ReALM-GEN
WORKSHOP

ICLR
International Conference On
Learning Representations

ReALM–GEN

Real-World Constrained and Preference-Aligned Flow- and Diffusion-based Generative Models

https://realm-gen-workshop.github.io

A Workshop Proposal for ICLR 2026

TL;DR

Diffusion and flow-based generative models power today's breakthroughs in Generative AI, showing impressive results in generating various types of data ranging from images and video to protein molecules and text. However, making them *respect real-world constraints* and *align with users' preferences* at post-training phase or at inference time, is still an unsolved challenge. ReALM–GEN at ICLR 2026 will bring together a diverse community of researchers spanning theoretical foundations of ML and generative models, vision, language, robotics, and scientific applications of AI, to explore bold ideas and practical tools for *adapting and/or steering pretrained flow- and diffusion-based models* toward real-world constraint satisfaction and alignment with user preferences.

Workshop Summary

Recent advances in diffusion and flow-based models for generation of various types of data e.g. images, text, molecules, raise an urgent question:

> **How can we systematically adapt, and steer such models so that they satisfy real–world constraints *and* align with user preferences?**

The ReALM–GEN Workshop explores the emerging frontier of *aligning* the generation process of diffusion- and flow-based models. We assume that the diffusion- and flow-based models are *already trained* to generate samples from a probability distribution $p_0(x)$. Aligned generative tasks can then be viewed through a unified framework i.e., as sampling from a tilted distribution $p_t(x)$,

$$p_t(x) \propto p_0(x) \exp(r(x)), \tag{1}$$

where $r(x)$ is a reward. This reward may represent, for example, preference alignment scores, Kim et al. [2024a], guidance objectives, Bansal et al. [2023], or real-world constraints—such as a data-fidelity term in inverse problems, $r(x) = -\frac{1}{2}\|y - \mathcal{A}(x)\|_2^2$.

This perspective places diverse controlled generative tasks into a unified framework: solving MRI reconstruction (an inverse problem), producing a storybook illustration (a conditional generation task), or designing a molecule that maximizes a preference score can all be interpreted as respecting externally provided evidence, directives, through maximizing a given reward. By unifying these perspectives, ReALM–GEN seeks to bring together researchers working on cross-domain theory with those developing new methods for *adapting* or *steering* diffusion- and flow-based models at inference time.

ReALM–GEN's Objectives

Diffusion and flow-based models have been at the heart of previous workshops focusing on generative modeling. However, all of them have mostly addressed *pretraining aspects of generative* *models*, such as architectural design, large-scale training, stability/convergence analyses, as well as the sampling efficiency. ReALM–GEN will be the first to focus solely on the fast-growing research area of *post–training* for diffusion and flow-based models and its connections to the emerging field of *alignment*.

The main objective of ReALM–GEN is to advance the *theory, methodology, and applications* of alignment for diffusion and flow-based models, as outlined below.

(a) **Theory**
- Distribution tilting and energy-based models, Havens et al. [2025].
- Reinforcement-learning–based post-training of diffusion models to optimize downstream rewards, Uehara et al. [2024].
- Stochastic optimal control view of diffusion and flow-based models, Schrödinger bridges, and links with Optimal Transport (OT), Khrulkov et al. [2022], Park et al. [2024].
- Steerability of diffusion and flow-based models e.g., representation disentanglement, identifiability and representation geometry, parsimony/sparsity and their roles in inverse problems and conditional generation, Shuai et al. [2024], Li et al. [2024b].

(b) **Methodology**
- Novel approaches for supervised, reward-based, and sampling-based fine-tuning, Anil et al. [2025], Uehara et al. [2024].
- Efficient conditional generation and constraint satisfaction across vision, language, and multimodal models (e.g., guidance, adapters, plug–and–play modules), Bansal et al. [2023], Epstein et al. [2023].
- Test-time scaling, continual test-time adaptation, inference-time alignment, steering, Uehara et al. [2025], Ramesh and Mardani [2025].
- Guidance mechanisms for discrete diffusion models, Schiff et al. [2024].

