FlowLensing: Simulating Gravitational Lensing with Flow Matching

Hamees Sayed[1,2] **Pranath Reddy**[3] **Michael W. Toomey**[4] **Sergei Gleyzer**[5]

[1]Smallest AI [2]Indian Institute of Technology Madras, India
[3]University of Florida, Gainesville, FL 32611, USA
[4]Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[5]Department of Physics and Astronomy, University of Alabama, Tuscaloosa, AL 35401, USA

hamees@smallest.ai kumbam.pranath@ufl.edu mtoomey@mit.edu sgleyzer@ua.edu

Abstract

Gravitational lensing is one of the most powerful probes of dark matter, yet creating high-fidelity lensed images at scale remains a bottleneck. Existing tools rely on ray-tracing or forward-modeling pipelines that, while precise, are prohibitively slow. We introduce FlowLensing, a Diffusion Transformer-based compact and efficient flow-matching model for strong gravitational lensing simulation. FlowLensing operates in both discrete and continuous regimes, handling classes such as different dark matter models as well as continuous model parameters ensuring physical consistency. By enabling scalable simulations, our model can advance dark matter studies, specifically for probing dark matter substructure in cosmological surveys. We find that our model achieves a speedup of over 200$\times$ compared to classical simulators for intensive dark matter models, with high fidelity and low inference latency. FlowLensing enables rapid, scalable, and physically consistent image synthesis, offering a practical alternative to traditional forward-modeling pipelines.

1 Introduction

Gravitational lensing [1] occurs when light from a distant galaxy or quasar is bent by the gravity of an intervening object, typically by a galaxy and its more massive dark matter halo. This phenomenon distorts and magnifies the background source, sometimes producing multiple images, and has become a powerful tool to probe the distribution of dark matter and test competing particle models.

Simulating realistic lensing images at scale, however, is computationally demanding. Existing tools like lenstronomy [2] and PyAutoLens [12] can generate high fidelity images by solving the lens equation via ray tracing or grid-based solvers, but their high cost for complex dark matter models makes them unsuitable for large statistical studies – in particular when attempting to study affects from dark matter substructure. Deep learning based generative models offer speedups, yet current approaches often struggle with fidelity, conditioning control, or slow inference due to long sampling chains.

To address these challenges, we propose FlowLensing, a flow-matching [11] model with diffusion transformer (DiT) [14] as the backbone that learns a direct mapping from astrophysical parameters to lensed images. Flow matching avoids iterative denoising, enabling faster and more stable sampling. Our method captures both broad dark matter scenarios and fine-grained lens properties, producing sharper, physically consistent images. As further discussed in Section 4.2, strong classification and regression results serve as indirect evidence of physics consistency. Overall, our approach dramatically reduces inference time, making it a practical alternative to classical simulators.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Machine Learning and the Physical Sciences.

2 Datasets

To train `FlowLensing`, we used two simulated datasets of gravitational lensing images, generated with `lenstronomy` to mimic Euclid-like survey observations. These datasets also capture dark matter substructure effects enabling tests of physical consistency and model performance. All images are 64×64 pixels and normalized to $[-1, 1]$ for training. Details are provided in the subsections below.

2.1 Dark Matter Model Conditioned (Discrete)

The first dataset is composed of simulated galaxy-galaxy strong lensing images that are generated using the publicly available simulation pipeline `lenstronomy`. There are 89,104 elements in the dataset, where every image is 64 x 64 pixels and is meant to resemble an observation typical of a Euclid-like survey. Furthermore, each simulated host lens is modeled as a sheared isothermal ellipse [4], while each source is described via a Sersic light profile [3]. Every image within the dataset falls into one of three categories, each defined by a different dark matter substructure class. The first class is a baseline that involves no simulated substructure and has only a CDM [19] host halo as the deflector, while the second class assumes CDM substructure modeled as truncated NFW haloes [7]. Finally, the third class models axionic dark matter [13] with $m \approx 10^{-23}$ eV as vortex-like defects.

2.2 Lensing Model Parameters (Continuous)

The second dataset has 30,000 elements and consists of only CDM images, with and without substructure, that were produced in the same way as the first dataset. So while all lensing simulations consist of the same dark matter type, a set of continuous parameters of the lens-source system were regressively extracted to define each one: einstein radius (θ_E), the coordinates of the source in the image plane with respect to the center of the image (x, y), and the slope of the subhalo mass function (β). Conditioning each simulated image in this way will allow for a more physically consistent performance from the flow matching model [20].

