The Amazing Stability of Flow Matching

Rania Briq[1,2] r.briq@fz-juelich.de **Michael Kamp**[2,3,4] **Ohad Fried**[5] **Sarel Cohen**[5]

Stefan Kesselheim[1,6,7]

[1]**Forschungszentrum Jülich**
[2]TU Dortmund
[3]Lamarr Institute
[4]Institute for AI in Medicine, UK Essen
[5]Reichman University
[6]Helmholtz AI
[7]University of Cologne

Abstract

The success of deep generative models in generating high-quality and diverse samples is often attributed to particular architectures and large training datasets. In this paper, we investigate the impact of these factors on the quality and diversity of samples generated by *flow-matching* models. Surprisingly, in our experiments on CelebA-HQ dataset, flow matching remains stable even when pruning 50% of the dataset. That is, the quality and diversity of generated samples are preserved. Moreover, pruning impacts the latent representation only slightly, that is, samples generated by models trained on the full and pruned dataset map to visually similar outputs for a given seed. We observe similar stability when changing the architecture or training configuration, such that the latent representation is maintained under these changes as well. Our results quantify just how strong this stability can be in practice, and help explain the reliability of flow-matching models under various perturbations.

1 Introduction

Diffusion models [29, 30] have driven tremendous advances in generative modeling over the past few years [27, 14, 4, 20]. Flow-matching (FM) methods, an alternative to diffusion models [13], promise several advantages in efficiency and simplicity while achieving competitive performance [18, 19, 31, 1, 20]. Yet, training these models remains demanding both in compute and data, therefore, tailoring a dataset to desired generative properties has the potential to significantly reduce the computational cost. In this paper, we study the stability of FM under data perturbation, and develop informed approaches to data pruning. Using standard metrics, we probe stability and find a surprising result: even under strong perturbation of data and model architecture, trajectories initialized with the same random noise evolve to visually similar samples.

For diffusion models, stability in generating high-dimensional realistic data has been observed in several works: Kadkhodaie et al. [15] observed that diffusion models produce similar outputs under the same seed when trained on two disjoint subsets of the data, arguing that different splits converge to a similar geometry-aligned basis that follows image contours. Mlodozeniec et al. [23] confirmed the stability phenomenon for diffusion models while studying data attribution, and report that the likelihood remains nearly constant across models trained on 50% random subsets. However, this was

EurIPS 2025 Workshop on Principles of Generative Modeling (PriGM).

not shown for FM models, nor was it shown how this invariance can be exploited to train diffusion models on less data. Furthermore, in diffusion models, the objective is related to entropic-transport optimization (Schrödinger bridge) [6], which is stable to data perturbations [11]. By contrast, the flow-matching objective fits a velocity field $u(x, t)$, whose ODE transports noisy latents towards a clean manifold. While diffusion's stability is theoretically grounded, the behavior of FM models under data and architectural perturbations remains largely underexplored.

We empirically analyze the behavior of FM models trained on subsets of data and with different architectures. We summarize our contributions as follows:

- We show that FM models are remarkably stable. The generated images are visually similar under a wide variety of perturbations, including training on disjoint subsets of the dataset, including random or informed methods, labael-based or label-agnostic clustering, swapping of the entire dataset, as well as model architecture shrinkage. Each perturbation affects the generated data semantically only minimally.
- Inspired by previous work in discriminative models, we introduce three informed data pruning methods to FM models and study their influence qualitatively and quantitatively.
- Our proposed cluster-based resampling method that balances the distribution between different clusters can even improve the evaluation metrics of the generated images.

2 Approach

We probe the stability of FM models under various perturbations to the training data distribution, as well as the model architecture. Data perturbations include dataset pruning, where given a dataset S, we find a subset $S' \subset S$ using the pruning methods proposed in [5]. We (i) use a *random* subset as a baseline, whose performance serves as a lower bound for methods that require computation; (ii) rank based on a sample's training signal: gradient norm or loss computed along shared noise paths and timesteps; and (iii) cluster samples using their semantic features in a pretrained embedding space. For each method, we also apply the inverse criterion, i.e. we select samples with the lowest scores instead of the highest ones, and denote it by the superscript -1.

