




ASTRORIM: JOINT TRAINING OF A CONDITIONAL FORWARD OPERATOR WITH A RECURRENT INFERENCE MACHINE FOR SOURCE OBJECT RECONSTRUCTION IN GRAVITATIONAL LENS SYSTEMS

JACK WALSH 
 Greystones Community College

JOHN BRENNAN 
 Maynooth University

JOHN REGAN 
 Maynooth University

CREIDHE O’SULLIVAN 
 Maynooth University

ABSTRACT

Gravitational lensing provides a magnified probe of distant galaxies and dark matter, but recovering unlensed source morphologies and mass profile diagrams from observed lensed images is an ill-posed inverse problem subject to model degeneracies and time consuming parametric modeling. We present **AstroRIM**, a compact ($\sim 650k$ parameters) iterative reconstruction system that *jointly* trains a Recurrent Inference Machine (RIM) with a differentiable, physics-parameterized forward operator whose *parameters are predicted per observation* by a LensParameterEncoder. The encoder maps each observed lens-plane image to nine physically-interpretable parameters $\{b, q, \phi, x_0, y_0, \gamma, \gamma_\phi, \kappa_s, r_s\}$ that condition the forward operator (SIE + external shear + NFW halo; PSF + noise). Conditioning the forward model on *per-lens* predictions rather than using a fixed global forward operator, which substantially reduces our forward-model mismatch and improves generalization across Einstein radii $0.3''$ – $10.0''$, ellipticities 0.0 – 0.7 , PSF families (Gaussian/Moffat/Airy) and diverse source morphologies. Joint optimisation uses a *forward fidelity* term, which enforces that predicted parameters reproduce the observation from the ground-truth source. We train on roughly 500k simulated pairs (six simulator variants) with differential learning rates and stabilization (gradient clipping, EMA, AMP). On held-out synthetic data AstroRIM achieves $\text{MSE} \approx 3.7 \times 10^{-4}$ and $\text{SSIM} \approx 0.951$, and provides promising initial demonstrations of both source-plane object recovery, and mass profiling on real HST/Castles/Euclid images (Muñoz *et al.* 1999). The code and simulation recipes are publicly available to encourage reproducibility.

20jwalsh@greystonescollege.ie
 John.Brennan@mu.ie
 John.Regan@mu.ie
 Creidhe.OSullivan@mu.ie

1. INTRODUCTION

1.1. Scientific context and the inverse problem

Strong gravitational lensing bends and magnifies light from background sources, producing highly informative yet geometrically distorted images that sensitively encode the foreground mass distribution and the unlensed source morphology. This astrophysical magnification has been used to study galaxy structure, dark matter substructure, and cosmology (e.g., stellar and dark mass decomposition, time-delay cosmography), and it will play an even larger role as near-future surveys (Euclid, Roman, Rubin, JWST follow-up) deliver orders of magnitude more lenses (Treu 2010; Schneider *et al.* 2006; Collett 2015; Weiner *et al.* 2020).

In layman’s terms, reconstructing the *source plane* image x from a noisy observed lens-plane image y is an inverse problem governed by the lens equation,

$$\beta(\theta) = \theta - \alpha(\theta), \quad (1)$$

where $\alpha(\theta)$ is the deflection field produced by the lens mass distribution and the mapping from x to y includes PSF convolution and detector effects. The inverse task is ill-conditioned and typically underdetermined: different combinations of source structure and lens potential can produce similar y (mass–source degeneracies), and observational nuisances (PSF mismatch, correlated noise, lens light residuals) further complicate reliable recovery. Traditional fully Bayesian forward-modeling approaches (e.g., pixelated Bayesian reconstructions, nested-sampling / MCMC of parametric lens models) provide principled uncertainty quantification, but are computationally expensive and scale poorly to the very large samples expected in modern surveys (Suyu *et al.* 2017; Vegetti and Koopmans 2010).

1.2. Machine learning approaches and limitations

Machine learning (ML) methods have been actively investigated as faster alternatives for lens parameter estimation and source reconstruction. Convolutional neural networks (CNNs) have been shown to regress parametric lens parameters from images with dramatic speedups (Hezaveh *et al.* 2017): such methods predict global lens parameters rapidly but typically assume a fixed forward model and often do not recover full pixelated source morphologies. Simulation-based inference and density-estimation approaches (including normalizing flows and likelihood-free inference) provide calibrated posteriors under a simulation prior, but can be sensitive to simulation mismatch and remain limited when the data-model gap is large (Legin *et al.* 2021; Perreault Levasseur *et al.* 2017). Iterative learned solvers, in particular the Recurrent Inference Machine (RIM), learn an iterative update operator that integrates forward-model gradients into recurrent updates and have shown strong performance on inverse imaging tasks (Putzky and Welling 2017; Morningstar *et al.* 2019).

Despite these advances there remains a consistent, practical gap: *most prior ML approaches treat the forward operator as fixed* (a simulator or analytic warping that is common to all training examples) or learn only the inverse mapping. Fixed forward models are brittle in the presence of PSF mismatch, unmodelled instrumental effects, or mass-model inaccuracies; conversely, purely black-box inverse networks can have millions of parameters and limited interpretability. A method that can adapt the forward operator per observation while preserving interpretable physics parameters and end-to-end trainability promises to reduce mismatch and combine speed with physical consistency.

1.3. *AstroRIM: conditional, per-lens forward operator + RIM*

We introduce **AstroRIM**, whose central novelty is a *conditional per-observation forward operator* whose physics parameters are predicted from each observed image by a *LensParameterEncoder*. The encoder outputs nine bounded, physically motivated parameters

$$\Theta(y_{\text{obs}}) = \{b, q, \phi, x_0, y_0, \gamma, \gamma_\phi, \kappa_s, r_s\},$$

which parametrize an SIE primary lens, external shear and an optional NFW halo, together with a parametric (Moffat + learned residual) PSF. These parameters condition a differentiable ray-tracing forward operator F_ϕ , and the RIM uses the *conditional adjoint* F_ϕ^\dagger at each iteration to compute gradient-like signals that guide iterative refinement of the source estimate. Critically, *the forward model is not fixed*: every observation receives its own forward parameters inferred from pixels, enabling AstroRIM to absorb instrument- and configuration-specific differences and reduce forward-model mismatch.

We train the encoder, forward operator and RIM jointly with a multi-component loss that includes a *forward-fidelity* term,

$$\mathcal{L}_{\text{fwd}} = \|F_\phi(x_{\text{GT}}, y_{\text{obs}}) - y_{\text{obs}}\|_2^2, \quad (2)$$

which acts as a self-consistency constraint: if the encoder predicts physically meaningful parameters for the observation, applying the forward operator to the true

source should reproduce the observation. This constraint, combined with the RIM reconstruction loss and PSF/parameter regularizers, encourages interpretable encoder outputs and reduces degenerate solutions.

1.4. *Contributions and roadmap*

Our main contributions are:

1. **Conditional per-lens architecture:** the *LensParameterEncoder* predicts physics parameters per observation that condition a differentiable forward operator, enabling a per-lens forward model instead of a single global operator.
2. **Forward fidelity loss and joint training:** a self-consistency term that enforces that predicted parameters reproduce observations from ground-truth sources and a stable training protocol (differential learning rates, gradient clipping, EMA, AMP).
3. **Simulation diversity and generalization:** training on 500k synthetic pairs generated from six simulator variants (PSF families, noise regimes, substructure probabilities and instrument regimes) to broaden domain coverage.
4. **Compact, interpretable design:** total model size $\sim 650\text{k}$ parameters (encoder $\sim 350\text{k}$, RIM $\sim 300\text{k}$, PSF small), while predicting physically-interpretable parameters rather than a huge black-box mapping.
5. **Comprehensive evaluation:** validation on held-out simulated data showing $\text{MSE} \approx 3.7 \times 10^{-4}$ and $\text{SSIM} \approx 0.951$, plus illustrative real-data runs and ablation studies exploring dark matter halo profiling, and real lensing data survey analysis.

The remainder of the paper is organized as follows. Section 2 reviews the necessary lensing physics and related ML approaches. Section 3 describes the AstroRIM architecture, loss terms and training schedule. Section 4 presents quantitative reconstruction and forward-operator diagnostics on simulated test sets and illustrative real-data examples; Section 5 discusses limitations and paths toward robust real-data deployment.

2. BACKGROUND

2.1. *Gravitational lensing physics*

We summarize the minimal strong-lensing formalism relevant for AstroRIM; a thorough treatment can be found in standard references (Schneider *et al.* 2006; Treu 2010).

2.1.1. *Lens equation and deflection*

Under the thin-lens approximation, photon deflections are described in angular coordinates. The lens equation maps image-plane coordinates θ to source-plane coordinates β ,

$$\beta(\theta) = \theta - \alpha(\theta), \quad (3)$$

where $\alpha(\theta) = \nabla\psi(\theta)$ is the deflection angle arising from the (projected) lensing potential $\psi(\theta)$. Image multiplicity, magnification, and distortions are controlled by the

Jacobian of the lens mapping. For smooth analytic mass models; the deflection and magnification can be written in closed form or evaluated numerically.

2.1.2. SIE (Singular/Softened Isothermal Ellipsoid) and external shear

Galaxy-scale lenses are commonly approximated by the Singular Isothermal Ellipsoid (SIE) or softened variations thereof. The SIE captures the approximately isothermal mass profiles observed in early-type galaxies and enables a simple mapping between the deflection scale parameter (commonly expressed via an Einstein radius θ_E or deflection strength b) and a one-dimensional velocity dispersion σ_v via

$$\theta_E = 4\pi \left(\frac{\sigma_v^2}{c^2} \right) \frac{D_{ls}}{D_s}, \quad (4)$$

where D_{ls} and D_s are angular-diameter distances (cosmology-dependent). Ellipticity is introduced by rescaling coordinates and rotating by a position angle ϕ ; external shear γ and orientation γ_ϕ account for tidal perturbations from groups or large-scale structure (Kormann *et al.* 1994; Keeton 2001).

2.1.3. NFW halo and substructure

Larger-scale dark-matter halos and group/cluster components are frequently modeled with Navarro–Frenk–White (NFW) profiles characterized by a scale radius r_s and characteristic density (or equivalently κ_s in convergence units). Small-scale perturbations (subhalos) are often modeled as truncated isothermal spheres or NFW subhalos and are of particular interest for dark-matter microphysics studies (Vegetti and Koopmans 2010). The combination of SIE + external shear + optional NFW component captures the dominant lensing physics at galaxy and group scales while keeping parameter dimensionality interpretable.

2.1.4. Observational effects: PSF and noise

Real images are not clean ray-traced maps: the lensed, convolved image is

$$y(\theta) = [K_{\text{PSF}} * x(\theta - \alpha(\theta))] + n(\theta) + b_{\text{sky}}, \quad (5)$$

where K_{PSF} denotes the PSF kernel (which may be approximated as Gaussian, Moffat or Airy with additional residual structure), $n(\theta)$ represents instrumental and photon noise (possibly correlated), and b_{sky} is the sky/background level. Accurate inversion therefore requires either excellent PSF/ noise modelling or methods robust to PSF/ noise mismatch; AstroRIM uses a hybrid parametric+residual PSF (global learnable FWHM, β , plus a small residual kernel) and the forward-fidelity term to reduce mismatch.

2.1.5. Choice of parameterization for conditioning

We parametrize the observation-conditioned forward operator by nine parameters,

$$\Theta = \{b, q, \phi, x_0, y_0, \gamma, \gamma_\phi, \kappa_s, r_s\},$$

which are compact yet sufficient to represent a broad class of galaxy-scale lens configurations encountered in surveys (Einstein radii $\sim 0.3\text{--}10.0''$, ellipticities $0\text{--}0.7$).

Predicting these parameters per-observation provides a physically-interpretable way for the forward operator to adapt to instrument- and lens-specific differences while keeping the conditioning dimensionality modest and the model compact.

2.2. Machine learning for lens inversion: prior art and taxonomy

This subsection classifies previous ML approaches and highlights their strengths and limitations relative to AstroRIM’s design goals.

2.2.1. CNN-based parameter regression

Convolutional neural networks trained to regress analytic lens parameters (SIE, power-law, shear) from images achieve dramatic speedups and reasonable accuracy for many tasks (e.g. Hezaveh *et al.* 2017 and follow-ups). These methods enable near-instant parameter estimates suitable for large survey pre-screening but often assume idealized simulations, require lens-light subtraction or multi-band pre-processing, and do not directly recover pixelated source morphologies (Hezaveh *et al.* 2017).

2.2.2. Variational / Bayesian / simulation-based inference

Variational methods, VAEs, normalizing flows and simulation-based inference (SBI) apply probabilistic modeling to provide posterior approximations and calibrated uncertainties for lens parameters (Legin *et al.* 2021; Karchev *et al.* 2022). These approaches are attractive for principled uncertainty quantification, but their performance depends on simulation realism and they can be computationally involved when jointly inferring high-dimensional pixelated sources and lens parameters. Recent works combine expressive priors (GPs, continuous neural fields, diffusion priors) with differentiable forward models to jointly infer lens and source posteriors (Karchev *et al.* 2022; Mishra-Sharma and Yang 2022).

2.2.3. Unrolled / learned iterative solvers and RIMs

Algorithm-unrolling (LISTA, Learned Primal–Dual, variational networks) and learned iterative solvers have become a powerful paradigm for inverse problems. Theoretical and empirical work (LISTA; Gregor & LeCun 2010) showed that structured unrolling can drastically accelerate inference; Learned Primal–Dual and variational networks extended this to non-linear forward operators and imaging modalities (Gregor and LeCun 2010; Adler and Öktem 2018; Hammernik *et al.* 2018). The Recurrent Inference Machine (RIM) framework (Putzky & Welling 2017) learns an iterative update rule driven by data-conditional gradient-like signals and has been successfully applied to lensing (Morningstar *et al.* 2019) where the RIM acts as a learned optimizer for pixelized sources (Putzky and Welling 2017; Morningstar *et al.* 2019).

