Chung et al., 2023; Song et al., 2023a; Boys et al., 2024), variational inference (Mardani et al., 2024; Zilberstein et al., 2025), or sequential Monte-Carlo methods (Cardoso et al., 2024; Dou and Song, 2024; Chen et al., 2025).

However, adapting these techniques to more computationally efficient *latent-space* diffusion models (Rombach et al., 2022) is not straightforward, as the model defines a prior $p(z_0)$ on a latent variable instead of the ground truth signal x_0. Approaches such as Rout et al. (2023); Song et al. (2024) adapt diffusion posterior sampling (Chung et al., 2023) to the latent space with data consistency regularization, while Zilberstein et al. (2025) builds on RED-Diff (Mardani et al., 2024) and simulate a particle system for an augmented distribution $q(x_0, z_0|y)$ defined on the product of pixel and latent space.

In this work, we depart from adapting existing pixel-space methods. Instead, we formulate the posterior sampling problem from first principles as a Wasserstein gradient flow of a KL

2nd Workshop on Frontiers in Probabilistic Inference: Sampling Meets Learning, 39th Conference on Neural Information Processing Systems (NeurIPS 2025).

divergence functional directly in the latent space. Using the diffusion prior, we develop a regularized gradient flow, which we term Diffusion-regularized Wasserstein Gradient Flow (DWGF). We derive a system of ordinary differential equations (ODEs) that approximates this flow, providing a principled method for solving inverse problems with latent diffusion priors.

Notations. We use $\mathsf{X} = \mathbb{R}^{d_x}$ to denote the ambient (pixel) space, $\mathsf{Y} \subseteq \mathbb{R}^{d_x}$ the observation space, and $\mathsf{Z} = \mathbb{R}^{d_z}$ the latent space. We denote the space of probability measures on $\mathbb{R}^d$ with finite q-th moments as $\mathcal{P}_q(\mathbb{R}^d)$; in this work, we focus on the cases where $q = 2$. We use $\delta\mathcal{F}[\mu]/\delta\mu$ to denote the first L^2-variation of functional $\mathcal{F}$ at μ.

2 A Wasserstein Gradient Flow Approach

Assume that we are given a pretrained latent diffusion model, which consists of a decoder $p_{\phi^-}(x_0|z_0)$, encoder $r_{\phi^-}(z_0|x_0)$, and a diffusion generative model $p_{\theta^-}(z_0)$ in the latent space. Given the inverse problem defined in (1), we would like to leverage this model to obtain an approximate posterior distribution, denoted here

$$q_\mu(x_0|y) = \int p_{\phi^-}(x_0|z_0)\mu(z_0|y)\mathrm{d}z_0. \tag{2}$$

In order to obtain this distribution, one needs to approximate the unknown distribution $\mu(z_0|y)$. We thus consider the following regularized optimization problem over $\mathcal{P}_2(\mathsf{Z})$:

$$\mu_\star(z_0|y) \in \arg\min_{\mu\in\mathcal{P}_2(\mathsf{Z})} \mathcal{F}[\mu(z_0|y)] + \gamma\mathcal{R}[\mu(z_0|y)], \tag{3}$$

where $\mathcal{F}[\mu] = D_{\mathrm{KL}}\left(q_\mu(x_0|y)\|p(x_0|y)\right)$ and $\mathcal{R} : \mathcal{P}_2(\mathsf{Z}) \to \mathbb{R}_+$ can be taken as any regularization that makes the minimizers of the gradient flow well-defined (Crucinio et al., 2024), and $\gamma > 0$ controls the strength of regularization. In order to solve the problem in (3), we adopt a gradient descent approach in the space of probability measures . The gradient flow of any functional $\mathcal{L}[\mu]$ on $\mathcal{P}_2(\mathsf{Z})$ starting at some $\bar{\mu}_0$ is given by (Figalli and Glaudo, 2021):

$$\frac{\partial\mu_t}{\partial t} = \nabla_{z_0}\cdot\left(\mu_t\nabla_{z_0}\frac{\delta\mathcal{L}[\mu]}{\delta\mu}\right), \qquad \mu_0 = \bar{\mu}_0. \tag{4}$$

The PDE above is the continuity equation corresponds to the following ODE (Ambrosio et al., 2005, Chapter 8):

$$\frac{\mathrm{d}z_{0,t}}{\mathrm{d}t} = -\nabla_{z_0}\frac{\delta\mathcal{L}[\mu]}{\delta\mu} = -\left(\nabla_{z_0}\frac{\delta\mathcal{F}[\mu]}{\delta\mu} + \gamma\nabla_{z_0}\frac{\delta\mathcal{R}[\mu]}{\delta\mu}\right), \qquad z_{0,0}\sim\bar{\mu}_0, \tag{5}$$

where we have set $\mathcal{L}[\mu] := \mathcal{F}[\mu] + \gamma\mathcal{R}[\mu]$. The solution to (5) defines a flow map $T_t(z_{0,0}) = z_t$ such that $\mu_t = (T_t)_\#\bar{\mu}_0$. To simulate (5), we need to compute the first variation $\delta\mathcal{L}[\mu]/\delta\mu$ and its gradient. In what follows, we will construct the regularization and derive the first variations.