(c) **Applications**
- Computer vision & robotics, Ajay et al. [2022], Wagenmaker et al. [2025].
- Scientific and medical imaging, Li et al. [2024a].
- Tabular data synthesis, Liu et al. [2024].
- Language and multimodal reasoning, Ye et al. [2024], Wang et al. [2025].
- Protein/molecule design, Gruver et al. [2023].

Workshop Scope

Our workshop spans post-training and inference-time alignment of diffusion and flow-based models, diffusion LLMs, and multimodal generators, with special emphasis on the following problems:

- **Inverse Problems:** Diffusion priors for inverse imaging problems, physics–guided inference, [Feng et al., 2023, Daras et al., 2024a,b].
- **Conditional Generation:** Text–to–image/video, [Zhang et al., 2023], style transfer, in-context editing, [Zhang et al., 2025], counterfactual generation, [Ma et al., 2024].
- **Steering:** Latent editing and representation engineering, score-based guidance, plug-and-play modules, and preference-optimization methods for alignment (DPO and recent variants) [Rafailov et al., 2023, Wallace et al., 2024, Nichol et al., 2021].
- **Safety alignment:** Concept engineering, [Luo et al., 2024], sparse-autoencoder (SAE) probes, [Kantamneni et al., 2025, Huang et al., 2024, Zhang and Nanda, 2023, Kim et al., 2024b].
- **Fine Tuning:** Supervised, Reward- and sampling-based fine-tuning, Tang [2024], Uehara et al. [2024], Fan et al. [2025].
- **Discrete Diffusion Models/Diffusion Language Models:** Guidance for discrete diffusion models, Schiff et al. [2024], Austin et al. [2021], Nie et al. [2025], Suresh et al.

[2025] Constrained decoding of diffusion LLMs, Mündler et al. [2025].

Community Engagement

Call for Papers and Review Process

We will issue a call for papers and accept submissions via OpenReview. Two tracks will be offered: (a) **Full Papers** with original contributions (up to 9 pages), and (b) **Short Papers** featuring preliminary ideas and late-breaking results (up to 4 pages).

We will follow ICLR Workshop standards and conduct a rigorous, double-blind review. To build a strong reviewer pool in diffusion- and flow-based models, we will circulate a call for reviewers and provide a Google Form for self-nominations. To further encourage engagement, we will recognize Outstanding Reviewers for exemplary service. Each submission will receive at least two reviews, and conflicts of interest will be handled in accordance with ICLR guidelines. Members of the organizing committee will serve as meta-reviewers to ensure consistency and quality across decisions. All accepted papers will be presented as posters, and the highest-rated papers will be invited for contributed spotlight talks.

Travel Awards & Inclusion

To broaden participation especially from low- and middle-income countries and underrepresented communities, we will actively secure sponsorship for travel awards from organizations such as Google DeepMind, Meta AI, NVIDIA, and OpenAI. Awards will prioritize student and early-career researchers with accepted work, assessed on both merit and demonstrated financial need. We will run a transparent, deadline-driven application process, publish clear selection criteria, and allocate a dedicated portion of funds to first-time attendees. Support may cover airfare, accommodation, visa fees, and registration; we will also offer childcare and accessibility stipends where needed.

Social Event, Post-Workshop Engagement

We plan to host a dinner at the end of the workshop to foster interaction among attendees and spark discussions that may lead to new collaborations. Our goal is to build a strong, collaborative community around the emerging topic of post-training and alignment for diffusion and flow-based models. To this end, we will launch a Discord channel and encourage all attendees to participate. We also plan to run a monthly online reading group, inviting leading experts in the field to present their work. Workshop attendees will be strongly encouraged to contribute to these activities.