2.2.4. Differentiable physics and forward models

Differentiable simulators and differentiable ray-tracing toolkits (e.g., Lenstronomy, differentiable JAX-based renderers) make it possible to propagate gradients through the forward model, enabling end-to-end learning

of components that interact with physics-based rendering (Birrer and Amara 2018; Biggio *et al.* 2023). Physics-informed methods (PINNs) incorporate governing equations directly into the loss to bias learning with physical constraints (Raissi *et al.* 2019); however, in lensing we usually prefer a *physics-parameterised differentiable forward operator* (warping + PSF + noise) rather than enforcing PDE residuals. Differentiable forward modeling enables joint optimization of encoder-forward-RIM systems, but a remaining gap is that very few past works simultaneously learn per-observation forward parameters from pixels while training an iterative inverse solver end-to-end — this is the niche AstroRIM targets (Birrer and Amara 2018; Raissi *et al.* 2019).

2.3. Gaps in prior work and motivation for conditional per-lens models

The literature shows three broad limitations that motivate AstroRIM’s design:

- **Fixed forward operators:** Many learned reconstructions assume a global forward model; this amplifies domain shift when real data deviate from simulator assumptions (PSF, noise correlations, small-scale structure).
- **Lack of interpretable conditioning:** Black-box image-to-image mappings may reconstruct plausible sources but do not provide directly interpretable physics parameters that can be reused for downstream science (mass models, substructure inference).
- **Training fragility of joint systems:** Jointly training inverse networks and forward models can be unstable without careful learning-rate schedules and regularization; yet when stabilized, joint training can reduce mismatch (the very thing fixed-forward methods struggle with).

AstroRIM addresses these gaps by coupling: (i) a compact LensParameterEncoder that predicts interpretable physics parameters from each observation, (ii) a differentiable forward operator conditioned on those parameters, and (iii) an iterative RIM inverse solver that receives conditional gradients at each iteration. The forward-fidelity loss provides explicit self-consistency that encourages the encoder to predict physically-meaningful parameters rather than abstract conditioning codes.

3. METHODS

3.1. Model Architecture

3.1.1. Overview

AstroRIM consists of two jointly-trained neural networks that work in tandem to solve the lensing inverse problem through a *conditional, per-lens* approach:

1. **The conditional physics-parameterized forward operator F_ϕ :** A differentiable model that *infers lens parameters from each observation* and uses these parameters to model gravitational warping, PSF convolution, and background addition. Critically, the lens parameters are *predicted per observation*, not fixed globally.

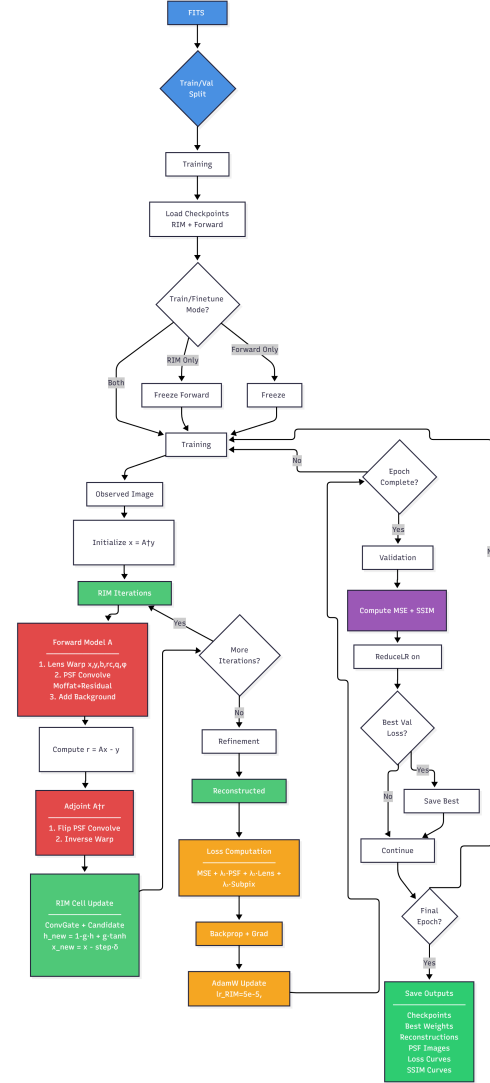


FIG. 1.— High-level training loop overview

2. The Recurrent Inference Machine (RIM):

A recurrent neural network parameterized by learnable weights θ that iteratively refines source estimates using gradient-like feedback from the *observation-conditioned* forward operator (Putzky and Welling 2017).

These components communicate at each iteration through a conditioning mechanism: the forward operator analyzes the observed lensed image to predict lens-specific parameters, then uses these parameters to project source estimates to observation space. The RIM processes residuals via the conditioned adjoint operator. Figure 1 illustrates this training loop.

The joint training paradigm means that both ϕ (including the lens parameter encoder) and θ are optimized simultaneously. The forward operator learns to predict physically meaningful lens parameters from observations while the RIM learns to exploit the forward model’s structure. A key innovation is the **forward fidelity loss**, which ensures the predicted parameters can accurately reproduce observations from ground-truth sources.

3.1.2. The Conditional Physics-Parameterized Forward Operator

The forward operator F_ϕ is the core innovation of AstroRIM: rather than using fixed or globally-shared lens parameters, it *predicts lens parameters from each observation* and uses these parameters to model the complete the lensing forward process. This per-observation conditioning enables the model to adapt to diverse lens configurations and instrumental effects without requiring explicit calibration.

The forward operator comprises three sequential stages: gravitational warping, PSF convolution, and background addition, all conditioned on observation-specific lens parameters predicted by a learned encoder; for a visual overview of this process, see Figure 2.

Stage 0: The Lens Parameter Encoder—The **LensParameterEncoder** is a convolutional neural network that maps an observed lensed image $\mathbf{y}_{\text{obs}} \in \mathbb{R}^{H \times W}$ to a compact set of nine physically-interpretable lens parameters:

$$\Theta(\mathbf{y}_{\text{obs}}) = \{b, q, \phi, x_0, y_0, \gamma, \gamma_\phi, \kappa_s, r_s\} \quad (6)$$

where:

- $b \in [0.01, 0.30]$: deflection strength (related to Einstein radius)
- $q \in [0.2, 1.0]$: axis ratio (ellipticity parameterization)
- $\phi \in [-\pi, \pi]$: position angle of lens major axis
- $x_0, y_0 \in [-0.2, 0.2]$: lens center (normalized coordinates)
- $\gamma \in [0.0, 0.2]$: external shear amplitude
- $\gamma_\phi \in [-\pi, \pi]$: external shear position angle
- $\kappa_s \in [0.005, 0.50]$: NFW characteristic convergence
- $r_s \in [0.05, 0.50]$: NFW scale radius (normalized coordinates)

The encoder architecture consists of four convolutional stages with progressively increasing channel dimensions, group normalization for training stability, and max pooling for spatial downsampling:

$$\mathbf{h}_1 = \text{Pool}(\text{ReLU}(\text{GN}_8(\text{Conv}_{3 \times 3}(\mathbf{y}_{\text{obs}}; 1 \rightarrow 32)))) \quad (7)$$

$$\mathbf{h}_2 = \text{Pool}(\text{ReLU}(\text{GN}_8(\text{Conv}_{3 \times 3}(\mathbf{h}_1; 32 \rightarrow 64)))) \quad (8)$$

$$\mathbf{h}_3 = \text{Pool}(\text{ReLU}(\text{GN}_8(\text{Conv}_{3 \times 3}(\mathbf{h}_2; 64 \rightarrow 96)))) \quad (9)$$

$$\mathbf{h}_4 = \text{Pool}(\text{ReLU}(\text{GN}_8(\text{Conv}_{3 \times 3}(\mathbf{h}_3; 96 \rightarrow 128)))) \quad (10)$$

where GN_8 denotes GroupNorm with 8 groups and Pool denotes 2×2 max pooling. After the convolutional stages, we apply adaptive average pooling to a fixed 8×8 spatial dimension, flatten, and pass through two fully-connected layers (Wu and He 2018):

$$\mathbf{z} = \text{AdaptivePool}_{8 \times 8}(\mathbf{h}_4) \in \mathbb{R}^{128 \times 8 \times 8} \quad (11)$$

$$\mathbf{f}_1 = \text{ReLU}(\text{FC}(\text{Flatten}(\mathbf{z}); 8192 \rightarrow 128)) \quad (12)$$

$$\mathbf{f}_2 = \text{ReLU}(\text{FC}(\mathbf{f}_1; 128 \rightarrow 64)) \quad (13)$$

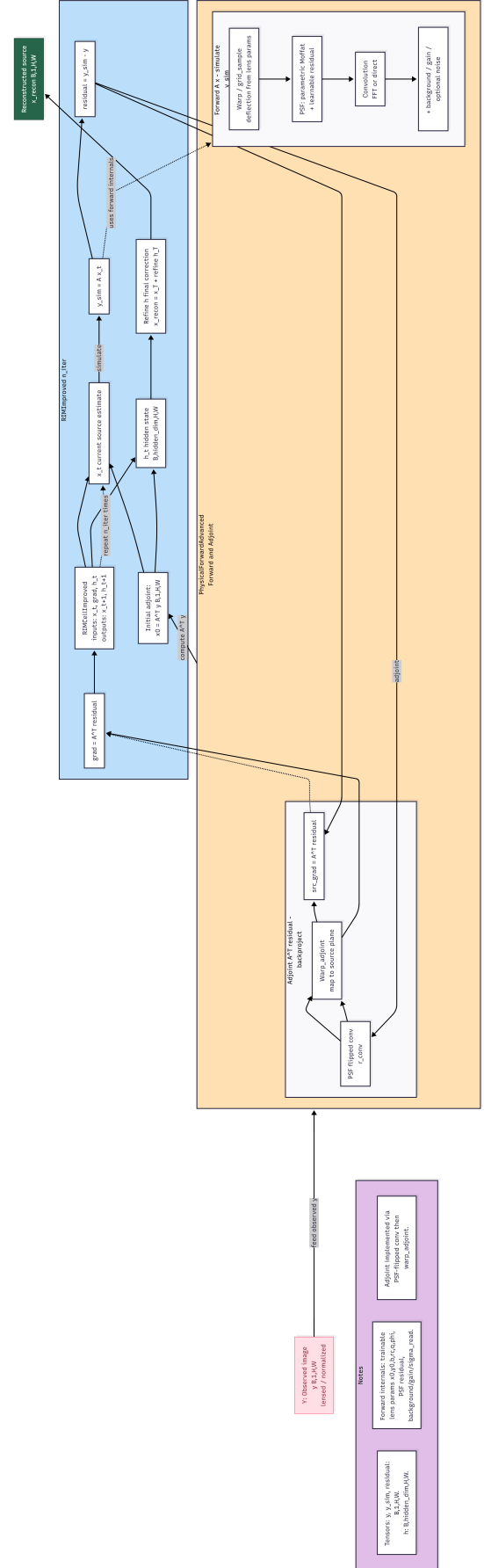


Fig. 2.— Physics-Parameterized Forward Operator structure

Parameter Prediction Heads with Physical Constraints—

From the final feature vector $\mathbf{f}_2 \in \mathbb{R}^{64}$, we predict each lens parameter through dedicated single-layer heads with parameter-specific activation functions that enforce physical bounds. For example, the deflection strength b is predicted as:

$$b = 0.01 + 0.29 \cdot \sigma(\text{FC}(\mathbf{f}_2; 64 \rightarrow 1)) \quad (14)$$

ensuring $b \in [0.01, 0.30]$ via the sigmoid function $\sigma(\cdot)$. Similarly, angular parameters use \tanh to constrain to $[-\pi, \pi]$:

$$\phi = \pi \cdot \tanh(\text{FC}(\mathbf{f}_2; 64 \rightarrow 1)) \quad (15)$$

The complete set of prediction heads follows the same pattern, with each activation function and range chosen to match the physical parameter distributions observed in strong lens surveys. The encoder contains approximately **350,000** learnable parameters, distributed across the convolutional backbone ($\sim 300\text{k}$) and the nine prediction heads ($\sim 50\text{k}$). This is remarkably compact compared to unconstrained black-box approaches while maintaining full physical interpretability.

Critically, the encoder is trained *jointly* with the RIM through the forward fidelity loss (Section 3.3.3), which ensures that predicted parameters are physically meaningful: they must be able to reproduce the observation when applied to the ground-truth source via the forward model.

Stage 1: Conditional Gravitational Warping— Given a source image \mathbf{x} and the observation-specific parameters $\Theta(\mathbf{y}_{\text{obs}})$, we compute the gravitational deflection field. The primary lens is modeled as a softened isothermal ellipsoid (SIE):

$$\alpha_{\text{SIE}}(\theta; \Theta) = \frac{b}{\sqrt{(q\theta'_x)^2 + (\theta'_y)^2 + r_c^2}} \begin{bmatrix} q\theta'_x \\ \theta'_y \end{bmatrix} \quad (16)$$

where $\theta' = R(-\phi) \cdot (\theta - \theta_0)$ applies rotation by position angle ϕ and translation by center $\theta_0 = (x_0, y_0)$ —all predicted from \mathbf{y}_{obs} . The core radius $r_c = 0.01$ (in normalized coordinates) is fixed for numerical stability.

External shear is added using the predicted amplitude γ and position angle γ_ϕ :

$$\alpha_{\text{shear}}(\theta; \gamma, \gamma_\phi) = \gamma \begin{bmatrix} \cos(2\gamma_\phi) & \sin(2\gamma_\phi) \\ \sin(2\gamma_\phi) & -\cos(2\gamma_\phi) \end{bmatrix} \theta \quad (17)$$

NFW Dark Matter Halo Component (Optional)— When enabled, we include a Navarro-Frenk-White (NFW) halo component parameterized by the predicted convergence κ_s and scale radius r_s . The NFW deflection angle is computed using a numerically stable formulation that handles three regimes ($x = r/r_s < 1$, $x > 1$, and $x \approx 1$) (Navarro *et al.* 1997; Wright and Brainerd 2000):

$$\alpha_{\text{NFW}}(r; \kappa_s, r_s) = \frac{4\kappa_s r_s}{x} \cdot F(x), \quad x = \frac{r}{r_s} \quad (18)$$

where $F(x)$ is evaluated as:

$$F(x) = \begin{cases} 1 - \frac{2}{\sqrt{1-x^2}} \tanh^{-1} \left(\sqrt{\frac{1-x}{1+x}} \right) & \text{if } x < 1 \\ 1 - \frac{2}{\sqrt{x^2-1}} \tan^{-1} \left(\sqrt{\frac{x-1}{1+x}} \right) & \text{if } x > 1 \\ 2/3 & \text{if } x \approx 1 \end{cases} \quad (19)$$

The \tanh^{-1} and \tan^{-1} functions are implemented with careful clamping to avoid numerical instabilities near domain boundaries. The total deflection field combines all components:

$$\alpha_{\text{total}}(\theta; \Theta) = \alpha_{\text{SIE}}(\theta) + \alpha_{\text{shear}}(\theta) + \alpha_{\text{NFW}}(\theta) \quad (20)$$

The warped image is computed via differentiable grid sampling using PyTorch’s `grid_sample` function:

$$\mathbf{x}_{\text{warp}}(\theta) = \mathbf{x}(\theta - \alpha_{\text{total}}(\theta; \Theta(\mathbf{y}_{\text{obs}}))) \quad (21)$$

This operation is fully differentiable with respect to both the source \mathbf{x} and the encoder parameters ϕ , enabling end-to-end gradient-based optimization.