2.1 Wasserstein gradient of $D_{\mathrm{KL}}\left(q_\mu(x_0|y)\|p(x_0|y)\right)$

Wasserstein-2 (W_2) gradient of the first term $\mathcal{F}[\mu] = D_{\mathrm{KL}}(q_\mu(x_0|y)\|p(x_0|y))$ can be derived similarly to Wang et al. (2023). In particular, the first variation is given by (see Appendix A.1 for the full derivation)

$$\frac{\delta\mathcal{F}[\mu]}{\delta\mu} = \mathbb{E}_{p(x_0|z_0)}[\log q_\mu(x_0|y) - \log p(x_0|y)]. \tag{6}$$

Then to obtain the first term of the drift in the ODE (5), we compute

$$\nabla_{z_0}\frac{\delta\mathcal{F}[\mu]}{\delta\mu} := \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\left[\nabla_{x_0}(\log q(g_{\phi^-}(\epsilon,z_0)|y) - \log p(g_{\phi^-}(\epsilon,z_0)|y))\frac{\partial\mathcal{D}_{\phi^-}(z_0)}{\partial z_0}\right], \tag{7}$$

where we have used the chain rule and the reparameterization trick $g_{\phi^-}(\epsilon,z_0) := \mathcal{D}_{\phi^-}(z_0) + \rho\epsilon$ for the Gaussian decoder. As noted by Wang et al. (2024a), the first term $\nabla_{x_0}\log q(g_{\phi^-}(\epsilon,z_0)|y)$ is zero due to reparameterization (cf. Appendix A.3).

We note that the only intractable term in (7) is the posterior score $\nabla_{x_0} \log p(x_0|y)$, which admits the decomposition $\nabla_{x_0} \log p(x_0|y) = \nabla_{x_0} \log p(y|x_0) + \nabla_{x_0} \log p(x_0)$. While the first term $\nabla_{x_0} \log p(y|x_0)$ is tractable for being the gradient of the Gaussian likelihood (1), the data score $\nabla_{x_0} \log p(x_0)$ requires approximation. Assuming regularity conditions that permit interchanging the gradient and integral, we can express the prior score as an expectation, which we approximate using the encoder of the pretrained latent diffusion model:

$$\nabla_{x_0} \log p(x_0) = \int_{\mathsf{Z}} [\nabla_{x_0} \log p(x_0|z_0)] p(z_0|x_0) \mathrm{d}z_0 \approx \int_{\mathsf{Z}} [\nabla_{x_0} \log p(x_0|z_0)] \tilde{r}_{\phi^-}(z_0|x_0) \mathrm{d}z_0, \quad (8)$$

where $\tilde{r}_{\phi^-}(z_0|x_0)$ is the approximate posterior distribution given by a Variational Autoencoder (VAE) (Kingma and Welling, 2013), which is part of the pretrained latent diffusion model.

2.2 Wasserstein gradient of $\mathcal{R}[\mu(z_0|y)]$

To fully leverage the information provided by the generative model, we first construct the prior regularization term $\mathcal{R}(\mu(z_0|y))$ based on the pretrained latent diffusion model. Analogous to Wang et al. (2023); Luo et al. (2023), we consider a weighted KL divergence along the diffusion process:

$$\mathcal{R}[\mu(z_0|y)] := D_{\mathrm{KL}}^{w,[0,T]}(\mu(z_0|y) \| p_{\theta^-}(z_0)) = \int_0^T w(s) D_{\mathrm{KL}}(\mu(z_s|y) \| p_{\theta^-}(z_s)) \mathrm{d}s, \quad (9)$$

where $w(s) : [0,T] \to \mathbb{R}_+$ is a time-dependent weighting term and the densities $\mu(z_s|y) := \int_{\mathsf{Z}} \mathcal{N}(z_s; \alpha_s z_0, \sigma_s^2 I) \mu(z_0|y) \mathrm{d}z_0$ are pushforwards of μ through the forward transition kernel of the diffusion model. We remark that the weighted KL divergence in (9) is a well-defined functional and admits favorable properties (cf. Appendix A.4) as summarized below.

Theorem 2.1. *The weighted KL divergence $D_{\mathrm{KL}}^{w,[0,T]}(\mu(z_0|y) \| p_{\theta^-}(z_0))$ (9) is i) nonnegative, ii) convex in the first component $\mu(z_0|y)$, and iii) is minimized if and only if the standard KL divergence $D_{\mathrm{KL}}(\mu(z_0|y) \| p_{\theta^-}(z_0))$ is minimized.*

Accordingly, we obtain the gradient of the first variation of $\mathcal{R}$ (cf. Appendix A.2):

$$\nabla_{z_0} \frac{\delta \mathcal{R}[\mu]}{\delta \mu} = \mathbb{E}_{s,\epsilon} \left[\tilde{w}(s) \left(\nabla_{z_s} \log \int_{\mathsf{Z}} p(z_s|z_0) \mu(z_{0,t}|y) \mathrm{d}z_0 - \nabla_{z_s} \log p_{\theta^-}(z_s) \right) \frac{\partial z_s}{\partial z_0} \right], \quad (10)$$

where the expectation is taken over $s \sim \mathcal{U}(0,T)$, $\epsilon \sim \mathcal{N}(0,I)$, we set $\tilde{w}(s) := T w(s)$, and $p(z_s|z_0) = \mathcal{N}(z_s; \alpha_s z_0, \sigma_s^2 I)$ is the forward kernel of the diffusion model. To approximate the integral in (10), we use a particle-based approach to form a Monte-Carlo approximation using $\{z_0^{(i)}\}_i \sim \mu(z_0|y)$ (Wang et al., 2023; Lim and Johansen, 2024).