Workshop Format and Tentative Schedule

ReALM–GEN will be an in-person event, with live-streaming and recordings available to maximise accessibility. All invited speakers and panelists have confirmed they will attend in person. However, contingency plans are in place to enable remote presentations if any speaker or panellist is unable to attend. ReALM–GEN will feature *invited talks* by leading experts, *contributed spotlight presentations* that will be nominated from strongest submissions to the workshop and two *poster sessions*. It will also include a *themed mini-panel* (with three of our invited speakers as participants), that will focus on safety/alignment at post-training phase, and *a larger panel* to debate on *diffusion and flow-based alignment in the era of Agentic AI*. A tentative detailed schedule of the workshop is presented in Table 1.

Invited speakers

- ✓ **(Confirmed) Yaron Lipman (Meta AI, Israel)**: Yaron Lipman is a Research Scientist at Meta Fundamental AI Research (FAIR). His work spans generative modeling and geometric

Time	Session	Details
08:45–09:00	Registration & Setup	Badge pickup, poster setup for Session 1
09:00–09:10	Opening Remarks	Welcome; goals & logistics (ReALM–GEN overview)
09:10–09:40	Yaron Lipman	25 min + 5 min Q&A
09:40–10:10	Contributed Spotlights I	4×7 min lightning talks (top papers), 2 min transitions
10:10–11:00	Poster Session 1 + Coffee	Accepted papers; authors at posters
11:00–11:30	Jong Chul Ye	25 min + 5 min Q&A
11:30–12:00	Thematic Mini-Panel	*Safety/Alignment via Post-training*: 3 panelists + 10 min audience Q&A
12:00–13:15	Lunch Break	Birds-of-a-Feather tables (alignment, controllability, SOC/OT)
13:15–13:45	Ruiqi Gao	25 min + 5 min Q&A
13:45–14:15	Contributed Spotlights II	4×7 min lightning talks, 2 min transitions
14:15–15:05	Poster Session 2 + Coffee	Accepted papers
15:05–15:35	Marta Skreta	25 min + 5 min Q&A
15:35–16:05	Volodymyr Kuleshov	25 min + 5 min Q&A
16:05–16:45	Panel Discussion	chair + 4 panelists (Alignment of Generative Models + Agentic AI)
16:45–17:20	Peter Holderieth	25 min + 5 min Q&A
17:20–17:30	Closing Remarks	Awards (best poster/spotlight); next steps & community resources

Table 1: Tentative schedule

learning, with pioneering contributions to continuous- and discrete-time flows and flow-matching methodologies and reward-based alignment of flow- and diffusion-based models. Yaron's work supports the workshop's focus on aligning diffusion- and flow-based generators for reliable, steerable behavior.

- ✓ **(Confirmed) Jong Chul Ye (KAIST, South Korea)**: Jong Chul Ye is a Professor at KAIST whose group bridges inverse problems, medical imaging, and modern generative modeling. He develops physics-guided and data-consistent diffusion/score-based methods that deliver controllable reconstructions and robust sampling. His focuses on incorporating domain constraints into generative flows for safe and aligned deployment. His contributions align tightly with the workshop's emphasis on trustworthy, guided diffusion and flow-based models for solving inverse problems.
- ✓ **(Confirmed) Ruiqi Gao (Google DeepMind, USA)**: Ruiqi Gao is a Research Scientist at Google DeepMind working on core generative modeling, including diffusion and flow-based approaches, scalable training, and efficient sampling. Her research explores objectives and dynamics that improve sample quality, stability, and controllability across modalities. She brings a rigorous perspective on how training losses and sampling algorithms affect alignment and reliability in practice.
- ✓ **(Confirmed) Volodymyr Kuleshov (Cornell University, USA):** Volodymyr Kuleshov is an Assistant Professor at Cornell Tech. His research centers on reliable, practical machine learning—spanning calibration, robustness, and rigorous evaluation of generative

models in real-world settings such as healthcare. Recently, he has advanced several themes central to our workshop, including diffusion language models, test-time scaling to improve diffusion/flow model behavior, and guidance methods for discrete-time diffusion models.

