Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **546,** 1–26 (2026) https://doi.org/10.1093/mnras/stag155
Advance Access publication 2026 February 6

Galactic alchemy: deep learning map-to-map translation in hydrodynamical simulations

Philipp Denzel [1]★ Yann Billeter [1,2] Frank-Peter Schilling [1] and Elena Gavagnin [1,3]

[1]*Centre for Artificial Intelligence, Zurich University of Applied Sciences ZHAW, Technikumstrasse 71, CH-8400 Winterthur, Switzerland*
[2]*Institute of Science Technology and Policy, ETH Zurich, Universitätstrasse 41, CH-8092 Zurich, Switzerland*
[3]*Institute of Business Information Technology, Zurich University of Applied Sciences ZHAW, Theaterstrasse 17, CH-8400 Winterthur, Switzerland*

Accepted 2026 January 16. Received 2026 January 16; in original form 2025 October 25

ABSTRACT

We present the first systematic study of multidomain map-to-map translation in galaxy formation simulations, leveraging deep generative models to predict diverse galactic properties. Using high-resolution magnetohydrodynamical simulation data (from the ILLUSTRIS TNG suite), we compare conditional generative adversarial networks (GANs) and diffusion models under unified pre-processing and evaluation, optimizing their U-Net architectures and attention mechanisms for physical fidelity on galactic scales. Our approach jointly addresses seven astrophysical domains – including dark matter, gas, neutral hydrogen, stellar mass, temperature, and magnetic field strength – while introducing physics-aware evaluation metrics that quantify structural realism beyond standard computer vision measures. We demonstrate that translation difficulty correlates with physical coupling, achieving near-perfect fidelity for mappings from gas to dark matter and mappings involving astro-chemical components such as total gas to H I content, while identifying fundamental challenges in weakly constrained tasks such as gas to stellar mass mappings. Our results establish GAN-based models as competitive counterparts to state-of-the-art diffusion approaches at a fraction of the computational cost (in training and inference), paving the way for scalable, physics-aware generative frameworks for forward modelling and observational reconstruction in the Square Kilometre Array (SKA) era.

Key words: hydrodynamics – software: machine learning – galaxies: structure – galaxies: stellar content – dark matter.

1 INTRODUCTION

The spatial matter distribution of galaxies is the result of a complex and chaotic interaction between its individual components, such as dark matter (DM), stellar populations, (predominantly hydrogen) gas, supermassive black holes (SMBHs), and their surrounding environment. Interactions between these components are governed by their mutual gravitational and electromagnetic forces, and hydrodynamical processes which collectively shape the structure and properties of galaxies over cosmic time (J. Binney & S. Tremaine 2011; C. J. Conselice 2014; M. D'Onofrio et al. 2016; B. M. Tinsley 2022). These interactions imprint subtle signatures in the phase-space distribution of a galaxy, retaining the various feedback mechanisms that have influenced its formation and evolution (J. Binney & E. Vasiliev 2023; L. Bassini et al. 2024). Thus, the physical components encode distinct aspects of galaxy evolution:

(i) **DM haloes** of galaxies dominate the gravitational potential into which baryonic matter flows and forms visible sub-structure (S. D. M. White & C. S. Frenk 1991; B. Moore et al. 1999; C. Frenk & S. White 2012).

(ii) **Stellar mass** reflects the cumulative outcome of star formation and feedback, but its distribution is temporally highly non-local and entropic (C. F. McKee & E. C. Ostriker 2007; R. C. Kennicutt & N. J. Evans 2012; H. S. Hwang, J. Shin & H. Song 2019), while star formation is spatially localized in H_2 clouds within the interstellar medium (ISM; T. Colman et al. 2024; E. Schinnerer & A. Leroy 2024).

(iii) **Gas** traces the baryonic backbone of galaxies, regulating cooling, heating, and star formation through radiative feedback cycles (E. Gavagnin et al. 2017; M. Luisi et al. 2021) and turbulence induced by active galactic nuclei (AGNs; P. Biernacki & R. Teyssier 2018; M. Valentini et al. 2019), supernovae (D. Fielding et al. 2017; D. Ibrahim & C. Kobayashi 2023), stellar winds (M. R. Krumholz et al. 2014; J. Bally 2016), galaxy–galaxy mergers (P. F. Hopkins et al. 2006; A. Cibinel et al. 2019), or interaction with the intergalactic medium (IGM; A. L. Muratov et al. 2017; B. M. Poggianti et al. 2019).

(iv) **Neutral hydrogen** and **21cm brightness** are key observational tracers of the ISM for low-redshift galaxies, critical for radio surveys with MEERKAT, ASKAP, or the upcoming SKA-M ID (such as WALLABY, MIGHTEE-H I, MHONGOOSE, or the MeerKAT Fornax Survey; N. Maddox et al. 2021; W. J. G. Blok et al. 2024; F. M. Maccagni & W. D. Blok 2024; F. M. Maccagni & P. Serra 2025; T. O'Beirne et al. 2025).

★ E-mail: philipp.denzel@zhaw.ch

(v) **Temperature** captures thermodynamic states determined by the virialization of gas within the dark matter potential, and further shaped by shocks, cooling, and AGN-driven outflows (S. R. Ward et al. 2024; K. Zubovas, M. Tartėnas & M. A. Bourne 2024).

(vi) **Magnetic fields** emerge from turbulent amplification and influence gas dynamics (e.g. R. Beck 2015; M. Rieder & R. Teyssier 2017), yet remain poorly constrained observationally.

Recovering these domains from limited information is challenging; observationally due to technical limitations: for most instruments, signals beyond the local Universe – especially from H I – become too faint due to intrinsic dimming (H. Messias et al. 2024), foreground contamination (MeerKLASS Collaboration 2025), and low-frequency radio-frequency interference (RFI) (S. E. Harper & C. Dickinson 2018; B. N. Engelbrecht et al. 2024).

Conversely, theoretical inference of these domains has computational challenges: the understanding of the distribution of matter in the Universe remains largely driven by numerical simulations. Among these, (magneto)hydrodynamical simulations present the most principled approach to model and capture the non-linear co-evolution of dark and baryonic matter fields across cosmological and astrophysical scales (for recent reviews, see R. A. Crain & F. Voort 2023). However, this quality comes at steep computational costs or forces detrimental trade-offs between resolution and volume.

To mitigate these challenges, simpler alternatives such as dark-matter-only (DMO) simulations (e.g. D. Potter, J. Stadel & R. Teyssier 2017; S. Cheng et al. 2020; T. Ishiyama et al. 2021) reproduce large-scale structure and halo statistics at reduced cost but omit baryonic physics. Semi-analytical models (SAMs) attempt to compensate by applying post de facto prescriptions to approximate baryonic effects on top of DMO outputs (e.g. A. A. Berlind et al. 2003; R. S. Somerville et al. 2008; A. Schneider et al. 2019; A. Obuljen et al. 2023). While these methods enable exploration of cosmological parameter space, they lack the fidelity needed to capture the full complexity of galaxy-scale feedback and morphology. In particular, their intrinsic post-hoc nature often ignores the gravitational back-reaction caused by the redistribution of baryons on the DM field.

With the proliferation of deep generative models, a complementary line of research has emerged that seeks to emulate aspects of these simulations rather than compute them from first principles. Recent efforts have explored enhancing simulations and augmenting galaxy models through scalable deep learning techniques in various ways. For instance, N. Perraudin et al. (2019) use scalable generative adversarial networks (GANs) (for details see Section 2.3.1) to produce entire *N*-body 3D cubes of the cosmic DM distribution in a multiscale approach. Still, techniques aiming for the full 3D reconstruction of cosmological simulations often face challenges in scaling to resolutions where individual galaxies can be resolved. Alternatively, M. Bernardini et al. (2021, 2025) employ Wasserstein-GANs (and later versions StyleGAN) to paint baryons on to thin slices of simulation boxes from the FIRE simulation suite. Y. Li et al. (2021) and A. Schanz, F. List & O. Hahn (2024) use StyleGAN and denoising diffusion models, respectively, to superresolve cosmic large-scale structure predictions. L. Thiele et al. (2020) applied a U-Net architecture (for details see Section 2.4) to infer observable thermal and kinematic Sunyaev–Zel'dovich maps of haloes from DMO simulations, explicitly linking theory to observations. Similarly, U. Chadayammuri et al. (2023) use a U-Net for image-to-image translation of ILLUSTRISTNG galaxy cluster haloes to the corresponding baryonic fields.

Most studies focus on a single aspect of a simulation's galaxy formation or feedback model and do not fully reproduce (or harness) all physical modes of simulated galaxies (for details see Section 2, equation 3).

In this paper, we introduce a novel application of deep generative models for map-to-map translation across multiple astrophysical domains in cosmological simulations on the galaxy-scale level. In contrast to other works, we propose a more comprehensive representation of a galaxy's physical state across multiple domains relevant to its formation and evolution, by training models on different combinations of properties (e.g. gas density, stellar mass, dark matter), without relying on explicit heuristics or phenomenological tuning. Using high-resolution magnetohydrodynamical simulation data from the ILLUSTRIS TNG suite (TNG50–1; D. Nelson et al. 2017; A. Pillepich et al. 2017; V. Springel et al. 2017; F. Marinacci et al. 2018; J. P. Naiman et al. 2018), we systematically compare conditional generative adversarial networks (GANs) and diffusion models under unified pre-processing and evaluation. Our approach goes beyond prior work by jointly addressing multiple domains and introducing physics-aware metrics – such as asymmetry, clumpiness, concentration, and power spectra – that assess structural realism and astrophysical fidelity beyond standard computer vision measures. We show that GAN-based models can achieve performance comparable to diffusion models at a fraction of the computational cost (in training and inference), in particular for map-to-map translations involving astrochemical components. Moreover, a set of deep generative models including all domain translations – translations between maps of different astrophysical properties – provides a comprehensive representation of a galaxy's formation scenario (for details see Section 2 and equation 3). Finally, the generative models establish a bridge between theory and observation by incorporating domains that are directly observable, such as 21-cm brightness, into the translation process. This is particularly relevant for upcoming large-scale radio surveys with the Square Kilometre Array (SKA; R. Braun et al. 2015; L. Staveley-Smith & T. Oosterloo 2015) telescopes, which will probe the cosmic distribution of H I through 21cm emission. By enabling the reconstruction of astrophysical quantities from observational proxies and forward modelling of instrument-specific effects, our approach provides a scalable pathway to interpret SKA data within the context of galaxy formation scenarios.

The remainder of this paper is structured as follows: Section 2 details our methodology, models, evaluation metrics, and data, Section 3 presents the results, and Section 4 discusses implications and future directions.

2 DATA AND METHODOLOGY

Our work aims to address the limitations identified above by leveraging high-resolution simulation data as the foundation for a generative modelling approach. To this end, we require a data set that captures the full complexity of baryonic and DM interactions (i.e. magnetohydrodynamics) at galaxy scales. In the following Section 2.1, we detail the selection criteria and pre-processing steps applied to construct our data set of galaxy maps.

2.1 Data set

The ILLUSTRISTNG project is a series of publicly released, cosmological magnetohydrodynamical simulations of galaxy formation, run with the AREPO (R. Weinberger, V. Springel & R. Pakmor 2020) moving-mesh code (D. Nelson et al. 2017; A. Pillepich et al. 2017; V. Springel et al. 2017; F. Marinacci et al. 2018; J. P. Naiman et al. 2018). Each simulation self-consistently solves the coupled evolution of DM, cosmic gas, luminous stars, and SMBHs. The TNG50–1 simulation was run with a total of 2×2160^3 resolution elements, a DM mass resolution of $3.1 \times 10^5\,\mathrm{M}_\odot\,h^{-1}$, and a baryon mass resolution of $5.7 \times 10^4\,\mathrm{M}_\odot\,h^{-1}$, providing a rare combination of large volume and fine resolution in a simulation released to the public. Galaxies were selected from snapshots between $z = 1$ and $z = 0$ with a required minimum number of resolution elements of 10^4, to ensure sufficient resolution even for larger satellite and dwarf galaxies.

Projections on to images for each selected galaxy were performed in multiple domains (galaxy properties ζ):

(i) DM mass (DM; column density)
(ii) stellar mass (STARS; column density)
(iii) total gas mass (GAS; column density)
(iv) H I gas mass (H I; column density)
(v) (mock) 21-cm brightness temperature (21CM)
(vi) gas temperature (TEMP)
(vii) magnetic field strength (BFIELD)

The projections extend to two half-mass radii of a galaxy's total gas mass, ensuring each domain image has the same spatial resolution for a given galaxy. All but the 21-cm brightness temperature maps are directly simulated quantities. The former were generated following F. Villaescusa-Navarro et al. (2018) where the H I density $\rho_{\mathrm{H\,I}}$ is gridded using the nearest-grid-point mass assignment scheme, sliced into chosen frequency bandwidths, and projected on to a 2D brightness temperature map using the transformation

$$T_b = 189h \left(\frac{H_0(1+z)^2}{H(z)} \right) \frac{\rho_{\mathrm{HI}}}{\rho_c}\,\mathrm{mK}, \tag{1}$$

where H_0 and $H(z)$ are the Hubble parameters (for current and given redshift z, respectively) and ρ_c the Universe's critical density. Finally, Gaussian smoothing is applied to mimic the telescope beam at the target angular resolution (here, a nominal angular resolution for an SKA-MID-like performance of 0.5 arcsec was used). The map projections were performed using an adapted version of the PYLIANS3 code (F. Villaescusa-Navarro 2018). The resulting data set of multiple domains counts 504 000 512×512 images in total (72 000 images per domain), produced from roughly 3000 galaxies per snapshot (6 in total), each galaxy randomly rotated (on all axes) four different ways before projection for data augmentation. Note that the first iteration of the data set contained fewer samples with a slightly higher average total halo mass; this data set was used where explicitly stated in Sections 2.7 and 3.

In summary, the data set contains a set of galaxy projections in multiple domains (different physical modes of a galaxy) which jointly approximate the fiducial TNG model, i.e. ILLUSTRISTNG's formation scenario.

Deep learning networks often work best on non-peaked data distributions, numerically standardized in intervals between $[0, 1]$ (for uniform priors) or $[-1, 1]$ (for Gaussian priors). Inspired by common transformation used in the high-energy physics domain (see e.g. T. Finke et al. 2021), we use the following scaling for all maps

Table 1. The pre-processing transformation parameters: c is the normalization constant (in the respective units of the corresponding maps) and γ the power scaling. The Boolean b deciding whether the transformation maps to a symmetric or non-negative interval was always 1 for diffusion models and 0 for GANs.

	DM	STARS	GAS	HI	21CM	TEMP	BFIELD
c	2×10^{10}	8×10^{11}	10^{10}	10^8	165	10^8	10^{-1}
γ	8	16	8	8	8	8	8

$$\tilde{x} = (b+1)\left(\frac{x}{c}\right)^{\frac{1}{\gamma}} - b \tag{2}$$

where $c \neq 0$ is the normalization constant (around the maximum of the data distribution), $\gamma \sim \mathcal{U}\{0, \mathcal{O}(10)\}$ the power scaling, and the Boolean parameter $b \in \{0, 1\}$ depending on whether the interval should map to $[0, 1]$ or $[-1, 1]$. The exact values for the γ parameters were found via grid search (such that the median of the data set distribution is ≥ 0.3 and ≤ 0.6) per domain as listed in Table 1; b was 0 for all models with a uniform prior (GANs) and 1 for all models with a Gaussian prior (diffusion models). This transformation normalizes the data ranges and stabilizes the variances in the data, making them more Gaussian-like.

2.2 Galaxy formation scenario

Capturing the complex interplay between baryonic components and DM distributions at the galaxy scale is computationally the most expensive task in any numerical simulation. Often, a trade-off between simulation size and resolution is required to make a hydrodynamical treatment even feasible. Additionally, surrogate techniques, so-called sub-grid models, are employed to capture effects of baryonic components below the resolution limit. The ill-constrained parameters of such sub-grid models are calibrated to match observed properties at the simulated scales, leading to degeneracy and difficulties in the interpretation of outcomes (cf. R. A. Crain & F. van de Voort 2023).

For this reason, different simulation suites produce similarly realistic galaxies with a wide variety of formation 'recipes'. Notable, publicly available (and thus for this work relevant) examples of such suites are the EAGLE (J. Schaye et al. 2014; R. A. Crain et al. 2015; S. McAlpine et al. 2016), HORIZON-AGN (Y. Dubois et al. 2014), ILLUSTRISTNG (D. Nelson et al. 2017; A. Pillepich et al. 2017; V. Springel et al. 2017; F. Marinacci et al. 2018; J. P. Naiman et al. 2018), and SIMBA (R. Davé et al. 2019) suites.

The summary of all these physical effects characterizing a simulated population of galaxies, we will abstractly describe as a galaxy *formation scenario*Φ. In Bayesian terms, a simulation describes galaxy samples from a *population*Γ_i by the marginalization

$$P(\Gamma|\Phi) = \sum_{\zeta \in \Omega} P(\Gamma|\zeta) P(\zeta|\Phi), \tag{3}$$

where $\zeta \in \Omega$ represents a *galaxy property* from the set of galaxy characteristics Ω. There will also be *nuisance parameters*ν which lead to the expression of a galaxy distribution but are not related to any physical galaxy property, such as orientation

$$P(\Gamma|\zeta) = \sum_{\nu} P(\Gamma|\zeta, \nu) P(\nu). \tag{4}$$

A major inconvenience of simulations is the impracticality of drawing new samples from the galaxy population $g \sim P(\Gamma|\Phi)$, as this would require re-running an entirely new simulation at repeated computational expense. We pose that one or a set of deep generative models can properly encapsulate a simulation's formation scenario ϕ by learning individual galaxy properties ζ, enabling in-painting a learnt formation scenario on to DMO simulations.

Recent advancements in deep learning techniques have demonstrated their efficacy in performing generative tasks that involve complex functional mappings between images. Given that simulated galaxies are typically reduced to 2D for comparison with observational data, this study will focus on image-based deep learning techniques.