2.3 Algorithmic Considerations

Final Gradient Flow. We can simply combine the gradients derived in (7) and (10) to obtain the final gradient flow in (5). As we have an intractable term, we will simulate N identical copies of this ODE, where the integral in (10) is approximated using these particles as mentioned. We mention some practical aspects of this implementation below.

Deterministic Encoding During our experiments, we observe that the diagonal log covariance matrix of the approximate Gaussian posterior only contains small values (≈ -17), thus we may view the encoding process as deterministic and take $\mathcal{E}_{\phi^-} : \mathsf{X} \to \mathsf{Z}$ as a map to the mean of the Gaussian. In this case, (8) reduces to $(1/\rho^2)[\mathbb{E}_{\tilde{r}_{\phi^-}(z_0|x_0)}[\mathcal{D}_{\phi^-}(z_0)] - x_0] \approx (1/\rho^2)[\mathcal{D}_{\phi^-}(\mathcal{E}_\phi(x_0)) - x_0]$, whose vector Jacobian product with the decoder gradient resembles the data consistency term $\nabla_{z_t} ||\mathbb{E}[z_0|z_t] - \mathcal{E}(\mathcal{D}(\mathbb{E}[z_0|z_t]))||^2$ introduced in Rout et al. (2023); Song et al. (2024).

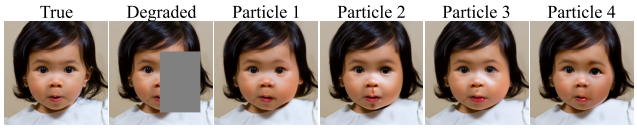
Adaptive Optimizer To accelerate convergence on the challenging optimization landscapes common in imaging inverse problems, we employ a non-standard discretization for the flow in (5). Instead of a simple Euler step, we treat the drift terms as gradients and apply the Adam optimizer (Kingma and Ba, 2014). We discuss this choice in Appendix C.

3 Experiments

We evaluate our method DWGF on the FFHQ dataset (Karras et al., 2019) downsampled to a resolution of 512×512. We compare with Posterior Sampling with Latent Diffusion (PSLD) (Rout et al., 2023), one of the established baselines for solving inverse problems with a latent diffusion prior and Repulsive Latent Score Distillation (RLSD) (Zilberstein et al., 2025), which is a recent method based on similar ideas from gradient flows. We adopt same the experimental settings as in Zilberstein et al. (2025) and report the best results therein. In this work, we only evaluate our method on box inpainting and super-resolution. We report standard metrics: Peak Signal-to-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), and Fréchet Inception Distance (FID) (Heusel et al., 2017).

	Inpainting (Box)			**SR (×8)**		
Method	**FID ↓**	**PSNR ↑**	**LPIPS ↓**	**FID ↓**	**PSNR ↑**	**LPIPS ↓**
PSLD (Rout et al., 2023)	57.70	22.72	0.082	81.31	24.82	0.314
RLSD (Zilberstein et al., 2025)	29.18	24.98	0.079	65.42	28.39	0.286
Ours	118.05	27.56	0.184	101.94	32.71	0.193

Table 1: Results for large box inpainting and 8× super-resolution, all with noise $\sigma_y = 0.001$ on the FFHQ-512 validation dataset.



(a) Diversity of the particles x_0 produced by DWGF on large box inpainting.

(b) Qualitative comparison of the DWGF, PSLD, and RLSD on inpainting and super-resolution (8×) tasks.

Figure 1: Qualitative results on the FFHQ dataset downsampled to 512×512 resolution.

As shown in the quantitative results (Table 1), our algorithm achieves comparable performance in terms of PSNR and LPIPS but suffers from poor FID. Examining the qualitative results (Figures 1a and 1b), we conjecture that this is caused by the mode-seeking behavior of the KL divergence, leading to blurry reconstructions. We point out that additional regularizations such as entropy (Wang et al., 2024b) or repulsive potentials (Corso et al., 2024; Zilberstein et al., 2025; Hu et al., 2025) can be incorporated into our functional in (3) to tackle mode collapse.

4 Conclusion

We proposed DWGF, a novel approach of solving inverse problems using latent diffusion models as priors. We note that our approach can be extended to the conditional setting where we jointly optimize the prompt embedding as in Spagnoletti et al. (2025) using Euclidean-Wasserstein gradient flows (Kuntz et al., 2023). Additionally, it would be interesting to explore control variates (Wang et al., 2024a) for variance reduction.

Acknowledgements

We thank the anonymous reviewers for their constructive comments. TW is supported by the Roth Scholarship from the Department of Mathematics, Imperial College London.

References

Akyildiz, O. D., Girolami, M., Stuart, A. M., and Vadeboncoeur, A. (2025). Efficient prior calibration from indirect data. *SIAM Journal on Scientific Computing*, 47(4):C932–C958.

Ambrosio, L., Gigli, N., and Savaré, G. (2005). *Gradient flows: in metric spaces and in the space of probability measures*. Springer.

Boys, B., Girolami, M., Pidstrigach, J., Reich, S., Mosca, A., and Akyildiz, O. D. (2024). Tweedie moment projected diffusions for inverse problems. *Transactions on Machine Learning Research*. Featured Certification.