- ✓ **(Confirmed) Marta Skreta (University of Toronto, Canada):** Marta Skreta is a PhD student at the University of Toronto. Her research spans generative modelling, with applications to chemistry and self-driving labs. She has published several works on controlling the inference-time behaviour of diffusion models, and her research aligns closely with the workshop's objective to explore inference-time alignment/steerability topics, and their scientific applications.
- ✓ **(Confirmed) Peter Holderieth (MIT, USA):** Peter Holderieth is 2nd-year PhD student at CSAIL at MIT working with Tommi Jaakkola. He works on machine learning algorithms, in particular generative modeling, as well connections to mathematics and science. He has recently awarded a Best Paper Award at Frontiers in Probabilistic Inference workshop at ICLR 2025. In his talk, he will present his recent work on a sampling-based approach for alignment of flow and diffusion models.

Panelists

- ✓ **(Confirmed) Arash Vahdat (NVIDIA Research, USA)**: Arash Vahdat is a Research Director, leading the fundamental generative AI research (GenAIR) team at NVIDIA Research. Before joining NVIDIA, he was a research scientist at D-Wave Systems, working on generative learning and its applications in label-efficient training. Before D-Wave, Arash was a research faculty member at Simon Fraser University (SFU), where he led deep learning-based video analysis research and taught master courses on machine learning for big data. Arash's current area of research is focused on generative learning with applications in multimodal training, accelerated generative models and gen AI for science.
- ✓ **(Confirmed) Karsten Kreis (NVIDIA Research, USA)**: Karsten Kreis is a Principal Research Scientist at NVIDIA Research focusing on generative AI. Karsten's research interests span both the development of foundational generative AI algorithms and their application across scientific and creative domains. Recently, he has been focusing on generative learning for molecular modeling and is leading efforts in generative modeling for protein design. Karsten is broadly excited by the intersection of generative AI and the natural sciences. In addition to scientific applications, he has also worked extensively on generative models for content creation and digital artistry—including image, video, and 3D generation.
- ⌛ **(awaiting response) Tali Dekel (Weizmann Institute, Israel)**: Tali Dekel is an Associate Professor at the Faculty of Mathematics and Computer Science at the Weizmann Institute of Science, and a Staff Research Scientist at Google DeepMind. Before that she was a Postdoctoral Associate with Prof. Bill Freeman, at CSAIL, MIT. She completed my Ph.D. in the school of Electrical Engineering of Tel-Aviv University, where she was supervised by Prof. Shai Avidan (TAU) and Prof. Yael Moses (IDC). Her main research interests include images and videos analysis, multi-view systems, 3D structure and motion estimation, image synthesize and rendering.
- ⌛ **(awaiting response) Chenlin Meng (Stanford University, USA):** Chenlin Meng is a co-founder of Pika (Pika Labs), the generative-AI startup building creative video tools. She studied AI at Stanford, focusing on diffusion and generative models, and has published widely cited work in machine learning. At Pika, she helps push the frontier of text-to-video and visual creation, bridging cutting-edge research with products used by a broad community of creators. Before Pika, Chenlin co-authored influential work on diffusion/score-based models—most notably DDIM (Denoising Diffusion Implicit Models)—that helped shape

today's state of the art.

Attendance

Expected size of the audience

Our primary audience includes researchers in the theory and algorithmic development of diffusion and flow-based models, sampling and fine-tuning, inverse problems, and reinforcement learning for alignment, as well as practitioners working across applications such as computational biology, computer vision, robotics, and language modelling. Given the strong and growing interest in these areas within the ML community, and attendance at recent related workshops, we project an in-person audience of 250–300 participants. We will support online attendance, adhering to ICLR live-streaming guidelines and using SlidesLive platform.

Outreach and Accessibility

We aim to promote the workshop across social media channels including Bluesky, X, LinkedIn. We will also send and digital flyers via e-mail to our industry and academic networks and ask the members of Progam Commitee to share these materials within their professional circles and institutions. This coordinated outreach is aimed at maximizing visibility, broadening participation, and ensuring strong engagement from both industry and academia. We have also created a website (https://realm-gen-workshop.github.io), that will help us share important details regarding the workshop. We plan to use the website to upload online material that will be accessible to researchers who could not join the workshop in person.

