# An Organism Starts with a Single Pix-Cell: A Neural Cellular Diffusion for High-Resolution Image Synthesis

Marawan Elbatel[1], Konstantinos Kamnitsas[2,4,5], and Xiaomeng Li[1,3⋆]

[1] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[2] Department of Engineering Science, University of Oxford, Oxford, UK
[3] HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Shenzhen, China
[4] Department of Computing, Imperial College London, London, UK
[5] School of Computer Science, University of Birmingham, Birmingham, UK

**Abstract.** Generative modeling seeks to approximate the statistical properties of real data, enabling synthesis of new data that closely resembles the original distribution. Generative Adversarial Networks (GANs) and Denoising Diffusion Probabilistic Models (DDPMs) represent significant advancements in generative modeling, drawing inspiration from game theory and thermodynamics, respectively. Nevertheless, the exploration of generative modeling through the lens of biological evolution remains largely untapped. In this paper, we introduce a novel family of models termed Generative Cellular Automata (GeCA), inspired by the evolution of an organism from a single cell. GeCAs are evaluated as an effective augmentation tool for retinal disease classification across two imaging modalities: Fundus and Optical Coherence Tomography (OCT). In the context of OCT imaging, where data is scarce and the distribution of classes is inherently skewed, GeCA significantly boosts the performance of 11 different ophthalmological conditions, achieving a 12% increase in the average F1 score compared to conventional baselines. GeCAs outperform both diffusion methods that incorporate UNet or state-of-the art variants with transformer-based denoising models, under similar parameter constraints. Code is available at: https://github.com/xmed-lab/GeCA.

**Keywords:** Generative Cellular Automata (GeCA) · Diffusion Models

## 1 Introduction

Retinal diseases rank among the leading causes of vision disabilities and blindness if they remain untreated. Medical imaging modalities such as fundus photography and Optical Coherence Tomography (OCT) are widely used for diagnosing retinal conditions. OCT, offering a comprehensive view of the retinal

⋆ Correspondence: eexmli@ust.hk

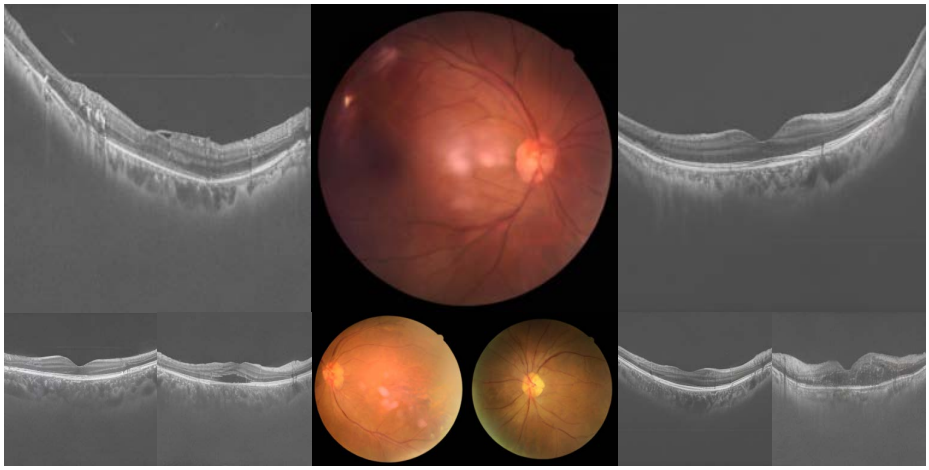


Fig. 1: Selected synthetic images from our GeCA trained on Fundus and OCT.

layers compared to the fundus, is the preferred modality for diagnosing specific diseases such as Diabetic Retinopathy (DR) and Age-related Macular Degeneration (AMD) [17]. Recently, deep learning approaches have been introduced for retinal disease screening, utilizing both fundus [15] and OCT [30]. Nevertheless, the development of these approaches is significantly hindered by the scarcity of publicly accessible datasets, particularly for OCT. Despite its advantages, OCT imaging is more costly and less employed than fundus photography, leading to a scarcity of OCT datasets. Therefore, it becomes crucial to develop a novel solution for retinal disease diagnosis using OCT imaging, especially considering its scarcity as well as its skewed disease distribution.