Cardoso, G., el idrissi, Y. J., Corff, S. L., and Moulines, E. (2024). Monte carlo guided denoising diffusion models for bayesian linear inverse problems. In *The Twelfth International Conference on Learning Representations*.

Chen, H., Ren, Y., Min, M. R., Ying, L., and Izzo, Z. (2025). Solving inverse problems via diffusion-based priors: An approximation-free ensemble sampling approach. *arXiv preprint arXiv:2506.03979*.

Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C. (2023). Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*.

Corso, G., Xu, Y., Bortoli, V. D., Barzilay, R., and Jaakkola, T. S. (2024). Particle guidance: non-i.i.d. diverse sampling with diffusion models. In *The Twelfth International Conference on Learning Representations*.

Crucinio, F. R., De Bortoli, V., Doucet, A., and Johansen, A. M. (2024). Solving a class of fredholm integral equations of the first kind via wasserstein gradient flows. *Stochastic Processes and their Applications*, 173:104374.

Dou, Z. and Song, Y. (2024). Diffusion posterior sampling for linear inverse problem solving: A filtering perspective. In *The Twelfth International Conference on Learning Representations*.

Engel, E., Dreizler, R. M., and Dreizler, R. M. (2011). *Density Functional Theory: An Advanced Course*. Theoretical and Mathematical Physics. Springer Nature.

Figalli, A. and Glaudo, F. (2021). *An invitation to optimal transport, Wasserstein distances, and gradient flows*. EMS textbooks in mathematics. European Mathematical Society.

Geng, Z., Deng, M., Bai, X., Kolter, J. Z., and He, K. (2025). Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, (25):723–773.

Hagemann, P., Hertrich, J., Altekrüger, F., Beinert, R., Chemseddine, J., and Steidl, G. (2024). Posterior sampling based on gradient flows of the MMD with negative distance kernel. In *The Twelfth International Conference on Learning Representations*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). 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*.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*.

Hu, Y., Mei, K., Sahraee-Ardakan, M., Kamilov, U. S., Milanfar, P., and Delbracio, M. (2025). Kernel density steering: Inference-time scaling via mode seeking for image restoration. *arXiv preprint arXiv:2507.05604*.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

Kuntz, J., Lim, J. N., and Johansen, A. M. (2023). Particle algorithms for maximum likelihood training of latent variable models. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*.

Lim, J. N. and Johansen, A. M. (2024). Particle semi-implicit variational inference. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Lim, J. N., Kuntz, J., Power, S., and Johansen, A. M. (2024). Momentum particle maximum likelihood. In *Forty-First International Conference on Machine Learning*.

Luo, W., Hu, T., Zhang, S., Sun, J., Li, Z., and Zhang, Z. (2023). Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Mardani, M., Song, J., Kautz, J., and Vahdat, A. (2024). A variational perspective on solving inverse problems with diffusion models. In *The Twelfth International Conference on Learning Representations*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rout, L., Raoof, N., Daras, G., Caramanis, C., Dimakis, A., and Shakkottai, S. (2023). Solving linear inverse problems provably via posterior sampling with latent diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Santambrogio, F. (2016). Euclidean, Metric, and Wasserstein gradient flows: an overview. *arXiv preprint arXiv:1609.03890*.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*.

Song, B., Kwon, S. M., Zhang, Z., Hu, X., Qu, Q., and Shen, L. (2024). Solving inverse problems with latent diffusion models via hard data consistency. In *The Twelfth International Conference on Learning Representations*.

Song, J., Vahdat, A., Mardani, M., and Kautz, J. (2023a). Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. (2023b). Consistency models. *International Conference on Machine Learning*.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Spagnoletti, A., Prost, J., Almansa, A., Papadakis, N., and Pereyra, M. (2025). LATINO-PRO: LAtent consisTency INverse sOlver with PRompt optimization. *arXiv preprint arXiv:2503.12615*.

Stuart, A. M. (2010). Inverse problems: a bayesian perspective. *Acta numerica*, 19:451–559.

Vadeboncoeur, A., Girolami, M., and Stuart, A. M. (2025). Efficient deconvolution in populational inverse problems. *arXiv preprint arXiv:2505.19841*.

Villani, C. (2003). *Topics in optimal transportation*. Graduate studies in mathematics ; v. 58. American Mathematical Society.

Wang, P., Fan, Z., Xu, D., Wang, D., Mohan, S., Iandola, F., Ranjan, R., Li, Y., Liu, Q., Wang, Z., and Chandra, V. (2024a). SteinDreamer: Variance reduction for text-to-3D score distillation via stein identity. *arXiv preprint arXiv:2401.00604*.

Wang, P., Xu, D., Fan, Z., Wang, D., Mohan, S., Iandola, F., Ranjan, R., Li, Y., Liu, Q., Wang, Z., and Chandra, V. (2024b). Taming mode collapse in score distillation for text-to-3D generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA.

Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., and Zhu, J. (2023). Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wu, H., He, L., Zhang, M., Chen, D., Luo, K., Luo, M., Zhou, J.-Z., Chen, H., and Lv, J. (2024). Diffusion posterior proximal sampling for image restoration. In *ACM Multimedia 2024*.