2.3 Deep generative modelling

As general function approximators, deep learning neural networks have proven extremely useful for data processing across various scientific disciplines (K. Hornik, M. Stinchcombe & H. White 1989; I. Goodfellow, Y. Bengio & A. Courville 2016). Their ability to beat the curse of dimensionality allows for extraction of subliminal signals from complex data, finding hidden patterns or concepts that are difficult to (manually) formalize. Especially deep learning generative models have demonstrated unparalleled results, creating high-quality synthetic data, modelling complex systems and processes (J. Whang 2023; S. Bengesi et al. 2024). The goal of deep generative models is to learn an implicit (true) data distribution from which a finite number of samples is available for training (cf. S. Bond-Taylor et al. 2022); this usually means fitting an overparametrized model $p_\theta(x) \approx p(x)$ (typically a neural network with parameters θ) such that new samples $\hat{x} \sim p_\theta(x)$ can be drawn and/or the likelihood $p_\theta(x)$ be evaluated. *Conditional* generative models additionally include control variables c which guide the generative process such that $p_\theta(x|c) \approx p(x|c)$. *Image-to-image translation* is a particular application of conditional generation, where an input image of one domain is transformed into a corresponding output image in a different domain (Y. Pang et al. 2022); in this study, the domains are defined by individual galaxy properties $c \equiv \zeta$ where $\zeta \in \Omega$ (see Sections 2.1 and 2.2). Examples of image-to-image tasks include style transfer, image colourization, denoizing, superresolution, or semantic segmentation. Within the sciences, such tasks have been adapted in and across many disciplines, showing impressive performance in modelling the distribution of atomistic systems, proteins, and biomolecules (e.g. A. Rives et al. 2021; J. B. Ingraham et al. 2023; N. Rønne, A. Aspuru-Guzik & B. Hammer 2024; A. Schneuing et al. 2024), particle jets (e.g. T. Golling et al. 2024; M. Leigh et al. 2024), or for medical imaging enhancements (e.g. J. Bullock, C. Cuesta-Lazaro & A. Quera-Bofarull 2019; M. Amirian et al. 2024).

The conditional probability distributions approximated by these models directly correspond with the terms in equation (3); thus such methods are particularly well-suited for this investigation. We examined GAN and diffusion-based approaches, as detailed in Sections 2.3.1 and 2.3.2. Both approaches are known to produce high-quality samples. While *diffusion models* are considered state-of-the-art in scientific applications of image generation, they are intrinsically inefficient in their inference process, even when applied in latent space, even more so in pixel space (cf. P. Dhariwal & A. Nichol 2021). On the other hand, *GANs* can efficiently generate samples with a single forward pass, but generally have poorer training stability and distribution coverage. Therefore, we investigated both approaches for this work's use case and compared their results, advantages, and challenges. As a secondary objective, we assess whether GAN-based models can achieve performance comparable to diffusion models, as this would substantially reduce computational costs and enable scalable deployment in large-scale simulation pipelines. Demonstrating such parity would not only accelerate inference but also substantially reduce the time required to iterate over all image translation directions, enabling more comprehensive exploration of domain mappings within practical computational budgets.

2.3.1 Generative adversarial networks

GANs are two-component models where a generative network, the *generatorG*, and a discriminative network, the *discriminatorD*, compete in an adversarial game; first introduced by I. J. Goodfellow et al. (2014). G aims to map an implicit distribution $p_G(z)$ from noise variables z, typically drawn from a normalized Gaussian $z \sim \mathcal{N}(0, \boldsymbol{I})$ (centred around 0, with unit variance), to samples indistinguishable from the true data distribution $p(x)$. At the same time, D is optimized to distinguish between generated samples from p_G and real samples y from the true data distribution. The adversarial game simultaneously invokes the minimization of the objective $\mathcal{L}_{\text{adversarial}}(G, D)$ by G and the maximization of the same by D. These seemingly diametrical goals give rise to an efficient mechanism which optimizes $G \to G^*$ leading to plausible, high-quality samples

$$G^* = \arg\min_G \max_D \mathcal{L}_{\text{adversarial}}(G, D)\,. \tag{5}$$

This effectively eliminates the need to formulate an explicit loss function, as the discriminator will take that role; in other words, the loss function is learnt.

P. Isola et al. (2016) furthermore demonstrated a conditional version (cGAN) of this adversarial game as a general-purpose solution to image-to-image translation dubbed *pix2pix*. Although very similar to the classical GAN formulation, both cGAN networks are additionally conditioned on an input image x. The generator learns a mapping from input to output image domain space $G : (x, z) \mapsto y$. The discriminator is also additionally shown the input image with the corresponding generator output $G(x, z)$. This subtly changes the interpretation of its task from originally judging the realness of generated images to a judgment on the plausibility of the domain mapping (see Fig. 1 for an illustration). Accordingly, the GAN optimization objective is constructed as follows

$$\mathcal{L}_{\text{cGAN}} = \mathbb{E}_{x,y}\left[\log D(x, y)\right] + \mathbb{E}_{x,z}\left[\log\left(1 - D(x, G(x, z))\right)\right], \tag{6}$$

where $\mathbb{E}_{x,y}$ and $\mathbb{E}_{x,z}$ denote the expectation value taken over the joint distribution of the corresponding random variables. The first term is the average prediction strength of the discriminator when the images are sampled from the data distribution. The second term establishes the actual adversarial game, describing the average discriminator's prediction strength when the images are sampled from the generator.

Moreover, P. Isola et al. (2016) proposed to mix the GAN objective with a traditional p-normed L_p loss term

$$\mathcal{L}_{L_p} = \mathbb{E}_{x,y,z}\left[||y - G(x, z)||_p\right], \tag{7}$$

where $p = 1$ (Manhattan norm) was found to be optimal by the authors whereas $p = 2$ (Euclidean norm) leads to blurriness in the predicted images.

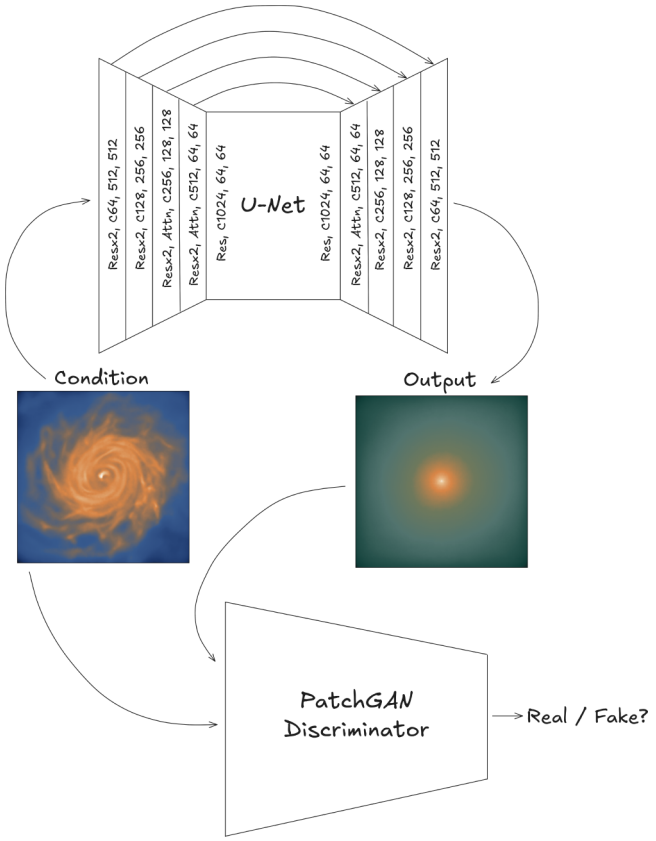


Figure 1. Conditional GAN scheme: An input image conditions the U-Net generator (including skip connections from the encoder to the decoder), while the discriminator uses the input image and generated output to judge the domain translation. The U-Net used in this work consists of four levels, each with two residual blocks (Resx2), optionally followed by attention layers (Attn), progressively increasing/decreasing the channel dimension (C), and downsampling/upsampling blocks for modifying image dimensions (starting from the original image size 512×512).

The final adversarial objective is then given by

$$\mathcal{L}_{\text{adversarial}}(G, D) = \mathcal{L}_{L_1}(G) + \lambda \cdot \mathcal{L}_{\text{cGAN}}(G, D), \tag{8}$$

where the objective weighting factor λ can be treated as a fixed hyperparameter or adaptively tuned similar to P. Esser, R. Rombach & B. Ommer (2020).

Finally, note that the noise variable z is necessary to learn a stochastic mapping, matching a distribution other than a delta function. However, P. Isola et al. (2016) have found noise input ineffective as cGAN models tend to simply ignore the noise and suggested to use dropout at test time instead to capture the full entropy of the modelled conditional distributions.

In practice, GANs are notoriously difficult to train despite their proven ability to generate high-quality samples. Two major challenges are *vanishing gradients* and *mode collapse*, which can be mitigated through architectural and objective modifications. Architectural strategies include residual skip connections to improve gradient flow (K. He et al. 2015), experimenting with normalization layers [batch S. Ioffe & C. Szegedy (2015), group Y. Wu & K. He (2018), layer J. L. Ba, J. R. Kiros & G. E. Hinton (2016), or none], and refining deconvolution operations near the generator output (A. Odena, V. Dumoulin & C. Olah 2016). Objective-based approaches involve alternative loss formulations for $\mathcal{L}_{\text{cGAN}}(G, D)$, such as DCGAN (A. Radford, L. Metz & S. Chintala 2015), LSGAN (X. Mao et al. 2016), or Wasserstein-GAN variants (WGAN, WGAN-GP; M. Arjovsky, S. Chintala & L. Bottou 2017; I. Gulrajani et al. 2017). Due to the minimax nature of GANs, losses often oscillate rather than converge, making diagnosis difficult. Overall, balancing generator and discriminator remains inherently unstable (cf. M. Arjovsky & L. Bottou 2017), requiring alternating gradient updates or separately scheduled learning rates.

In this study, we closely followed the implementation of the Pix2Pix model by P. Isola et al. (2016), including the aforementioned techniques and best practices. The generator is implemented as a standard U-Net (O. Ronneberger, P. Fischer & T. Brox (2015); architecture modifications are detailed in Section 2.4), paired with a *PatchGAN* discriminator which evaluates the plausibility of an image in sub-regions rather than a classical full-image discrimination. The discriminator can be restricted to enforce the correctness in local patches because the L_1 loss in equation (8) motivates the model to correctly predict low-frequency features in images, ultimately leading to more details in generated samples.

2.3.2 *Diffusion-based models*

Diffusion models have emerged as the de facto state of the art in computer vision (CV), surpassing – in stability, distribution coverage, and arguably in sample quality – models like GANs, normalizing flows or variational autoencoders (VAEs). They, colloquially speaking, learn to iteratively denoise a corrupted version of the data. More precisely, diffusion models include a *forward* (noising) process which is designed to push samples off the data manifold and a *backward* (denoising) process for which a model is trained to produce trajectories back to that data manifold, generating plausible samples. There are various framings for diffusion models, leading to slightly different expressions for these forward and backward processes. Here, we give a high-level overview of the formalisms relevant to this study.

Following J. Ho, A. Jain & P. Abbeel (2020)'s description of Denoising Diffusion Probabilistic Models (DDPMs), the forward and backward processes take the form of Markov chains (of length T). The forward process starts from the input x_0 and step-wise transitions to latent variables $\{x_1, \ldots, x_T\}$ (and vice versa for the backward process). Each forward transition at a particular time-step t only depends on the previous step and its probability is parametrized as a normalized, diagonal Gaussian $\mathcal{N}$

$$q(x_t|x_{t-1}) \colon = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t \boldsymbol{I}), \tag{9}$$

where the variance is $\beta_t \in (0, 1)$ and typically scheduled as $\beta_{t-1} < \beta_t$. In the limit of infinitesimal step sizes, the true reverse process has the same functional form as the forward process, a well-known fact from Brownian diffusion in physics (see equations 76 and 77 in W. Feller 1949). Thus, learning to approximate the backward process for small (enough) step sizes becomes feasible and can be analogously parametrized as

$$p_\theta(x_{t-1}|x_t) \colon = \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \tag{10}$$

where the mean μ_θ and variance Σ_θ are modelled using a neural network. Like the forward process, the backward process is a Markov chain for which its joint probability is given by the product of individual step conditionals

$$p_\theta(x_{0:T}) \colon = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{11}$$

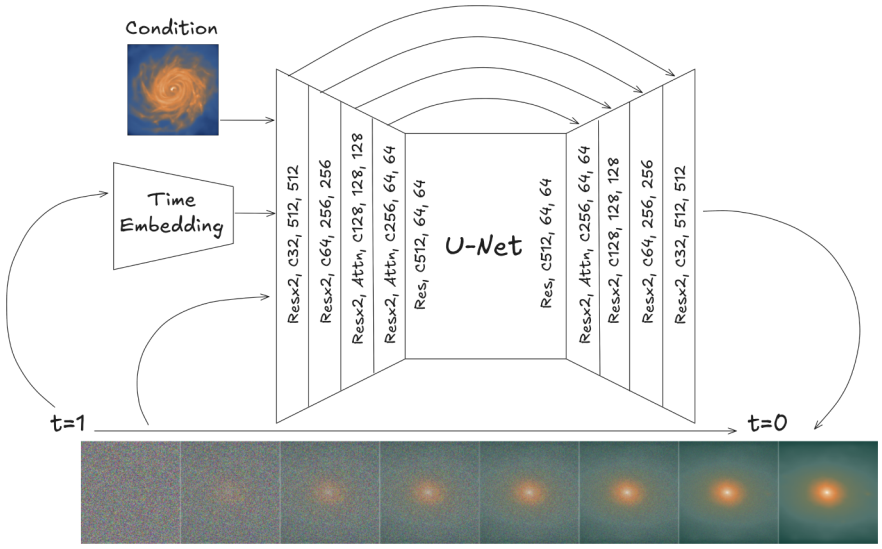


Figure 2. DDPM scheme: A noised input image is iteratively denoised from $t = 1$ to $t = 0$ in multiple steps by a U-Net generator, given the time step and the input image as condition. The U-Net architecture is the same as the one used for the GAN model (cf. Fig. 1) with two subtle differences: the architecture is narrower (in channel dimension) to optimize performance during inference and the additional (sinusoidal) time embedding input is forwarded to each residual block in the U-Net.

where the marginal probability is a pure Gaussian $p(x_T) = q(x_T) = \mathcal{N}(0, \boldsymbol{I})$.

The reverse step $p_\theta(x_{t-1}|x_t)$, that is the neural network, has various implementations. J. Ho et al. (2020) observed more stable training when the network only predicted μ_θ and assumed the variances to be time-dependent constants $\Sigma_\theta(t) = \beta_t \boldsymbol{I}$. Through the *reparametrization trick*, it is also possible to predict the added noise ϵ through a neural network ϵ_θ rather than the mean of the Gaussian, typically implemented as U-Nets (see Fig. 2 for an illustration). Other forms of diffusion models directly predict the original data point x_0, or some combination of both (T. Salimans & J. Ho 2022).

In any case, the diffusion loss can be shown to generally reduce to

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I}),t\sim\mathcal{U}_{[0,T]}}\left[\gamma'_\eta(t)||\epsilon - \epsilon_\theta(x_t, t)||^2\right], \tag{12}$$

where $\gamma'_\eta(t)$ is an optional weighting pre-factor (with learnable bounds) evaluated via automatic differentiation. The noise schedule $\gamma_\eta(t)$ also has various forms with the simplest schedule linearly increasing between two extremal bound hyperparameters $\eta = \{\gamma_{\min}, \gamma_{\max}\}$.

For conditional generative tasks, conditioning variables c (here images of the original domain) are fed as additional inputs to the network during training $\epsilon_\theta(x_t, t, c)$ (besides the noised image x_t and the time-step t). The conditioning can be further enforced by guiding the diffusion process, pushing the backward process in the direction of the gradient of the target condition probability (J. Ho & T. Salimans 2022). *Classifier-free diffusion guidance* achieves this through a modified training procedure by linearly combining null-labelled $\emptyset$ diffusion and conditioned diffusion $\tilde{\epsilon}_\theta(x_t, t, c) = \epsilon_\theta(x_t, t, \emptyset) + s(\epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, \emptyset))$ given a guidance strength s. At inference time, samples can be artificially pushed towards the conditional direction by increasing the guidance strength $s \geq 1$.

In this study, various noise schedules and objective variations have been optimized in hyperparameter searches, see Section 2.7 and Appendices A, B, and C for details.

2.4 Neural network architectures

All network implementations can be found in our CHUCHICHAESTLI package[1] published on PyPI and publicly available on https://github.com/CAIIVS/chuchichaestli. Here, we give an overview of their architecture, but for details we refer to the correspondingly listed sources.

2.4.1 U-Net

GAN as well as diffusion models implement their generative networks using the U-Net convolutional architecture, first introduced by O. Ronneberger et al. (2015). It was initially designed for segmentation of biomedical images, but has since been adapted to generative tasks for many other scientific fields (e.g. W. Yao et al. 2018; J. Andersson, H. Ahlström & J. Kullberg 2019; M. Bianco et al. 2025). Its basic structure consists of a contracting (encoder) and an expansive (decoder) path, resulting in characteristically U-shaped graphs. While the architectural blocks in a U-Net have seen various updates since its inception, the basic encoder level follows the typical convolutional network structure with repeated $3\,x\,3$ convolutional layers each followed by activations (LeakyReLU or ReLU) and a downsampling layer (a convolutional layer with stride 2); more recent versions additionally include residual block connections to improve gradient flow (K. He et al. 2015). With multiple levels, this leads to image compression, feature extraction, and ultimately representational learning. For image-to-image domain translation tasks, the structure of the decoder blocks is typically mirrored using deconvolutional layers to recover the input image resolution. Due to the repeated application of downsampling convolutional operations, spatial information is lost in deeper levels of the encoder. To this end, U-Nets additionally include skip connections between the corresponding levels which directly pass the encoder output information, concatenated to the output from lower decoder levels, and

[1] Release version `v0.2.13`.

effectively integrate spatial information in the expansive path of the U-Net.