Stage 2: PSF Convolution (Global Parameters)— Unlike the lens parameters which are predicted per observation, the PSF model uses *globally-shared learnable parameters* that are common across all training examples. This design choice reflects the assumption that PSF characteristics are instrument-specific rather than lens-specific.

The PSF is modeled as a hybrid combination of a parametric Moffat profile and a learned residual kernel (Moffat 1969):

$$K_{\text{PSF}} = K_{\text{Moffat}}(\text{FWHM}, \beta) + K_{\text{residual}} \quad (22)$$

The Moffat profile is parameterized by two global learnable parameters, FWHM (full width at half maximum) and β (power-law index) (Moffat 1969):

$$K_{\text{Moffat}}(r; \text{FWHM}, \beta) = \left[1 + \left(\frac{r}{\alpha} \right)^2 \right]^{-\beta}, \quad \alpha = \frac{\text{FWHM}}{2\sqrt{2^{1/\beta} - 1}} \quad (23)$$

where FWHM and β are represented as exponentials of unconstrained parameters to ensure positivity:

$$\text{FWHM} = \exp(\log \text{FWHM}_{\text{raw}}), \quad \text{initialized to } 3.0 \text{ pixels} \quad (24)$$

$$\beta = \exp(\log \beta_{\text{raw}}), \quad \text{initialized to } 4.5 \quad (25)$$

The residual kernel $K_{\text{residual}} \in \mathbb{R}^{21 \times 21}$ is a small learned correction that captures systematic deviations from the parametric form (e.g., optical aberrations, diffraction effects). It is initialized near zero and constrained to remain small through regularization (Section 3.3.3).

Convolution is implemented efficiently via FFT for images larger than 64×64 pixels:

$$\mathbf{y}_{\text{conv}} = \mathcal{F}^{-1} [\mathcal{F}[\mathbf{x}_{\text{warp}}] \odot \mathcal{F}[K_{\text{PSF}}]] \quad (26)$$

where \mathcal{F} denotes the 2D Fourier transform and \odot denotes element-wise multiplication. For smaller images, direct spatial convolution via `torch.nn.functional.conv2d` is used.

Stage 3: Background Addition and Noise (Global Parameters)— A globally-shared uniform background level b_{bg} is added to the convolved image:

$$\mathbf{y} = \mathbf{y}_{\text{conv}} + b_{\text{bg}} \quad (27)$$

where $b_{\text{bg}} = \exp(\log b_{\text{bg,raw}})$ is parameterized to ensure positivity. This represents the sky background level in the observations.

During training, we optionally add realistic instrumental noise to match the simulator’s noise characteristics:

$$\mathbf{N}_{\text{counts}} = \text{Poisson}(\text{gain} \cdot \mathbf{y}) \quad (28)$$

$$\mathbf{N}_{\text{read}} \sim \mathcal{N}(0, \sigma_{\text{read}}^2) \quad (29)$$

$$\mathbf{y}_{\text{noisy}} = \frac{\mathbf{N}_{\text{counts}} + \mathbf{N}_{\text{read}}}{\text{gain}} \quad (30)$$

where gain and σ_{read} are also global learnable parameters, initialized to physically plausible values ($\text{gain} = 1.0 \text{ e}^-/\text{ADU}$, $\sigma_{\text{read}} = 1 \times 10^{-3} \text{ e}^-$). However, in practice we found that noise addition during forward-operator training was not necessary, as the RIM learns to handle realistic noise through its data-driven iterative updates.

The Conditional Adjoint Operator F_{ϕ}^{\dagger} — To compute gradient-like signals for the RIM, we require an adjoint (transpose) operator. Critically, this adjoint *also depends on the observation* \mathbf{y}_{obs} through the predicted lens parameters, ensuring that the gradient signal respects the lens configuration inferred from each specific observation:

$$\mathbf{g} = F_{\phi}^{\dagger}(\mathbf{r}; \mathbf{y}_{\text{obs}}) = \text{Warp}^T(\text{PSF}^T * \mathbf{r}; \Theta(\mathbf{y}_{\text{obs}})) \quad (31)$$

The transpose warping operation uses the same deflection field $\alpha_{\text{total}}(\theta; \Theta(\mathbf{y}_{\text{obs}}))$ but applies it in the *opposite direction* via grid sampling with inverted coordinates, and the PSF transpose is implemented as convolution with the flipped kernel.

Summary: Observation-Conditional Forward Modeling— The key innovation is that *every observation receives its own set of lens parameters*, predicted from pixels by the encoder. This per-observation conditioning allows the forward operator to adapt to:

- Different Einstein radii (0.3”–2.0” in our training data)
- Varying ellipticities (0.0–0.7)
- Diverse lens centers and position angles
- Different external shear configurations
- Presence or absence of NFW halos

Meanwhile, PSF and noise parameters remain global, capturing instrument-level characteristics shared across observations. This hybrid design—per-lens physics, global instrument model—balances flexibility with parameter efficiency.

The complete forward operator (encoder + physics model) contains approximately **350,000** learnable parameters: $\sim 350\text{k}$ in the encoder and ~ 450 in the global PSF/noise model. While this is larger than a simple

parametric forward model, it remains compact compared to typical black-box image-to-image networks (which often exceed 10 million parameters) while maintaining full physical interpretability through the explicit prediction of nine lens parameters per observation.

3.2. Synthetic Training Dataset

3.2.1. Simulation Philosophy

To train AstroRIM, we require paired observations: lensed images \mathbf{y} and corresponding ground-truth unlensed sources \mathbf{x} . Real gravitational lens observations do not come with ground truth, so we must rely on synthetic data generated from a forward simulator. Our simulation strategy aims to produce **diverse, realistic training examples** that span the range of lensing configurations and observational conditions AstroRIM might encounter.

We deliberately generated data from **multiple simulation scripts** (versions v1 through v6, described below) with varying parameter ranges, noise models, and complexity. This diversity prevents the model from overfitting to a single simulator’s quirks and encourages learning of robust features. Each simulation version introduces slightly different:

- **Lens configurations:** varying ellipticity ranges, inclusion/exclusion of substructures (NFW halos, subhalos), shear strengths
- **Source morphologies:** different Sérsic parameter ranges, clump statistics, spiral structure probabilities
- **Observational realism:** Gaussian vs. Moffat PSFs, varying noise levels, additional detector effects (PRNU, CTI, correlated noise) in later versions

By training on this heterogeneous dataset, we increase the likelihood that AstroRIM generalizes to real observations, which inevitably differ from any single simulation.

3.2.2. Simulation Versions

SimgenV1: Foundation baseline with strong-arc emphasis— The first simulation version establishes the core pipeline and emphasizes clearly visible gravitational arcs. Sources are positioned near lens caustics to ensure prominent lensing features, with Einstein radii restricted to $\theta_E \in [0.9, 1.5]$ arcsec. The PSF is exclusively Gaussian with narrow FWHM range (0.06–0.10 arcsec), representing space-like observations. Noise is deliberately kept mild (read noise 0.8–1.8 e^- , sky 0.06–0.25 ADU/pixel) to maximize arc visibility for initial training. Source brightness is tuned such that arcs are not overwhelmed by lens light (source magnitudes 21.0–23.2, lens offset -0.3 to $+1.2 \text{ mag}$). This version serves as the cleanest training signal, with minimal detector artifacts and no substructure beyond the primary SIE lens and external shear.

SimgenV2: Enhanced physics and moderate complexity— Version 2 introduces additional physical realism while maintaining good signal quality. The Einstein radius range is slightly expanded to (0.9, 1.5) arcsec with soft clipping that allows occasional extreme values (8% probability). Source morphologies are enriched with more

aggressive clumping (up to 6 clumps with 75% probability) and occasional spiral structure (35% probability). Lens complexity increases with the introduction of NFW group halos (18% probability) and subhalo populations (70% probability, Poisson(1.0) count). Minor detector effects are added: 1% PRNU, small sky gradients (2% max), and occasional cosmic ray hits (2% probability). The PSF remains Gaussian but with added ellipticity (up to 8%). Brightness contrast is actively managed through iterative adjustment to ensure source peaks remain at least 35% of lens peaks, preventing arc washout.

SimgenV3: Multi-instrument diversity— Version 3 introduces **instrument heterogeneity** by sampling from three distinct observational regimes: space-like (45% probability: FWHM 0.06–0.12 arcsec, pixel scale 0.04 arcsec), ground-like (45%: FWHM 0.5–1.2 arcsec, pixel scale 0.20 arcsec), and intermediate (10%: FWHM 0.12–0.35 arcsec, pixel scale 0.08 arcsec). This variation forces the model to generalize across different angular resolutions and sampling scales. The PSF model expands to include both Gaussian and Moffat profiles (with $\beta \in [2.8, 4.5]$), with increased ellipticity (up to 35%). New realism features include sub-pixel shifts (85% probability), PSF mismatch between simulation and measurement (60% probability), correlated noise (60% probability, $\sigma = 0.8$ –4.0 pixels), and sky-subtraction residuals (60% probability). Source positioning becomes more sophisticated, with 28% placed near caustics for strong arcs and 72% distributed broadly for weaker configurations (STScI et al. 2024).

SimgenV4: Expanded parameter space and reduced noise ceiling— Version 4 broadens the physical parameter space while reducing maximum noise levels. The velocity dispersion range expands to (80, 380) km/s, and ellipticity extends to (0.0, 0.7), covering nearly circular to highly elongated lenses. Source magnitudes become fainter (20.0–25.0) with larger lens brightness offsets (−2.0 to +3.0 mag), creating more challenging reconstruction scenarios. Shear strength increases to 0.16, and NFW halo probability rises to 22%. Critically, noise parameters are **reduced**: read noise capped at 3.0 e[−] (down from 8.0), sky limited to 1.0 ADU/pixel (down from 3.0), PRNU reduced to 0.4% (from 1.0%), and cosmic ray intensity lowered to 15–200 ADU. This “reduced maximum noise” configuration improves training stability while maintaining realistic statistical variation. Double main lenses (galaxy pairs) are introduced with 8% probability.

SimgenV5: Weak-lensing emphasis and rejection sampling— Version 5 shifts focus toward **weak lensing regimes** by deliberately suppressing strong multiple imaging. Velocity dispersions are restricted to (80, 180) km/s, producing smaller Einstein radii ($\theta_E \sim 0.3$ –1.0 arcsec). A rejection-sampling procedure uses Lenstronomy’s `LenEquationSolver` to test whether a source at multiple trial positions produces multiple images; if so, parameters are resampled (up to 10 attempts). This produces a dataset dominated by single-imaged, weakly-magnified sources. Substructure probabilities are reduced (group halos 8%, subhalos 25% with Poisson(0.7)) to minimize small-scale perturbations. Source morphologies are refined with improved bulge-disk decomposition and more

realistic spiral structure. Field galaxy contamination increases (90% probability, 3–9 extra sources). Noise remains moderate, and vignetting effects are occasionally added (12% probability). This version challenges the model to recover subtle shape distortions rather than dramatic arc features.

SimgenV6: Maximum observational realism— Version 6 represents the most physically sophisticated simulation, incorporating state-of-the-art detector and noise modeling. Source morphologies are categorized by type (50% spiral, 30% elliptical, 20% irregular), with morphology-specific Sérsic parameter distributions. The PSF includes an **Airy pattern option** for diffraction-limited cases, alongside Gaussian and Moffat. Optical aberrations are added via coma-like asymmetries (40% probability). Noise modeling reaches new fidelity with a **realistic correlation kernel** (measured from Hyper Suprime-Cam) applied via FFT convolution, **charge transfer inefficiency (CTI)** simulating CCD readout charge trapping (70% probability), and improved cosmic ray modeling affecting multiple pixels. The cosmology is updated to FlatLambdaCDM ($H_0 = 70$, $\Omega_m = 0.3$) for consistency. Photometric sampling uses more realistic magnitude distributions (sources centered at mag 22.5 with Gaussian scatter). Sky gradients become 2D (previously 1D), and vignetting is more pronounced (15% probability, 8% max intensity variation). This version produces the most challenging training data, with noise characteristics closely matching real survey images (Aihara et al. 2018; Bosch et al. 2018; Miyazaki et al. 2018).

3.2.3. Simulation Protocol

Each synthetic pair (source, lensed image) is generated using a physics-based forward simulator built on **Lenstronomy**, a widely-used Python package for gravitational lens modeling. The simulation proceeds in five stages (Birrer and Amara 2018; Birrer et al. 2021):

Stage 1: Source Galaxy Generation— Source galaxies are modeled as multi-component systems comprising a bulge and disk, with optional additional substructure: **Bulge-Disk Decomposition**: The bulge and disk are each parameterized as elliptical Sérsic profiles (Sérsic 1968; Graham and Driver 2005):

$$I(R) = I_e \cdot \exp \left[-b_n \left(\left(\frac{R}{R_e} \right)^{1/n} - 1 \right) \right], \quad (32)$$

where:

- I_e is the surface brightness at the effective radius R_e
- n is the Sérsic index ($n \approx 4$ for bulges, $n = 1$ for exponential disks)
- $b_n \approx 2n - 0.331$ ensures half the light is within R_e
- $R = \sqrt{x'^2 + (y'/q)^2}$ is the elliptical radius with axis ratio $q = 1 - e$

The coordinates (x', y') are rotated by position angle ϕ .