Diversity and Inclusion Commitment

We intentionally structured the organizing team, and invited speakers to reflect diversity across seniority , gender, and geography, and gender balance among both organizers and keynote speakers.

Diversity of Organizing Team

Our organizing team spans multiple continents, institutions, and career stages, from PhD student to junior and senior faculty to industry researchers, ensuring a range of perspectives in planning and decision-making. Team members are affiliated with EPFL, MIT, the University of Warwick, Imperial College London, and NVIDIA Research, among others, reflecting diversity across academia and industry as well as geography. We intentionally pursued gender balance and representation across seniority in assembling the committee.

Diversity of Invited Speakers and Panelists

The confirmed speakers bring expertise across foundational generative modeling, alignment, and applied domains (e.g., medical imaging, molecules, language/multimodal systems) and represent universities, research institutes, and industry labs. Their geographic distribution (e.g., USA, Canada, Israel, South Korea) and varied seniority levels (from senior researchers to rising scholars) further broadens the set of viewpoints featured in the program.

Diversity of topics

The workshop centers on post-training and inference-time alignment for diffusion and flow-based models, covering theory (e.g., distribution tilting, stochastic optimal control), methods (e.g., reward-guided generation, test-time scaling, discrete diffusion guidance), and applications (e.g., robotics, medical/scientific imaging, tabular data, language and multimodal reasoning, protein/molecule design). This breadth invites participation from applied mathematics, statistics, computer science, and domain sciences, encouraging cross-disciplinary interaction and new collaborations.

Inclusion

We are committed to a welcoming, harassment-free environment for all participants regardless of gender identity, sexual orientation, disability, physical appearance, race, nationality, or religion. We will support remote engagement via live streaming (SlidesLive platform) and public materials shared via our website, for those unable to attend in person. We also aim to seek support from industry companies, to establish a travel and registration assistance program for attendees facing financial barriers, directly offsetting expenses such as the conference registration fee.

Funding and Sponsorship

We have already secured support from NVIDIA to provide GPUs to the best spotlight/poster awardees. We plan to conctat more companies to secure funding for travel awards and registration fee coverage.

Previous related workshops

In recent years, several workshops have tackled advances in generative modeling and probabilistic inference that intersect with our theme. Notably, the Workshop on Diffusion Models at NeurIPS 2023 centered around core advances across images, video, audio, 3D, and science. A recurring series, Structured Probabilistic Inference and Generative Modeling (SPIGM) (ICML 2024, NeurIPS 2025) has addressed theory, methodology, and applications at the intersection of inference and generative modeling. The Deep Generative Model in Machine Learning: Theory, Principle and Efficacy workshop at ICLR 2025 emphasized foundations-expressivity, reliability, and efficiency of deep generative models. Finally, the Frontiers in Probabilistic Inference: Sampling Meets Learning (FPI) at ICLR 2025 focused on modern sampling and learning-augmented inference-explicitly including sampling for weighted/targeted generative modeling.

These previous related workshops have touched upon topics adjacent to our workshop. Namely, NeurIPS 2023's diffusion workshop tracked general progress in diffusion modeling rather than post-training constraint satisfaction or preference alignment. SPIGM's 2024/2025 editions cover structured inference and generative modeling but do not specifically center on test-time tilting/steering of pretrained diffusion and flow models. Likewise, ICLR 2025's FPI workshop focused on scalable sampling and learning-meets-sampling methods across domains, offering valuable tools but not a dedicated forum for constraint-driven, post-training alignment in diffusion/flow generators.

