Dr. Campagne Jean-Eric
Université Paris-Saclay, CNRS/IN2P3, IJCLab
Bât. 100, 15 rue Georges Clémenceau
Orsay, France, 91405
jean-eric.campagne@ijclab.in2p3.fr
+33 01 69 15 41 00

Editor-in-Chief
Monthly Notices of the Royal Astronomical Society
17 mars 2025

Dear Editor-in-Chief,

I sincerely thank the Assistant Editor and the Referee for their careful reading of my manuscript and for their valuable comments and suggestions to improve the text.

In the document below, you will find my responses labeled as *"***Answer***"*, detailing the elements and arguments that motivated the changes in the revised version of the manuscript. To facilitate their identification in the latex/pdf version, these modifications are highlighted in red.

I hope this revised submission meets the referee's requirements.

Sincerely yours,
Jean-Eric Campagne
Senior Research Director
Laboratoire de Physique des 2 Infinis Irène Joliot-Curie (IJCLab)

Assistant Editor's Comments*:*
Please remove the references from your abstract.
Answer: I have removed the references.

Reviewer comments/suggestions*:*

Section2:
Reviewer*: The section provides a thorough and theoretically rigorous overview of generative models, including VAEs, GANs, normalizing flows, and diffusion models, showcasing a strong grasp of their principles and comparative strengths. This depth of discussion, supported by extensive references, highlights the authors' expertise and provides valuable context for readers interested in the broader methodological landscape. That said, it may be worth considering whether the level of theoretical detail is fully necessary for the paper's primary audience, especially given its focus on astronomy. Streamlining the discussion to emphasize the practical relevance of these models to astronomy—for instance, how they address challenges like data sparsity, noise, or resolution—could make the section more impactful. Additionally, including a high-level summary of the models and their relevance for readers less familiar with deep learning, as well as expanding on practical implementation details such as data preprocessing, training setup, and computational requirements, would enhance accessibility and ensure the discussion remains tightly aligned with the paper's goals.*

Answer: We thank the referee for recognizing our effort in presenting the theoretical context, and we pay close attention to their suggestions, even though we do not have the same level of expertise in optical galaxy surveys. The mathematical description has been condensed (nb. the figure total number has been reduced). A new section, "Some Comments", has been added to motivate the transition from GANs to diffusion score-based models. We present arguments supporting our choice of models and provide a practical rule of thumb that estimates the necessary ratio of the total number of available pixels to model size for effective learning. These arguments serve as general guidelines, and we hope they will be useful to anyone interested in studying other generative models.
As a reminder, Sec. 3.2 provides further details on the model optimization.

Section3:
***Reviewer**: Page 6, 2nd column, line 5- Mentioning no data compression in diffusion and flow based compared to GANs or VAEs makes the point about latent space size more immediately accessible.*

Answer: I fully agree and reformulate accordingly the sentence.

Section 3.1:
***Reviewer**: Not sure why details of Kadkhodai et al sample matter in this study, they can be removed.*

The cited requirement of 10^5 images for transitioning from memorization to generalization in datasets like CelebA and LSUN Bedroom may not directly translate to galaxy images. Galaxy datasets might exhibit different variability and structural complexity, which could lower the necessary sample size. The relationship between sample size and data complexity has been studied, such as in Liu et al. (2021), where it is demonstrated that less complex distributions require fewer samples to approximate effectively.

A bit more description of the data (detailed in Smith et al. 2022) would be helpful. For instance what is the sample size that 10^5 can be used for training, what are the redshift and brightness limits.

Answer: We thank the referee for the suggestions, which have led to a reformulation of the text and a clearer emphasis on the possible outcomes of our experiments. In response, we propose to swap the order of the "The Dataset" and "The Models" sections and first expand the "The Dataset" section according to the referee's recommendations. We have also provided additional details about the SDSS DR9 dataset (noting that the mention of DR7 was an error, which we take this opportunity to correct). This description is largely reproduced
from Smith et al. (2022). Furthermore, we explicitly emphasize that our choice of dataset is by no means restrictive, and we hope that readers will be encouraged to replicate the experiment using the materials provided in the companion GitHub repository (which was already referenced in the initial submission).