Xie, S., Xiao, Z., Kingma, D. P., Hou, T., Wu, Y. N., Murphy, K. P., Salimans, T., Poole, B., and Gao, R. (2024). EM distillation for one-step diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., and Freeman, W. T. (2024a). Improved distribution matching distillation for fast image synthesis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., and Freeman, W. T. (2024b). Improved distribution matching distillation for fast image synthesis. *arXiv preprint arXiv:2405.14867*.

Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W. T., and Park, T. (2024c). One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *arXiv preprint arXiv:1801.03924*.

Zhu, Y., Zhang, K., Liang, J., Cao, J., Wen, B., Timofte, R., and Van Gool, L. (2023). Denoising Diffusion Models for Plug-and-Play Image Restoration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Zilberstein, N., Mardani, M., and Segarra, S. (2025). Repulsive latent score distillation for solving inverse problems. In *The Thirteenth International Conference on Learning Representations*.

A Derivations

Notations. We denote the space of probability measures on $\mathbb{R}^d$ with finite q-th moments as $\mathcal{P}_q(\mathbb{R}^d)$. We equip the space $\mathcal{P}_2(\mathbb{R}^d)$ with the usual Wasserstein-2 scalar product (Figalli and Glaudo, 2021) and denote its tangent space as $T\mathcal{P}_2(\mathbb{R}^d)$. For simplicity, we assume all distributions considered in this work admit differentiable densities with respect to the Lebesgue measure.

A.1 Derivation of the first variation of $\mathcal{F}$

We first state a standard result on the first variation of the KL divergence (cf. Villani (2003); Santambrogio (2016); Figalli and Glaudo (2021)):

Lemma A.1. *For $p, q \in \mathcal{P}_2(\mathbb{R}^d)$, the first variation of the KL divergence is given by:*

$$\frac{\delta D_{\mathrm{KL}}(q\|p)}{\delta q}(x) = \log q(x) - \log p(x) + 1, \quad \forall x \in \mathbb{R}^d \tag{11}$$

Proof. Using the equivalent definition of first variation, we can write for $m \in T\mathcal{P}_2(\mathbb{R}^d)$ and $t > 0$:

$$D_{\mathrm{KL}}(q + tm\|p) = D_{\mathrm{KL}}(q\|p) + t\left\langle m, \frac{\delta D_{\mathrm{KL}}(q\|p)}{\delta q}\right\rangle + o(t), \tag{12}$$

where the inner product is defined $\langle m, f\rangle := \int_{\mathbb{R}^d} f(z)m(z)\mathrm{d}z$ for all $f, m \in T\mathcal{P}_2(\mathbb{R}^d)$. Using Taylor expansion of $(z+t)\log(z+t) = z\log z + t(\log z + 1) + o(t)$, we can write $D_{\mathrm{KL}}(q+tm\|p)$ as:

$$\begin{aligned} D_{\mathrm{KL}}(q + tm\|p) &= \int (q(x) + tm(x))[\log(q(x) + tm(x)) - \log p(x)]\mathrm{d}x & (13)\\ &= \int q(x)\log q(x)\mathrm{d}x - \int q(x)\log p(x)\mathrm{d}x & (14)\\ &\quad + t\int (\log q(x) - \log p(x) + 1)m(x)\mathrm{d}x + o(t), & (15)\end{aligned}$$

which shows the desired result by matching the terms. □

Lemma A.2. *For $\mathcal{F} : \mathcal{P}_2(\mathsf{Z}) \to \mathbb{R}_+$ with $\mathcal{F}[\mu] := \int w(z)\mu(z)\mathrm{d}z$ (cf. Section 2) for any fixed $w(z) : \mathbb{R}^d \to \mathbb{R}$, we have:*

$$\frac{\delta \mathcal{F}[\mu]}{\delta \mu}(z) = w(z) \tag{16}$$

Proof. Similar to the proof above, for any $m \in T\mathcal{P}_2(\mathsf{Z})$ and $t \in \mathbb{R}$, we have:

$$\begin{aligned} \mathcal{F}[\mu + tm] &= \int w(z)(\mu(z) + tm(z))\mathrm{d}z & (17)\\ &= \int w(z)\mu(z)\mathrm{d}z + t\int w(z)m(z)\mathrm{d}z. & (18)\end{aligned}$$

From the definition of the functional derivative, this implies the desired result. □

Proposition A.3. For $q_\mu(x_0|y) = \int p(x_0|z_0)\mu(z_0|y)\mathrm{d}z_0$ and $p(x_0|y)$ both in $\mathcal{P}_2(\mathbb{R}^d)$ we have:

$$\frac{\delta D_{\mathrm{KL}}(q_\mu(x_0|y)\|p(x_0|y))}{\delta \mu} = \mathbb{E}_{p(x_0|z_0)}[\log q_\mu(x_0|y) - \log p(x_0|y)] \tag{19}$$

Proof. Using the chain rule of functional derivatives (Engel et al., 2011, Appendix A.3) and the two lemmas above, we obtain:

$$\begin{aligned} \frac{\delta D_{\mathrm{KL}}(q_\mu(x_0|y)\|p(x_0|y))}{\delta \mu} &= \int \frac{\delta D_{\mathrm{KL}}(q_\mu(x_0|y)\|p(x_0|y))}{\delta q}(x) \cdot \frac{\delta q_\mu(x_0|y)}{\delta \mu}(z_0)\mathrm{d}x & (20)\\ &= \int [\log q_\mu(x_0|y) - \log p(x_0|y)]p(x_0|z_0)\mathrm{d}x, & (21)\end{aligned}$$

where we have set $w(z_0) = p(x_0|z_0)$ in Lemma A.2 as the Gaussian decoder distribution. □

We now derive the gradient of the first variation.