For each source, we randomly sample:

- **Bulge:** $R_{e,b} \in [0.03, 0.14]$ arcsec, $n_b \in [1.5, 4.5]$
- **Disk:** $R_{e,d} = R_{e,b} \times \mathcal{U}(1.4, 3.6)$, $n_d = 1.0$ (exponential)
- **Flux ratio:** $f_{\text{bulge}} \sim \text{Beta}(1.6, 3.0)$ (most sources disk-dominated)
- **Ellipticity:** $e \in [0.0, 0.6]$, with slight misalignment between bulge and disk

Additional Complexity: To increase realism beyond smooth profiles:

- **Extra sources** (companions): With probability 1.0, we add $N_{\text{extra}} \in [0, 3]$ additional faint Sérsic components at random positions, representing merging companions or satellite galaxies.
- **Clumps** (star-forming regions): With probability 0.75, we add $N_{\text{clump}} \in [1, 6]$ Gaussian clumps:

$$I_{\text{clump}}(x, y) = A \cdot \exp \left[-\frac{(x - x_c)^2 + (y - y_c)^2}{2\sigma_{\text{clump}}^2} \right], \quad (33)$$

where $A \in [0.02, 0.8] \times I_{\text{max}}$ and $\sigma_{\text{clump}} \in [0.4, 4.0] \times \Delta x_{\text{over}}$ (super-pixel units). These represent HII regions or star-forming knots.

- **Spiral structure:** With probability 0.35, we multiply the surface brightness by:

$$S(\theta, r) = 1 + 0.06 \cdot \sin(3\theta + 5r/r_{\text{max}} + \phi_0), \quad (34)$$

where $\theta = \arctan 2(y, x)$ and $\phi_0 \sim \mathcal{U}(0, 2\pi)$. This adds spiral arms without explicitly modeling spiral density waves.

Stage 2: Lens Mass Distribution— The foreground lens is modeled as a **Singular Isothermal Ellipsoid (SIE)** with optional additional mass components: **SIE Deflection:** The primary lens produces a deflection angle:

$$\alpha(\theta) = \theta_E \cdot \frac{\theta_{\text{ell}}}{|\theta_{\text{ell}}| + r_c}, \quad (35)$$

where the Einstein radius is:

$$\theta_E = 4\pi \left(\frac{\sigma_v^2}{c^2} \right) \left(\frac{D_{ls}}{D_s} \right), \quad (36)$$

relating the velocity dispersion σ_v to the angular lensing strength. D_s and D_{ls} are angular diameter distances to the source and lens-source, computed using **Planck15 cosmology** ($H_0 = 70$ km/s/Mpc, $\Omega_m = 0.3$). The elliptical coordinates are:

$$\theta_{\text{ell}} = R(\phi_{\text{lens}}) \cdot [q\theta_x, \theta_y]^T, \quad (37)$$

with rotation matrix $R(\phi_{\text{lens}})$ and axis ratio q_{lens} .

We sample velocity dispersions $\sigma_v \in [80, 380]$ km/s, corresponding to Einstein radii spanning weak ($\theta_E \sim 0.3''$) to strong ($\theta_E \sim 2.0''$) lensing regimes.

External Shear: With probability 1.0, we add external shear (from nearby groups):

$$\alpha_{\text{shear}} = \gamma \begin{bmatrix} \cos(2\phi_\gamma) & \sin(2\phi_\gamma) \\ \sin(2\phi_\gamma) & -\cos(2\phi_\gamma) \end{bmatrix} \theta, \quad (38)$$

with $\gamma \in [0.0, 0.16]$ and random position angle ϕ_γ .

NFW Halo: With probability 0.05–0.22 (depending on simulation version), we add a Navarro-Frenk-White dark matter halo representing a group or cluster:

$$\alpha_{\text{NFW}}(r) = \frac{4\rho_s R_s^3}{r} \left[\ln \left(\frac{r}{2R_s} \right) + F(r/R_s) \right], \quad (39)$$

with scale radius $R_s \in [10, 30]$ arcsec.

Subhalos: With probability 0.15–0.75, we add $N_{\text{sub}} \sim \text{Poisson}(0.5\text{--}1.4)$ SIS subhalos:

$$\alpha_{\text{SIS}}(r) = \theta_{E,\text{sub}}/r, \quad (40)$$

with $\theta_{E,\text{sub}} \in [0.003, 0.02]$ arcsec. These represent dark matter substructure or satellite galaxies.

Stage 3: Ray-Shooting and Super-Sampling— To accurately model lensing without aliasing artifacts, we super-sample each detector pixel by a factor of 4, giving 384×384 rays per 96×96 image. For each super-pixel at image-plane coordinate θ , we:

1. Compute the source-plane coordinate via the lens equation: $\beta = \theta - \alpha(\theta)$
2. Use bilinear interpolation to sample the source surface brightness at β
3. Accumulate the result in a super-sampled lensed image

This oversampling ensures smooth gradients and prevents numerical artifacts in the lensing transformation.

Stage 4: Photometry and Normalization— Surface brightness profiles are converted to observed counts using:

$$C_{\text{total}} = t_{\text{exp}} \times 10^{-0.4(m-m_0)}, \quad (41)$$

where $t_{\text{exp}} \in [900, 1800]$ seconds is the exposure time, m is the AB magnitude, and $m_0 = 25.94$ is the zero-point (corresponding to HST/ACS F814W). Source magnitudes are sampled from $m_{\text{src}} \in [21.0, 24.0]$, and lens magnitudes are offset by $\Delta m \in [-0.2, 1.5]$ (allowing the lens to be brighter or fainter than the source).

Counts are distributed according to the normalized surface brightness profile, then converted to counts per super-pixel using the super-pixel area ($\Delta\theta_{\text{over}} = 0.04/4 = 0.01$ arcsec) (STScI et al. 2024).

Stage 5: Observational Effects— The super-sampled lensed image (source + lens light) is convolved with a PSF, downsampled to detector resolution, and corrupted with realistic noise: **PSF Convolution:** The PSF is modeled as Gaussian, Moffat, or (in v6) Airy:

$$\text{PSF}_{\text{Moffat}}(r; \alpha, \beta) = \frac{\beta - 1}{\pi\alpha^2} \left[1 + \left(\frac{r}{\alpha} \right)^2 \right]^{-\beta}, \quad (42)$$

with FWHM $\in [0.06, 0.15]$ arcsec (space-to-ground seeing), $\beta \in [2.5, 4.8]$, and optional ellipticity $e_{\text{PSF}} \in [0.0, 0.12]$.

Downsampling: The convolved super-sampled image is averaged over 4×4 blocks to produce the final 96×96 detector image.

Noise Simulation: We apply a sequence of realistic noise sources:

1. **Photon noise:** Poisson statistics on counts converted to electrons: $N_e = \text{Poisson}(C \times g)$
2. **Read noise:** $N_e \leftarrow N_e + \mathcal{N}(0, \sigma_{\text{read}}^2)$ with $\sigma_{\text{read}} \in [0.8, 2.2] \text{ e}^-$
3. **PRNU (Photo-Response Non-Uniformity):** Multiplicative fixed pattern: $C \leftarrow C \times [1 + \mathcal{N}(0, 0.01^2)]$
4. **Sky gradient:** Linear gradient with max variation 0.02–0.03 across the image
5. **Cosmic rays:** With probability 0.02–0.03, add 1–5 hits of intensity $\mathcal{U}(40, 350)$ ADU
6. **Correlated noise (v5–v6):** Gaussian smoothing of white noise to simulate drizzle/resampling artifacts (Fruchter and Hook 2002)
7. **Charge Transfer Inefficiency (CTI) (v6 only):** Simulated charge trapping along readout direction

The complete noise model ensures the synthetic data exhibits the full complexity of real CCD observations.

3.2.4. Dataset Statistics

Our training dataset comprises:

- **Total simulations generated:** 500,000 source-lensed pairs (250,000 initial + 250,000 for finetuning)
- **Simulation scripts used:** Versions v1, v2, v3, v4, v5, and v6, each contributing different parameter ranges and realism levels
- **Train/validation split:** 80% / 20% (200,000 train / 50,000 val per generation phase)
- **Test set:** A held-out set from independent simulator runs (evaluation ongoing)

The dataset spans a wide range of physical configurations:

- **Einstein radii:** $0.3''$ to $2.0''$ (weak to strong lensing)
- **Ellipticities:** 0.0 to 0.7 (circular to highly elliptical)
- **Source morphologies:** early-type to late-type spirals, mergers, clumpy irregulars
- **Signal-to-noise:** peak SNR from ~ 5 to ~ 100

This diversity is intentional: by exposing AstroRIM to a broad distribution of lensing scenarios, we aim to produce a model that generalizes beyond the specifics of any single simulator.

3.3. Training Procedure

3.3.1. Initial Training Phase

The initial training phase was conducted on 250,000 synthetic pairs split 80/20 into training and validation sets (200,000 train, 50,000 validation). Training proceeded for 30 epochs, at which point validation metrics plateaued, indicating convergence. During this phase, both the RIM and forward operator parameters were optimized jointly from random initialization (except for the parametric PSF components, which were initialized to reasonable physical values as described in Section 3.1.2). The forward operator’s PSF parameters (FWHM, β) were initialized to match typical ground-based seeing conditions (FWHM ≈ 0.10 arcsec, $\beta \approx 4.0$), while the residual PSF kernel was initialized near zero. Lens parameters were initialized to mild values ($b \approx 0.08$, $r_c \approx 0.01$, $q \approx 0.7$) corresponding to a moderately elliptical lens. The RIM’s convolutional weights were initialized using PyTorch’s default Kaiming uniform initialization, which is appropriate for ReLU-activated networks (STScI *et al.* 2024).

Training was monitored using both validation MSE and SSIM (Structural Similarity Index). We observed steady improvement in both metrics over the first 15 epochs, followed by slower refinement through epoch 30. The learning rate schedule automatically reduced the learning rates three times during training (at epochs 12, 21, and 27) in response to validation plateaus.

3.3.2. Finetuning Phase

Following initial convergence, we conducted a finetuning phase to further improve generalization. We generated an additional 250,000 synthetic pairs using the same simulation protocols but with independently sampled random seeds, ensuring the finetuning data was drawn from the same distribution but consisted of entirely new examples. This finetuning set was again split 80/20 (200,000 train, 50,000 validation). For finetuning, we reduced both learning rates by a factor of 10 relative to the initial training phase ($\eta_{\text{RIM}} = 5 \times 10^{-6}$, $\eta_{\text{forward}} = 5 \times 10^{-7}$) to avoid disrupting the learned representations. Finetuning proceeded for 7 epochs, with early stopping triggered when validation MSE showed no improvement for 5 consecutive epochs.

The final model achieved a validation MSE of approximately 3.7×10^{-4} and SSIM of 0.951 on the held-out validation set from the finetuning phase. These metrics represent the average performance across diverse lensing configurations, noise realizations, and source morphologies present in our synthetic dataset.

3.3.3. Loss Function and Regularization

The primary training objective minimizes reconstruction error while ensuring the predicted lens parameters yield physically accurate forward models. The total loss comprises four components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{fwd}} \mathcal{L}_{\text{forward}} + \lambda_{\text{PSF}} \mathcal{L}_{\text{PSF}} + \lambda_{\text{param}} \mathcal{L}_{\text{param}} \quad (43)$$

Reconstruction Loss— The reconstruction loss measures how well the RIM recovers ground-truth sources:

$$\mathcal{L}_{\text{recon}} = \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}_{\text{final}}^i - \mathbf{x}_{\text{GT}}^i\|_2^2 \quad (44)$$

Forward Fidelity Loss— The forward fidelity loss ensures that the predicted lens parameters can accurately reproduce the observed lensed image from the ground-truth source. This enforces that the encoder learns physically meaningful parameters:

$$\mathcal{L}_{\text{forward}} = \frac{1}{N} \sum_{i=1}^N \|F_{\phi}(\mathbf{x}_{\text{GT}}^i, \mathbf{y}_{\text{obs}}^i) - \mathbf{y}_{\text{obs}}^i\|_2^2 \quad (45)$$

This term acts as a *self-consistency constraint*: if the encoder has correctly identified the lens parameters for observation i , then applying the forward model to the true source with those parameters should reproduce the observation. We use $\lambda_{\text{fwd}} = 0.5$ by default, giving this term substantial weight.

PSF Smoothness Regularization— To prevent high-frequency noise in the residual PSF kernel, we apply total-variation regularization (unchanged from original):

$$\mathcal{L}_{\text{PSF}} = \|\nabla K_{\text{residual}}\|_2^2 = \sum_{i,j} [(K_{i+1,j} - K_{i,j})^2 + (K_{i,j+1} - K_{i,j})^2] \quad (46)$$

with $\lambda_{\text{PSF}} = 10^{-3}$.

Lens Parameter Regularization— To encourage physically plausible lens parameters and prevent degenerate solutions, we apply soft constraints on the predicted parameters:

$$\mathcal{L}_{\text{param}} = \mathcal{L}_{\text{param}}^{\text{range}} + \mathcal{L}_{\text{param}}^{\text{prior}} \quad (47)$$

Range penalties discourage extreme values:

$$\mathcal{L}_{\text{param}}^{\text{range}} = \frac{1}{N} \sum_{i=1}^N \left[\max(0, -b_i)^2 + \max(0, b_i - 0.5)^2 + (x_0^i)^2 + (y_0^i)^2 + (\gamma_i - 0.05)^2 \right] \quad (48)$$

This penalizes negative deflection strengths, extremely large b values, lens centers far from the origin, and deviations from typical shear values.

Prior penalties encourage parameters near physically expected values:

$$\mathcal{L}_{\text{param}}^{\text{prior}} = \frac{1}{N} \sum_{i=1}^N \left[(b_i - 0.08)^2 + (q_i - 0.7)^2 + (\kappa_s^i - 0.05)^2 + (r_s^i - 0.2)^2 \right] \quad (49)$$

These terms gently bias parameters toward typical values observed in our training data while allowing flexibility for diverse lens systems. We use $\lambda_{\text{param}} = 10^{-3}$.

3.3.4. Optimization

We optimize the combined parameter set $\{\theta, \phi\}$ (RIM and forward operator including encoder) using the **AdamW** optimizer with *separate learning rates* for the two components (Loshchilov and Hutter 2019):

- $\eta_{\text{RIM}} = 5 \times 10^{-5}$ for RIM parameters θ
- $\eta_{\text{forward}} = 5 \times 10^{-6}$ for forward operator parameters ϕ (including encoder)

The $10\times$ slower learning rate for the forward operator is essential for stable joint training. The encoder must learn to extract physically meaningful lens parameters from observations—a more constrained, physics-informed task than the RIM’s flexible image-to-image mapping. Rapid parameter updates would destabilize the conditioning mechanism before the RIM adapts.