We also appreciate the referee's comments regarding image complexity. This question arose in our minds prior to conducting the experiment, and we initially expected—perhaps naively—that the transition would occur below the 10^5 threshold determined by Kadkhodaie et al. However, our experimental results indicate that this is not the case, as all models exhibit suboptimal optimization when trained on a smaller number of images. This realization motivated the optimization tests developed in Section 3.3.4 ("Flow-Based Models Tests") and Section 3.3.5 ("GAN-Based Models Tests").

To keep the complexity estimation straightforward and within the scope of our framework, we propose a simple approach based on the JPEG compression ratio as a first-order proxy for image complexity. We hope that this fulfils the referee requirement. I have also made a refinement of the last paragraph of Section 3.3.3 (“Two Diffusion-Based Models Test”).

Concerning Liu et al. (2021), the authors do not explicitly conclude that less complex distributions require fewer samples for effective approximation. However, they discuss the ability to generate images with few-shot learning. The key point here is that, while generating well-formed images is important, my focus is on how the model can effectively capture the full data density distribution. Therefore, I chose Light-weight GAN from Liu et al. (2021) for its ability to perform well with small datasets, and because the hinge loss can be used for optimization testing.

Section 3.2
***Reviewer**: The section provides a clear summary of architectures and training configurations, but some choices could use clarification. Were the batch size differences (e.g., 10 for GANs vs. 32 for Glow) and resizing to 128×128 enforced by the existing architectures? If so, would adjusting these configurations or building tailored architectures better suited to the data help ensure that such choices do not affect the comparability of experimental results?*

Answer: The section 'The models' as mentioned above, has now been moved before the 'The dataset' section. I have also transferred the description of the GPU infrastructure used to the end of this new section.

I understand the referee's concern, and I would like to clarify my approach. It is not my intention to compare the three chosen architectures in terms of 'which one is the best,' as this question is too general. Rather, my focus is on investigating how one can perform tests, particularly with the two-model scenario, to ensure that the model behaves as intended. I hope the sentence I have added at the end of the introduction to the 'Experiment' section clarify this point.

Addressing the referee's question more directly, I indeed tested the models with various settings. However, due to resource limitations, my goal was not to replicate all the numerical tests described in the articles that detail these architectures. The results presented in Section 3.3 are intended to be general guidelines. They are meant to help anyone who wants to use alternative settings for the models I employed or test entirely different model architectures. I have explicitly mentioned this in the section.

Section 3.3.2
***Reviewer**: column1 line 44- two u-net networks is a bit confusing, maybe mention right at the beginning the difference in training size of the two. Also mention whether this is a u-net in a DDPM or some other denoiser not mentioned before?*

Answer: I agree that the reader may perceive this as a disconnected subject. We have added a clarification to establish a connection with Equation 40 of Section 2.4, which discusses the backward diffusion process for generation.

***Reviewer**: last paragraph, "significant implications", maybe mention the implication.*

Answer: The clarification up on the connection with Equations 20,21 of Section 2.4 at the beginning of this section was certainly the missing piece. However, I have reformulated the last sentence of the last paragraph.

***Reviewer**: More detail on PSNR might be helpful. For instance, are the galaxy cutouts normalized? Is the PSNR calculated over the entire cutout or specifically on the galaxy region? Could the model learning the background alone significantly affect the PSNR? Finally, is PSNR the most appropriate measure for evaluating the quality of denoising.*

Answer: I thank the referee for allowing me to clarify this point. The task of the denoiser is to compute the score as described in Equation 40 of Section 2.4. We could not be satisfied with a denoiser that is only effective at denoising the pixels corresponding to the galaxy. As for the evaluation, PSNR is related to the MSE metric, which is the loss function used to optimize the denoiser, so it is an appropriate choice. I agree, however, that this point should have been explicitly mentioned.