The basic U-Net structure in this work resembles the implementation by P. Isola et al. (2016) with a few notable updates:

(i) we opted for Swish activation functions (P. Ramachandran, B. Zoph & Q. V. Le 2017) instead of ReLU and LeakyReLU,
(ii) each block optionally includes a self-attention (A. Vaswani et al. 2017) or convolutional self-attention layer (B. Yang et al. 2019),
(iii) dropout regularization in hidden layers (with a probability of 0.2).

Self-attention enables the handling of global interactions between pixels regardless of their relative position in the image and nicely complements the inherently local convolutional pixel treatment. Originally applied to language tasks, it quickly became an essential ingredient of any state-of-the-art neural network for image processing. However, since attention increases the computational complexity quadratically with sequence length, transformer networks become quickly infeasible, especially for high-dimensional data like images. N. Parmar et al. (2018), J. Ho et al. (2019), and D. Weissenborn, O. Täckström & J. Uszkoreit (2019) proposed various solutions to this problem which usually entail reducing the receptive field and long-range interactions as compromise. We tested such *convolutional self-attention layers* [2] in our U-Nets, but have not noticed any significant improvements in performance or efficiency over classical self-attention (see Section 3).

Moreover, for the use in diffusion models the U-Net additionally contains a sinusoidal time embedding (aka positional embedding) to keep track of the time-step in the diffusion process. The time embedding is injected in all residual blocks via linear projection layers whose outputs are added to the blocks' first convolutions.

2.4.2 PatchGAN

As introduced in Section 2.3.1 a PatchGAN is a patch-based discriminator which models an image as a Markovian field where each probability depends on neighbouring patches within a patch diameter. This concept was initially explored by C. Li & M. Wand (2016) in the context of texture synthesis and later implemented for general image-to-image translation by P. Isola et al. (2016). Our discriminator networks for adversarial training were adapted from P. Isola et al. (2016) with a 70 x 70 pixel receptive field. The implementation follows a simple convolutional block pattern consisting of batch normalization, activation (LeakyReLU), and two-dimensional downsampling convolutional layers.

2.5 Image-based evaluation metrics

To evaluate the similarity and quality of generated galaxy maps during and after training these neural networks, we first employ a set of widely used metrics from the CV domain. These metrics provide a baseline for assessing pixel-level accuracy (distortion), perceptual fidelity, and statistical realism in image synthesis tasks. While they are not tailored to astrophysical data, they offer valuable insights into the generative performance of deep learning models and can be used for initial hyperparameter tuning.

[2]Where key, query, and value representations are mapped using two-dimensional convolutions instead of fully connected linear layers.

2.5.1 Mean squared error (MSE)

MSE quantifies the average squared difference between corresponding pixels in two images

$$\mathrm{MSE}\,(x, \hat{x}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2\,, \tag{13}$$

where x_i and $\hat{x}_i$ are pixel values in the reference and generated images, respectively, and N is the total number of pixels.

It is sensitive to small pixel-level deviations and is often used to measure reconstruction accuracy. However, it does not account for perceptual or structural similarity.

2.5.2 Peak signal-to-noise ratio (PSNR)

PSNR expresses the ratio between the maximum possible pixel value and the power of the error signal

$$\mathrm{PSNR}\,(x, \hat{x}) = 10 \cdot \log_{10}\left(\frac{\mathrm{c}^2}{\mathrm{MSE}\,(x, \hat{x})}\right), \tag{14}$$

where c is the maximum pixel value range (typically 1 for normalized images, or 2 if the data range from –1 to 1).

Higher PSNR values indicate better fidelity. It is commonly used in image compression and denoising tasks. Due to the logarithmic scaling, the metric is often stated in the decibel unit (dB).

2.5.3 Structural similarity index (SSIM)

SSIM evaluates perceptual similarity by comparing luminance, contrast, and structural information between two images (Z. Wang et al. 2004)

$$\mathrm{SSIM}\,(x, \hat{x}) = \frac{(2\mu_x\mu_{\hat{x}} + k_1)\,(2\sigma_{x\hat{x}} + k_2)}{\left(\mu_x^2 + \mu_{\hat{x}}^2 + k_1\right)\left(\sigma_x^2 + \sigma_{\hat{x}}^2 + k_2\right)} \tag{15}$$

where μ, σ, and $\sigma_{x\hat{x}}$ are the means, variances, and covariances of the images, and k_1, k_2 are stabilizing constants.

SSIM ranges from 0 to 1, with higher values indicating greater structural similarity. It is more aligned with human visual perception than MSE or PSNR.

2.5.4 Fréchet inception distance (FID)

FID measures the distance between the distributions of real and generated images in a feature space extracted by a pre-trained neural network (such as C. Szegedy et al. 2015):

$$\mathrm{FID} = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right) \tag{16}$$

where (μ_r, Σ_r) and (μ_g, Σ_g) are the mean and covariance of real and generated image set features.

Lower FID scores indicate that the generated images are statistically similar to real ones in terms of feature distribution. FID is widely used to evaluate generative models such as GANs and diffusion models. However, since it is evaluated with model backbones typically pre-trained on ImageNet (J. Deng et al. 2009, an extensive data set consisting of three-channel, natural images), its application on scientific maps may be problematic. In our use case, each map is replicated on three-channels before its features

are extracted and thus does not exhibit the same colour variation as for natural images. Moreover, critics argue that FID's reliance on ImageNet-trained embeddings and its assumption of Gaussianity in high-level feature space render it ill-suited for domains with drastically different image statistics, such as scientific or medical imaging (T. Kynkäänniemi et al. 2022; S. Jayasumana et al. 2023). In contrast, some studies have shown that using ImageNet-trained features can still correlate better with human perception than domain-specific feature extractors, even in, e.g. medical image synthesis tasks (M. Woodland et al. 2024), which is why we decided to include it in our evaluation despite its controversy; correlation tests on the results of this work confirmed its validity (outlined in the Appendix D).

The metrics mentioned above serve as a foundational measures for evaluating image-to-image translation tasks. While they offer general-purpose assessments of distortion and perceptual image quality, they do not capture the domain-specific physical properties of galaxies. To address this, we complement them with a set of astrophysical metrics tailored to the structural and morphological characteristics of galaxy maps.

2.6 Astrophysical evaluation metrics

To assess the physical plausibility of the generated galaxy maps beyond pixel-wise similarity, we introduce a set of astrophysically motivated metrics. These metrics are designed to quantify structural, morphological, and distributional properties of galaxies, enabling a more rigorous comparison between generated and ground truth samples. Each metric captures a distinct aspect of galaxy morphology and mass distribution, reflecting the underlying formation scenario.

2.6.1 Asymmetry error (AE)

Asymmetry evaluates the rotational symmetry of a galaxy map by comparing it to its 180° rotated counterpart (centred on the galaxy). This is a standard morphological indicator in observational astronomy (first introduced by D. Schade et al. 1995), and often used to identify signs of mergers, tidal interactions, or structural disturbances. It is typically defined as part of the concentration–asymmetry–smoothness parameter system (CAS; C. J. Conselice, M. A. Bershady & A. Jangren 2000; C. J. Conselice 2003). Here, we define the AE in a slightly simplified adaptation as the difference between the normalized asymmetry of a ground truth I_r and generated map I_g

$$\mathrm{AE}(I_r, I_g) = \frac{\sum_{ij} |I_{r,ij} - I_{r,ij}^{180^\circ}|}{\sum_{ij} I_{r,ij}} - \frac{\sum_{ij} |I_{g,ij} - I_{g,ij}^{180^\circ}|}{\sum_{ij} I_{g,ij}}, \tag{17}$$

where I^{180° are the 180°-rotated map correspondents. Higher asymmetry errors with respect to generated maps indicate discrepancies in structural symmetry which may indicate unrealistic morphology or artefacts.

2.6.2 Smoothness/clumpiness error (SCE)

The so-called clumpiness quantifies the presence of small-scale structures such as star-forming regions or dense gas clumps (cf. CAS parameter system; C. J. Conselice et al. 2000; C. J. Conselice 2003). We calculate a proxy by subtracting a smoothed version of the map from the original and measuring the positive residuals, to avoid biasing the result through smoothing artefacts and removal of diffuse regions.

$$\mathrm{SCE}(I_r, I_g) = \frac{\sum_{ij} \max(I_{r,ij} - S_{r,ij}, 0)}{\sum_{ij} I_{r,ij}} - \frac{\sum_{ij} \max(I_{g,ij} - S_{g,ij}, 0)}{\sum_{ij} I_{g,ij}}, \tag{18}$$

where S is a smoothed (Gaussian blurred) version of the corresponding map I; here, a Gaussian kernel size of 10 per cent of the half-mass radius was used. A high SCE value may indicate excessive noise or unrealistic fragmentation, while a low error suggests smooth, well-resolved distributions. This metric is particularly relevant for evaluating the realism of baryonic, frictional components like gas and stars and substructure in DM haloes. Just as for the AE metric (equation 17), it should be noted that equation (18) can assume negative values, which signify more clumpiness (or asymmetry) in the generated sample. For the purpose of metric aggregation, however, the absolute magnitude (or squared) of these values should be taken.

2.6.3 Centre-of-mass distance (COMD)

COMD measures the Euclidean distance between the centre of mass of the generated map and that of the ground truth. The centre of mass of a galaxy reflects the spatial alignment of its component distribution.

$$\mathrm{COMD}(I_r, I_g) = \left\| \frac{\sum_{ij} p_{ij} I_{r,ij}}{\sum_{ij} I_{r,ij}} - \frac{\sum_{ij} p_{ij} I_{g,ij}}{\sum_{ij} I_{g,ij}} \right\|_2, \tag{19}$$

where p_{ij} are the spatial coordinates of distribution elements (pixels). To our knowledge, this type of COM-based evaluation introduced here, has never been used for deep learning before. Misalignment may indicate translation artefacts, structural inconsistencies, or failure to preserve spatial coherence. This metric is particularly important for tasks involving domain translation where positional accuracy is critical.

2.6.4 (Cumulative) Radial curve errors (CRCE/RCE)

This metric compares the radial intensity or mass profile of the generated map to that of the ground truth. The radial profile is computed by calculating the mass fraction in concentric radial bins centred on the galaxy's centre of mass

$$\mathrm{RCE}(I_r, I_g) = \frac{1}{K} \sum_{k}^{K} \left| \frac{\sum_{ij \in \Psi_k} I_{r,ij}}{\sum_{ij} I_{r,ij}} - \frac{\sum_{ij \in \Psi_k} I_{g,ij}}{\sum_{ij} I_{g,ij}} \right|, \tag{20}$$

where K is the number of concentric bins and Ψ_k the set of pixels located within the k-th radial annulus centered on the galaxy's centre of mass. This treats the radial distribution as a normalized probability density function multiplied by the bin width, ensuring the metric is scale-invariant. Thus, *Radial Curve Errors* capture relative deviations in the spatial distribution of matter rather than total intensity. It is essential for validating the structural integrity of generated galaxies; by quantifying the concentration at different radii, it is a strict test of morphology. As a complement, the analogous comparison of cumulative radial distributions of generated and ground truth maps measures radial scale length and how centrally concentrated the mass distribution is. Errors from cumulative profiles are sensitive to morphological compactness

and spatial allocation. Essentially, this metric is a generalization of the concentration parameter from the CAS system introduced by (cf. CAS parameter system; C. J. Conselice et al. 2000; C. J. Conselice 2003).

2.6.5 Power spectrum errors (PSE)

PSE compares the radially averaged 2D power spectrum shapes (i.e. squared magnitude of the Fourier coefficients at each frequency) of two maps. For each map, we compute the 2D discrete Fourier transform $\mathcal{F}$ and derive the power spectrum as the squared modulus of the Fourier coefficients. The resulting 2D power spectrum is then radially averaged in Fourier space to obtain a 1D power spectrum curve which characterizes the distribution of power as a function of spatial frequency (the scales correspond to each map's extent). Finally, the normalized power spectrum curve residuals can be reduced by means of summation or averaging.

$$\mathrm{PSE}(I_r, I_g) = \sum_k^K \left| \frac{\sum_{uv \in \tilde{\Psi}_k} |\mathcal{F}(I_r)_{uv}|^2}{K \sum_{uv} |\mathcal{F}(I_r)_{uv}|^2} - \frac{\sum_{uv \in \tilde{\Psi}_k} |\mathcal{F}(I_g)_{uv}|^2}{K \sum_{uv} |\mathcal{F}(I_g)_{uv}|^2} \right| , \quad (21)$$

where $\tilde{\Psi}_k$ is the set of pixels located within the kth radial annulus in 2D (u, v) Fourier space. This approach assesses similarity in spatial structure, texture, and particularly characteristic, second-order (filamentary) scales, independent to normalization, translation, or rotation. The PSE is especially useful when validating a model's accurate reproduction of multiscale spatial features. Although power spectra are often utilized in cosmological and astrophysical studies, to our knowledge, this work represents the first adaptation as a metric for evaluating deep generative models.

The aforementioned metrics collectively provide a robust framework for evaluating the astrophysical realism of generated galaxy maps. They complement traditional image similarity metrics from Section 2.5 by incorporating domain-specific knowledge and physical constraints, thereby enabling a more meaningful assessment of generated samples.

Finally, beyond pixel-level and morphological assessments, we also perform an inter-model consistency analysis based on integrated physical quantities (such as e.g. average magnetic field strength or total mass content). By substituting individual components with model predictions while keeping others at ground truth, we quantify biases and scatter, as well as cross-source disagreement for a fixed target domain. These metrics reveal whether different translation models preserve masses and energy globally and maintain physically plausible component fractions, independent of local image fidelity. Furthermore, they test whether models can be chained in cycles, potentially avoiding the need to train all model translation permutations if the goal is to complete a physical model from an arbitrary galaxy property. This approach provides a complementary, physically grounded perspective on model performance, ensuring that generated maps respect fundamental conservation principles and astrophysical scaling relations.

2.7 Experiments

Given that both generative methodologies described in Sections 2.3.1 and 2.3.2 employ U-Net architectures as their backbone, it is essential to optimize the architectural hyperparameters for the specific characteristics of the data set. However, exhaustive hyperparameter searches across all possible configurations are computationally prohibitive, especially for generative models. We therefore constrain our ablation studies to architectural components that have demonstrated the most significant impact on conditional image generation performance; detailed results of these architectural studies can be found in the Appendices A, B, and C. For these ablations, a coarse grid search across diverse optimizers, learning rates, and loss term weights have been carried out beforehand to find good values/choices. The hyperparameters for the final network architectures used in Section 2.8 have been optimized using the Optuna (T. Akiba et al. 2019) and Ray Tune (R. Liaw et al. 2018) frameworks.

Based on these hyperparameter searches, we adopt a single backbone configuration for all domain-translation experiments to ensure comparability across tasks and model families. For GAN-based models, the generator is a four-level U-Net (base width 64 channels) with residual blocks, Swish activations, and dropout ($p = 0.2$), augmented with three self-attention blocks placed in the lowest-resolution stages near the bottleneck (starting at the third encoder level) to capture long-range structure at moderate computational cost (illustrated in Figs 1 and 2). For diffusion-based models, we use the same U-Net design and attention placement, but with a narrower channel width (base 32 channels) and a sinusoidal time embedding injected into each residual block. This architecture is kept fixed across all domain pairs, and only task-specific training dynamics (e.g. conditional inputs and objective weights) follow the tuned settings above.

All models were implemented in PYTORCH. The experiments were conducted on Nvidia V100/A100/H100/H200 GPUs depending on the specific VRAM requirements. Adam optimizers (separate ones in the case of GAN-based models) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and weight decay of 10^{-5} were used. Unless otherwise stated, the maximum learning rate was set to 5×10^{-5} for generators and 1×10^{-5} for discriminators (where used), with a one-cycle policy schedule (following L. N. Smith & N. Topin 2017). It provides a smoother warm-up phase at lower learning rates, a ramp up to the maximum learning rate, and a cosine-annealing phase to 10^{-4} of the maximum value. When attention layers are included, this schedule was found to lead to less instabilities during GAN training. DDPM models used a cosine noise schedule with $T = 300$ iteration steps (for training and sampling), classical DDPM sampling, and a classifier-free guidance strength of 1. All model experiments were trained for 30 epochs with a batch size of 8. The data sets were split into 85 per cent training, 10 per cent validation, and 5 per cent test sets, ensuring that galaxies from the same halo did not appear in multiple splits. All reported metrics for experiments in Section 2.7 and Appendices A, B, and C were evaluated on the validation set whereas astrophysical validation (in Section 2.8) was performed on the test set, and reported after the final epoch.

2.8 Domain translations

With all model components conservatively optimized, the final stage of experiments extended the map-to-map translation task to encompass all available domains. Given the combinatorial nature of the data set, exhaustively exploring all 5040 possible domain translations is infeasible. However, it is reasonable to expect that the complexity of translations tasks varies between the astrophysical interactions between the components. For instance, domain translations such as GAS → HI or 21CM → GAS are likely to be less complex, as they represent information completion or reduction. On the other hand, mappings like STARS → DM are inherently

more challenging due to the 'weak' coupling of these components in the simulation.

To capture this diversity, we selected a representative sub-set of domain pairs that span a broad range of translation difficulties; the translations are centred around GAS due to its close relation to observable quantities and, thus, astronomical relevance (as mentioned in Section 1). These included the following mappings:

(i) within baryonic components

  (a) GAS → HI,
  (b) GAS → 21CM,
  (c) 21CM → GAS,
  (d) GAS → STARS

(ii) baryonic-to-DM translations

  (a) GAS → DM,
  (b) DM → GAS

(iii) thermodynamic transformations

  (a) GAS → TEMP

(iv) magnetic field strength reconstructions

  (a) GAS → BFIELD.