Learning Rate Schedule— Training proceeds in two phases:

1. **Warmup (5 epochs)**: Learning rates increase linearly from 0 to target values to prevent large early gradient steps.
- Plateau reduction**: After warmup, we monitor validation loss. If no improvement occurs for $P = 6$ consecutive epochs, we reduce both learning rates by factor $\rho = 0.5$.

Stabilization Techniques— Joint training of coupled conditional models requires careful stabilization:

Gradient clipping: We clip the global gradient norm to $\tau_{\text{clip}} = 5.0$ to prevent exploding gradients when encoder predictions change rapidly:

$$\nabla \mathcal{L} \leftarrow \min \left(1, \frac{\tau_{\text{clip}}}{\|\nabla \mathcal{L}\|} \right) \nabla \mathcal{L} \quad (50)$$

Exponential Moving Average (EMA): We maintain shadow parameters with decay $\alpha_{\text{EMA}} = 0.999$:

$$\phi_{\text{EMA}} \leftarrow \alpha_{\text{EMA}} \cdot \phi_{\text{EMA}} + (1 - \alpha_{\text{EMA}}) \cdot \phi \quad (51)$$

The EMA parameters provide a smoothed version of the model that is more stable for evaluation.

Automatic Mixed Precision (AMP): We use FP16 arithmetic with dynamic loss scaling, reducing memory usage by $\sim 40\%$ and accelerating training by $\sim 2\times$ with negligible accuracy impact (Micikevicius et al. 2018).

Data Augmentation— To improve generalization, we apply stochastic augmentations (probability 0.5 each):

- Horizontal/vertical flips
- 90° rotations ($k \in \{0, 90, 180, 270\}$)

These transformations are applied consistently to both source and observation pairs, preserving physical correspondence. They are appropriate for gravitational lensing as the physics is rotationally symmetric.

3.4. Computational Requirements

The complete training pipeline (initial training + fine-tuning) consumed approximately 37 GPU-hours on an NVIDIA RTX 2080 Ti. Memory requirements remained well within the 11GB VRAM budget, with peak usage around 8GB for batch size 8. Training could potentially be accelerated by using larger batch sizes on GPUs with more memory, though we did not explore this systematically. Inference is considerably more efficient. Processing a single 96×96 observation through 10 RIM iterations requires approximately 300 ms on CPU (Intel i7-12700H), or roughly 3 samples per second. On GPU, inference accelerates to approximately 50-100 ms per sample, though the exact timing depends on batch size and GPU model. This makes AstroRIM practical for processing large lens surveys, where thousands of systems may need to be analyzed.

The model checkpoint files are compact: the RIM state (300k parameters) requires approximately 1.2 MB storage in FP32 format, while the forward operator (14k parameters) requires only 56 KB. Together with optimizer states and training metadata, a complete checkpoint is approximately 10 MB. This compact representation facilitates model distribution and deployment.

3.5. Evaluation Metrics

For quantitative evaluation of reconstruction quality, we employ two complementary metrics:

Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{HW} \sum_{i,j} (x_{\text{final},ij} - x_{\text{GT},ij})^2, \quad (52)$$

which measures pixel-wise reconstruction accuracy. MSE is sensitive to absolute intensity errors and provides a direct measure of the loss being optimized. **Structural Similarity Index (SSIM)** (Wang *et al.* 2004; van der Walt *et al.* 2014):

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \quad (53)$$

where μ denotes local mean, σ^2 denotes local variance, σ_{xy} denotes local covariance, and c_1, c_2 are small constants for numerical stability. SSIM is designed to correlate with human perception of image quality and is particularly sensitive to structural content (edges, textures) rather than absolute intensity. It ranges from 0 (no similarity) to 1 (perfect similarity).

We compute SSIM using a sliding 11×11 Gaussian window with $\sigma = 1.5$ pixels, as implemented in scikit-image. Both metrics are computed on normalized images (scaled to $[0, 1]$ range) to ensure comparability across different brightness scales.

For evaluation on real data (where ground truth is unavailable), we assess reconstruction quality through:

- **Visual inspection:** Examining whether reconstructed sources exhibit plausible galaxy morphology (smooth profiles, coherent structure, no obvious artifacts)
- **Forward consistency:** Computing $\|F_\phi(\mathbf{x}_{\text{recon}}) - \mathbf{y}_{\text{obs}}\|^2$ to verify that the reconstruction, when forward-projected, matches the observation

- **Comparison to parametric models:** Where available, comparing AstroRIM reconstructions to those obtained from traditional lens modeling software (e.g., Lenstronomy-based MCMC inversions) (Birrer and Amara 2018; Birrer *et al.* 2021)

3.6. Code and Data Availability

All code for AstroRIM, including the model architecture, training scripts, simulation pipeline, and inference utilities, is publicly available at <https://github.com/Mad-At-Line/AstroRim>. The repository includes:

- Model definitions (`trainer.py`, `finetuning.py`)
- Simulation scripts (`simgen.v1.py` through `simgen.v6.py`)
- Data loading and preprocessing utilities
- Trained model checkpoints
- Example inference notebooks demonstrating application to real data
- Documentation and usage instructions

The synthetic training dataset (500,000 pairs, approximately 12 GB uncompressed) is available upon request. Due to storage constraints, we do not host this dataset publicly, but we provide the simulation scripts that allow exact reproduction of the training data given appropriate computational resources (estimated 100 CPU-hours for full dataset generation). Model checkpoints from the final fine-tuned model (pre-finetuning) are included in the repository, allowing immediate application of AstroRIM without retraining. We provide checkpoints in PyTorch native format (`.pt`).

4. RESULTS

We evaluate AstroRIM using (i) aggregate comparisons to published/approximate reference values across a small real-lens set, (ii) detailed diagnostic panels for two representative real lenses (RXJ1131 and SDSSJ1004), and (iii) simulated benchmarks where ground truth is available (reported here in structured form; figures are in the Appendix for clarity).

4.1. Real-lens case studies

The full multi-panel diagnostic figures for RXJ1131 and SDSSJ1004 are present in the Appendix B (Figures 4 and 5), along with per-lens summary tables of inferred parameters and reconstruction metrics (Tables 3 and 4). Here we focus on interpreting what the diagnostics imply physically and where model limitations show up.

4.1.1. RXJ1131: stable geometry and internally consistent mass diagnostics

RXJ1131 is a comparatively clean galaxy-scale lens. In the diagnostic panel (Appendix Figure 4), the forward model reproduces the dominant ring/arc morphology and concentrates residual structure into localized regions rather than producing large coherent arc-shaped mismatches. This is consistent with recovering the correct global lensing geometry, with remaining discrepancies attributable to fine-scale effects such as PSF mismatch, imperfect lens-light removal, or small-scale structure not captured by the parametric lens form.

TABLE 1
ASTRO-RIM MODEL VS. PUBLISHED/APPROXIMATE REFERENCE VALUES. M_E^{ref} IS THE EINSTEIN MASS ENCLOSED WITHIN θ_E^{ref} COMPUTED FROM (z_l, z_s, θ_E) IN FLAT Λ CDM ($H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\Omega_m = 0.3$). σ^{ref} IS PUBLISHED WHEN AVAILABLE, OTHERWISE SIS-EQUIVALENT FROM (z_l, z_s, θ_E) .

Lens	z_l	z_s	θ_E^{model} [arcsec]	θ_E^{ref} [arcsec]	$ \% \Delta \theta_E $	σ^{model} [km/s]	σ^{ref} [km/s]	$ \% \Delta \sigma $	M_E^{model} [M_\odot]	M_E^{ref} [M_\odot]	$ \% \Delta M_E $
B1608	0.630	1.390	0.853	1.135	24.8	251.0	247.0	1.6	3.14×10^{11}	4.99×10^{11}	37.0
B2045	0.870	1.280	1.174	1.076	9.1	404.4	391.4	3.3	1.09×10^{12}	9.29×10^{11}	16.9
B2108	0.365	0.670	2.171	2.285	5.0	344.8	325.0	6.1	1.17×10^{12}	1.64×10^{12}	28.5
CFRS03	0.938	2.941	2.120	2.100	0.9	384.5	387.0	0.6	1.81×10^{12}	1.76×10^{12}	2.8
HST01247†	0.500	2.000	1.115	1.085	2.8	247.1	243.3	1.5	3.16×10^{11}	2.86×10^{11}	10.2
HST14113	0.460	2.810	1.017	0.900	12.9	210.8	208.9	0.9	1.55×10^{11}	1.67×10^{11}	7.3
RXJ1131	0.295	0.658	1.804	1.900	5.1	352.7	360.7	2.2	7.67×10^{11}	7.95×10^{11}	3.5
SDSSJ1004	0.680	1.734	5.914	7.995	26.0	650.7	754.5	13.8	2.58×10^{13}	2.35×10^{13}	10.0
Mean abs. % error (MAPE)					10.8			3.8			14.5

† Adopted (z_l, z_s) from external and inferred (metadata) sources

The inferred convergence κ field is compact and only mildly elongated, consistent with the near-round inferred axis ratio ($q \approx 0.885$; Appendix Table 3). The radial $\kappa(r)$ profile decreases smoothly and crosses the critical regime near $\kappa \sim 1$ at $\theta_E \approx 1.804''$, while the cumulative mass profile rises monotonically to the enclosed Einstein-scale mass reported in the diagnostics. Overall, RXJ1131 behaves as a “stable” case: (i) forward consistency is strong, and (ii) derived diagnostics are mutually compatible across the κ map, radial profile, and enclosed-mass curve.

4.1.2. SDSSJ1004: strong ellipticity and more structured residuals in a complex field

SDSSJ1004 is a more challenging system in both scale and complexity. The model captures the dominant arc placement and overall geometry (Appendix Figure 5), but the residual map exhibits more structured positive/negative patches and point-like outliers. This is expected in real fields where PSF/background mismatch, unmodeled foreground light, and deviations from the assumed noise model can produce coherent residuals even when the global lensing solution is broadly correct.

The inferred κ map is strongly elongated, consistent with the lower axis ratio ($q \approx 0.589$; Appendix Table 4). The radial profile crosses the critical regime at a much larger Einstein scale ($\theta_E \approx 5.914''$), and the azimuthal κ variation at $r = \theta_E$ reflects significant angular structure (ellipticity and/or tidal contributions). Despite increased residual structure compared to RXJ1131, the diagnostics remain qualitatively self-consistent: the elongated κ morphology, the $\kappa(r)$ critical crossing near θ_E , and the rising enclosed-mass trend agree with a large-scale lensing configuration.

4.2. Simulated benchmark performance (ground truth available)

We evaluate AstroRIM on a held-out set of synthetic strong-lensing images for which the unlensed source-plane ground truth is available. Following the evaluation protocol, each lensed/ground-truth pair is scaled by a per-example normalization factor so that metrics are computed on $[0, 1]$ images, enabling fair comparison across simulations spanning different intrinsic brightness scales. We report structural similarity (SSIM) and mean squared error (MSE) on these normalized images.

Across $N = 900$ simulated inputs, AstroRIM achieves an average reconstruction quality of $\text{SSIM} = 0.927 \pm$

0.067 and $\text{MSE} = 5.95 \times 10^{-4} \pm 7.37 \times 10^{-4}$. Median performance is $\text{SSIM} = 0.950$ and $\text{MSE} = 3.00 \times 10^{-4}$, with the central 10–90 percentile ranges spanning 0.831–0.990 in SSIM and 6.65×10^{-5} – 1.58×10^{-3} in MSE. Overall, 72.9% of reconstructions achieve $\text{SSIM} \geq 0.90$ and 49.6% achieve $\text{SSIM} \geq 0.95$, while only 6.0% fall below $\text{SSIM} < 0.80$.

To assess robustness across simulator variants, we stratify results by the simulation-family label encoded in filenames (Table 2). Performance is consistently high across all subsets, with the best aggregate results obtained on `simgen5-6` ($\text{SSIM} = 0.972 \pm 0.025$; $\text{MSE} = 3.20 \times 10^{-4} \pm 4.38 \times 10^{-4}$). The widest performance distribution occurs for `simgen4`, which contains the lowest-SSIM outliers in the full set (minimum $\text{SSIM} = 0.611$). Without additional simulator metadata, we do not attribute these degradations to a single factor; however, they are consistent with a subset containing more challenging observation conditions or source/lens configurations (e.g., reduced effective constraints from the arcs, increased blur/noise, or more complex morphologies), which primarily manifests as mild over-smoothing and suppression of faint extended structure in the reconstruction.

Representative reconstructions are provided in Appendix D (Figure 7), showing the lensed input, ground-truth source, and AstroRIM reconstruction for a curated subset of simulations.

4.3. Failure example

4.3.1. LRG-3-757: failure likely dominated by non-survey image conversion and missing calibration metadata

Figure 6 (Appendix C) shows our primary failure case, LRG-3-757. Unlike the other real-lens examples, this input was produced by taking a `.png` image, converting it to FITS, and then normalizing it. Crucially, the image is therefore *not* tied to a survey-calibrated data product: there is no reliable pixel-to-arcsecond scale in the header, no PSF model, no background/noise model, and no guarantee that the pixel values remain linearly proportional to flux. These missing ingredients break key assumptions in our forward operator and in the physical conversion layer.

Observed failure signature.— This case is not a subtle degradation: the residual panel exhibits a strong ring-like structure that is spatially coherent around the

TABLE 2

ASTRO-RIM RECONSTRUCTION METRICS ON SIMULATED DATA, COMPUTED ON PER-EXAMPLE NORMALIZED IMAGES IN $[0, 1]$. WE REPORT MEAN \pm STANDARD DEVIATION, AND MEDIAN WITH THE 10–90 PERCENTILE RANGE IN BRACKETS.