Proof of (7). By Proposition A.3 above, we see that

$$\frac{\delta D_{\mathrm{KL}}(q_\mu(x_0|y)\|p(x_0|y))}{\delta \mu} = \mathbb{E}_{p(x_0|z_0)}[\log q_\mu(x_0|y) - \log p(x_0|y)]. \tag{22}$$

Assume standard regularity assumptions, we can exchange the expectation and the gradient. An application of the reparameterization trick $g_{\phi^-}(\epsilon, z_0) := \mathcal{D}_{\phi^-}(z_0) + \rho\epsilon$ yields:

$$\nabla_{z_0} \frac{\delta \mathcal{F}[\mu]}{\delta \mu} = \nabla_{z_0} \mathbb{E}_{p(x_0|z_0)}[\log q_\mu(x_0|y) - \log p(x_0|y)] \tag{23}$$

$$= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}[\nabla_{z_0}(\log q_\mu(g_{\phi^-}(\epsilon, z_{0,t})|y) - \log p(g_{\phi^-}(\epsilon, z_{0,t})|y))] \tag{24}$$

$$= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\left[\nabla_{x_0}(\log q_\mu(g_{\phi^-}(\epsilon, z_{0,t})|y) - \log p(g_{\phi^-}(\epsilon, z_{0,t})|y)) \frac{\partial \mathcal{D}_{\phi^-}(z_{0,t})}{\partial z_{0,t}}\right] \tag{25}$$

where we have used the chain rule in the last step. □

A.2 Derivation of the first variation of $\mathcal{R}$

Similar to Lemma A.1 above, we can derive the first variation of the weighted KL divergence (9). Recall that the weighted KL divergence for $p, q \in \mathcal{P}_2(\mathbb{R}^d)$ is given by:

$$D_{\mathrm{KL}}^{w,[0,T]}(q\|p) = \int_{[0,T]} w(s) D_{\mathrm{KL}}(q_s\|p_s)\mathrm{d}s, \tag{26}$$

where both q_s and p_s are defined as pushforwards through a Markov kernel: $q_s(z_s) := \int p(z_s|z_0)q(z_0)\mathrm{d}z_0$, where we define $p(z_s|z_0) := \mathcal{N}(z_s; \alpha_s z_0, \sigma_s^2 I)$.

Proposition A.4. For $p, q \in \mathcal{P}_2(\mathbb{R}^d)$, the first variation of the weighted KL divergence is given by:

$$\frac{\delta D_{\mathrm{KL}}^{w,[0,T]}(q\|p)}{\delta q}(x_0) = \int_{[0,T]} w(s)\mathbb{E}_{p(z_s|z_0)}[\log q_s(z_s) - \log p_s(z_s)]\mathrm{d}s, \quad \forall z_0 \in \mathbb{R}^d \tag{27}$$

Proof. Applying the chain rule for functional derivatives again, we obtain:

$$\frac{\delta D_{\mathrm{KL}}^{w,[0,T]}(q\|p)}{\delta q}(z_0) = \int \frac{\delta D_{\mathrm{KL}}^{w,[0,T]}(q\|p)}{\delta D_{\mathrm{KL}}(q_s\|p_s)}(s) \frac{\delta D_{\mathrm{KL}}(q_s\|p_s)}{\delta q}(z_0)\mathrm{d}s \tag{28}$$

$$= \int w(s)\mathbb{E}_{p(z_s|z_0)}[\log q_s(z_s) - \log p_s(z_s)]\mathrm{d}s \tag{29}$$

where we have used Lemma A.2 to obtain the first variation of the first term. To obtain the second term $\delta D_{\mathrm{KL}}(q_s\|p_s)/\delta q$, we use the proof of Proposition A.3 by replacing $p(x_0|z_0)$ with $p(z_s|z_0)$, $q_\mu(x_0|y)$ with $q_s(z_s)$, and $p(x_0|y)$ with $p(z_s)$. □

We can now derive the gradient of the first variation in (10).

Proof of (10). Substituting the definition for $p = p_{\theta^-}(z_0), q = \mu(z_0|y)$ with $p(z_s|z_0) = \mathcal{N}(z_s; \alpha_s z_0, \sigma_s^2 I)$ into Proposition A.4 and taking the gradient yields:

$$\nabla_{z_0} \frac{\delta \mathcal{R}[\mu]}{\delta \mu} = \nabla_{z_0} \int_{[0,T]} w(s)\mathbb{E}_{p(z_s|z_0)}\left[(\log \mu(z_s|y) - \log p_{\theta^-}(z_s))\right]\mathrm{d}s \tag{30}$$

$$= \int_{[0,T]} w(s)\nabla_{z_0}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\left[(\log \mu(g_s(\epsilon, z_0)|y) - \log p_{\theta^-}(g_s(\epsilon, z_0)))\right]\mathrm{d}s \tag{31}$$

$$= \int_{[0,T]} w(s)\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\left[(\nabla_{z_0}\log \mu(g_s(\epsilon, z_0)|y) - \nabla_{z_0}\log p_{\theta^-}(g_s(\epsilon, z_0)))\right]\mathrm{d}s \tag{32}$$

$$= \int_{[0,T]} w(s)\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\left[(\nabla_{z_s}\log \mu(g_s(\epsilon, z_0)|y) - \nabla_{z_s}\log p_{\theta^-}(g_s(\epsilon, z_0)))\frac{\partial z_s}{\partial z_0}\right], \tag{33}$$

where the second line follows from the reparameterization trick $g_s(\epsilon, z_0) = \alpha_s z_0 + \sigma_s \epsilon$, the third line follows from an exchange of the gradient and the integral, and we used the chain rule in the final line.