For each selected pair, models were trained using the optimized U-Net configuration identified in previous experiments, with attention layers placed near the bottleneck.

Both GAN-based and diffusion-based models were evaluated, and their outputs compared using the full suite of CV and astrophysical metrics. This strategy allowed us to assess not only the fidelity of individual translations but also the consistency of physical quantities across domains. In particular, we investigated whether certain domain pairs exhibit systematic biases or structural artefacts, and whether translation difficulty correlates with the intrinsic entropy or sparsity of the source domain. The results of these experiments are summarized in the following Section 3.

3 RESULTS

In this section, we present a comprehensive evaluation of the proposed generative models across multiple galaxy-domain translation tasks. Section 3.1 provides a qualitative assessment of representative samples to illustrate visual fidelity and structural realism. Next, we report quantitative performance metrics for all tested domain mappings, including both traditional computer vision measures and astrophysically motivated indicators in Section 3.2, followed by an interpretation of these metrics and their implications for morphological plausibility in Section 3.3. We then compare the relative strengths and trade-offs between GAN-based and diffusion-based approaches in terms of accuracy, stability, and computational efficiency (Section 3.4). Finally, we examine the global consistency of inferred physical quantities across the unseen test population to assess whether models preserve integrated properties such as total mass and energy (Section 3.5). Together, these analyses provide a multifaceted view of model performance, highlighting correlations between translation difficulty and physical coupling (a detailed correlation analysis is provided in Appendix D).

3.1 Qualitative assessment of samples

Fig.3 shows representative samples of map-to-map translations across the (unseen) test set of domain pairs. Each triplet shows the input map (left), the ground truth target (middle), and the model prediction (right). For strongly coupled domains such as GAS → DM, both GAN and DDPM reproduce global morphology and substructures with high fidelity across various scales and mass ranges. In some cases, smaller satellite haloes are either missing or were generated without any counterpart in the ground truth maps. When present, they are typically plausible domain translations of the input map.

Also, the translations GAS → HI, GAS → 21CM, and 21CM → GAS are consistently in excellent agreement for both models, with only mild over or underestimation in some systems.

For thermodynamic and field-like targets (GAS → TEMP, GAS → BFIELD), DDPM predictions better preserve global gradients, whereas GANs sometimes sharpen local contrast and slightly overemphasize smaller map features.

The arguably most challenging inverse mappings (e.g. DM → GAS) reveal residual artefacts and misaligned substructures for both models, underscoring the difficulty of inferring baryonic components from DM alone.

Similarly, both models struggle to faithfully reproduce translations involving the weakly correlated components GAS → STARS. Samples from this task exhibit noticeable deviations: predicted stellar maps fail to capture the clumpy, centrally concentrated structures, reflecting the intrinsic (temporal) non-locality and higher entropy of the stellar distribution.

Overall, these examples illustrate that translation quality correlates strongly with the physical coupling between source and target domain, and that GAN models and DDPM reproduce very similar samples and differ mostly in the details and high-frequency features: GANs often excel in structural sharpness for tightly coupled mappings, whereas DDPMs better maintain global coherence in more weakly constrained tasks.

3.2 Overall performance across translation tasks

The measured model performance varies systematically with the physical coupling between source and target domains (Tables 2 and 3). As in previous experiments, the SSIM metric saturates quickly during training and is less discriminative than the other image-based metrics.

Table 2 lists image-based (traditional CV) metric evaluations for all domain translation tasks, grouped in pairs of GAN and DDPM. The best mean value of each metric across all tasks and models is listed in bold. Similarly, Table 3 shows the set astrophysical metric evaluations in the same order and grouping.

Among all tested translations, GAS → DM attains the highest overall fidelity: GAN and DDPM models reach best FID scores of 1.56 ± 0.36 and 2.03 ± 0.08, respectively, with PSNR values above 35 dB and SSIM $\gtrapprox 0.997$ (see Table 2). The astrophysical metric evaluations listed in Table 3 confirm this trend for GAS → DM: asymmetry and clumpiness errors are among the smallest, COM offsets are negligible, and cumulative total mass deviations (evaluated at R_{50}) and power-spectrum errors remain modest.

Translations within the baryonic sector also perform strongly when the target is closely tied to the gas morphology. GAS → HI and GAS → 21CM achieve low FID values of 4–6 and competitive PSNR/MSE. These models show low morphological errors (AE and SCE), minimal COM drift, excellent recovery of the radial profiles, and reproduce the expected near-monotonic relations between the total gas mass, neutral hydrogen mass, and 21-cm brightness temperature.

Moreover, the inverse mapping 21CM → GAS remains tractable with similar FID values up to 7.6, competitive ranges for the other

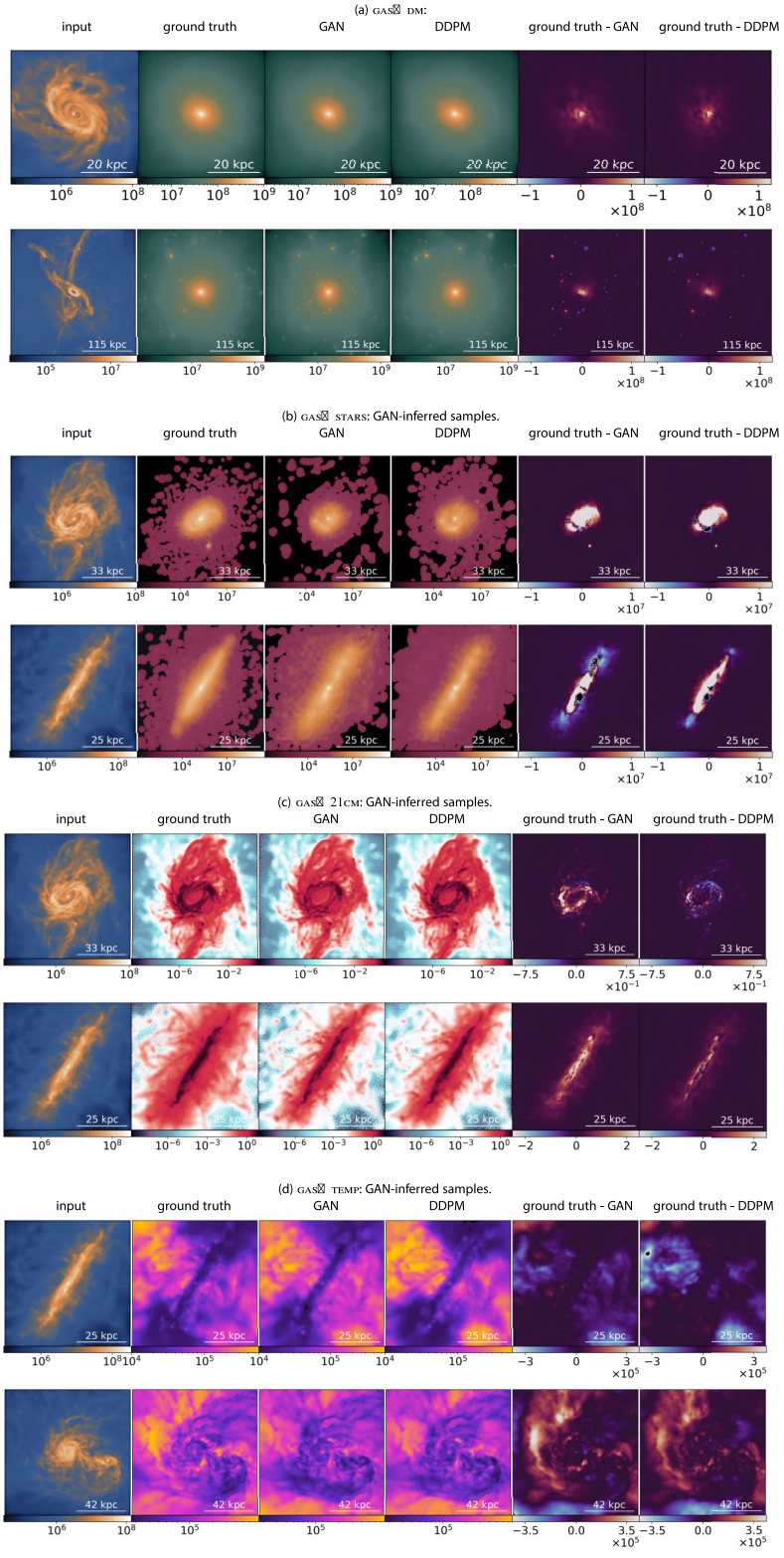


Figure 3. Samples from selected models and tasks (more in the Appendix E). Each panel shows a model input map on the left, the corresponding ground truth, and prediction from GANs and DDPMs on the right. The rightmost two maps show residuals between ground truth and GAN, and DDPM, respectively. Qualitative comparison confirms the alignment of astrophysical plausibility and human perception with astrophysical metrics and FID (see Tables 2 and 3). The comparison of residual maps indicates that GANs, while exhibiting subtle error biases shifting the mean of the error distribution away from zero at times (particularly evident in, e.g. the first sample of GAS→21CM, Fig. 3c), have lower absolute errors but higher cumulative errors. A comparison with quantitative results confirms this observation. Furthermore, FID (and other astrophysical metrics) seem to be most sensitive to structural errors rather than absolute error magnitudes, whereas the opposite is the case for metrics like PSNR (for which a small number of high-magnitude error pixels can dominate the metric).

Table 2. Extensive results for the entire suite of map-to-map translation models with image-based metrics (see Section 2.5). The values listed are mean of the respective metrics from the last five epochs (duration chosen as patience parameter when testing for convergence), as metric values for GANs fluctuate more. Note that the data ranges differ for GAN and DDPM models, which inherently biases the metric towards DDPMs by ∼6.02 dB for the same MSE value. Thus, the PSNR values for DDPM models were implicitly unbiased in the discrimination analysis. The aggregate scores are RMS (root mean square) of all other metrics normalized to their 5th–95th percentile.

Translation	Model	PSNR ↑	SSIM ↑	MSE ($\times 10^{-4}$) ↓	FID ↓	Aggregate score ↑
GAS→DM	GAN	35.31	**0.9974**	3.82	**1.56**	**0.9702**
GAS→DM	DDPM	41.17	0.9970	4.24	2.03	0.9673
GAS→STARS	GAN	18.55	0.5738	324.12	60.56	0.0092
GAS→STARS	DDPM	23.34	0.5577	324.60	56.17	0.0099
GAS→HI	GAN	33.27	0.9739	15.31	4.57	0.9113
GAS→HI	DDPM	39.99	0.9749	17.11	5.86	0.9122
GAS→21CM	GAN	31.60	0.7958	17.90	3.57	0.8066
GAS→21CM	DDPM	38.55	0.8133	17.95	5.78	0.8123
GAS→TEMP	GAN	37.05	0.9973	5.04	9.91	0.9558
GAS→TEMP	DDPM	41.56	0.9967	3.99	7.86	0.9459
GAS→BFIELD	GAN	**38.76**	0.9964	**2.75**	9.80	0.9658
GAS→BFIELD	DDPM	43.39	0.9955	3.60	8.38	0.9663
DM→GAS	GAN	31.28	0.9853	12.18	36.36	0.7876
DM→GAS	DDPM	36.96	0.9845	10.62	22.87	0.8227
21CM→GAS	GAN	35.95	0.9904	4.46	7.60	0.9478
21CM→GAS	DDPM	42.08	0.9900	3.75	5.63	0.9580

Table 3. Extensive results for the entire suite of map-to-map translation models with mean-averaged astrophysical metrics (see Section 2.6; R_{50} denotes the half-mass radius). The values listed are mean of the respective metrics from the last five epochs (duration chosen as patience parameter when testing for convergence). The aggregate scores are RMS (root mean square) of all other metrics normalized to their 5th–95th percentile.

Translation	Model	AE ↓	SCE ↓	COMD ↓	CRCE (at R_{50}) ↓	PSE ↓	Aggregate score ↑
GAS→DM	GAN	0.0655	0.0027	0.0211	0.2132	0.0788	0.8483
GAS→DM	DDPM	0.0746	0.0032	0.0215	0.2196	0.0856	0.8352
GAS→STARS	GAN	0.7460	0.0975	0.0657	1.3772	0.0690	0.1657
GAS→STARS	DDPM	0.4466	0.0812	0.0297	1.2875	0.0596	0.3719
GAS→HI	GAN	0.0839	0.0207	0.0128	0.2684	**0.0307**	**0.9292**
GAS→HI	DDPM	0.0885	0.0219	0.0136	0.2948	0.0363	0.9019
GAS→21CM	GAN	0.0713	0.0186	**0.0109**	0.2192	0.0452	0.9103
GAS→21CM	DDPM	0.0813	0.0210	0.0120	0.2765	0.0524	0.8686
GAS→TEMP	GAN	0.0901	0.0024	0.0561	0.1754	0.0568	0.8029
GAS→TEMP	DDPM	0.0793	**0.0019**	0.0484	**0.1605**	0.0597	0.8083
GAS→BFIELD	GAN	0.0822	0.0093	0.0371	0.2209	0.1000	0.7602
GAS→BFIELD	DDPM	**0.0647**	0.0072	0.0294	0.1928	0.0875	0.8047
DM→GAS	GAN	0.1093	0.0224	0.0367	0.3143	0.0328	0.8191
DM→GAS	DDPM	0.1085	0.0184	0.0357	0.2946	0.0333	0.8327
21CM→GAS	GAN	0.0891	0.0161	0.0148	0.3483	0.0641	0.8229
21CM→GAS	DDPM	0.0705	0.0131	0.0124	0.3231	0.0621	0.8597

pixel-wise metrics. The aligned performance with its counterpart across all astrophysical metrics suggests that reconstructing gas maps from observational 21-cm inputs is feasible.

In contrast, mapping DM → GAS is substantially harder, only scoring within an FID range between 22 and 45, and slightly but consistently worse results across all astrophysical metrics.

However, the most challenging mapping is clearly GAS → STARS, which yields FID scores above well above 50, PSNR below 20 dB, and SSIM values well below the normal saturation levels. The large morphological errors, especially in asymmetry, reflect the models' inability to capture the alignment and ellipticity of the mass distributions, and the clumpiness errors indicate the models' difficulty to cope with the high non-locality of the stellar components.

We note that hyper-parameters (U-Net size/attention and PatchGAN settings) were primarily optimized on the GAS → DM task (Section 2.7); this may confer a slight advantage to GAS → DM in cross-task comparisons. Thus, we repeated a small hyper-parameter sweep for a balanced set of tasks (GAS → STARS, GAS → HI, and DM → GAS) to assess possible task-selection bias and performed a regret analysis based on the average FID score. The resulting task ranking was unchanged and the regret of the optimal configuration (as in Section 2.7) remained small across tasks, indicating that the results reflect intrinsic task difficulty rather than tuning alone.

3.3 Metric interpretation

Direct comparison of PSNR and MSE across models require care because of the different preprocessing ranges (see equation 2): GAN inputs/outputs are mapped [0,1], while DDPMs use [−1, 1]. The DDPM pixel value range is a factor of 2 larger, which biases PSNR by roughly 6.02 dB for the same MSE. Thus, throughout

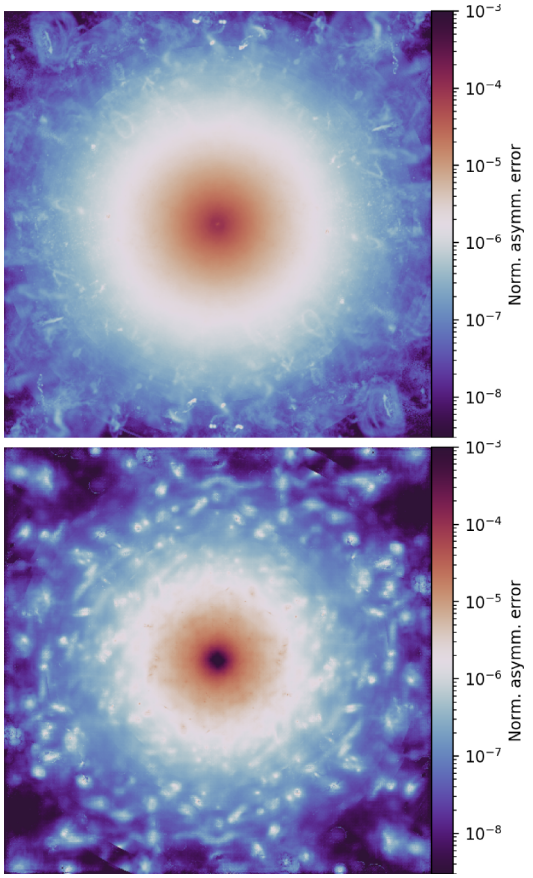


Figure 4. Examples of normalized asymmetry error maps for the mappings GAS→21CM (top) and GAS→STARS (bottom) in the test set, inferred by GANs. The overall mean error is around an order of magnitude larger for GAS→STARS and relatively uniform but exhibits a slight chequerboard pattern, indicating the difficulty to model the fine-grained structure of the stellar mass distribution. GAS→21CM exhibits smaller irregular errors which are noticeable due to overall lower average error.