Subset	N	SSIM (mean \pm std)	SSIM (median [P10,P90])	MSE _{norm} (mean \pm std)	MSE _{norm} (median [P10,P90])
All sims	900	0.927 \pm 0.067	0.950 [0.831, 0.990]	$5.95 \times 10^{-4} \pm 7.37 \times 10^{-4}$	3.00×10^{-4} [6.65×10^{-5} , 1.58×10^{-3}]
simgen1	150	0.900 \pm 0.058	0.905 [0.821, 0.972]	$9.43 \times 10^{-4} \pm 8.68 \times 10^{-4}$	7.21×10^{-4} [1.66×10^{-4} , 1.92×10^{-3}]
simgen2	150	0.905 \pm 0.066	0.925 [0.803, 0.974]	$7.61 \times 10^{-4} \pm 7.80 \times 10^{-4}$	4.53×10^{-4} [1.24×10^{-4} , 1.83×10^{-3}]
simgen3	150	0.915 \pm 0.065	0.931 [0.821, 0.984]	$5.36 \times 10^{-4} \pm 6.67 \times 10^{-4}$	2.67×10^{-4} [4.92×10^{-5} , 1.42×10^{-3}]
simgen4	150	0.896 \pm 0.089	0.920 [0.770, 0.983]	$6.91 \times 10^{-4} \pm 8.77 \times 10^{-4}$	3.58×10^{-4} [8.31×10^{-5} , 1.80×10^{-3}]
simgen5–6	300	0.972 \pm 0.025	0.981 [0.944, 0.995]	$3.20 \times 10^{-4} \pm 4.38 \times 10^{-4}$	1.75×10^{-4} [4.94×10^{-5} , 6.82×10^{-4}]

Einstein-ring radius, and the χ^2 map highlights substantial misfit concentrated on the ring. The diagnostic metrics (Appendix Table 5) indicate high mismatch (e.g. $\chi^2 \sim 2.6 \times 10^3$) despite the model producing a plausible-looking reconstruction. In other words: AstroRIM finds a lens solution that *looks* reasonable, but the forward prediction cannot reproduce the observed intensity distribution.

Most likely root causes (ranked).— Based on the failure morphology and the data provenance, the dominant causes are likely the following.

- 1. Nonlinear intensity mapping from PNG (gamma/sRGB) and dynamic-range loss.** PNG images are typically stored in sRGB (gamma-compressed) space and may have undergone contrast stretching, clipping, denoising, sharpening, or lossy rescaling prior to export. Converting such an image into FITS preserves pixel values but *not* the original linear flux field. Because AstroRIM’s likelihood implicitly assumes approximately linear photometry (model prediction \rightarrow observed counts plus noise), gamma/compression can create systematic ring-wide discrepancies that manifest exactly as coherent residual rings.
- 2. Unknown PSF and implicit convolution mismatch.** Survey images are blurred by an instrument PSF; our forward model is only as accurate as the assumed PSF treatment. A PNG-to-FITS workflow often bakes in additional smoothing/sharpening and resampling kernels that are *not* modeled. The result is a characteristic “too-sharp vs. too-blurry” mismatch along high-gradient structures like Einstein rings, producing alternating positive/negative residual bands around the ring.
- 3. Pixel scale is not known (arcsec/pixel assumed), so geometric scale can be wrong.** Our pipeline requires a pixel scale to convert from normalized coordinates to arcseconds and to report θ_E and $M(< \theta_E)$. Here, the run used a pixel scale of $\sim 0.074''/\text{pix}$ (Appendix Table 5), but with no validated WCS this number may be wrong by factors of a few. Even a modest scale error pushes the inferred Einstein radius and the effective ring radius out of alignment with the true geometry, which again appears as a coherent ring residual.
- 4. Noise/background model mismatch (no variance map).** Without a background estimate and a per-pixel variance model, the optimization may overweight bright ring pixels or underweight diffuse

structure, steering the solution toward a visually plausible but statistically inconsistent fit. This inflates χ^2 and concentrates misfit where the signal is highest.

- 5. The image may not be a true strong lens (or contains strong non-lensing structure).** Because the input is not survey-verified, the “ring” could be driven partly by intrinsic galaxy structure (e.g. ring galaxy features, processing artifacts, or compositing), not gravitational lensing. AstroRIM can still return a lens-like parameter set, but the forward model cannot reproduce non-lensing morphology, leaving structured residuals.

Why the reported physical mass is not reliable here.— The run reports $\theta_E \approx 1.04''$ and $M(< \theta_E) \approx 3.3 \times 10^{11} M_\odot$ under assumed (z_l, z_s) and an assumed pixel scale (Appendix Table 5). However, when the pixel scale and photometric linearity are uncertain, *both* θ_E and the Σ_{crit} -based conversion become systematically unstable. For this reason we treat LRG-3-757 as a diagnostic failure case and do not interpret its physical masses as astrophysically meaningful.

Practical fixes / next tests.— This failure mode is actionable. The highest-impact fixes are: (i) re-run using a *native* linear FITS data product (no PNG step), (ii) supply or estimate a PSF and convolve the forward model consistently, (iii) use a verified pixel scale (from WCS or instrument metadata) or explicitly fit a scale factor, and (iv) include a background/noise model or variance map to stabilize residual weighting. If the system is not a confirmed lens, a quick sanity check is to fit with a more flexible lens-light + source model (or apply a lens-candidate classifier) before interpreting lens parameters.

5. DISCUSSION

AstroRIM is designed to close a practical gap in strong-lensing inference: methods that are fast enough for survey-scale use are often either (i) highly parametric and therefore expensive when run fully Bayesian, or (ii) purely data-driven and therefore fragile under forward-model mismatch and difficult to interpret physically. The core idea explored here is that *speed and interpretability need not be traded against robustness*, provided the forward operator itself is allowed to adapt per observation in a physically structured way. In AstroRIM, a LensParameterEncoder predicts a compact set of bounded lens and halo parameters for each system, which condition a differentiable forward operator. A Recurrent Inference Machine (RIM) then performs iterative source-plane refinement using the *conditional adjoint* of this operator at each step. Joint training is stabilized by a forward-fidelity/self-consistency term that requires the predicted

parameters to reproduce the observation when applied to the ground-truth source.

5.1. *What the simulated benchmarks say (and what they do not)*

On held-out synthetic data with ground truth available, AstroRIM achieves high structural fidelity across a broad distribution of conditions. The median SSIM and median normalized MSE indicate that, for typical simulated systems, the model recovers source morphology with small pixel-wise error, while the mean and standard deviation quantify a meaningful but controlled tail of more difficult cases. Two points are important for interpretation:

1. **The distribution matters more than the mean.** The reported percentile ranges show that performance is not dominated by a small subset of easy examples. Instead, the method maintains good reconstructions across most of the test distribution, with a minority of outliers consistent with cases that are intrinsically less constrained by the data (e.g., weak arcs, heavy blur/noise, or complex morphologies). This is the behavior expected from a learned iterative solver: when the lensing geometry provides strong constraints, the RIM can converge to a high-fidelity solution; when constraints are weak, the learned prior biases reconstructions toward smoother, lower-variance morphologies.
2. **Generalization across simulator variants supports the “conditional forward” motivation.** Stratifying by simulation-family label shows consistently strong performance across multiple realism regimes, including the most observationally complex variants. While the best-performing subsets likely reflect regimes closer to the model’s effective inductive biases (or where constraints are strongest), the key result is that no single simulator family dominates the success cases. This supports the premise that per-observation conditioning helps absorb instrument/configuration differences (PSF family, noise regime, Einstein-radius range) without requiring a single fixed operator to match all conditions.

At the same time, these synthetic benchmarks should be interpreted with two explicit caveats:

- **Metrics are computed on per-example normalized images.** This is appropriate for comparing structural recovery across diverse brightness scales, but it partially decouples evaluation from absolute photometric calibration. In real deployments where flux calibration, background subtraction, and variance modeling matter for downstream physical inference (e.g., mass normalization, uncertainty calibration), further validation on calibration-consistent data products is required.
- **Simulation realism is necessary but not sufficient for real-data robustness.** Even very sophisticated simulators cannot fully reproduce all survey artifacts (PSF anisotropy, correlated noise,

detector systematics, lens-light subtraction residuals, masking, neighboring objects). As a result, strong sim performance is best read as evidence that the *inverse solver is capable* under matched forward assumptions, not as a guarantee that every real system will be recovered without additional calibration and domain adaptation.

5.2. *Behavior on real lenses: forward consistency as a diagnostic*

The real-lens examples illustrate two complementary properties of AstroRIM that are valuable for practical use.

First, in cleaner galaxy-scale systems such as RXJ1131, AstroRIM can produce reconstructions that are not only visually plausible but also *forward-consistent*: when the reconstruction is passed through the inferred forward operator, residuals remain relatively localized and the derived mass-profile diagnostics (e.g., the qualitative structure of the inferred convergence field and its radial trends) are mutually consistent. This is precisely where a conditional forward operator is most beneficial: small mismatches in PSF or lens geometry that would otherwise accumulate into structured residuals can be partially compensated by per-observation parameter inference.

Second, in more complex configurations such as SDSSJ1004, AstroRIM can still capture the global lensing geometry and recover coherent source structure, but the residual and diagnostic panels highlight the limits of the current forward model. In practice, these limits are not merely failure points; they are *actionable diagnostics*. Spatially coherent residuals are informative: they indicate specific model-mismatch modes (PSF convolution mismatch, background misestimation, lens-light contamination, or noise/variance mismatch) rather than unstructured reconstruction noise. In other words, AstroRIM can function as both an inference engine and a *model-checking tool* for determining what forward ingredients are missing for a given dataset.

5.3. *Failure analysis and why it is scientifically useful*

The explicit failure case (LRG-3-757) is important because it demonstrates a realistic failure mode that can occur even when the reconstruction appears superficially plausible. The core issue is not that the RIM “cannot reconstruct”; rather, the pipeline is forced to operate outside the assumptions that tie pixel values to a physically meaningful forward model. Nonlinear intensity mapping from PNG workflows, unknown/implicit PSF and resampling kernels, and uncertain pixel scale can all produce ring-like residual structures that are coherent on the Einstein-ring scale. In such a setting, AstroRIM may still find a lens-like parameter set and a visually credible source, but forward projection cannot match the observation, and derived physical quantities (e.g., θ_E and mass conversions) become unreliable.

Practically, this motivates a deployment principle: AstroRIM should be run on *native, linear, survey-calibrated FITS products* whenever possible, with a verified pixel scale (WCS/instrument metadata), a PSF estimate or prior, and (ideally) a variance/background model. When those ingredients are absent, AstroRIM can still serve as

a qualitative reconstruction tool, but outputs should not be interpreted as physically calibrated measurements.

5.4. *Why conditional per-lens forward modeling matters for the field*

Most ML lens inversion pipelines struggle with sim-to-real gaps because the forward operator is effectively “baked in” globally: either implicitly (an end-to-end mapping learns a simulator-specific inverse) or explicitly (a fixed differentiable operator is assumed). AstroRIM’s per-lens conditioning is a middle path: the forward model remains physics-structured and interpretable, but its parameters are adapted for each observation. This has several downstream implications:

- **Interpretability and reusability.** The inferred parameters are directly useful as initial conditions for traditional modeling pipelines, including MCMC or nested sampling approaches, and can also support fast survey triage (e.g., approximate Einstein radius, ellipticity/shear trends).
- **Reduced forward-model mismatch.** Allowing $\Theta(y_{\text{obs}})$ to vary per lens helps absorb heterogeneity in lens configurations and observation conditions without introducing a high-dimensional, opaque conditioning code. The forward-fidelity loss further encourages these parameters to remain physically meaningful rather than merely predictive.
- **A natural path to hybrid workflows.** AstroRIM can be used as a fast “first pass” to produce a reconstruction and lens-parameter estimate, then refined by a smaller set of expensive Bayesian runs on scientifically high-value systems. This division of labor is well matched to the coming regime of tens-to-hundreds of thousands of lenses from Euclid/Roman/Rubin-era discovery pipelines.

5.5. *Computational practicality and survey-scale relevance*

A key contribution of AstroRIM is that it is lightweight enough for real use: inference on a single 96×96 observation over 10 RIM iterations can be executed in $\mathcal{O}(10^2)$ ms on GPU and a few $\times 10^2$ ms on CPU, and the checkpoint sizes are small enough to distribute and run broadly without specialized infrastructure. This matters operationally: a method that is scientifically compelling but computationally prohibitive cannot be a survey pipeline component. AstroRIM’s reported compute profile suggests it can realistically be integrated into lens-candidate processing streams where thousands of systems must be analyzed repeatedly.

5.6. *Limitations and next steps*

The results here also clarify what is *not* yet solved. The most important limitations and high-impact improvements are:

1. **Calibration dependence of physical conversion.** Derived physical masses and Einstein-radius conversions depend on pixel scale and redshift assumptions. Robust real-data usage therefore requires instrument-calibrated priors (or explicit fitting of scale) and well-defined redshift inputs.

2. **Lens-light contamination and background/variance modeling.**

Foreground lens light can imprint structured residuals that the source model may partially absorb, biasing inferred parameters. Joint modeling of lens light (or robust subtraction) and inclusion of a variance map/background model would improve statistical consistency and reduce ring-scale residual structure.

3. **Uncertainty quantification.** As presented, AstroRIM produces point estimates. For scientific inference (especially substructure and cosmological applications), the next step is uncertainty-aware inference over both Θ and derived mass profiles. Practical approaches include ensembles, approximate Bayesian inference for the encoder, or hybridization with simulation-based inference (SBI) that treats AstroRIM outputs as proposal/summary statistics.

4. **Sim-to-real adaptation.** Despite diverse simulation variants, domain shift remains the core challenge. Finetuning on small sets of survey-calibrated lenses (or self-supervised adaptation using forward-consistency objectives on real data) is a promising direction, as is explicit modeling of correlated noise and PSF anisotropy matched to the target instrument.

Taken together, these points suggest a clear roadmap: (i) make the observational calibration interface explicit (pixel scale, PSF, variance), (ii) incorporate lens light into the forward model or preprocessing, (iii) add uncertainty quantification, and (iv) use domain adaptation to close the remaining sim-to-real gap.

5.7. *Broader significance*

AstroRIM’s broader significance is that it demonstrates a viable, compact pattern for integrating learned iterative inference with physics-structured, observation-conditioned forward modeling. The approach preserves the central scientific virtues of traditional lens modeling (explicit forward operator, interpretable parameters, and self-consistency checks) while providing the speed required for the next generation of lens samples. Even when outputs are not yet fully calibrated for high-precision cosmology, the method is immediately useful as a rapid delensing and diagnostic tool, and as a strong initializer for more expensive inference pipelines.