3 Methodology

3.1 Flow Matching with Diffusion Transformer

Flow matching is a continuous-time generative modeling process that learns to transform samples from a simple prior distribution (typically Gaussian noise) to a target data distribution through a continuous flow. Unlike traditional diffusion models that rely on a fixed noising schedule, flow matching directly learns the vector field that guides the transformation process. Compared to score-based diffusion methods [15], flow matching avoids the need for stochastic differential equations and instead directly estimates the velocity field that transports particles along deterministic paths.

Given a pair of data $x_1 \sim p_{\text{data}}$ and noise $x_0 \sim \mathcal{N}(0, I)$, the interpolant at time $t \in [0, 1]$ is defined as

$$x_t = (1 - t)x_0 + tx_1.$$

The corresponding optimal target velocity is

$$v(x_t, t) = x_1 - x_0,$$

and the training objective is to minimize the mean squared error between the predicted velocity $v_\theta(x_t, t, c)$ and the ground truth $v(x_t, t)$:

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0,x_1,t,c}\left[\|v_\theta(x_t, t, c) - (x_1 - x_0)\|^2\right],$$

where c represents the conditioning signal.

We implement v_θ using a DiT, which operates on image patches with self-attention [17] and integrates conditioning information through adaptive layer normalization (adaLN) [5]. To improve controllability, we apply classifier-free guidance [10] by randomly dropping conditioning during training with probability $p_{drop} = 0.1$. At inference, conditional and unconditional predictions are combined as

$$\tilde{v}_\theta(x_t, t, y) = v_\theta(x_t, t, \emptyset) + w \cdot (v_\theta(x_t, t, y) - v_\theta(x_t, t, \emptyset)),$$

where w is the guidance weight and $\emptyset$ denotes the unconditional case.

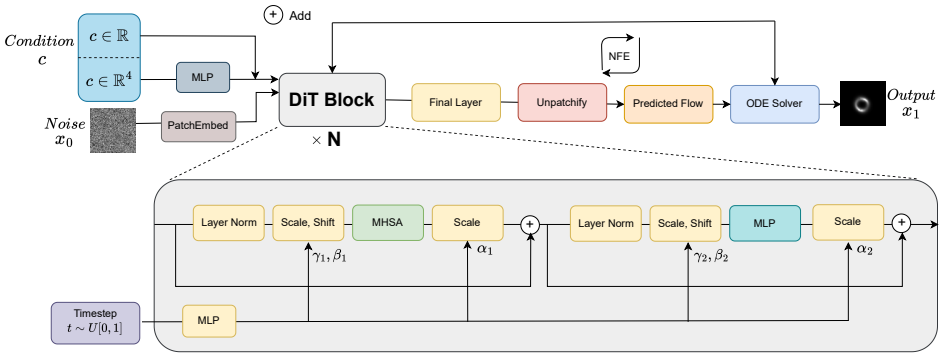

Figure 1: Schematic of `FlowLensing` inference.

3.2 Experiment 1: Dark Matter Model Conditioned Generation

In our first experiment, we condition `FlowLensing` on discrete classes representing different dark matter models. The conditioning variable c takes one of three categorical values (CDM,axion,no substructure) mentioned in Section 2.1. The conditioning is implemented through a learned embedding table that maps discrete class labels to the model's hidden dimension. During training, we apply classifier-free guidance by randomly replacing class labels with a special *'null'* token, enabling unconditional generation and guidance during sampling.

3.3 Experiment 2: Lensing Parameter Conditioned Generation

This experiment extends conditioning to continuous lensing parameters for fine-grained control and interpolation. The conditioning vector $c \in \mathbb{R}^4$ is projected into the model's hidden space using a multi-layer perceptron (MLP). Classifier-free guidance is applied by randomly masking the conditioning vector to zeros during training, with a learned null embedding used for unconditional cases. This setup enables the model to smoothly traverse parameter space and generate physically plausible lensing systems.