Pruning methods

Gradient-based scoring (Grad) . Under this strategy, we train a small surrogate model $\approx 7\%$ of the full training schedule, and use it to estimate the gradient magnitude for each sample using $M = 2$ fixed random noisy samples and $T = 8$ timesteps, creating shared noise paths for all the samples and decreasing the variance stemming from randomness. The gradient norms are then averaged over M and T using exponential moving average (EMA) estimate per $t \in T$ to remove large scale bias of larger noise bands.

$$s_i^{\text{grad}} = \frac{1}{T}\sum_{k=1}^{T}\frac{1}{M}\sum_{m=1}^{M}\frac{\left\|\nabla_\theta\, \ell\big(x_i;\, t_k, x_0^{(m)}\big)\right\|_2^2}{\mu_g(t_k)}, \tag{1}$$

where $\mu_g(t_k)$ is the EMA estimate at timestep t_k over samples and noise endpoints of the squared gradient norm $\|\nabla_\theta \ell(x_i; t_k, x_0^{(m)})\|_2^2$ with respect to the model parameters θ, computed inside the loop, and $x_0^{(m)}$ are the shared noise endpoints. The computation is easily parallelizable, however, this is an expensive method and we only apply it to gain insights into the effect of high-gradient samples on the model. Since samples with a large gradient influence the learned velocity field, we expect retaining them has a positive impact on the model.

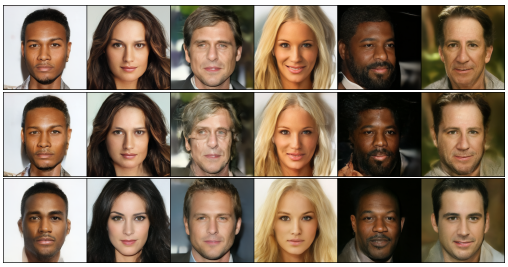
Loss-based scoring (Loss). We apply the same setup used in *Grad* and define s_i^{Loss} similarly, replacing $\|\nabla_\theta \ell\|_2^2$ by ℓ and μ_g by μ_ℓ.

Cluster-based scoring (Clust). We extract the image embeddings using the pretrained visual model CLIP [26]). We then use k-means [22] to cluster the samples, producing groups that share similar semantic characteristics. There are two criteria to consider here, (i) how many samples to select from a cluster, and (ii) which samples. For (i), we select either a number *proportional* to the cluster size or a *balanced* number, i.e. selecting an equal number of samples from each cluster. The first inherits the underlying distribution imbalance, while the latter balances skewed datasets. For (ii), we score a cluster's population based on their distance from the cluster center, and select either those located

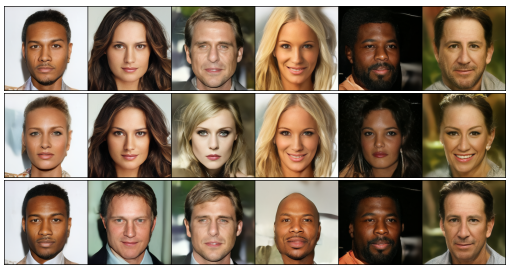
nearest to its center or furthest. The nearest samples form a representative subset retaining the core characteristics of the distribution, while the furthest samples cover more difficult and scarce samples. We refer to these variants as $Clust_{p/b}^{1/-1}$, indicating nearest/furthest and proportional/balanced. In our experiments, we choose $k = 24$ based on analysis of the clusters' inertia.

3 Experiments

(a) Two disjoint random subsets



(b) DiT-XL/4 → DiT-S/2 → U-Net



(c) Both genders/Female/Male

(d) CelebHQ → FFHQ

Figure 1: Stability of the generated images. (a) We train the model on two disjoint random subsets of the data, and obtain visually very similar images. (b) The data is split into two sets based on zero-shot classification as male/female. Images we visually interpret as belonging to the retained partition are semantically preserved, while images of the complementary class swap the semantic interpretation. (c) Model capacity change from DiT-XL to DiT-S retains high similarity, while switching to a U-Net architecture retains similarity to a lesser degree. (d) Changing the training dataset from CelebHQ to FFHQ, while still using CelebHQ VAE, retains similarity too.

	Unpruned	Random	Grad	Grad^{-1}	Loss	Loss^{-1}	Clust_p	Clust_p^{-1}	Clust_b	Clust_b^{-1}
FID ↓	24.24	25.25±0.38	24.62	29.75	33.92	23.49	25.19	27.96	**22.80**	26.77

(a) FID on CelebHQ with 4k generated samples at $pr = 0.5$. Random is averaged over 3 seeds.

Random	Grad	Grad^{-1}	Loss	Loss^{-1}	Clust_p	Clust_p^{-1}	Clust_b	Clust_b^{-1}
0.83 ± 0.11	0.79 ± 0.12	0.80 ± 0.12	0.80 ± 0.12	0.80 ± 0.13	0.80 ± 0.12	0.81 ± 0.12	0.81 ± 0.12	0.79 ± 0.13

(b) ArcFace cosine similarity between each pruned model and *Unpruned* for $pr = 0.5$. $N = 4k$ pairs matched by seed are evaluated. Here, $\pm$ denotes standard deviation over image pairs. Unmatched pairs yield 0.37 ± 0.11.