6. CONCLUSION

We presented AstroRIM, a compact (~ 650 k-parameter) end-to-end lens inversion system that jointly trains (i) a Recurrent Inference Machine for iterative source reconstruction and (ii) a differentiable, conditional physics-parameterized forward operator whose lens parameters are predicted *per observation* by a LensParameterEncoder. A forward-fidelity/self-consistency loss couples these components and encourages the inferred parameters to remain physically meaningful while reducing forward-model mismatch.

On held-out simulated benchmarks with ground truth, AstroRIM achieves high reconstruction fidelity across

a broad range of lens configurations and observation regimes, demonstrating that learned iterative solvers can recover complex source morphologies when guided by conditional, physics-based gradients. On real lenses, AstroRIM produces forward-consistent reconstructions and interpretable mass-profile diagnostics in cleaner systems, while residual structure in more complex cases highlights specific forward-model ingredients that must be improved for robust deployment. The explicit failure case further emphasizes the importance of calibration-consistent data products and forward-consistency checks when interpreting physical quantities.

The central takeaway is that *conditional, per-lens forward modeling* is a practical and scientifically interpretable way to improve robustness to heterogeneous real-world data, while retaining the speed needed for survey-scale analysis. With improvements in PSF/background handling, lens-light modeling, uncertainty quantification, and domain adaptation to survey-calibrated data, AstroRIM offers a promising path toward rapid delensing and mass-profile characterization for the large lens samples expected from upcoming surveys.

ACKNOWLEDGMENTS

We thank the ML4Astro community, Kevin Nolan, Morgan Fraser, and the open-source scientific Python ecosystem.

APPENDIX
GITHUB CODE REPOSITORY

The complete codebase with detailed instructions for operation and method replication is available on our project's GitHub page:
<https://github.com/Mad-At-Line/AstroRim>



FIG. 3.— GitHub repository landing page.

REAL LENS DIAGNOSTIC FIGURES AND STATISTICS
RXJ1131-1231

TABLE 3
RXJ1131-1231: ASTRO RIM INFERRED LENS PARAMETERS AND DERIVED DIAGNOSTICS.

Quantity	Value	Quantity	Value
b	0.1837	q	0.8846
ϕ [rad]	-3.1287	(x_0, y_0)	(0.0163, 0.0046)
γ	0.1621	γ_ϕ [rad]	0.0573
κ_s	0.0573	r_s	0.1113
θ_E [arcsec]	1.804	σ_v (SIS) [km/s]	353
$M(< \theta_E)$ [M_\odot]	7.67×10^{11}	Pixel scale [arcsec/pix]	0.2045
RMS residual	0.0357	SNR	1.36
(z_l, z_s) (run)	(0.295, 0.654)	Source max/mean	0.032/0.002

SDSS J1004+4112

TABLE 4
SDSS J1004+4112: ASTRO RIM INFERRED LENS PARAMETERS AND DERIVED DIAGNOSTICS.

Quantity	Value	Quantity	Value
b	0.2996	q	0.5887
ϕ [rad]	-3.1416	(x_0, y_0)	(0.1507, 0.1789)
γ	0.1994	γ_ϕ [rad]	0.4986
κ_s	0.4986	r_s	0.3705
θ_E [arcsec]	5.914	σ_v (SIS) [km/s]	569
$M(< \theta_E)$ [M_\odot]	1.71×10^{13}	Pixel scale [arcsec/pix]	0.4112
RMS residual	0.0481	SNR	1.50
(z_l, z_s) (run)	(0.5, 2.0)	Source max/mean	0.116/0.004

FAILURE CASE: LRG-3-757
SIMULATION RESULTS

Gravitational Lens Mass Profile Analysis with Source Reconstruction

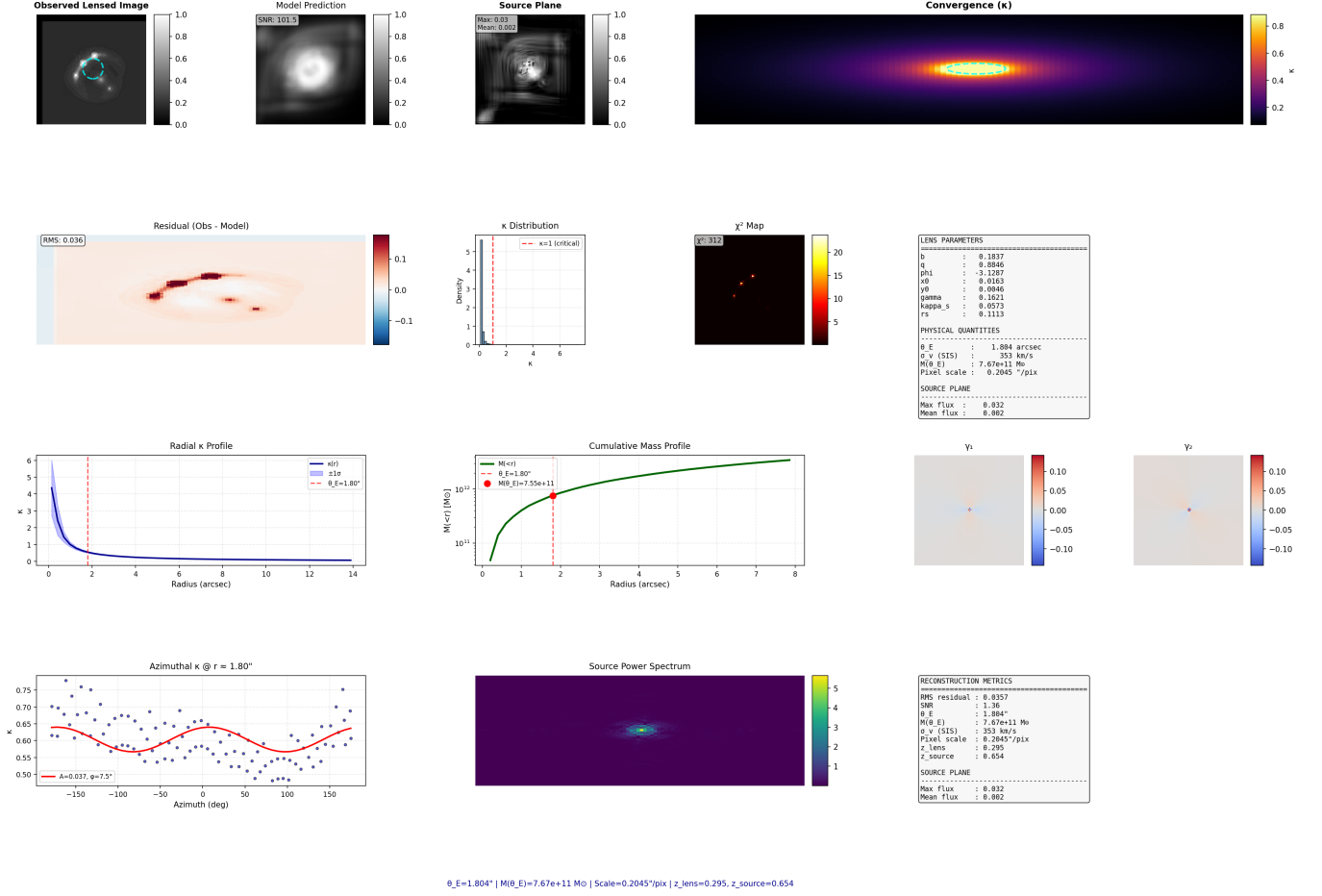


FIG. 4.— RXJ1131-1231: Full diagnostic panel showing mass reconstruction, residual maps, and parameter distributions.

TABLE 5

LRG-3-757 FAILURE CASE: INFERRED PARAMETERS AND DERIVED QUANTITIES. VALUES ARE REPORTED FOR COMPLETENESS, BUT PHYSICAL QUANTITIES SHOULD BE TREATED AS *illustrative only* DUE TO UNCERTAIN PIXEL SCALE AND NON-SURVEY PNG→FITS CONVERSION.

Quantity	Value	Quantity	Value
Lens ID	LRG-3-757	z_l (assumed)	0.44
z_s (assumed)	2.37	Pixel scale used [arcsec/pix]	0.074214
b	0.292593	q	0.95634
ϕ [rad]	-3.10997	(x_0, y_0)	(-0.001812, -0.001348)
γ	0.035337	γ_ϕ [rad]	-3.05895
κ_s	0.470615	r_s (norm)	0.242910
r_s [arcsec] (derived)	0.865315	σ_v (SIS) [km/s]	227.18
θ_E [arcsec]	1.04230	$M(< \theta_E)$ [M_\odot]	3.34×10^{11}
Σ_{crit} [kg/m ²]	4.1014	N_{pix} (aperture)	616
SIS mass (norm)	0.30497	NFW mass (norm)	0.35238
Total mass (norm)	0.65735	χ^2 (map label)	2598

Gravitational Lens Mass Profile Analysis with Source Reconstruction

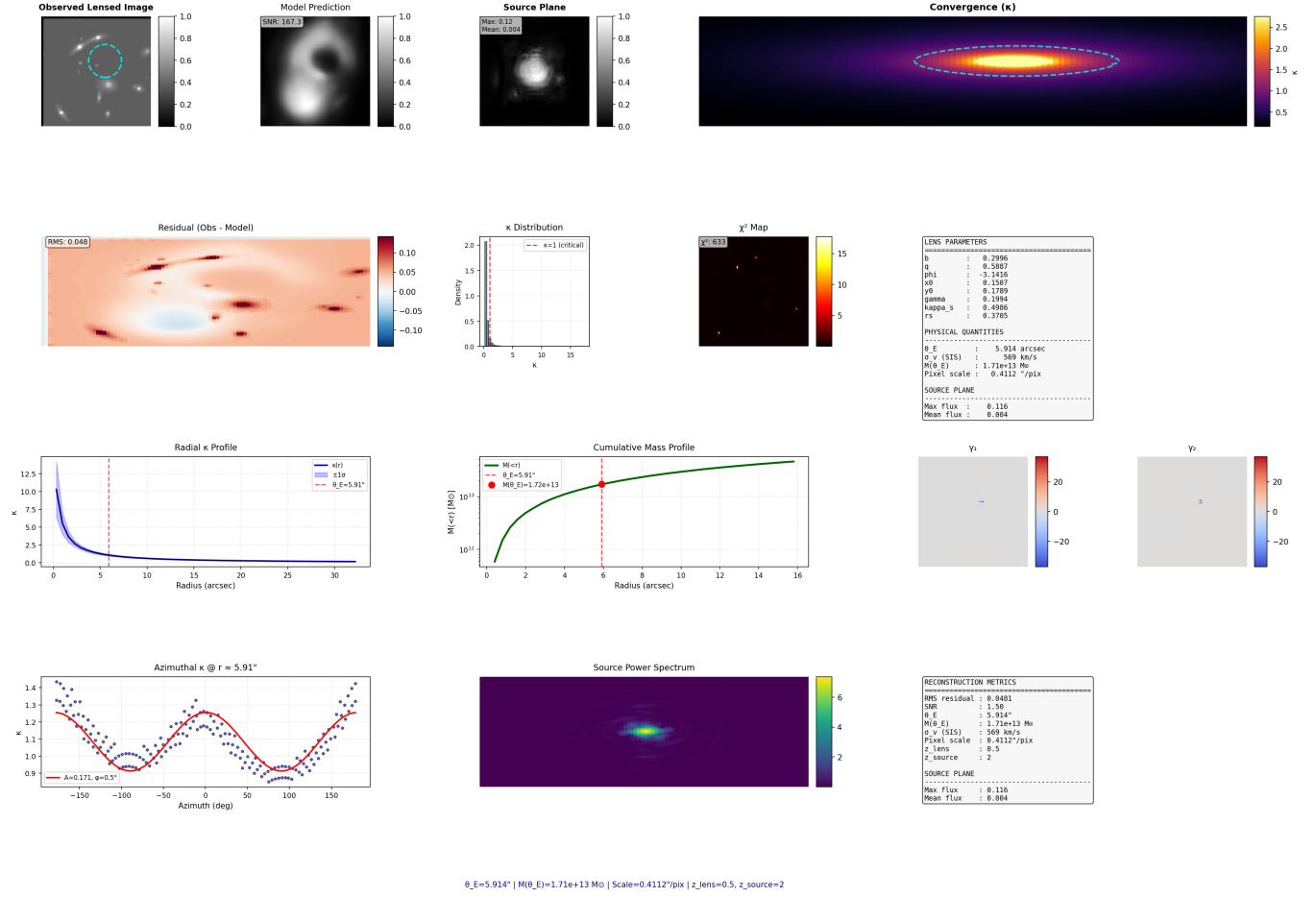


FIG. 5.— SDSS J1004+4112: Full diagnostic panel showing mass reconstruction, residual maps, and parameter distributions.