4 Results

We evaluate `FlowLensing` on two simulated datasets: one with discrete dark matter classes (Section 2.1) and another with continuous lensing parameters (Section 2.2). Performance is assessed using classification metrics (AUC) for discrete classes and regression metrics (R^2) for continuous parameters, alongside image quality metrics to quantify reconstruction fidelity. We also benchmark inference latency to highlight computational efficiency. All evaluations use a 30M-parameter model with a classifier-free guidance (CFG) scale of 2 and 100 denoising steps via an Euler ODE solver.

4.1 Image Quality Metrics

To quantitatively evaluate the fidelity of generated images, we report standard reconstruction and perceptual metrics. Mean Squared Error (MSE) measure pixel-level accuracy, Peak Signal-to-Noise Ratio (PSNR) captures overall signal quality, and Structural Similarity Index (SSIM) [18] quantifies perceptual similarity. We additionally report Fréchet Inception Distance (FID) [8] for completeness, although we note that the Inception model [16] used in FID is not trained on our domain-specific data and therefore may not be an ideal indicator of astrophysical realism. Inference efficiency is also critical: our model requires only 100 denoising steps versus 1000 for the baseline, achieving a $\sim$13.3$\times$ speedup. Table 1 summarizes reconstruction quality and single-sample generation latency and in Figure 2, we provide a side-by-side comparison of real and generated images to qualitatively assess fidelity.

The baseline model for comparison is a DDPM [9] with a U-Net [5] backbone, evaluated with 1000 NFE steps at inference, whereas our method achieves competitive performance using only 100 NFE steps. This highlights both the efficiency and effectiveness of our approach.

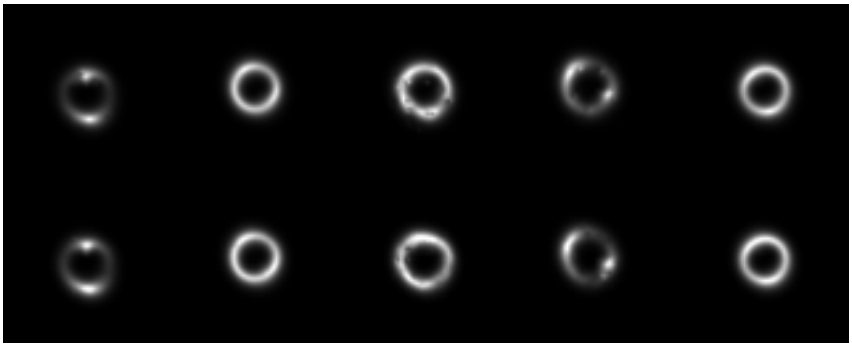

Figure 2: Real (top) vs. generated (bottom) images from `FlowLensing`.

Model	MSE ↓	FID ↓	Latency (s) ↓	PSNR ↑	SSIM ↑
Ours	**0.0108**	**1.614**	**0.36**	**68.68**	**0.9993**
Baseline	0.0110	87.312	4.8	30.78	0.8870

Table 1: Comparison of reconstruction quality metrics against the baseline.

4.2 Downstream Evaluation: Classification and Regression

To evaluate the utility of the learned representations, we assess two downstream tasks: classification and regression. For classification, a ResNet18 [6] classifier trained on the original simulated dataset is evaluated on images produced by our model. For regression, the final fully connected layer of ResNet18 is replaced with a 4-dimensional linear layer to predict astrophysical parameters, serving primarily as a sanity check for parameter recoverability. The strong results across both tasks provide indirect evidence that our model preserves underlying physical consistency. The outcomes are summarized in Table 2.

Class. (AUC)			**Reg. (R^2)**			
Class	**Ours**	**Base.**	$\boldsymbol{x}$	$\boldsymbol{y}$	$\boldsymbol{\beta}$	$\boldsymbol{\theta_E}$
CDM	**1.00**	0.92	0.945	0.940	0.833	Constant (1.281)
Axion	**1.00**	0.91				
No Substructure	**1.00**	0.75				

Table 2: Downstream evaluation: classification (AUC) and regression (R^2).