Expanding datasets with synthetic images through generative modeling has been shown to significantly enhance diagnostic accuracy in medical imaging, particularly in scenarios where data is scarce and class distribution is skewed [6,32,20]. Current generative models primarily utilize diffusion-based optimization [8], relying heavily on architectures such as UNet [24,6] and transformers [3,23]. Despite their effectiveness, these models require a vast number of parameters, training on large-scale datasets, and often segmentation priors [36]. These inefficiencies present considerable challenges, particularly in medical imaging, where datasets, annotations, and computational resources are often scarce. Inspired by biological processes, Neural Cellular Automata (NCA) [18] emerge as a promising alternative, offering advancements in diverse tasks with fewer parameters [27,21,13,11]. While NCA have shown promise in enabling medical image segmentation tasks under resource-constrained settings [13,11], their application in generative tasks results in low-resolution outputs [22,26,12] and lacks comprehensive performance comparisons, particularly in the evaluation of downstream tasks, where NCA's efficiency for image generation remains an unresolved challenge.

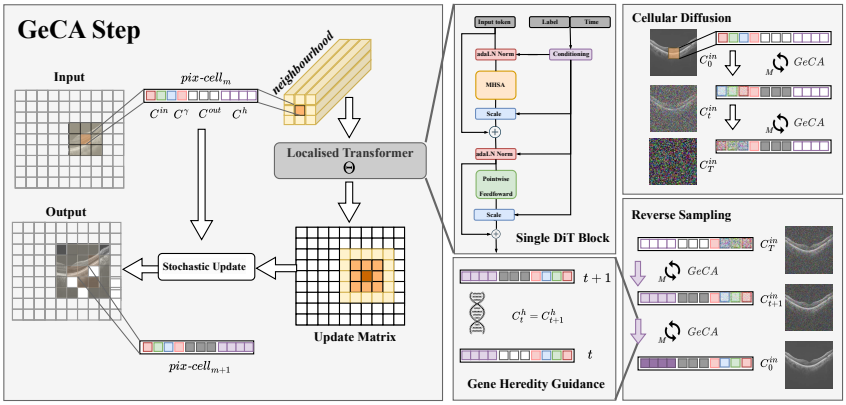



Fig. 2: GeCA overall framework.

To address these challenges, we propose a novel approach for incorporating NCA in image generation by integrating diffusion objectives specifically devised for NCA's unique structure. Operating in the latent space, scaling Neural Cellular Automata (NCA) with transformers, and introducing a novel Gene Heredity guidance method for enhanced reverse sampling, we present Generative Cellular Automata (GeCA). GeCA surpasses the state-of-the-art Diffusion Transformers (DiTs) [23] in image generation across two modalities: Fundus and OCT. By extending the application of GeCA to dataset expansion, we augment the scarce OCT dataset with synthetic images, resulting in a 12% improvement in the average F1-score for multi-label retinal disease classification compared to conventional baselines. Our contributions can be summarized as:

- We introduce Generative Cellular Automata (GeCA), a novel model that integrates Neural Cellular Automata (NCA) with diffusion objectives, tailored specifically for NCA's unique structure.
- We propose Gene Heredity Guidance (GHG) to improve GeCA's image sampling. *GHG enabled GeCA to surpass SOTA DiT in image generation and retinal disease classification* with half of DiT's parameters.
- Through a detailed examination of diffusion models in OCT image generation, we demonstrate their capability to augment training datasets with synthetic images, boosting OCT's multi-label retinal disease classification.

## 2 Generative Cellular Automata

### 2.1 An Organism Starts With a Single Pix-Cell

NCA [18] model an input image with height $H$ and width $W$ as a grid comprising $H \times W$ entities, which we refer to as *pix-cells* in our methodology. Each *pix-cell* represents a time-dependent state space representation, facilitating dynamic

evolution akin to cellular processes in an organism, *i.e.*, image. We parameterize the state of each *pix-cell* at step $m$ as a vector of scalars, defined as:

$$\textit{pix-cell}_m = \{C^{in}, C^{\gamma}, C^{out}, C^{h}\}, \tag{1}$$

where $C^{in}$ denotes the input values of the image (e.g., one scalar for grayscale and three for RGB input images), $C^{\gamma}$ represents a positional encoding, defined by a continuous and smooth sinusoidal function facilitating spatial awareness within the grid [5,29], $C^{out}$ refers to the output values indicating a *pix-cells*'s targeted state (equivalent to $C^{in}$ in dimension), and $C^{h}$ represents the hidden state variables reflecting the *pix-cell*'s internal state during evolution.