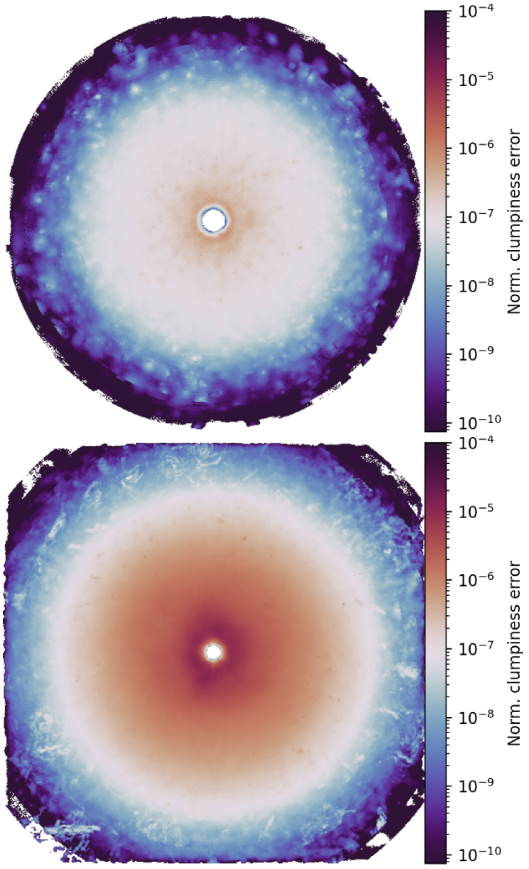


Figure 5. Examples of normalized clumpiness error maps for the mappings GAS→DM (top) and DM→GAS (bottom) in the test set, inferred by GANs. To keep numerical stability, the inner regions of the error maps have been masked to 5 per cent of the map's respective half-mass radius. The overall mean error is around an order of magnitude larger for DM→GAS, indicating the increased difficulty of predicting baryonic components from DM compared to the inverse mapping. Moreover, due to the collisionless nature of DM, its distributions tend to be smoother, which also contributes to the lower mean error. For GAS→DM, errors mainly arise due to the wrong estimate of DM substructure in the haloes, whereas errors for DM→GAS indicate unrealistic fragmentation in small-scale structures.

the cross-model PSNR comparisons in this Section 3 we implicitly remove this bias.

On these data domains, pixel distortion metrics PSNR, MSE, and especially SSIM are near-ceiling for several tasks (e.g. GAS → DM) and can underdiscriminate subtle morphological differences in smooth, high-resolution simulation maps. Conversely, the suite of astrophysically motivated metrics (AE, SCE, COMD, CRCE, and PSE) remains sensitive to structural realism and are model-agnostic. Moreover, we observed a strong correlation of FID with AE, SCE, and CRCE (Pearson $r \geq 0.88$; for details, see in Appendix D).

Figs 4 and 5 illustrate how global errors manifest spatially. Both metrics measure important features (e.g. structural symmetry and fine-structure resolution) indicative of morphological realism and overall plausibility of generated samples. Fig. 4 shows the average asymmetry error map for GAS → 21CM versus GAS → STARS. Errors of the latter task are roughly an order of magnitude larger with a slight chequerboard pattern, indicating unresolved fine-structure and adversarial artefacts. For GAS → DM and DM → GAS (Fig. 5) the harder inverse mapping (latter) exhibits higher small-scale residuals consistent with unrealistic fragmentation.

Centre-of-mass drift errors can also be decomposed in more detail (Fig. 6). While the COMD only measures the scalar global drift, higher values may have different causes: the upper panel shows a near-uniform distribution of COM drifts (good positional agreement), whereas the lower panel exhibits a noticeable angular bias, signalling a systematic vectorial drift of the inferred mass centroid.

3.4 Model types: performance and trade-offs

There is no universal winner between GANs and DDPMs across all tasks. GANs tend to achieve lower FIDs when the target is tightly tied to the gas morphology (e.g. GAS → DM, GAS → HI, and GAS → 21CM), while DDPMs often deliver more favourable astrophysical fidelity (lower AE, SCE, and COMD) for less strongly related quantities such as GAS → TEMP, or GAS → BFIELD. Moreover, GANs inherently exhibit more quality fluctuations even long into training due to the adversarial nature of their objective; this is evidenced by the typically higher standard deviations of the metric results from the last five epochs. These complementary behaviours suggest that adversarial training sharpens struc-

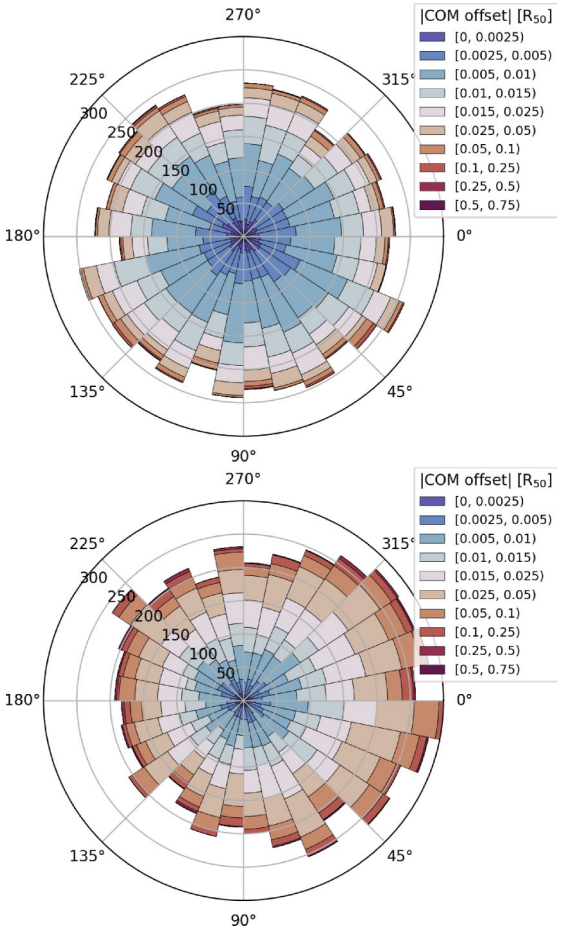


Figure 6. Examples of the angular distribution of COM drifts for the mappings GAS→HI (top) and GAS→STARS (bottom) in the test set, inferred by DDPMs. While the upper wind rose diagram shows a uniform distribution for GAS→HI COM drifts, GAS→STARS exhibits an angular bias towards 0°. The concentration of these errors in lower offset bins (in units of the half-mass radius R_{50}), as shown for GAS→HI, indicates low overall drift and typically good agreement with the ground truth.

tural realism in strongly coupled mappings, whereas diffusion-based modelling better preserves global morphology for thermodynamic and field-like targets. From a resource perspective, the GAN models in this work required ∼140 kWh training energy versus ∼520 kWh for DDPMs in our set-up (sum over all training run read-outs from Nvidia telemetry; excluding ablation tests, inference, and CPU/network). Both approaches are orders of magnitude more energy-efficient than re-running comparable hydrodynamical simulations $\mathcal{O}$(GWh) (cf. table 1in D. Nelson et al. 2019), but the ∼ 4× advantage of GANs can be decisive when many map-to-map translation models need to be trained.

3.5 Global consistency of inferred quantities

Fig.7 compares integrated inferred properties against ground truth for the unseen test population. For strongly coupled mappings such as GAS → HI and GAS → 21CM, both GAN and DDPM models recover total masses with minimal bias and scatter, indicating robust conservation of global properties. Notably, 99.9 per cent of errors for all listed mappings, including GAS → DM, 21CM → GAS, GAS → TEMP, GAS → BFIELD, are within a factor of 10 (see also Table 3). In contrast, the mapping GAS → STARS is exceptionally challenging for both models and presents large scatter and bias patterns (Fig. 7b): DDPMs overpredict at low masses and under-predict at the high-mass end, while GANs exhibit smaller mean bias but extreme scatter reaching beyond two orders of magnitude. These outcomes mirror the expected entropy and non-local differences among target domains and underline the task difficulty ordering observed in the other astrophysical metrics.

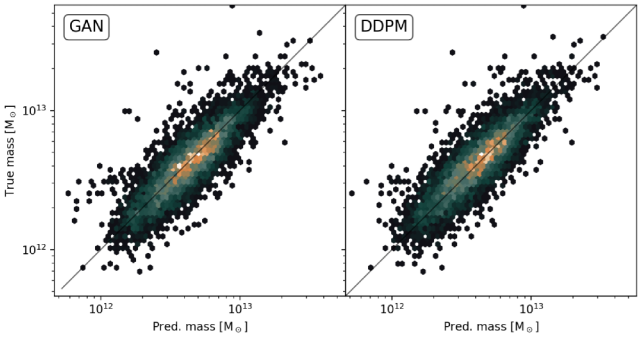
4 DISCUSSION AND CONCLUSIONS

We presented the first systematic study of multidomain map-to-map translations for galaxy formation simulations, introducing deep generative models as scalable data-driven alternatives that map between seven physical domains (DM, stellar mass, gas mass, neural hydrogen mass, 21-cm mock brightness, temperature, and magnetic field strength), comparing adversarial (GAN) and diffusion (DDPM) deep learning approaches under unified preprocessing and evaluation. Both approaches are able to learn physically plausible solutions to these domain translations, demonstrated on a data set of galaxy maps extracted from the IllustrisTNG suite (TNG50–1). Across extensive ablations and metrics – distortion (MSE, PSNR, SSIM), perceptual (FID) and astrophysical metrics (asymmetry, clumpiness, centre-of-mass drift, radial/cumulative curves, power spectra) – we find that translation difficulty strongly correlates with the physical coupling of source and target: GAS → DM achieves the best fidelity measured by image-based metrics (FID ≈ 2.0), GAS → HI, GAS → 21CM, and 21CM → GAS are likewise strong and conserve integrated quantities, while DM → GAS is substantially harder but still produces plausible results. GAS → STARS remains the most challenging across all measures. GANs tend to excel for tightly coupled targets with sharper structure and lower FID, whereas DDPMs better preserve global morphology and thermodynamic or field-like structure; this complementarity comes with a ∼ 4× difference in training energy in our setup (∼140 kWh versus ∼520 kWh). These results demonstrate the feasibility of learnt representations that encapsulate aspects of a simulation's formation scenario Φ from different observationally motivated inputs, while underscoring the need for domain-ware metrics and physics-informed inductive biases to tackle weakly constrained mappings. Notably, despite the controversy around the use of FID in scientific domains (cf. 2.5), it correlated surprisingly strongly with the astrophysical metrics which capture structural realism, suggesting it is an appropriate discriminator for our use case.

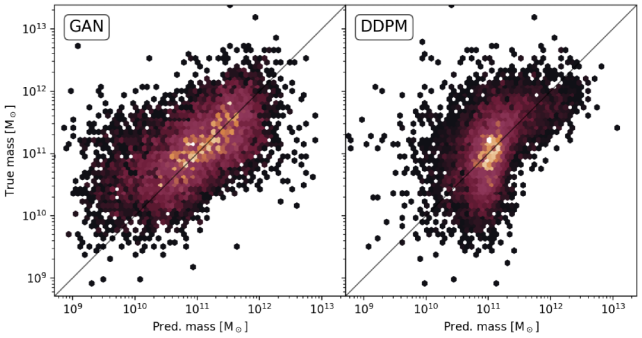
4.1.1 *Physical couplings*

The empirical task ordering we observe follows the expected information coupling among galaxy components. Gas traces the gravitational potential well and interacts collisionally, so GAS → DM is comparatively well-posed: large-scale morphology and substructure are strongly correlated, enabling excellent astrophysical veracity (low FID and morphological errors; see Tables 2 and 3 and Fig. 5). In contrast, the inverse mapping DM → GAS is underconstrained: while the DM halo delineates the potential, baryon distributions are additionally set by feedback, heating, and cooling; our models thus exhibit more clumpiness residuals and centre-of-mass drift (Fig. 5), consistent with fragmentation artefacts. The most difficult case, GAS → STARS, reflects the intrinsically non-local nature and higher entropy of stellar mass

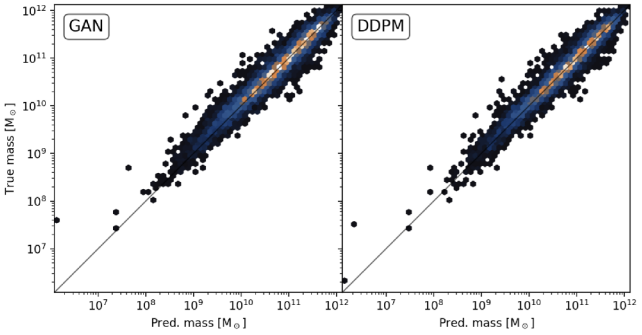
(a) GAS→DM: Inferred total DM mass of all galaxies in the test set.



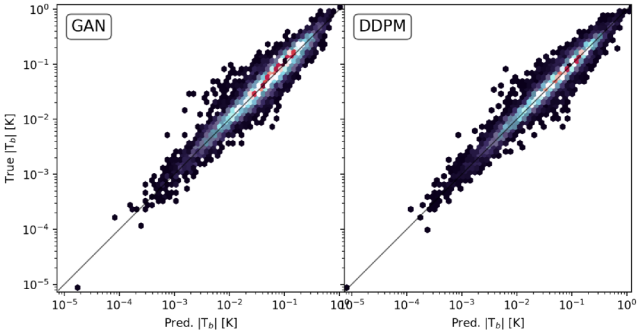
(b) GAS→STARS: Inferred total stellar mass of all galaxies in the test set.



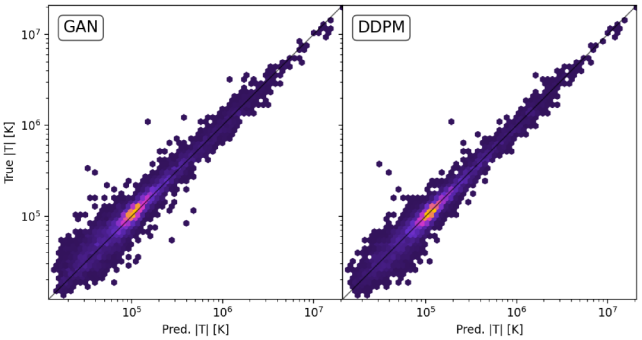
(c) GAS→HI: Inferred total neutral hydrogen gas mass of all galaxies in the test set.



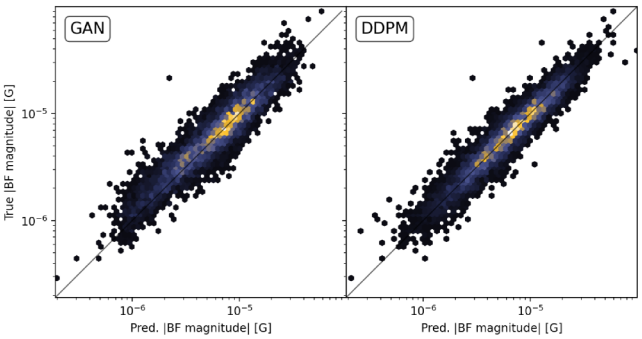
(d) GAS→21CM: Inferred average 21-cm brightness temperature of all galaxies in the test set.



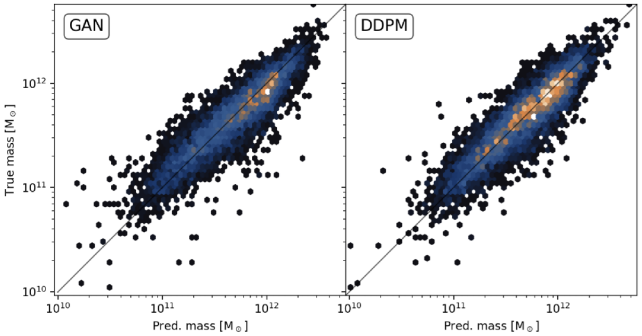
(e) GAS→TEMP: Inferred average temperature of all galaxies in the test set.



(f) GAS→BFIELD: Inferred average magnetic field strength of all galaxies in the test set.



(g) DM→GAS: Inferred total gas mass of all galaxies in the test set (from DM maps).



(h) 21CM→GAS: Inferred total gas mass of all galaxies in the test set (from 21-cm maps).

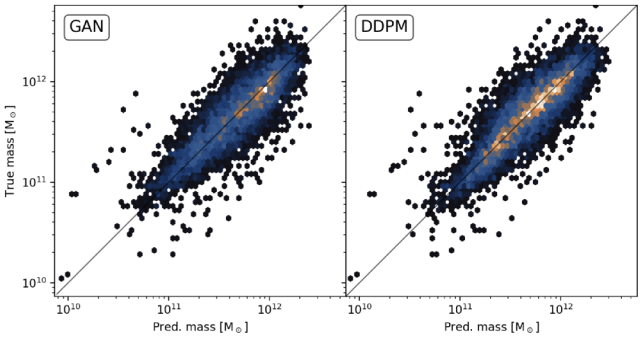


Figure 7. Global statistics of inferred versus true integrated quantities. The colour scheme qualitatively indicates histogram density and matches the task assignment analogous to Fig. 3. In general, GANs and DDPMs show no biases and minimal scatter of integrated quantities (except for GAS→STARS).

assembly: star formation depends on history and feedback cycles only weakly encoded in a single gas snapshot, leading to large asymmetry and clumpiness errors, poor FID, and strong biases in integrated stellar mass (see Tables 2 and 3 and Figs 4 and 6). Altogether, the results corroborate the conceptual view in equations (3) and (4): learning conditional terms is easier when nuisance parameters are few and the conditional entropy of the target given the source is low.

Integrated quantities provide an orthogonal check of global physical plausibility. We find minimal bias and scatter for most mappings. The outlier is GAS → STARS, which shows systematic bias and large scatter for both model types (Fig. 7b). These findings imply that for a sub-set of domains with strong coupling, chaining of models (e.g. 21CM → GAS → DM) may be feasible without much loss of information (without explicitly training for cycle-consistency).

4.1.2 Model choice guidance

No single model type dominates across all translations. Adversarial training yields good high-frequency results (at times mildly exaggerated), especially for targets strongly coupled to the gas morphology, but exhibits larger epoch-to-epoch variability – a hallmark of the minimax optimization game (cf. Tables 2, 3, and Section 2.3.1, equation 5). Diffusion models tend to preserve global gradients and in more complex couplings and often improve astrophysical plausibility at the cost of slower sampling and higher training time and energy. From a practitioner's standpoint:

(i) choose GANs for tight, morphology-driven mappings where fast inference is worth the small trade-off in accuracy;
(ii) choose DDPMs when the target encodes smoother or more complex fields, when robustness in astrophysical accuracy has highest priority.