The marginal density $\mu(g_s(\epsilon, z_0)|y) = \mu(z_s|y)$ can be approximated using an integral $\int_{\mathsf{Z}} p(z_s|z_0)\mu(z_{0,t}|y)\mathrm{d}z_0$, whence (10) follows. □

A.3 Decoder Reparameterization

We now show that the gradient $\nabla_{z_0}\mathbb{E}_{p(x_0|z_0)}[\log q_\mu(x_0|y)]$ is in fact zero with a reparametrizable Gaussian decoder $p(x_0|z_0)$ (Wang et al., 2024a). Assuming standard regularity conditions, we have:

$$\nabla_{z_0}\mathbb{E}_{p(x_0|z_0)}[\log q_\mu(x_0|y)] = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\left[\nabla_{z_0}\log\mathcal{N}(x_0; \mathcal{D}_{\phi^-}(z_0), \rho^2 I)\right] \tag{34}$$

$$= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\left[\nabla_{z_0} - \frac{1}{2\rho^2}\|\mathcal{D}_{\phi^-}(z_0) + \rho\epsilon - \mathcal{D}_{\phi^-}(z_0)\|^2\right] = 0. \tag{35}$$

A.4 Proof to Theorem 2.1

We restate the theorem here for convenience:

Theorem A.5. *The weighted KL divergence $D_{\mathrm{KL}}^{w,[0,T]}(\mu(z_0|y)\|p_{\theta^-}(z_0))$ (9) is i) nonnegative, ii) convex in the first component $\mu(z_0|y)$, and iii) is minimized if and only if the standard KL divergence $D_{\mathrm{KL}}(\mu(z_0|y)\|p_{\theta^-}(z_0))$ is minimized.*

Proof. We note that the first two properties follows from those of the KL divergence. For the third property, we note that the weighted KL is zero if and only if $D_{\mathrm{KL}}(\mu(z_t|y)\|p_{\theta^-}(z_t)) = 0$ for every t, since the weights $w(t)$ are nonnegative. Now using the argument in Wang et al. (2023), we see that $\mu(z_t|y) = p_{\theta^-}(z_t)$ if and only if their characteristic functions are equal $\varphi_{\mu(z_t|y)}(s) = \varphi_{p_{\theta^-}(z_t)}(s)$. But we have:

$$\varphi_{\mu(z_t|y)}(s) = \varphi_{\mu(z_0|y)}(\alpha_t s)\cdot\varphi_{\mathcal{N}(0,I)}(\sigma_t s) \tag{36}$$

$$\varphi_{p_{\theta^-}(z_t)}(s) = \varphi_{p_{\theta^-}(z_0)}(\alpha_t s)\cdot\varphi_{\mathcal{N}(0,I)}(\sigma_t s), \tag{37}$$

since we have from the forward process $z_t = \alpha_t z_0 + \sigma_t\epsilon$ for $\epsilon\sim\mathcal{N}(0,I)$. The third property thus follows from the positivity of the KL divergence. □

B Algorithm

We now give the pseudocode for simulating the gradient flow in Algorithm 1 below; we use the notation $z_{s,k}$ to denote the particles diffused to the time $s\in[0,T]$ and at step $k\in\mathbb{N}$ of the Euler discretization of the ODE (5). Note that instead of approximating the drift corresponding to the diffusion regularization in (10) by sampling time s uniformly from $[0,T]$, we follow Mardani et al. (2024) and adopt a deterministic schedule for sampling s to be $s\in\{T, T-1, \cdots, 0\}$. In our experiments, we take $w(s) = c\sigma_s^2/\alpha_s$ with a constant $c\in(0,1)$ similarly to Mardani et al. (2024) .

C Experimental Details

Implementation In practice, we use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of $lr = 1.0$ and default hyperparameters $(\beta_1, \beta_2) = (0.9, 0.999)$ to solve the ODE as in Algorithm 1; this is done by viewing $u(z_{0,k}^{(i)}) + \gamma v(z_{0,k}^{(i)})$ as the gradient of a loss function to be optimized. This approach effectively introduces momentum and an adaptive diagonal preconditioning at each step, which helps mitigate ill-conditioning and flat minima. While unconventional for numerical solutions to ODEs, this heuristic is supported by recent work that formally incorporates momentum (Lim et al., 2024) and preconditioning (Lim and Johansen, 2024) into the simulation of gradient flows.