To evolve a single *pix-cell* to a more complex organism—an image, we follow traditional NCA conventional that adopts a stochastic rule [18]. This means a *pix-cell* is updated at step $m$ randomly with a probability $p$, reflecting the non-simultaneous nature of cellular updates in self-organizing organisms. The update of a *pix-cell* focuses on updating only $C^{h}$ and $C^{out}$, given that $C^{in}$ and $C^{\gamma}$ are constant. This process, illustrated as GeCA step in Fig. 2, is defined as:

$$\textit{pix-cell}_{m+1} = \Theta(\textit{pix-cell}_m, \text{Neighborhood}_8) + \{0, 0, C^{out}_m, C^{h}_m\} \tag{2}$$

Departing from the hierarchical modeling with $M$ layers in the SOTA Diffusion Transformer (DiT), we parameterize $\Theta$ as a *single DiT block* featuring a localized self-attention mechanism, specifically computed across the 8 closest neighboring *pix-cells*. The localized attention strategy, implemented similarly to those in localized transformer-based methods [33,2,27], allows each *pix-cell* to grow independently by applying Eq. (2) for $M$ times, using the same $\Theta$. GeCA's approach shifts the focus in image generation towards local spatial interactions, moving away from the global context reliance observed in traditional models such as UNet [25] and standard transformers [29]. Nevertheless, GeCA achieves global coherence by accumulating long-term state-space representation via $C^{h}$, aligning with the foundational concepts documented in NCA [18,26,27,11,13,21], Mamba [7], universal transformers [4], and MLP-mixers [28].

### 2.2 Cellular Diffusion: Evolving Cells into Organisms

To train our model parameters $\Theta$, we utilize the well-established diffusion process first introduced in [8] with specific modifications in the forward and reverse steps. During the forward diffusion process, we initialize $C^{out}$ and $C^{h}$ with zeros, except for a single *pix-cell* located at the center of the $H \times W$ grid, which is initialized with random scalars to serve as the starting point for the cellular process. $C^{\gamma}$ is initialized with a sinusoidal positional encoding. $C^{in}$ can be described in the forward diffusion process on a per *pix-cell* level as:

$$C^{in}_t = \sqrt{\alpha_t} C^{in}_0 + \sqrt{1 - \alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{3}$$

where $\epsilon$, following a normal distribution, represents the noise added at each step, and $\alpha_t$, which is part of a pre-defined variance schedule, takes values within the interval $(0, 1)$ for each time step $t = 1$ to $T$.

We then perform $M$ cellular updates with Eq. (2) to developing $C_t^{out}$ and $C_t^h$. When $T \to \infty$, $C_T^{in}$ becomes equivalent to an isotropic Gaussian distribution [8]. Thus, the optimization process is simplified from a theoretical formulation to predict the noise $\epsilon$ from a *pix-cell* as:

$$L = \mathbb{E}_{t\sim[1,T],C_{0,t}}\left[\|\epsilon - C_t^{out}\|^2\right] \tag{4}$$

This formulation allows reverse sampling from a Gaussian noise $C_T^{in} \sim \mathcal{N}(0, \mathbf{I})$. Additionally, it allows adjusting $M$ during sampling to control the intensity of generation, from undergrowth to overgrowth; see Fig. 6 in the appendix

### 2.3 Improved Reverse Sampling via Gene Heredity

Representing an input image with *pix-cells*, a time-dependent state space representation, GeCA preserves long-term information within its internal hidden states, $C^h$, analogous to genetic material. Thus, we propose leveraging $C^h$ at time $t+1$ to guide the reverse generation of time $t$, mirroring the inheritance of genetic traits. Specifically, we modify each step in the reverse process to initiate the *pix-cell* hidden states, $C^h$, as:

$$C_t^h = \begin{cases} \epsilon \sim \mathcal{N}(0, I) & \text{if } t = T \text{ and grid-center } \textit{pix-cell}, \\ C_{t+1}^h & \text{otherwise.} \end{cases} \tag{5}$$

Simultaneously, $C^{out}$, for the grid-center *pix-cell* at each timestep is defined as:

$$C_t^{out} \sim \mathcal{N}(0, I) \tag{6}$$

Our proposed process, termed *Gene Heredity Guidance (GHG)*, sets the stage for denoising $C_t^{in}$ and refining $C_t^h$ from a plausible starting point. Following GHG, the denoising process to sample a synthetic *pix-cell*, $C_0^{in}$, adheres to traditional diffusion steps till $t \to 0$ as:

$$C_{t-1}^{in} = \frac{1}{\sqrt{\alpha_t}}\left(C_t^{in} - \frac{1-\alpha_t}{\sqrt{1-\alpha_{t-1}}} C_t^{out}\right), \tag{7}$$

Note that without our proposed *GHG*, the application of NCA in generative modeling is suboptimal (See Fig. 5).

### 2.4 Retinal Disease Classification

Classifying retinal disease from OCT images presents significant challenges due to data scarcity and skewed class distributions. In light of these challenges, we leverage generative modeling to augment the dataset effectively, a strategy proven to significantly enhance downstream classification tasks compared to conventional augmentation techniques [6,35].