Importantly, through targeted architectural and training optimizations, including U-Net depth/width tuning, attention placement near the bottleneck, and discriminator sizing (Section 2.7), we demonstrate that GAN-based models can achieve performance on par with state-of-the-art DDPMs for most mappings. This parity, combined with GANs' lower training energy and single-pass inference, positions them as a competitive and computationally sustainable alternative for large-scale deployment. Moreover, a hybrid approach which draws from each methods advantages while mitigating their disadvantages, could be an promising avenue for future work.

4.1.3 Implications for observations

This work offers a direct path to observational validation by incorporating domains that are measurable in practice, such as 21cm brightness and neutral hydrogen, into the translation process. This capability is particularly critical for the SKA, which will probe the distribution of H I in nearby galaxies to unprecedented precision. Here, two practical applications of our models emerge:

(i) *Forward modelling*: predicting 21-cm brightness from simulated gas maps and pass through an instrument response pipeline (such as Karabo; R. Sharma et al. 2025) for high-realism mock observations including SKA-like systematics.
(ii) *Reconstruction*: inferring gas distributions and related galactic properties from observed 21-cm maps of nearby galaxies to support feedback and morphological studies.

By embedding observational proxies and incorporating, e.g. beam smoothing, thermal noise, and foreground residuals into the generative framework (during training or via data augmentation), domain-shift robustness is increased; our astrophysical metrics are naturally suited to quantify degradation after instrumental effects. This provides a scalable pathway to interpret SKA data within the context of galaxy formation scenarios.

4.1.4 Limitations

Our models learn by design conditional slices of a simulation's formation scenario Φ. Because Φ depends on sub-grid physics and calibration, generalization across suites (e.g. IllustrisTNG, SIMBA, FIRE, or EAGLE) and redshift evolution must be demonstrated rather than assumed.

Furthermore, perceptual metrics such as FID carry domain-mismatch assumptions; fine-tuning feature extractors on domain-specific (astrophysical) data could provide an even better measure for astrophysical veracity. More flexible alternatives to FID such as LPIPS (Learned Perceptual Image Patch Similarity; R. Zhang et al. 2018) could improve evaluation fidelity even further.

Translation with weak couplings could be improved with additional constraints. Models for the GAS → STARS mapping lack sufficient mutual information between input and target domains, making the task particularly challenging.

4.1.5 Outlook

Future work will focus on addressing these limitations.

Weakly constrained mappings could be improved by further extending the data set domains with intermediates. For instance, since H_2 is more closely tied to star formation, it should provide better constraints for the stellar mass prediction via GAS → H_2 → STARS.

Alternatively, various inductive biases could also provide stronger constraints during training:

(i) *Regularization of the objective function*: directly physics-informed networks through, e.g. constraining mass within aperture, or penalties on radial-profile mismatch.
(ii) *Structure-aware discriminators*: adversarial heads operating on radial profiles, power spectra, or multiscale losses.
(iii) *Equivariant architecture*: SO(2)-aware U-Nets can reduce sample complexity, and inherently enforce symmetries, thereby explicitly handling nuisance parameters.
(iv) *Multidomain training*: predicting several targets at once in multiple channels would increase cross-domain robustness but increase processing time.
(v) *Cross-suite transfer learning*: cross-suite transfer learning and domain adaptation avoid re-training models on other simulation suites from scratch, requiring only a small amount of fine-tuning on the target simulation.
(vi) *Redshift conditioning*: redshift introduces temporal information to models and helps capture the true galaxy evolution through cosmic time.

Our findings demonstrate that learnt generative surrogates can transform galaxy formation research by bridging simulations and observations, reducing reliance on costly, repeated hydrodynamical runs. By coupling our blueprint for domain-aware assessment of physical realism with computational scalability, this work

marks a significant step towards efficient, next-generation modelling pipelines, automated survey interpretation, and managing the ensuing data deluge in the SKA era.

ACKNOWLEDGEMENTS

PD, FS, EG acknowledge support from the SKACH consortium through funding by SERI. Open access funding provided by ZHAW Zurich University of Applied Sciences. We would also like to thank the ZHAW Centre for Artificial Intelligence's science cluster admin M. Stadelmann for his support and management of the high resources this project demanded.

DATA AVAILABILITY

The original source of the data set is publicly released by the IllustrisTNG project. The extracted data set and model weights can be shared upon reasonable request. Our PYTORCH-based code used for the training of the presented deep learning models is publicly released on GitHub under a GPLv3 license (https://github.com/CAIIVS/chuchichaestli) and (https://github.com/phdenzel/skais-mapper), including scripts and *hydra* configurations (O. Yadan 2019) to re-create the results in this work.

REFERENCES

Akiba T., Sano S., Yanase T., Ohta T., Koyama M., 2019, in Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA, p. 2623
Amirian M., Barco D., Herzig I., Schilling F.-P., 2024, IEEE Access, 12, 10281
Andersson J., Ahlström H., Kullberg J., 2019, Magn. Reson. Med., 82, 1177
Arjovsky M., Bottou L., 2017, 5th International Conference on Learning Representations. ICLR, Toulon, France
Arjovsky M., Chintala S., Bottou L., 2017, Proceedings of the 34th International Conference on Machine Learning. JMLR, Sydney NSW, Australia, p. 214
Ba J. L., Kiros J. R., Hinton G. E., 2016, preprint (arXiv:1607.06450)
Bally J., 2016, ARA&A, 54, 491
Bassini L., Feldmann R., Gensior J., Faucher-Giguère C.-A., Cenci E., Moreno J., Bernardini M., Liang L., 2024, MNRAS, 532, L14
Beck R., 2015, A&AR, 24, 4
Bengesi S., El-Sayed H., Sarker M. K., Houkpati Y., Irungu J., Oladunni T., 2024, IEEE Access, 12, 69812
Berlind A. A. et al., 2003, ApJ, 593, 1
Bernardini M., Feldmann R., Anglés-Alcázar D., Boylan-Kolchin M., Bullock J., Mayer L., Stadel J., 2021, MNRAS, 509, 1323
Bernardini M. et al., 2025, MNRAS, 538, 1201
Bianco M. et al., 2025, MNRAS, 541, 234
Biernacki P., Teyssier R., 2018, MNRAS, 475, 5688
Binney J., Tremaine S., 2011, Galactic Dynamics. Princeton Univ. Press, Princeton, NJ USA, p.1
Binney J., Vasiliev E., 2023, MNRAS, 520, 1832
Blau Y., Michaeli T., 2018, in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Salt Lake City, UT, USA, p. 6228
Bond-Taylor S., Leach A., Long Y., Willcocks C. G., 2022, IEEE Trans. Pattern Analy. Mach. Intell., 44, 7327
Braun R., Bourke T. L., Green J. A., Keane E., Wagg J., 2015, in Proc. Advancing Astrophysics with the Square Kilometre Array–PoS(AASKA14). SISSA, Trieste, Italy, p. 174
Bullock J., Cuesta-Lazaro C., Quera-Bofarull A., 2019, in Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging. Proc. SPIE, Bellingham, WA, USA, p. 69
Chadayammuri U., Ntampaka M., ZuHone J., Bogdán Á., Kraft R. P., 2023, MNRAS, 526, 2812
Cheng S., Yu H.-R., Inman D., Liao Q., Wu Q., Lin J., 2020, in 2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID). IEEE, Melbourne, VIC, Australia, p. 685
Cibinel A. et al., 2019, MNRAS, 485, 5631
Colman T. et al., 2024, A&A, 686, A155
Conselice C. J., 2003, ApJS, 147, 1
Conselice C. J., 2014, ARA&A, 52, 291
Conselice C. J., Bershady M. A., Jangren A., 2000, ApJ, 529, 886
Crain R. A., van de Voort F., 2023, ARA&A, 61, 473
Crain R. A. et al., 2015, MNRAS, 450, 1937
D'Onofrio M. et al., 2016, The Physics of Galaxy Formation and Evolution. Springer International Publishing, Cham, Switzerland, p. 585
Davé R., Anglés-Alcázar D., Narayanan D., Li Q., Rafieferantsoa M. H., Appleby S., 2019, MNRAS, 486, 2827
de Blok W. J. G. et al., 2024, A&A, 688, A109
Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L., 2009, in 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Miami, FL, USA, p. 248
Dhariwal P., Nichol A., 2021, Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates, Inc., Red Hook, NY, USA, p. 8780
Dubois Y. et al., 2014, MNRAS, 444, 1453
Engelbrecht B. N. et al., 2024, MNRAS, 536, 1035
Esser P., Rombach R., Ommer B., 2020, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Nashville, TN, USA, p. 12868
Feller W., 1949, in Neyman J. ed., First Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, CA, USA, p. 403
Fielding D., Quataert E., Martizzi D., Faucher-Giguère C.-A., 2017, MNRAS, 470, L39
Finke T., Krämer M., Morandini A., Mück A., Oleksiyuk I., 2021, J. High Energy Phys., 2021, 161
Frenk C., White S., 2012, Annalen der Physik, 524, 507
Gavagnin E., Bleuler A., Rosdahl J., Teyssier R., 2017, MNRAS, 472, 4155
Golling T., Heinrich L., Kagan M., Klein S., Leigh M., Osadchy M., Raine J. A., 2024, Mach. Learn. Sci. Technol., 5, 035074
Goodfellow I., Bengio Y., Courville A., 2016, Deep Learning. MIT Press, Cambridge, MA, USA
Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y., 2014, Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. MIT Press, Cambridge, MA, USA, p. 2672
Gulrajani I., Ahmed F., Arjovsky M., Dumoulin V., Courville A., 2017, Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, p. 5769
Harper S. E., Dickinson C., 2018, MNRAS, 479, 2024
He K., Zhang X., Ren S., Sun J., 2015, 2016 IEEE Conference on ComputerVision and Pattern Recognition (CVPR). IEEE, Las Vegas, NV, USA, p. 770
Ho J., Salimans T., 2022, preprint (arXiv:2207.12598)
Ho J., Kalchbrenner N., Weissenborn D., Salimans T., 2019, preprint (arXiv:1912.12180)
Ho J., Jain A., Abbeel P., 2020, Proceedings of the 34th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, p. 6840
Hopkins P. F., Hernquist L., Cox T. J., Matteo T. D., Robertson B., Springel V., 2006, ApJS, 163, 1
Hornik K., Stinchcombe M., White H., 1989, Neural Networks, 2, 359
Hwang H. S., Shin J., Song H., 2019, MNRAS, 489, 339
Ibrahim D., Kobayashi C., 2023, MNRAS, 527, 3276
Ingraham J. B. et al., 2023, Nature, 623, 1070
Ioffe S., Szegedy C., 2015, Proceedings of the 32nd International Conference on Machine Learning. PMLR, Lille, France, p. 448
Ishiyama T. et al., 2021, MNRAS, 506, 4210

Isola P., Zhu J.-Y., Zhou T., Efros A. A., 2016, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Honolulu, HI, USA, p. 5967
Jayasumana S., Ramalingam S., Veit A., Glasner D., Chakrabarti A., Kumar S., 2024, 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Seattle, WA, USA, p. 9307
Kennicutt R. C., Evans N. J., 2012, ARA&A, 50, 531
Krumholz M. R. et al., 2014, Star Cluster Formation and Feedback. Univ. Arizona Press, Tucson, AZ, USA, p. 243
Kynkäänniemi T., Karras T., Aittala M., Aila T., Lehtinen J., 2023, The Eleventh International Conference on Learning Representations. ICLR, Kigali, Rwanda
Leigh M., Sengupta D., Quétant G., Raine J. A., Zoch K., Golling T., 2024, SciPost Physics, 16, 018
Li C., Wand M., 2016, Computer Vision - ECCV 2016. Springer International Publishing, Cham, Switzerland, p. 702
Li Y., Ni Y., Croft R. A. C., Matteo T. D., Bird S., Feng Y., 2021, Proc Natl. Acad. Sci., 118, e2022038118
Liaw R., Liang E., Nishihara R., Moritz P., Gonzalez J. E., Stoica I., 2018, preprint (arXiv:1807.05118)
Loshchilov I., Hutter F., 2016, 5th International Conference on Learning Representations. ICLR, Toulon, France
Luisi M. et al., 2021, Sci. Adv., 7, eabe9511
Maccagni F. M., Blok W. D., 2024, in Proc. 4th URSI Atlantic RadioScience Conference–AT-RASC 2024. IEEE, Meloneras, Spain, p. 1
Maccagni F. M., Serra P., 2025, preprint (arXiv:2507.18109)
Maddox N. et al., 2021, A&A, 646, A35
Mao X., Li Q., Xie H., Lau R. Y. K., Wang Z., Smolley S. P., 2016, 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, Venice, Italy, p. 2813
Marinacci F. et al., 2018, MNRAS, 480, 5113
McAlpine S. et al., 2016, Astron. Comput., 15, 72
McKee C. F., Ostriker E. C., 2007, ARA&A, 45, 565
MeerKLASS Collaboration, 2025, MNRAS, 537, 3632
Messias H., Guerrero A., Nagar N., Regueiro J., Impellizzeri V., Orellana G., Vioque M., 2024, MNRAS, 533, 3937
Moore B., Quinn T., Governato F., Stadel J., Lake G., 1999, MNRAS, 310, 1147
Muratov A. L. et al., 2017, MNRAS, 468, 4170
Naiman J. P. et al., 2018, MNRAS, 477, 1206
Nelson D. et al., 2017, MNRAS, 475, 624
Nelson D. et al., 2019, MNRAS, 490, 3234
O'Beirne T. et al., 2025, PASA, 42, e087
Obuljen A., Simonović M., Schneider A., Feldmann R., 2023, Phys. Rev. D, 108, 083528
Odena A., Dumoulin V., Olah C., 2016, Distill
Pang Y., Lin J., Qin T., Chen Z., 2022, IEEE Trans. Multimedia, 24, 3859
Parmar N., Vaswani A., Uszkoreit J., Kaiser L., Shazeer N., Ku A., Tran D., 2018, Proceedings of the 35th International Conference on Machine Learning. ICML, Stockholm, Sweden, p. 4055
Perraudin N., Srivastava A., Lucchi A., Kacprzak T., Hofmann T., Réfrégier A., 2019, Comput. Astrophys. Cosmol., 6, 5
Pillepich A. et al., 2017, MNRAS, 475, 648
Poggianti B. M. et al., 2019, ApJ, 887, 155
Potter D., Stadel J., Teyssier R., 2017, Comput. Astrophys. Cosmol., 4, 2
Radford A., Metz L., Chintala S., 2015, 4th International Conference on Learning Representations. ICLR, San Juan, Puerto Rico
Ramachandran P., Zoph B., Le Q. V., 2017, 6th International Conference on Learning Representations. ICLR, Vancouver, BC, Canada
Rieder M., Teyssier R., 2017, MNRAS, 472, 4368
Rives A. et al., 2021, Proc. Natl Acad. Sci., 118, e2016239118
Rønne N., Aspuru-Guzik A., Hammer B., 2024, Phys. Rev. B, 110, 235427
Ronneberger O., Fischer P., Brox T., 2015, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, Switzerland, p. 234
Salimans T., Ho J., 2022, 10th International Conference on Learning Representations. ICLR
Schade D., Lilly S. J., Crampton D., Hammer F., Fèvre O. L., Tresse L., 1995, ApJS, 451, L1
Schanz A., List F., Hahn O., 2024, Open J. Astrophys., 7
Schaye J. et al., 2014, MNRAS, 446, 521
Schinnerer E., Leroy A., 2024, ARA&A, 62, 369
Schneider A., Teyssier R., Stadel J., Chisari N. E., Brun A. M. L., Amara A., Refregier A., 2019, J. Cosmol. Astropart. Phys., 2019, 020
Schneuing A. et al., 2024, Nat. Comput. Sci., 4, 899
Sharma R. et al., 2026, Astronomy and Computing, 54, 101004
Smith L. N., Topin N., 2019, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications. Proc. SPIE, Baltimore, MD, USA, p. 1100612
Somerville R. S., Hopkins P. F., Cox T. J., Robertson B. E., Hernquist L., 2008, MNRAS, 391, 481
Springel V. et al., 2017, MNRAS, 475, 676
Staveley-Smith L., Oosterloo T., 2015, in Proc. Advancing Astrophysics with the Square Kilometre Array–PoS(AASKA14). SISSA, Trieste, Italy, p. 167
Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z., 2015, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Las Vegas, NV, USA, p. 2818
Thiele L., Villaescusa-Navarro F., Spergel D. N., Nelson D., Pillepich A., 2020, ApJ, 902, 129
Tinsley B. M., 2022, preprint (arXiv:2203.02041)
Valentini M. et al., 2019, MNRAS, 491, 2779
Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I., 2017, Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, p. 6000
Villaescusa-Navarro F., 2018, Astrophysics Source Code Library, record ascl:1811.008
Villaescusa-Navarro F. et al., 2018, ApJ, 866, 135
Wang Z., Bovik A., Sheikh H., Simoncelli E., 2004, IEEE Trans. Image Processing, 13, 600
Ward S. R., Costa T., Harrison C. M., Mainieri V., 2024, MNRAS, 533, 1733
Weinberger R., Springel V., Pakmor R., 2020, ApJS, 248, 32
Weissenborn D., Täckström O., Uszkoreit J., 2019, 8th International Conference on Learning Representations. ICLR, Addis Ababa, Ethiopia
Whang J., 2023, PhD thesis, Computer Science
White S. D. M., Frenk C. S., 1991, ApJ, 379, 52
Woodland M. et al., 2024, Feature Extraction for Generative Medical Imaging Evaluation: New Evidence Against an Evolving Trend. Springer Nature, Switzerland, p. 87
Wu Y., He K., 2018, Computer Vision – ECCV 2018. Springer International Publishing, Cham, Switzerland, p. 3,
Yadan O., 2019, Hydra–A framework for elegantly configuring complex applications, Github
Yang B., Wang L., Wong D., Chao L. S., Tu Z., 2019, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. Association for Computational Linguistics, Minneapolis, MN, USA, p. 4040
Yao W., Zeng Z., Lian C., Tang H., 2018, Neurocomputing, 312, 364
Zhang H., Goodfellow I., Metaxas D., Odena A., 2019, Proceedings of the 36th International Conference on Machine Learning. PMLR, Long Beach, CA, USA, p. 7354,
Zhang R., Isola P., Efros A. A., Shechtman E., Wang O., 2018, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Salt Lake City, UT, USA, p. 586
Zubovas K., Tartėnas M., Bourne M. A., 2024, A&A, 691, A151

APPENDIX A: U-NET EXPERIMENTS

Previous work has identified model capacity, defined by depth (number of U-Net levels) and width (number of hidden feature channels) as a key factors of generative fidelity (O. Ronneberger et al. 2015; P. Isola et al. 2016; J. Ho et al. 2020). Additionally,

the choice of normalization layers (P. Dhariwal & A. Nichol 2021), and the inclusion of residual and attention mechanisms (H. Zhang et al. 2018; P. Dhariwal & A. Nichol 2021), have consistently shown to enhance both training stability and output quality. In contrast, other design choices, such as the specific up-sampling scheme or minor variations in skip connections, tend to yield marginal improvements and diminishing returns. To systematically assess these factors, we first examine the impact of model capacity on generative performance, limiting experiments to high-impact components to keep computational costs tolerable.