Experimental setup

We use the transformer-based architecture DiT [25], and replace diffusion with flow-matching transport [9] (we name it FM-DiT), training a velocity field $u_\theta(x, t)$ along linear interpolants between Gaussian noise and the data. We also train a vector-quantized variational autoencoder (VQ-VAE) [32] using the same target dataset to encode the images, similar to Stable Diffusion [27][1]. DiT is based on a ViT-style transformer [8], which operates on image patches with global self-attention. For the architectural change experiment, we additionally train a U-Net backbone [28], following multi-scale convolutional encoder–decoder with skip connections as done in diffusion models [13, 27].

For quantitative evaluation, we report FID [12], which measures the Fréchet distance between feature embeddings of the generated and training distributions. For quantifying FM stability, we measure

[1]Code and experimental details are available at https://github.com/briqr/fm_stability.

ArcFace pairwise cosine similarity for faces [7], a standard embedding model for face identification. *Unpruned* refers to the model trained on the full dataset. $N = 4096$ denotes the number of generated samples. All experiments are carried out by training the respective models on the standard CelebHQ dataset [16], which is based on CelebA dataset [21]. We acknowledge the imbalance in the dataset, which can affect qualitative judgments and subgroups performance. We further emphasize that features and attributes of human faces are subjective. All reported images are chosen from the same sequence of random $x_0 \sim \mathcal{N}(0, I)$. The images were not hand-picked; we only selected a range inside a longer sequence.

Stability tests

We investigate FM stability using several stress tests, including substantial data perturbation through pruning and data swapping, and architectural changes either in the model capacity or architecture design.

Disjoint subsets. In Fig. 1a, we train two FM model instances on two disjoint random subsets of the data. When integrating the velocity field starting from the same random points x_0, we observe that the outputs are nearly identical. We quantify this consistency using ArcFace pairwise similarity, and obtain $sim = 0.69 \pm 0.12$ between $N = 4096$ sample pairs generated by both models, where $\pm$ denotes the standard deviation over pairs. For comparison, unrelated pairs yield $sim = 0.34 \pm 0.10$.

Cluster removal. Fig. 1c depicts another experiment that alters the training data substantially. The first FM model is trained on images classified as female by PaliGemma VLM [3], while the second on images classified as male, yielding ArcFace similarity 0.76 ± 0.17 and 0.58 ± 0.16 respectively. This experiment is analogous to dropping an entire cluster or mode of the distribution. The results show that apart from the removed cluster, the models continue to generate similar outputs, demonstrating FM stability to mode removal. We want to acknowledge that we performed the binary split on the gender attribute as a technical experiment, and that the societal concept and implications of gender are clearly much more complex.

Data swapping. The experiment depicted in Fig. 1d is the most extreme form of data alteration. The FM model is trained on a different but same-domain dataset, FFHQ [17], which also comprises human faces. Even then, the outputs retain resemblance and we obtain ArcFace similarity $sim = 0.58 \pm 0.15$ (unrelated pairs yield $sim = 0.30 \pm 0.10$), indicating that with a fixed latent space, a different but same-domain dataset such as FFHQ lies on the same manifold as CelebA-HQ, allowing FM models to continue to learn similar trajectories with a matching seed.

Figure 2: Stability of the generated images under different pruning strategies and their inverse, $pr = 0.5$. In order: $Grad^{1/-1}$, $Loss^{1/-1}$, $Clust_b^{1/-1}$.

Architectural change. Fig. 1b illustrates a different type of stability tests. Instead of perturbing the distribution, we train three model variants that differ in their capacity or architecture. The first two variants share the same transformer DiT architecture but differ in their size: DiT/XL-2 (675M parameters, 24 layers) and DiT/S-4 (33M, 12 layers). The third variant is based on U-Net architecture. The outputs were consistently similar under model capacity change, retaining ArcFace similarity 0.81 ± 0.12, indicating identity preservation. With the full architectural change, the similarity drops to 0.55 ± 0.13. While the drop is clearly more visible compared to capacity shrinkage, we observe that the coarse attributes are preserved, which again hints at global stability.