While these workshops established crucial foundations in diffusion, generative modeling, and probabilistic inference, our proposed workshop directly addresses the rapid shift of the field toward *the use of large pretrained generative models, where researchers/practitioners must meet safety, preference, and hard-constraint requirements at post-training phase and at inference time.* To this end, this workshop addresses an emerging need by focusing on distribution tilting, steering and guidance, and reward-/preference-based alignment of pretrained diffusion and flow models, providing a timely complement to the broader foundational workshops above.

Organizers and biographies

Our organizing team includes both established researchers and rising scholars in generative modeling, diffusion and flow-based models, inverse problems, and alignment. Our organizing team members (sorted alphabetically) are:

- **Charlotte Bunne (EPFL, Switzerland)**: Charlotte Bunne is an assistant professor at EPFL in the Computer Science and Life Sciences departments. Before, she was a PostDoc at Genentech and Stanford with Aviv Regev and Jure Leskovec and completed a PhD in Computer Science at ETH Zurich working with Andreas Krause and Marco Cuturi. Her research aims to advance biomedical science and personalized medicine through the application of machine learning and large-scale biomedical data. Prof. Charlotte Bunne has extensive experience in developing optimal transport and Schrödinger bridge–based methods and applying them to model temporal biological data. Prof. Charlotte Bunne has been co-organizer of workshops at ICLR, ICML and NeurIPS in the past.

- **Giannis Daras (MIT, USA)**: Giannis Daras is a Post-doctoral Associate at MIT supervised by Prof. Costis Daskalakis and Antonio Torralba. Giannis obtained his Ph.D. from the Computer Science department of UT Austin under the supervision of Prof. Alexandros Dimakis. Giannis works on practical and theoretical questions around deep generative models, with a focus on developing algorithms that enable training and sampling in challenging data regimes. Giannis has published multiple papers on this topic, and this line of work has been recognized with a Best Contribution Award at BASP, an Oral presentation and two Spotlight presentations at NeurIPS.
- **Paris Giampouras (University of Warwick, UK)**: Paris Giampouras is an Assistant Professor of Machine Learning/AI in the Department of Computer Science at the University of Warwick and an Affiliated Researcher at the Archimedes AI Research Center. Previously, he was Research Faculty at the Mathematical Institute for Data Science (MINDS) at Johns Hopkins University, where he also held a Marie Skłodowska-Curie postdoctoral fellowship (2019–2022). He has served as a Program Committee member of workshops at NeuRIPS and ICLR, and Area Chair for top-tier ML venues, including ICLR (2025, 2026), AAMAS 2025, and CPAL (2024, 2025). His recent research focuses on solving inverse problems with pretrained deep generative models, representational alignment, continual learning, and robustness.
- **Yingzehn Li (Imperial College, UK)**: Yingzhen Li is an Associate Professor in Machine Learning at the Department of Computing, Imperial College London, UK. She has worked extensively on approximate inference methods with applications to Bayesian deep learning and deep generative models. Her work has been recognized by e.g., AAAI 2023 New Faculty Highlights and invited tutorials at NeurIPS 2020 and UAI 2025. She was a co-organizer of the Advances in Approximate Bayesian Inference (AABI) symposium in 2020-2023, as well as many NeurIPS/ICML/ICLR workshops on topics related to probabilistic learning. She is a Program Chair for AISTATS 2024 and a General Chair for AISTATS 2025 and 2026.
- **Morteza Mardani (NVIDIA Research, USA)**: Morteza Mardani is a Principal Scientist at NVIDIA Research, where he leads work in generative AI. He is also a visiting researcher at Stanford University, where he previously served as a postdoctoral scholar and research associate. His research bridges theory and practice in generative modeling, with a current focus on diffusion and flow models. His contributions to generative modeling and statistical learning have been recognized with several awards, including the 2017 IEEE Signal Processing Society Young Author Best Paper Award. He is also an IEEE Distinguished Industry Speaker.
- **Johann Wenckstern (EPFL, Switzerland)**: Johann Wenckstern is a PhD student in Computer Science at EPFL, supervised by Charlotte Bunne. He previously earned his Master's degree in Mathematics from ETH Zurich and worked as a visiting graduate researcher at the Broad Institute of MIT and Harvard with Gad Getz and David Sontag. His current research focuses on developing generative models to forecast the dynamics of biological systems by integrating spatial, temporal, and multi-modal data.