5 Conclusion and Future Work

Our work introduced `FlowLensing`, a flow-matching model with a diffusion transformer backbone that generated high-fidelity gravitational lensing images over 200× faster than traditional simulators (0.36s vs. 4.8s per sample). Conditioned on dark matter models (CDM, axion, no substructure) and parameters like the subhalo mass function slope β, it achieved superior image quality (PSNR: 68.68, SSIM: 0.9993) and physical accuracy, with perfect classification AUC (1.00) and strong regression R^2 scores (0.833–0.945; Section 4.2). By enabling scalable, realistic simulations, `FlowLensing` advances dark matter studies, recovering β to distinguish CDM from axion models and probe small-scale structures in surveys like Euclid.

Looking ahead, we aim to weave lensing equations into the model's architecture for deeper physical fidelity, reducing reliance on post hoc checks. We also plan to benchmark `FlowLensing` against GANs and VAEs to explore its strengths across generative approaches.

6 Acknowledgements

We acknowledge useful conversations with Pranath Reddy. H.S. was a participant in the Google Summer of Code 2025 program. S.G. was supported in part by U.S. National Science Foundation award No. 2108645. Portions of this work were conducted in MIT's Center for Theoretical Physics and partially supported by the U.S. Department of Energy under grant Contract Number DE-SC0012567. M.W.T is supported by the Simons Foundation (Grant Number 929255).

References

[1] Matthias Bartelmann. Gravitational lensing. *Classical and Quantum Gravity*, 27(23):233001, November 2010.

[2] Simon Birrer and Adam Amara. lenstronomy: Multi-purpose gravitational lens modelling software package. *Physics of the Dark Universe*, 22:189–201, December 2018.

[3] Vincenzo Cardone. The lensing properties of the sersic model. *Astronomy and Astrophysics*, 415, 11 2003.

[4] Matthew R Gomer and Liliya L R Williams. Galaxy-lens determination of h0: the effect of the ellipse + shear modelling assumption. *Monthly Notices of the Royal Astronomical Society*, 504(1):1340–1354, 04 2021.

[5] Yunhui Guo, Chaofeng Wang, Stella X. Yu, Frank McKenna, and Kincho H. Law. Adaln: A vision transformer for multidomain learning and predisaster building information extraction from images. *J. Comput. Civ. Eng.*, 36(5), 2022.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pages 770–778. IEEE, June 2016.

[7] Felix M Heinze, Giulia Despali, and Ralf S Klessen. Not all subhaloes are created equal: modelling the diversity of subhalo density profiles in tng50. *Monthly Notices of the Royal Astronomical Society*, 527(4):11996–12015, 12 2023.

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[11] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[12] J W Nightingale, S Dye, and Richard J Massey. Autolens: automated modeling of a strong lens's light, mass, and source. *Monthly Notices of the Royal Astronomical Society*, 478(4):4738–4784, 05 2018.

[13] Ciaran A. J. O'Hare. Cosmology of axion dark matter. *PoS*, COSMICWISPers:040, 2024.

[14] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

[15] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[16] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4278–4284. AAAI Press, 2017.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[18] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[19] David H. Weinberg, James S. Bullock, Fabio Governato, Rachel Kuzio de Naray, and Annika H. G. Peter. Cold dark matter: Controversies on small scales. *Proceedings of the National Academy of Sciences*, 112(40):12249–12255, February 2015.

[20] Jonas Bernhard Wildberger, Maximilian Dax, Simon Buchholz, Stephen R Green, Jakob H. Macke, and Bernhard Schölkopf. Flow matching for scalable simulation-based inference. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

A Training Setup

We trained our flow-matching model using grayscale lensing images, normalized to the range $[-1, 1]$. The model operates in a latent space of dimension 512, with patch-based processing and a transformer backbone. An exponential moving average (EMA) of model weights was maintained throughout training to improve stability and sample quality. Table 3 summarizes the key architectural and optimization choices.

Table 3: Model and training hyperparameters.

Parameter	**Value**
Latent dimension	512
Patch size	2
Model depth	6
Attention heads	8
Optimizer	AdamW
Learning rate	1×10^{-5}
Batch size	128
Training epochs	300
EMA	Enabled

Training was performed on a single NVIDIA RTX A6000 Ada GPU using PyTorch. Each training run spanned 300 epochs with a batch size of 128. The AdamW optimizer was employed with a constant learning rate of 1×10^{-5}. All experiments were conducted in mixed precision to balance efficiency and numerical stability.