Table A1 summarizes the U-Net configurations tested in this initial set of targeted experiments. Each experiment varies only the architectural parameters under investigation, while all other training settings are held constant. For bench-marking, we selected the GAS → DM translation task, which exhibited intermediate difficulty across all domain pairs in preliminary tests.

All models were trained with adversarial loss for 30 epochs using the standard discriminator configuration (as described in Section 2.4), with a warm restart technique for stochastic gradient descent including a cosine annealing learning rate schedule (I. Loshchilov & F. Hutter 2016); the learning rate is attenuated according to a cosine and periodically reset. Initial learning rates were set to 10^{-4} for the generator and 5×10^{-5} for the discriminator.

The results of the U-Net size ablation study are summarized in Table A2. Among the tested configurations, MEDIUMU (64 channels, 4 levels) consistently achieved the best overall performance across most evaluation metrics. Notably, the SSIM metric seemed to saturate in all tests quickly, indicating most U-Net configurations yield structurally similar outputs to the ground truth, but may lack sensitivity with smooth, high-resolution distributions like those from simulations and may not capture subtle differences fine-grained textures and localized features.

Deeper and larger U-Net variants started exhibiting artefacts and overfitting that degraded perceptual quality of generated samples evident by higher PSNR values, but at the cost of increased FID. Moreover, larger models exhibited signs of mode collapse, with unreliable metric results. Conversely, the shallower and smaller performed comparably or worse in PSNR and SSIM but suffered from a substantially worse FID, suggesting insufficient capacity to model the full complexity of the domain mapping.

Based on these findings, we identified the MEDIUMU configuration as the optimal U-Net configuration for this task. It offers a favourable trade-off between performance and computational cost, avoids overfitting, and maintains stable training dynamics. This configuration is therefore used as the default architecture in all subsequent experiments unless stated otherwise.

Note that for diffusion models, spot tests resulted in comparable distortion metrics, however, to offset the substantial increase in computational cost, the U-Net width was reduced to 32 channels, prioritizing the inclusion of attention layers.

Table A1. Various U-Net configurations with varying sizes in depth and width that were tested. 'Width' refers to the base number of feature channel in the first U-Net layer, whereas 'Depth' is the number of levels between down- and up-sampling layers. 'Levels' indicates the number of stages (spatial down- and up-sampling layers) in the U-Net encoder and decoder. '# Params' is the total number of trainable parameters in the U-Net.

Designation	Width	Depth	Levels	# Params
TINYU	16	4	4	4010 369
SMALLU	32	4	4	16 024 833
MEDIUMU	64	4	4	64 066 049
MEDIUMU_L3	64	3	3	15 814 657
MEDIUMU_L5	64	5	5	257 037 825
LARGEU	128	4	4	256 197 633

Table A2. Evaluation results of various U-Net size configurations from Table A1 after training for 30 epochs. All experiments are based on the translation GAS→DM, adversarially trained with the same discriminator configuration. Model results in bold are optimal values, and those marked with † exhibit mode collapse and are not reliable.

Designation	PSNR ↑	SSIM ↑	MSE ↓	FID ↓
TINYU	35.31	0.9954	5.5×10^{-4}	12.02
SMALLU	39.12	0.9966	6.6×10^{-4}	12.75
MEDIUMU	39.76	**0.9977**	4.2×10^{-4}	**9.71**
MEDIUMU_L3	39.57	0.9967	6.8×10^{-4}	30.44
MEDIUMU_L5	**48.01**	0.9972	$\mathbf{2.9 \times 10^{-4}}$	18.31
LARGEU	†65.28	†0.9978	$^{\dagger}3.1 \times 10^{-3}$	†270.0

Table B1. U-Net configurations with various attention blocks positioning (encoder levels numbered top to bottom, continuing in the decoder bottom up; the number after 'U' indicates the block positioning, the prefixed numbers the total number of attention blocks if different from 1 or all.). The second column 'Loc' indicates where the attention layers are placed; if more than one attention layers were added, then it indicates the position of the first layer. The third column '# Enc' indicates how many attention layers are included in the encoder, the forth '# Dec', how many in the decoder. '# Params' is the total number of trainable parameters in the U-Net. The right-most column is the average time for a forward pass with a single batch.

Designation	Loc	# Enc	# Dec	# Params	Forward pass
ATTNU1	1	1	0	64 082 817	15.3765 s
ATTNU3	3	1	0	64 329 729	2.0411 s
ATTNU4	4	1	0	65 117 697	1.9542 s
ATTNUMID	5	1	0	68 266 497	1.9796 s
ATTNU5	5	0	1	68 266 497	1.9904 s
ATTNU6	6	0	1	65 117 697	2.2751 s
ATTNU8	8	0	1	64 132 353	27.8642 s
ATTN3XU3	3	2	1	69 581 825	5.7335 s
ATTNALL	1	4	4	71 046 529	48.2022 s

APPENDIX B: ATTENTION LAYER PLACEMENT EXPERIMENTS

Having established an optimal baseline configuration, we investigated the impact of attention layer placement within the U-Net architecture in a subsequent series of experiments (Table B1). While prior work suggests that attention mechanisms can enhance global context modelling (A. Vaswani et al. 2017), their effectiveness and efficiency may depend on the resolution level at which they are applied. To this end, we varied the position of self-attention blocks across encoder and decoder stages, including configurations with attention in early layers (high-resolution features), late layers (low-resolution, high-semantic features), and hybrid placements spanning multiple levels. While attention layers are expected to yield superior results no matter the placement, the main purpose of these experiments was to assess the relative performance differences of less computationally demanding

Table B2. Evaluation results of U-Net attention layer placement experiments from Table B1 after training for 30 epochs. All experiments are based on the translation GAS→DM, adversarially trained with the same discriminator configuration. Model results in bold are optimal values, and those marked with † exhibit mode collapse and are not reliable.

Designation	PSNR ↑	SSIM ↑	MSE ↓	FID ↓
ATTNU1	37.23	0.9968	3.4×10^{-4}	6.93
ATTNU3	37.10	0.9972	3.5×10^{-4}	6.64
ATTNU4	36.70	0.9968	3.6×10^{-4}	6.59
ATTNUMID	35.27	0.9983	3.7×10^{-4}	7.47
ATTNU5	35.76	**0.9983**	3.6×10^{-4}	6.60
ATTNU6	36.26	0.9974	3.6×10^{-4}	4.57
ATTNU8	†22.55	†0.9	$^{\dagger}1.2 \times 10^{-2}$	†253.2
ATTN3XU3	37.71	0.9970	3.4×10^{-4}	**4.04**
ATTNALL	**42.55**	0.9934	$\mathbf{3.2 \times 10^{-4}}$	8.70

placement in late layers to those in high-resolution features. All other architectural and training settings were kept identical to those in the optimal configuration from the U-Net size experiments. Only the learning rate update schedule was changed to a one-cycle policy (L. N. Smith & N. Topin 2017) due to instabilities in the generator-discriminator dynamics and to keep learning rate comparably high. The evaluation focused on image-based metrics to determine whether attention placement influences fine-grained structural fidelity. These experiments aim to identify the most effective strategy for leveraging attention without incurring unnecessary computational overhead.

The results of these experiments (Table B2) reveal that the benefits of self-attention layers in U-Net blocks are indeed dependent on both the number and placement within the network. While adding attention universally across all levels (ATTNALL) improved pixel-wise metrics such as PSNR, it comparatively degraded distributional consistency as measured by FID, indicating overparametrization. Conversely, a moderate number of self-attention layers in the deepest levels seems to generally improve distributional, perceptual fidelity compared to the previous experiments, in trade for distortion (cf. Y. Blau & T. Michaeli 2018). In particular, adding attention layers near the bottleneck (ATTN3XU3; attention layers in consecutive blocks starting in the third level of the U-Net) yielded the best FID scores while maintaining competitive PSNR and the other distortion metrics.

These findings suggest that for this data set global interactions are most effectively modelled when attention is applied to low-resolution, high-semantic feature maps, whereas attention in high-resolution layers may lead to equally or better performance but introduces unnecessary computational cost and instability. Moreover, all experiments have been repeated using the convolutional attention variant, with nearly identical results in each run, and minimally shorter forward pass timings. Based on these results, we adopt the configuration with three deep convolutional attention layers in the lower levels (ATTN3XU3) for all subsequent experiments, as it offers the best balance of generative fidelity, stability, and efficiency.

APPENDIX C: MODEL SPECIFICS

With the attention configuration fixed, we proceed to model-specific refinements. In this stage, we tuned discriminator architectures for GAN-based models and evaluate noise scheduling strategies for diffusion models.

Table C1. PatchGAN configurations of various sizes. 'Width' refers to the base number of feature channel in the first hidden layer, whereas 'Depth' is the number of hidden layers in the network. '# Params' is the total number of trainable parameters in the PatchGAN network.

Designation	Width	Depth	# Params
PGAN_SMALL	64	3	2765 505
PGAN_MEDIUM	64	4	11 165 377
PGAN_LARGE	64	5	44 742 337
PGAN_WIDE	128	3	178 875 777
PGAN_NARROW	32	3	11 197 281

Table C2. Evaluation results of PatchGAN size experiments from Table C1 after training for 30 epochs. All experiments are based on the translation GAS→DM, adversarially trained with the same generator configuration. Model results in bold are optimal values.

Designation	PSNR ↑	SSIM ↑	MSE ↓	FID ↓
PGAN_SMALL	32.46	0.9894	8.3×10^{-4}	18.04
PGAN_MEDIUM	**37.16**	0.9901	$\mathbf{4.3 \times 10^{-4}}$	**4.62**
PGAN_LARGE	36.56	**0.9960**	5.7×10^{-4}	8.47
PGAN_WIDE	34.77	0.9952	7.1×10^{-4}	8.65
PGAN_NARROW	35.16	0.9909	6.0×10^{-4}	9.75

For GAN-based models, the PatchGAN discriminator's width, depth, and number of hidden layer configurations were spot tested (see Table C1 for configurations). Based on the results in Table C2, the PGAN_MEDIUM configuration was used for the final model training. While there were no clear differences between the various configurations, smaller networks had the tendency to impose checker-board artefacts in the generator outputs and larger, wider ones lead to instabilities during training due to mismatched sizes between discriminator and generator.

For diffusion models, the U-Net includes a sinusoidal time-embedding with 32 channels in each block (as described in Section 2.4). Moreover, linear, quadratic, and cosine noise schedules have been tested, and cosine clearly improved image quality (with a consistent 2–3 dB improvement in PSNR, 0.05–0.1 difference in SSIM), convergence, and provided smoother denoizing transitions.

APPENDIX D: CORRELATION ANALYSIS BETWEEN FID AND DISTORTION/ASTROPHYSICAL METRICS

To quantify how well the FID metric tracks distortion and astrophysical fidelity, we computed pair-wise correlations between FID and (i) the astrophysical metrics (AE, SCE, COMD, CRCE, and PSE), and (ii) the image-based CV metrics (PSNR, SSIM, and MSE). Correlations are reported as Pearson's r (linear association) and Spearman's ρ (rank-based, monotonic association). As noted in Section 3.3, PSNR is biased across model types due to different pre-processing ranges, which introduces an ∼6.02 dB offset for identical MSE; we therefore de-biased PSNR for the correlation analysis. The reported test-set results include all eight translation tasks and two model types (GAN, DDPM).

By Pearson's r, FID correlates strongly with several astrophysical indicators of morphology and structure:

(i) CRCE ($r = 0.901$, $p = 2 \cdot 10^{-6}$)
(ii) SCE ($r = 0.898$, $p = 2 \cdot 10^{-6}$), and

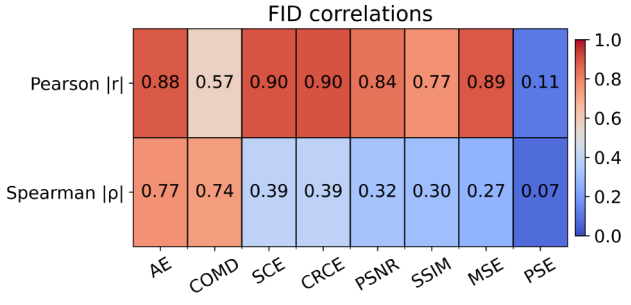


Figure D1. Pearson (top) and Spearman (bottom) correlations between FID and both CV distortion (MSE, PSNR, SSIM) and astrophysical metrics (AE, SCE, COMD, CRCE, PSE) across all tasks and models. FID tracks physics-aware morphology (AE, SCE, CRCE) and MSE most strongly; weaker association with PSE and COMD.

(iii) AE ($r = 0.883$, $p = 6 \cdot 10^{-6}$), where p is the corresponding probability value associated with the null hypothesis.

Among CV metrics, metrics that show a relatively strong absolute correlation with FID are

(i) MSE ($r = 0.888$, $p = 4 \cdot 10^{-6}$), while
(ii) PSNR ($r = -0.838$, $p = 5 \cdot 10^{-5}$) and
(iii) SSIM ($r = -0.766$, $p = 5.5 \cdot 10^{-4}$) correlate negatively as expected.

As for the remaining metrics,

(i) COMD ($r = 0.569$, $p = 0.021$) shows a moderate relationship, whereas
(ii) PSE ($r = -0.108$, $p = 0.692$) is weakly correlated with FID.

The upper row in Fig. D1 illustrates these results.

Pearson's r captures linear magnitude relationships and is therefore sensitive to the clear performance separation we observe between 'easier' translations (e.g. GAS → DM, GAS ↔ HI) and 'harder' ones (e.g. GAS → STARS). This separation yields high absolute r where FID improves parallel to astrophysical morphology scores and CV distortion metrics. Spearman's ρ, computed on ranks, is more conservative in our setting: rank consistency is partially reduced by metric saturation (notably SSIM on smooth, high-resolution simulation maps), small-sample rank fluctuations within the well-performing cluster, and mixed GAN/DDPM clusters where the linear gaps are large but the within-group ordering is noisier. Consequently, Spearman ρ is highest for

(i) AE ($\rho = 0.768$, $p = 5.2 \cdot 10^{-4}$) and
(ii) COMD ($\rho = 0.741$, $p = 0.001$),

but more modest (between 0.27 and 0.39 in magnitude) for metrics like SCE, CRCE, PSNR, SSIM, and MSE. The PSNR debias primarily affects Pearson and has a much smaller effect on Spearman, as expected for a near-uniform offset across one model family. Fig. D1 shows the Spearman's ρ in the lower row, ordered according to its absolute magnitude.

We examined eight correlations: a simple Bonferroni threshold is $\alpha/8 \approx 0.003$. Under this criterion, Pearson-level evidence remains strong for all metrics, except COMD and PSE fall short. For Spearman correlations, AE (and to a lesser extent COMD) remain significant, consistent with the above interpretation. In conclusion, the convergence between FID and multiple astrophysical morphology metrics suggests that, in this data set, FID is a suitable proxy for domain-aware realism, especially when augmented with physics-informed metrics in model selection.

APPENDIX E: DOMAIN TRANSLATION SAMPLES

This appendix contains Fig. E1 with the remaining domain translations not shown in Fig. 3.