The prior experience of our organizing team members in running workshops, conferences, and other research activities positions us well for a successful workshop. Charlotte Bunne will steer the program committee and link the agenda to computational biology applications, capitalizing on her SB/OT expertise and prior workshop-organizing experience. Giannis Daras will contribute to the technical program in the areas of inverse problems, and diffusion algorithms. Paris

Giampouras will coordinate reviewer recruitment and meta-reviewing to ensure a rigorous process, and will contribute to the technical program in the areas of conditional generation and safety alignment. Yingzhen Li will guide review policy, mentoring for junior reviewers, and session design, leveraging her leadership across AABI and AISTATS. Morteza Mardani will lead industry engagement and sponsorship outreach and help align evaluation with real-world practice, including coordinating the NVIDIA GPU awards for best posters/spotlights. He will also lead the technical program in the fields of diffusion/flow alignment and RL-based methods. Johann Wenckstern will contribute to the reviewing process and assist with organizational logistics.

References

Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making? *arXiv preprint arXiv:2211.15657*, 2022.

Gautham Govind Anil, Shaan Ul Haque, Nithish Kannen, Dheeraj Nagaraj, Sanjay Shakkottai, and Karthikeyan Shanmugam. Fine-tuning diffusion models via intermediate distribution shaping. *arXiv preprint arXiv:2510.02692*, 2025.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 843–852, 2023.

Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083*, 2024a.

Giannis Daras, Weili Nie, Karsten Kreis, Alex Dimakis, Morteza Mardani, Nikola Kovachki, and Arash Vahdat. Warped diffusion: Solving video inverse problems with image diffusion models. *Advances in Neural Information Processing Systems*, 37:101116–101143, 2024b.

Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023.

Jiajun Fan, Shuaike Shen, Chaoran Cheng, Yuxin Chen, Chumeng Liang, and Ge Liu. Online reward-weighted fine-tuning of flow matching with wasserstein regularization. In *The Thirteenth International Conference on Learning Representations*, 2025.

Berthy T Feng, Jamie Smith, Michael Rubinstein, Huiwen Chang, Katherine L Bouman, and William T Freeman. Score-based diffusion models as principled priors for inverse imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10520–10531, 2023.

Nate Gruver, Samuel Stanton, Nathan Frey, Tim GJ Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew G Wilson. Protein design with guided discrete diffusion. *Advances in neural information processing systems*, 36:12489–12517, 2023.

Aaron Havens, Benjamin Kurt Miller, Bing Yan, Carles Domingo-Enrich, Anuroop Sriram, Brandon Wood, Daniel Levine, Bin Hu, Brandon Amos, Brian Karrer, et al. Adjoint sampling:

Highly scalable diffusion samplers via adjoint matching. *arXiv preprint arXiv:2504.11713*, 2025.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. Ravel: Evaluating interpretability methods on disentangling language model representations. *arXiv preprint arXiv:2402.17700*, 2024.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? a case study in sparse probing. *arXiv preprint arXiv:2502.16681*, 2025.

Valentin Khrulkov, Gleb Ryzhakov, Andrei Chertkov, and Ivan Oseledets. Understanding ddpm latent codes through optimal transport. *arXiv preprint arXiv:2202.07477*, 2022.

Minu Kim, Yongsik Lee, Sehyeok Kang, Jihwan Oh, Song Chong, and Se-Young Yun. Preference alignment with flow matching. *Advances in Neural Information Processing Systems*, 37: 35140–35164, 2024a.

Sanghyun Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. Safety alignment backfires: Preventing the re-emergence of suppressed concepts in fine-tuned text-to-image diffusion models. *arXiv preprint arXiv:2412.00357*, 2024b.