Pruning strategies. In Fig. 2, we apply the proposed pruning methods and their inverse. Even though $Grad^{-1}$ (row 2) and *Loss* (row 3) produce artifacts, perceptually the images are similar to the inverse method. We apply the pruning methods using pruning fraction $pr = 0.5$ and report the

results in table 1a. *Random*'s FID deteriorates slightly from *Unpruned*. Using *Grad* almost does not change the FID, while its inverse (selecting lowest-grad samples) deteriorates significantly, as expected when dropping samples most influencing the model's weights. Selecting the highest-loss samples (*Loss*) substantially worsens the FID, compared to discriminative models [24]. In FM, these samples' predicted velocity fields deviate from the target flow, which is typical of samples present in low-density regions. Increasing these samples' representation therefore would lead to adversely impacting the flow. This explains why $\textit{Loss}^{-1}$ has the opposite effect. $\textit{Clust}_b$ even improves the FID, thanks to its uniform coverage across clusters, indicating that performance does not only depend on the sheer amount of data, but also on how balanced the data is. Across pruned variants versus their inverse, we obtain ArcFace cosine similarity in the range $0.72 - 0.74 \pm 0.13$, compared to 0.37 ± 0.11 for randomly shuffled pairs, indicating that matched outputs remain much closer than unmatched ones even when the training subsets are disjoint.
In table 1b, we quantify the stability of FM by comparing each pruned variant with *Unpruned*. We compare $N = 4096$ pairs matched by seed and observe that all methods maintain high similarity (above 0.79), compared to unrelated pairs (0.37 ± 0.12). This suggests FM models are very robust to perturbation in their training set: even methods that degraded the performance in FID, such as *Loss* and $\textit{Grad}^{-1}$, maintained high similarity with *Unpruned*.

4 Discussion and Outlook

We interpret our observation in terms of how well our model learns to approximate the true velocity field. We observe that for models trained under various perturbations, when starting from the same initial point x_0, the trajectories obtained by integrating the flow ODE end in points x_1 that are very close, and decode to perceptually similar images.

Recent works have begun to investigate FM models' ability in generalization, for example, Bertrand et al. [2] show that learning using the derived closed-form of the velocity field [10] in the finite data regime yields a similar performance as when using stochastic target $u(x, t)$, suggesting that stochasticity is averaged out and is therefore not the source of generalization. Our experiments on stability are complementary to this view; despite extreme data perturbations and architectural changes, trajectories starting from the same noise converge to nearby endpoints, suggesting their generalization does not stem from a single factor.

We studied removing entire clusters within the data distribution. For this model, trajectories starting from x_0 that would have ended up in endpoints in this cluster for a model trained on the full dataset, were rerouted to different endpoints. The corresponding images are clearly different, while images from retained clusters remain similar. In particular, trajectories sufficiently far away from the ones influenced by excluded clusters are only weakly impacted. We interpret this as a global stability: The flow field is only adjusted locally where necessary, while the global structure remains unaltered. This stability when removing data systematically allows enables training models with less data.

We also show that dataset pruning can be performed with little negative impact, or even with positive impact when done correctly. Some methods exhibit strong adverse effects, which hints at an intricate interaction of the dataset choice and generalization of FM models. We believe that understanding this interplay better is of high relevance for future powerful generative models trained on very large amounts of data, and could improve their efficiency substantially.

Acknowledgments

This work is funded by the German Federal Ministry for Economic Affairs and Energy within the project "nxtAIM". Additionally, the authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time on the GCS Supercomputer JUWELS Booster at Jülich Supercomputing Centre.

References

[1] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.

[2] Quentin Bertrand, Anne Gagneux, Mathurin Massias, and Rémi Emonet. On the closed-form of flow matching: Generalization does not arise from target stochasticity. *arXiv preprint arXiv:2506.03719*, 2025.

[3] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023.

[5] Rania Briq, Jiangtao Wang, and Stefan Kesselheim. Data pruning in generative diffusion models. *arXiv preprint arXiv:2411.12523*, 2024.

[6] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in neural information processing systems*, 34:17695–17709, 2021.

[7] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[8] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[10] Weiguo Gao and Ming Li. How do flow matching models memorize and generalize in sample data subspaces? *arXiv preprint arXiv:2410.23594*, 2024.

[11] Promit Ghosal, Marcel Nutz, and Espen Bernton. Stability of entropic optimal transport and schrödinger bridges. *Journal of Functional Analysis*, 283(9):109622, 2022.

[12] 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. *Advances in neural information processing systems*, 30, 2017.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[14] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[15] Zahra Kadkhodaie, Florentin Guth, Eero P Simoncelli, and St'ephane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representation. In *The Twelfth International Conference on Learning Representations*, 2024.

[16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[17] 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 (CVPR)*, 2019.

[18] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[19] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[20] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: a review on background, technology, limitations, and opportunities of large vision models (2024). *URL https://arxiv. org/abs/2402.17177*.

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[22] Stuart Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982.

[23] Bruno Mlodozeniec, Runa Eschenhagen, Juhan Bae, Alexander Immer, David Krueger, and Richard Turner. Influence functions for scalable data attribution in diffusion models. *arXiv preprint arXiv:2410.13850*, 2024.

[24] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607, 2021.

[25] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[29] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[30] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[31] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

[32] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.