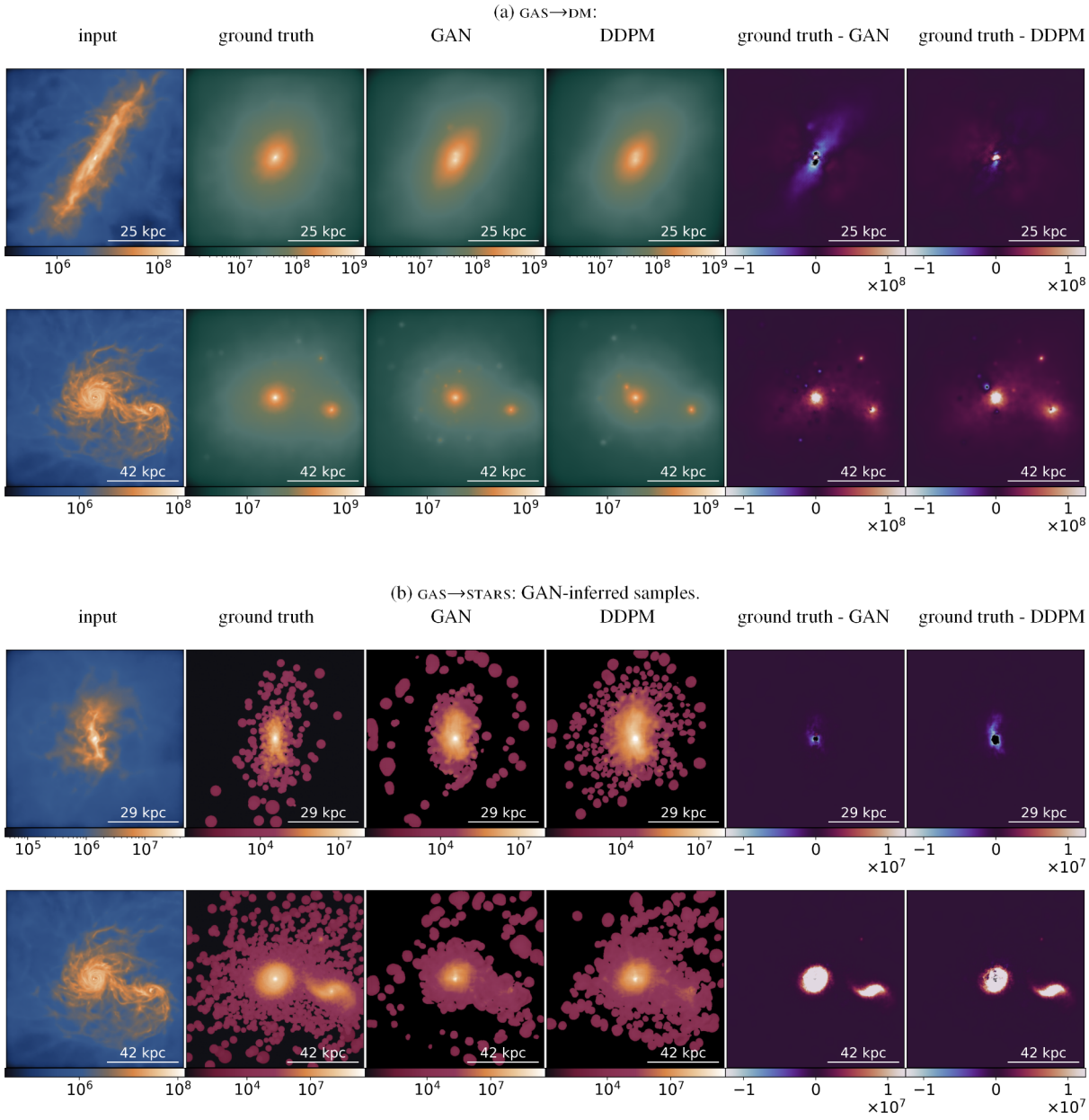


Figure E1. Remaining samples from models and tasks not shown in Fig. 3. Each panel shows a model input map on the left, the corresponding ground truth, and prediction from GANs and DDPMs on the right. The right-most two maps show residuals between ground truth and GAN, and DDPM, respectively. Qualitative comparison confirms the alignment of astrophysical plausibility and human perception with astrophysical metrics and FID (see Tables 2 and 3); see discussion in Fig. 3.

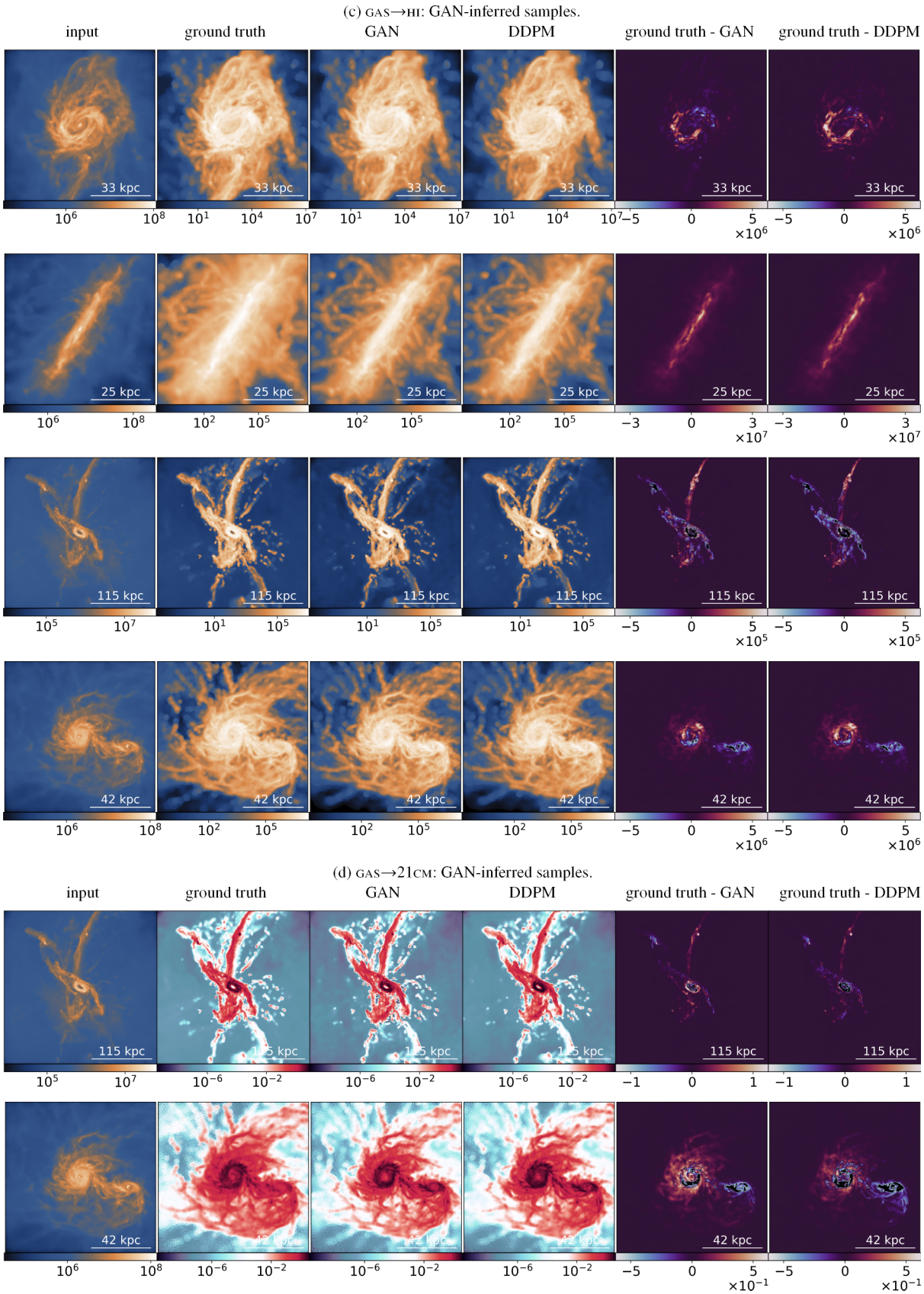


Figure E1. *Continued.*

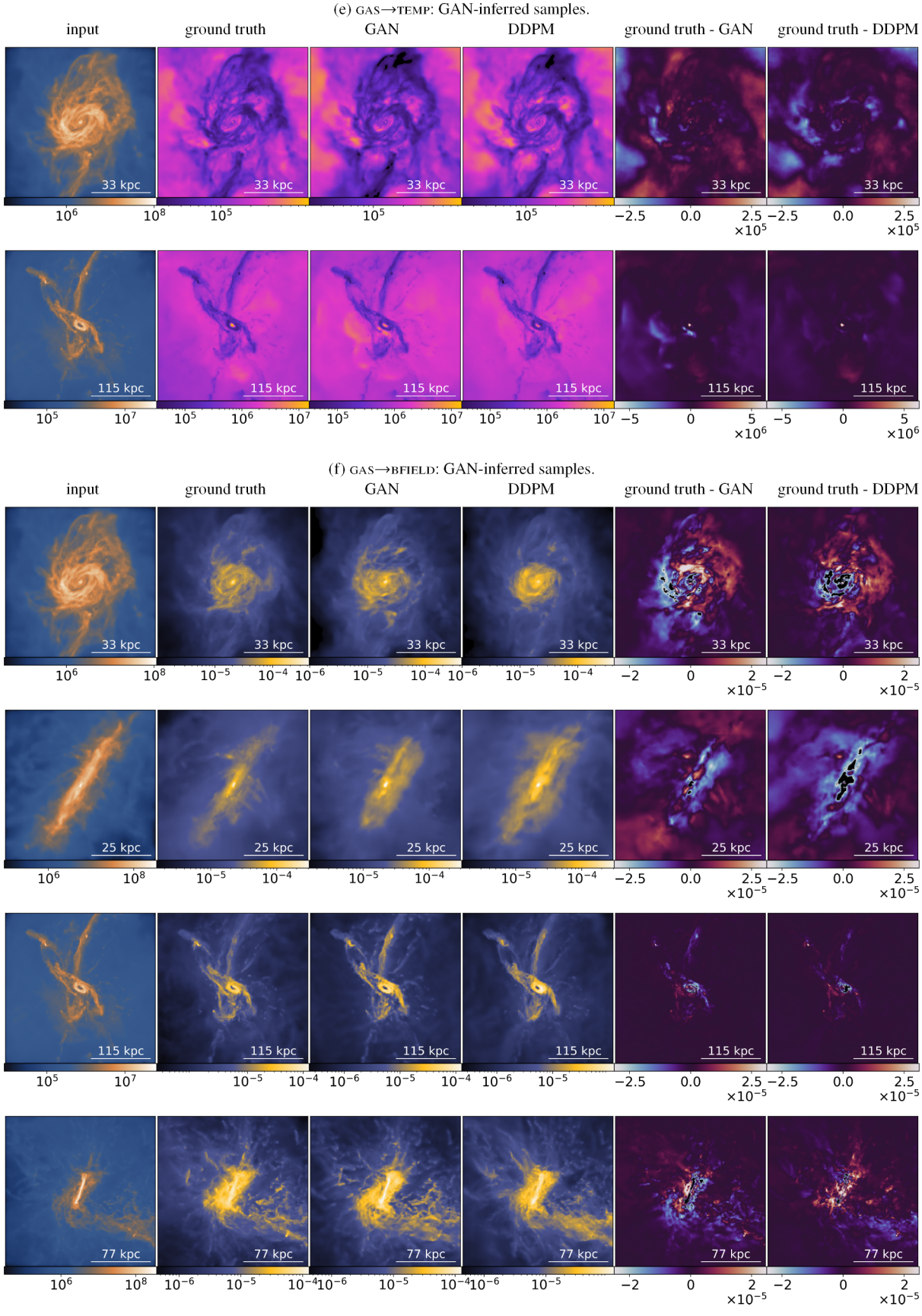


Figure E1. *Continued.*

(g) DM→GAS: GAN-inferred samples.

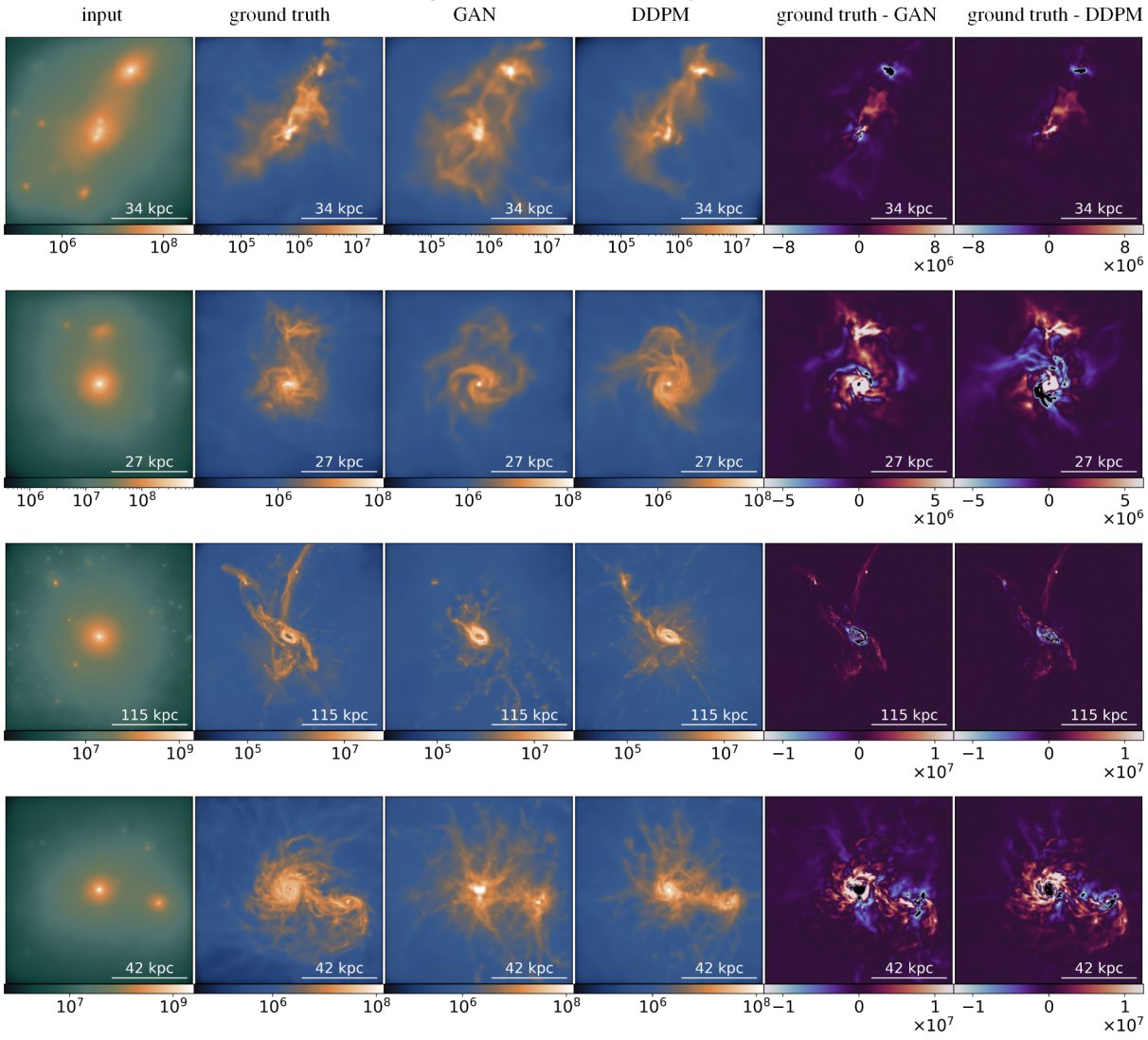


Figure E1. *Continued.*

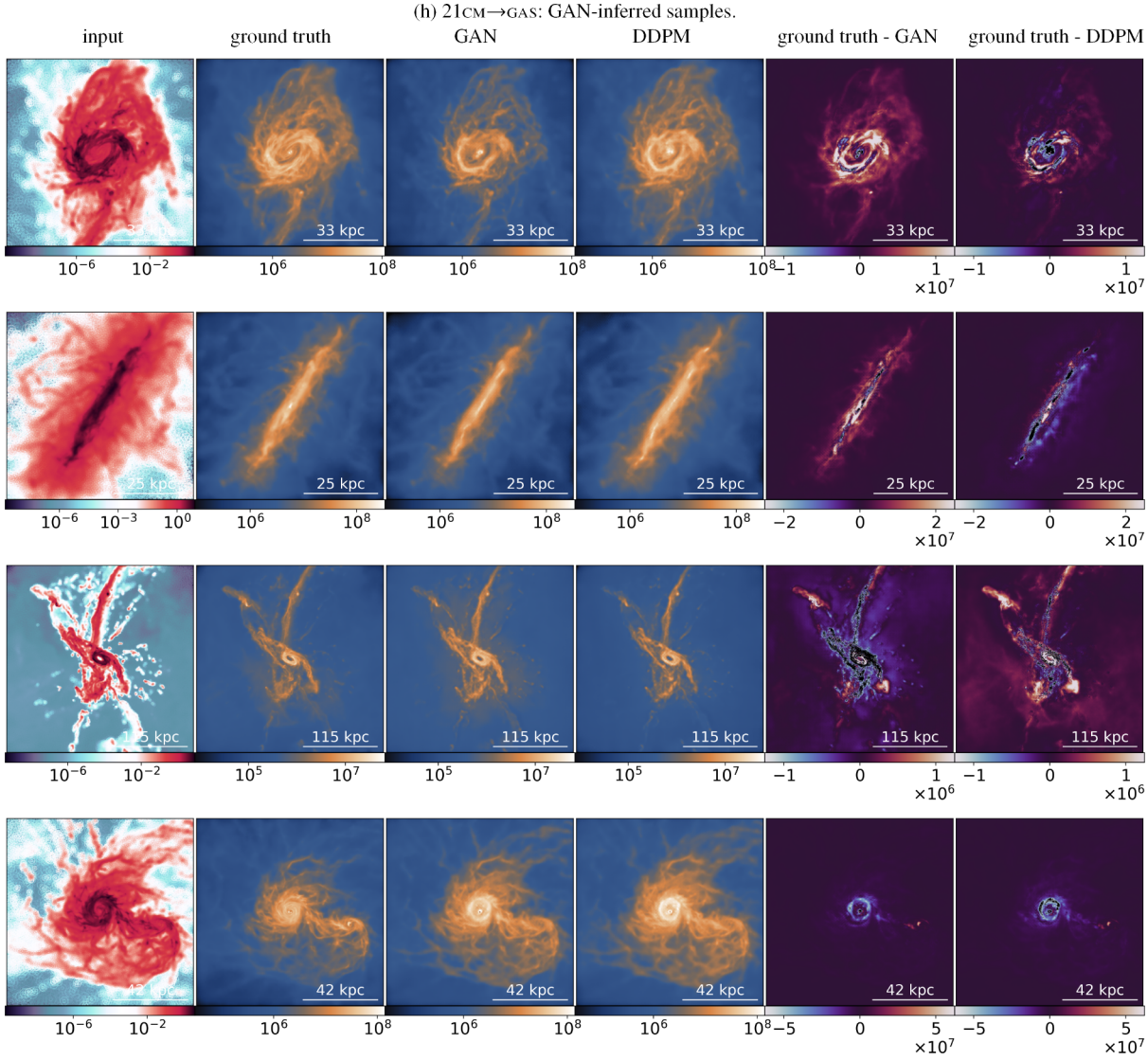


Figure E1. *Continued.*

This paper has been typeset from a TEX/LATEX file prepared by the author.