Guangyuan Li, Chen Rao, Juncheng Mo, Zhanjie Zhang, Wei Xing, and Lei Zhao. Rethinking diffusion model for multi-contrast mri super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11365–11374, 2024a.

Xirui Li, Charles Herrmann, Kelvin CK Chan, Yinxiao Li, Deqing Sun, Chao Ma, and Ming-Hsuan Yang. A simple approach to unifying diffusion-based conditional generation. *arXiv preprint arXiv:2410.11439*, 2024b.

Tongyu Liu, Ju Fan, Nan Tang, Guoliang Li, and Xiaoyong Du. Controllable tabular data synthesis using diffusion models. *Proceedings of the ACM on Management of Data*, 2(1):1–29, 2024.

Jinqi Luo, Tianjiao Ding, Kwan Ho Ryan Chan, Darshan Thaker, Aditya Chattopadhyay, Chris Callison-Burch, and René Vidal. Pace: Parsimonious concept engineering for large language models. *Advances in Neural Information Processing Systems*, 37:99347–99381, 2024.

Yuchen Ma, Valentyn Melnychuk, Jonas Schweisthal, and Stefan Feuerriegel. Diffpo: A causal diffusion model for learning distributions of potential outcomes. *Advances in Neural Information Processing Systems*, 37:43663–43692, 2024.

Niels Mündler, Jasper Dekoninck, and Martin Vechev. Constrained decoding of diffusion llms with context-free grammars. *arXiv preprint arXiv:2508.10111*, 2025.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Byoungwoo Park, Jungwon Choi, Sungbin Lim, and Juho Lee. Stochastic optimal control for diffusion bridges in function spaces. *Advances in Neural Information Processing Systems*, 37: 28745–28771, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Vignav Ramesh and Morteza Mardani. Test-time scaling of diffusion models via noise trajectory search. *arXiv preprint arXiv:2506.03164*, 2025.

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo P de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. *arXiv preprint arXiv:2412.10193*, 2024.

Zitao Shuai, Chenwei Wu, Zhengxu Tang, Bowen Song, and Liyue Shen. Latent space disentanglement in diffusion transformers enables precise zero-shot semantic editing. *arXiv preprint arXiv:2411.08196*, 2024.

Tarun Suresh, Debangshu Banerjee, Shubham Ugare, Sasa Misailovic, and Gagandeep Singh. Dingo: Constrained inference for diffusion llms. *arXiv preprint arXiv:2505.23061*, 2025.

Wenpin Tang. Fine-tuning of diffusion models via stochastic control: entropy regularization and beyond. *arXiv preprint arXiv:2403.06279*, 2024.

Masatoshi Uehara, Yulai Zhao, Tommaso Biancalani, and Sergey Levine. Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review. *arXiv preprint arXiv:2407.13734*, 2024.

Masatoshi Uehara, Yulai Zhao, Chenyu Wang, Xiner Li, Aviv Regev, Sergey Levine, and Tommaso Biancalani. Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review. *arXiv preprint arXiv:2501.09685*, 2025.

Andrew Wagenmaker, Mitsuhiko Nakamoto, Yunchu Zhang, Seohong Park, Waleed Yagoub, Anusha Nagabandi, Abhishek Gupta, and Sergey Levine. Steering your diffusion policy with latent space reinforcement learning. *arXiv preprint arXiv:2506.15799*, 2025.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

Jin Wang, Yao Lai, Aoxue Li, Shifeng Zhang, Jiacheng Sun, Ning Kang, Chengyue Wu, Zhenguo Li, and Ping Luo. Fudoki: Discrete flow-based unified understanding and generation via kinetic-optimal velocities. *arXiv preprint arXiv:2505.20147*, 2025.

Jiacheng Ye, Shansan Gong, Liheng Chen, Lin Zheng, Jiahui Gao, Han Shi, Chuan Wu, Xin Jiang, Zhenguo Li, Wei Bi, et al. Diffusion of thought: Chain-of-thought reasoning in diffusion language models. *Advances in Neural Information Processing Systems*, 37:105345–105374, 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*, 2025.