Gravitational Lens Mass Profile Analysis with Source Reconstruction

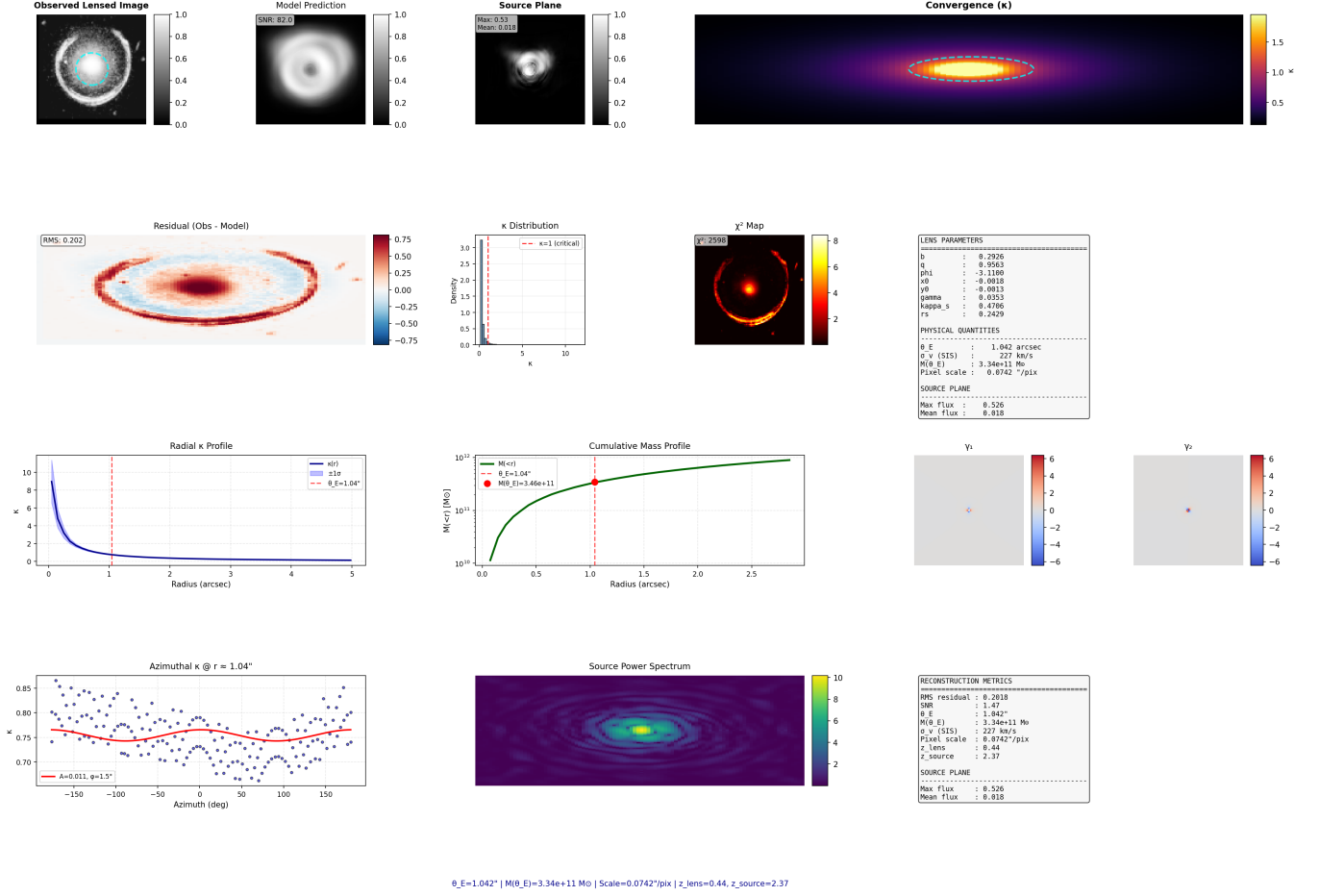


FIG. 6.— LRG-3-757 failure-case diagnostic panel. This input was generated via PNG→FITS conversion and normalization without survey calibration metadata, likely causing nonlinear intensity scaling, PSF mismatch, and uncertain pixel scale.

Reconstructions - Epoch 10

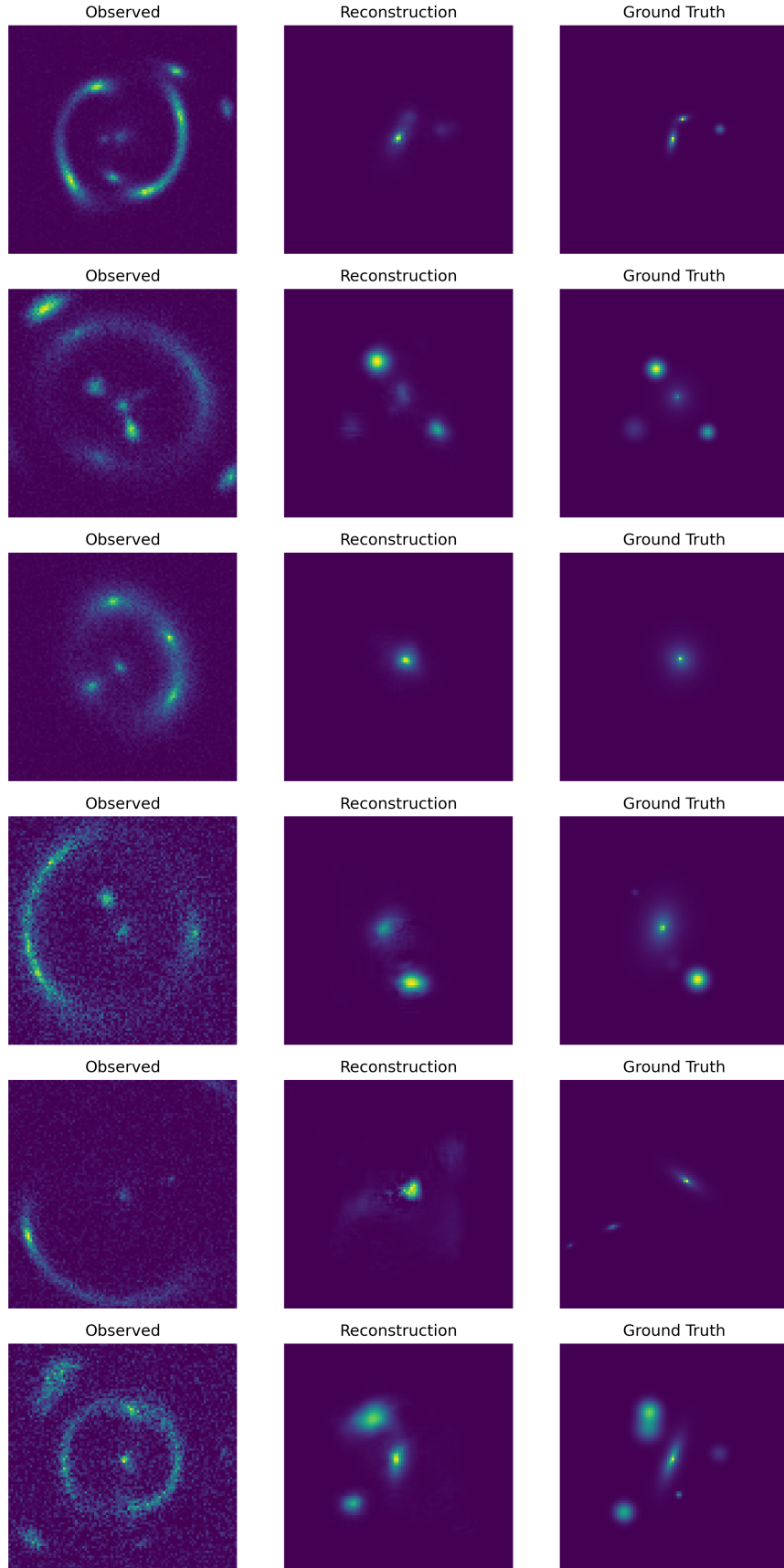


FIG. 7.— Varied simulation results (Random V1–V6) showing lensing reconstructions across different parameter configurations.

REFERENCES

- J. A. Muñoz, C. S. Kochanek, E. E. Falco, and et al. (1999) [arXiv:astro-ph/9902131](#).
- T. Treu, *Annual Review of Astronomy and Astrophysics* **48**, 87 (2010), reviews fundamentals and applications of strong gravitational lensing, emphasising its power to study mass distributions and cosmology[32616982963506†L48-L56].
- P. Schneider, C. S. Kochanek, and J. Wambsganss, eds., *Gravitational Lensing: Strong, Weak and Micro*, Saas-Fee Advanced Course, Vol. 33 (Springer, Berlin/Heidelberg, 2006) comprehensive textbook covering strong, weak and microlensing[287658190976042†L33-L42].
- T. E. Collett, *The Astrophysical Journal* **811**, 20 (2015), [arXiv:1507.02657 \[astro-ph.CO\]](#).
- C. Weiner, S. Serjeant, and C. Sedgwick, *Research Notes of the American Astronomical Society* **4**, 190 (2020), [arXiv:2010.15173 \[astro-ph.GA\]](#).
- S. H. Suyu, T. Treu, S. Hilbert, C. Blake, D. J. Lagattuta, P. J. Marshall, M. W. Auger, *et al.*, *Monthly Notices of the Royal Astronomical Society* **468**, 2590 (2017), describes the HOLICOW program to measure the Hubble constant with $< 3.5\%$ uncertainty using five time-delay lenses[382364073823731†L69-L92].
- S. Vegetti and L. V. E. Koopmans, *Monthly Notices of the Royal Astronomical Society* **408**, 1969 (2010).
- Y. D. Hezaveh, L. Perreault Levasseur, and P. J. Marshall, *Nature* **548**, 555 (2017), demonstrates that CNNs can estimate lens parameters millions of times faster than traditional maximum-likelihood models[280385048789853†L51-L71].
- R. Legin, Y. D. Hezaveh, L. Perreault Levasseur, and B. D. Wandelt, *arXiv preprint* (2021), uses density-estimation based simulation-based inference to predict lens parameters, anticipating hundreds of thousands of lenses from upcoming surveys[613462166998676†L49-L60], [2112.05278](#).
- L. Perreault Levasseur, Y. D. Hezaveh, and R. H. Wechsler, *The Astrophysical Journal Letters* **850**, L7 (2017), uses variational inference to approximate Bayesian neural network posteriors and quantify uncertainties in lensing parameters[201579821778691†L51-L67], [1708.08843](#).
- P. Putzky and M. Welling, *arXiv preprint* (2017), introduces recurrent inference machines (RIMs) that learn the inference algorithm itself via an RNN, achieving state-of-the-art results on image restoration tasks[142188878708325†L48-L62], [1706.04008](#).
- W. R. Morningstar, L. Perreault Levasseur, Y. D. Hezaveh, *et al.*, *arXiv preprint* (2019), applies a recurrent inference machine and a convolutional neural network to simultaneously reconstruct lensed sources and infer lens parameters[299943374073160†L40-L54], [1901.01359](#).
- R. Kormann, P. Schneider, and M. Bartelmann, *The Astrophysical Journal* **422**, L63 (1994), derives explicit lens equations and deflection angles for singular isothermal ellipsoids[382033800913973†L55-L76].
- C. R. Keeton, *arXiv preprint* (2001), describes algorithms for solving the lens equation and constraining lens models using positions, fluxes and time delays[324097635568390†L48-L57], [astro-ph/0102340](#).
- K. Karchev, A. Coogan, C. Weniger, *et al.*, *Monthly Notices of the Royal Astronomical Society* **512**, 661 (2022), employs a pre-trained galaxy diffusion model (AstroDDPM) and a denoising diffusion restoration model to reconstruct lensed sources[150273249541944†L51-L62].
- S. Mishra-Sharma and J. Yang, Neural Information Processing Systems (NeurIPS) Machine Learning and the Physical Sciences Workshop (2022), introduces continuous neural fields for non-parametric reconstruction of lensed sources and lenses, motivated by upcoming high-resolution surveys[74028827045285†L48-L60].
- K. Gregor and Y. LeCun, in *Proceedings of the 27th International Conference on Machine Learning (ICML)* (2010) pp. 399–406, trains feedforward networks to approximate sparse codes, reducing inference time by roughly a factor of ten[353180808265067†L6-L30].
- J. Adler and O. Öktem, *IEEE Transactions on Medical Imaging* **37**, 1322 (2018), unrolls a primal–dual optimization algorithm with learned proximal operators for tomographic reconstruction, [1707.06474](#).
- K. Hammernik, F. Knoll, *et al.*, *Magnetic Resonance in Medicine* **79**, 3055 (2018), embeds compressed-sensing MRI reconstruction in an unrolled gradient descent scheme with learned filters and nonlinearities[359404195080872†L51-L69].
- S. Birrer and A. Amara, *arXiv preprint* (2018), introduces an open-source Python package that reconstructs lens mass and surface brightness distributions accurately and flexibly[359182500222017†L48-L60], [1803.09746](#).
- L. Biggio, O. Bieri, J. Mutter, A. Amara, M. Tewes, S. Birrer, *et al.*, *Astronomy & Astrophysics* **673**, A132 (2023), introduces a continuous neural field for the lensing potential that can recover smooth and perturbed mass distributions without pre-training[277001616240138†L50-L69].
- M. Raissi, P. Perdikaris, and G. E. Karniadakis, *Journal of Computational Physics* **378**, 686 (2019), introduces physics-informed neural networks (PINNs) that embed physical laws into neural network training[190675837697164†L26-L46].
- Y. Wu and K. He, in *European Conference on Computer Vision (ECCV)* (2018) [arXiv:1803.08494](#).
- J. F. Navarro, C. S. Frenk, and S. D. M. White, *The Astrophysical Journal* **490**, 493 (1997).
- C. O. Wright and T. G. Brainerd, *The Astrophysical Journal* **534**, 34 (2000).
- A. F. J. Moffat, *Astronomy and Astrophysics* **3**, 455 (1969).
- STScI, L. Watkins, M. Peebles, and I. N. Reid (Space Telescope Science Institute (STScI)), *The Hubble Space Telescope Primer for Cycle 32*, Baltimore, MD, USA (2024), version 1.0.
- H. Aihara and et al., Publications of the Astronomical Society of Japan **70**, S4 (2018), [arXiv:1704.05858](#).
- J. Bosch and et al., Publications of the Astronomical Society of Japan **70**, S5 (2018).
- S. Miyazaki and et al., Publications of the Astronomical Society of Japan **70**, S1 (2018).
- S. Birrer, Y. Dong, A. J. Shajib, Y. Hezaveh, *et al.*, *arXiv preprint* (2021), describes the expansion of the lenstronomy package into a community-driven ecosystem[886659464932671†L53-L67], [2106.05976](#).
- J. L. Sérsic, *Atlas de galaxias australes* (Observatorio Astronómico, Universidad Nacional de Córdoba, 1968).
- A. W. Graham and S. P. Driver, Publications of the Astronomical Society of Australia **22**, 118 (2005).
- A. S. Fruchter and R. N. Hook, Publications of the Astronomical Society of the Pacific **114**, 144 (2002), [arXiv:astro-ph/9808087](#).
- I. Loshchilov and F. Hutter, in *International Conference on Learning Representations (ICLR)* (2019) [arXiv:1711.05101](#).
- P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, in *International Conference on Learning Representations (ICLR)* (2018) poster, [arXiv:1710.03740 \[cs.LG\]](#).
- Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, *IEEE Transactions on Image Processing* **13**, 600 (2004).
- S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, and et al., *PeerJ* **2**, e453 (2014)

provides fast and easy peer review for new papers in the [astro-ph](#) section of the arXiv, making the reviewing process simpler for authors and referees alike. Learn more at <http://astro.theoj.org>.

This paper was built using the Open Journal of Astrophysics L^AT_EX template. The OJA is a journal which