We set the balancing coefficient $\gamma = 0.15$ and the data consistency weight to $\lambda = 0.1\rho^2$ (cf. Section 2.3), where we choose the decoder standard deviation to be $\rho = 10^{-3}$. We

Algorithm 1 Diffusion-regularized Wasserstein Gradient Flow (DWGF)

Inputs: Observation $y \in \mathsf{Y}$, differentiable forward operator $\mathcal{A} : \mathsf{X} \to \mathsf{Y}$, initial particles $\{z_{0,0}^{(i)}\}_{i=1}^N$, diffusion weights $w(t)$, regularization strength γ, data consistency weight λ, pretrained latent diffusion model with score $s_{\theta^-}(\cdot,\cdot) : \mathsf{Z} \times \mathbb{R}_+ \to \mathsf{Z}$ and its VAE with encoder-decoder pair $(\mathcal{E}_{\phi^-}, \mathcal{D}_{\phi^-})$
Set counter $k \leftarrow 0$
for $s \in \{T, T-1, \cdots, 0\}$ **do**
  Sample $\epsilon^{(i)} \sim \mathcal{N}(0, I)$ and decode $x_0^{(i)} \leftarrow g_{\phi^-}(\epsilon, z_{0,k}^{(i)})$ for $i \in [N]$
  Data likelihood $\nabla_{x_0^{(i)}} \ell(x_0^{(i)}) \leftarrow \nabla_{x_0^{(i)}} (-\frac{1}{2\sigma_y^2} \|y - \mathcal{A}(x_0^{(i)})\|_2^2)$

$$u(z_{0,k}^{(i)}) \leftarrow \left(-\frac{\lambda}{\rho^2} [\mathcal{D}_{\phi^-}(\mathcal{E}_{\phi^-}(x_0^{(i)}))] - x_0^{(i)}] - \nabla_{x_0^{(i)}} \ell(x_0^{(i)}) \right) \frac{\partial \mathcal{D}_{\phi^-}(z_{0,k}^{(i)})}{\partial z_{0,k}^{(i)}} \tag{38}$$

  Sample $\nu^{(i)} \sim \mathcal{N}(0, I)$ for all $i \in [N]$
  Compute $z_{s,k}^{(i)} \leftarrow \alpha_s z_{0,k}^{(i)} + \sigma_s \nu^{(i)}$

$$v(z_{0,k}^{(i)}) \leftarrow w(s) \left(\nabla_{z_{s,k}^{(i)}} \log \frac{1}{N} \sum_{j=1}^{N} \mathcal{N}(z_{s,k}^{(i)}; \alpha_s z_{0,k}^{(j)}, \sigma_s^2 I) - s_{\theta^-}(z_{s,k}^{(i)}, s) \right) \frac{\partial z_{s,k}^{(i)}}{\partial z_{0,k}^{(i)}} \tag{39}$$

  Simulate $z_{0,k+1}^{(i)} \leftarrow \text{OptimizerStep}(z_{0,k}^{(i)}, u(z_{0,k}^{(i)}) + \gamma v(z_{0,k}^{(i)}))$ for all $i \in [N]$
  $k \leftarrow k + 1$
end for
return Particles decoded $\{\mathcal{D}_{\phi^-}(z_{0,k}^{(i)})\}_{i=1}^N$

choose Stable Diffusion v2.1 (Rombach et al., 2022) as the base model and set the number of sampling steps $T = 999$, which are the same as Zilberstein et al. (2025).

Evaluation We follow the experimental setups in Zilberstein et al. (2025) and evaluate our methods on the first 100 images from the validation set of FFHQ (Karras et al., 2019) using 4 particles. Due to limited computational budget, we do not reproduce the experiments in Zilberstein et al. (2025) but choose to report their best results for each task therein, namely the non-repulsive version of RLSD.

D Further Discussion and Related Works

Further Discussions Despite potentials of DWGF in inverse problems, it has some significant limitations. Firstly, our method does not yet match the performance of state-of-the-art in terms of FID (Section 3), which we attribute to the lack of proper regularization and insufficient hyperparameter tuning. Secondly, analogous to other variational approaches (Mardani et al., 2024; Zilberstein et al., 2025), DWGF requires a large number of sampling steps. A promising research direction is thus to integrate our framework with recent advances in few-step and consistency models (Song et al., 2023b; Luo et al., 2023; Yin et al., 2024a; Geng et al., 2025), which are emerging contenders to traditional iterative models. Finally, our method uses a particle cloud to approximate the intractable integral in solving the Wasserstein gradient flow, which may lead to considerable memory requirements.

Related Works Gradient flow approaches have been adopted in the context of general inverse problems in recent approaches (Akyildiz et al., 2025; Vadeboncoeur et al., 2025). In imaging, Hagemann et al. (2024) uses a gradient flow of the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012), which is parametrized as a sequence of pushforward maps. Zilberstein et al. (2025) uses an interacting particle system approach with repulsive potential, but they compute the marginal score of the latent distribution $\nabla_{z_t} \log q(z_t^{(i)}|y)$ as $-(z_t^{(i)} -$

$\alpha_t z_0^{(i)})/\sigma_t^2 = -\epsilon^{(i)}/\sigma_t$, which only holds when $q(z_0|y)$ is Gaussian, hence their approach reduces (approximately) to a Bures-Wasserstein gradient flow, similar to RED-Diff (Mardani et al., 2024). Our work is most related to Wang et al. (2023), which treat the parameters of an 3D representation MLP as particles to be optimized via Wasserstein gradient flow. However, their method operates in a conditional setting and involves training another network to approximate the score. Also related to our approach are works on score-distillation, which involve the training of a generator p_θ to approximate the output of a diffusion model (Yin et al., 2024c,b; Luo et al., 2023; Xie et al., 2024).