Section 3.3
***Reviewer**: In some of the experiments where increasing N from 100 to 100,000 improves generalization, it increases the likelihood of less probable data points being included in the training set. However, instead of attempting to approximate the full data density by brute force, a more efficient approach might involve strategically sampling the data distribution to ensure coverage of the parameter space (e.g., not just morphological parameters shown in figure 5/9, but sizes, ellipticities, noise levels in the background, etc). This could improve generalization without requiring massive datasets, focusing on capturing the range of the distribution rather than replicating its density. Figure 10, is there a sharp cut at the left side of the histograms or is xrange chosen for a reason.*

Answer: Concerning the former Figure 10, now Figure 9: the total number of entries is 1,000, but I have kept the same histogram range as in Figure 7 (new numbering), which results in a sharp cutoff on the left side of the histograms. However, the information is preserved: in the left histogram, the two models do not produce the same images, as indicated by the absence of a peak at cosine 1. In the right histogram, the model does not reproduce training images, which is also reflected by the absence of a peak at cosine 1. This is what I have put in other words after "We can draw some key observations from these results:…" at the beginning of section 3.3.4 ("Flow-Based Models Tests"). However, I have added a short comment.
The same comments apply for figure 13 (new numbering).

Regarding the morphological distribution figures (previously Figs. 5 & 9, now Figs. 4 & 8), I would first like to acknowledge that the referee is likely aware that in the ML domain, the Fréchet Inception Distance (FID) is commonly used as a quantitative measure to assess how well generated images match the validation dataset. However, I have deliberately chosen not to use FID (nor do I mention it in the article), as I find it somewhat artificial to rely on an Inception model trained on ImageNet, extract feature maps from one of its layers, and then compute reduced statistics to compare generated images to the dataset.

That said, I use morphological distributions because they are more natural and certainly well-known in the field of astronomy. However, they are not sufficient for assessing the quality of the model optimization, as discussed with respect to the new Figures 8, 12, and 15, which were generated with diffusion, Glow, and GAN models, respectively. The need to go beyond morphological distributions is also emphasized in the conclusion (Section 4)."

I greatly appreciate the referee's thoughtful reflection on how to « *improve generalization without requiring massive datasets* ». In this regard, I have considered using pretrained models on general-purpose images to explore their potential impact on the generalization process. However, this topic may fall outside the scope of the present article and could, of course, be addressed in future research. I have included a sentence along this line in the conclusion.

Section 4
Reviewer*: It is mentioned that the three models are compared while they are not really compared to each other directly. For example, are certain models (e.g., Glow or DDPM) inherently better suited to capturing galaxy-specific features such as faint structures, noise properties, or morphological diversity?*

Answer: I agree that this is not a direct comparison of the different types of models. While the main results had already been stated, I have slightly modified the last paragraph (takeaway) to incorporate remarks on possible ways to improve generalization without requiring massive datasets, as suggested by the referee. My focus was on facilitating the assessment of result robustness regarding optimization, which impacts the probability density function of the generator. Although I do not aim to provide a definitive evaluation, I tend to favor diffusion models, as they allow for a more direct assessment of whether the model operates in the generalization regime.

Miscellaneous (Author): I take the opportunity of the revision to add two references in the mathematical section

Song Y., Durkan C., Murray I., Ermon S., 2021b, in Beygelzimer A., Dauphin Y., Liang P., Vaughan J. W., eds, Advances in Neural Information Processing Systems. https://openreview. net/forum?id=AklttWFnxS9

Song Y., Sohl-Dickstein J., Kingma D. P., Kumar A., Ermon S., Poole B., 2021a, in International Conference on Learning Representations. https://openreview.net/forum?id=PxTIG12RRHS

and I have cleaned two cases of duplicate references (ie same article but with two disteinct Bibtex entries).