Following [35], we synthesize a training set expanded five-fold, mirroring the original training set's distribution. Given the original dataset's class distribution

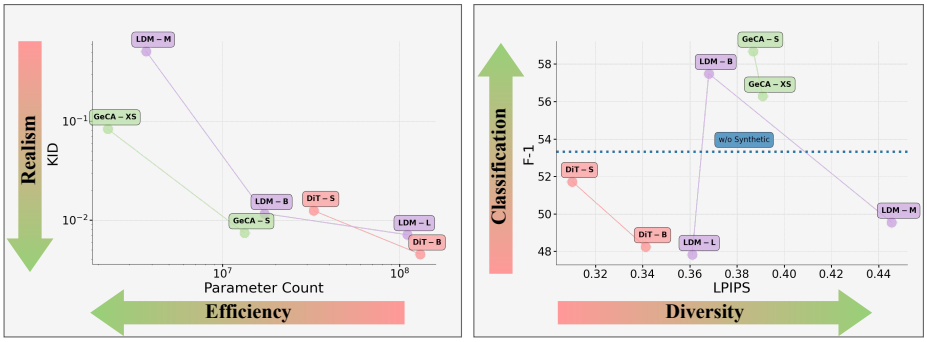



Fig. 3: Results summary on a public fundus dataset.

$p_{\text{orig}}(y)$, with $y$ representing the dataset labels and $N_{\text{orig}}$ as the original dataset size, the objective is to expand the dataset five-fold to $N_{\text{aug}} = 5 \times N_{\text{orig}}$, while preserving $p_{\text{orig}}(y)$. This is achieved by ensuring that the count of each label $y$ in the augmented dataset, $\text{Count}_{\text{aug}}(y)$, is five times its original count as:

$$p_{\text{aug}}(y) = p_{\text{orig}}(y), \quad \text{where} \quad \text{Count}_{\text{aug}}(y) = 5 \times \text{Count}_{\text{orig}}(y) \tag{8}$$

By preserving the original label distribution $p_{\text{orig}}(y)$ in the augmented dataset, we maintain the dataset's inherent distribution to avoid any potential bias.

## 3 Experiments

**Datasets.** We evaluate our model using two different datasets: OCT and Fundus. The multi-label OCT dataset, OCT-ML, is an in-house dataset consisting of 1435 samples from 369 eyes of 203 patients considering multiple diseases including normal, dry age-related macular degeneration (dAMD), wet age-related macular degeneration (wAMD), diabetic retinopathy (DR), central serous chorioretinopathy (CSC), pigment epithelial detachment (PED), macular epiretinal membrane (MEM), fluid (FLD), exudation (EXU), choroid neovascularization (CNV), and retinal vascular occlusion (RVO). Additionally, we provide the code necessary for both the generation process and the classification task, applied to DeepDRiD [16], *a publicly available fundus imaging dataset* encompassing five grading classes and follow the MedMnist split [31] (1,080 train, 120 val, 400 test). For the OCT-ML dataset, we adopt a five-fold cross-validation.

**Baselines.** Compared to previous NCA approaches [22,12] which exhibited suboptimal performance and did not compare with SOTA generative benchmarks, we compare our *GeCA* against DiT [23], *state-of-the-art diffusion transformers*, as well as the U-Net-based diffusion models from LDM [24], modifying the label embedding to support multi-label OCT generation. Training and inference for all baseline models adhere to the same hyperparameters with Classifier Free Guidance (CFG) [9] to facilitate conditional generation on downstream tasks.

Table 1: Quantitative image quality evaluation for two datasets. KID values are expressed in terms of $10^{-3}$ for each model. All baselines are trained and evaluated with *classifier free guidance (CFG)* [9] and $T = 250$.

| Method | # Param. (↓) | Fundus Dataset | | | OCT Dataset | | |
|---|---|---|---|---|---|---|---|
| | | KID (↓) | LPIPS (↑) | GG (> 0) | KID (↓) | LPIPS (↑) | GG (> 0) |
| LDM-B [24] | 17.3 M | $11.64_{\pm 2.1}$ | $0.37_{\pm 0.09}$ | −10.67 | $64.5_{\pm 10}$ | $0.39_{\pm 0.16}$ | -2.31 |
| DiT-S [23] | 32.7 M | $12.45_{\pm 2.8}$ | $0.31_{\pm 0.09}$ | −14.55 | $62.3_{\pm 5.9}$ | $0.37_{\pm 0.14}$ | -0.44 |
| **GeCA-S (ours)** | **13.3 M** | $\mathbf{7.42}_{\pm 1.6}$ | $\mathbf{0.39}_{\pm 0.11}$ | **2.02** | $\mathbf{49.1}_{\pm 8.0}$ | $\mathbf{0.53}_{\pm 0.16}$ | **0.34** |

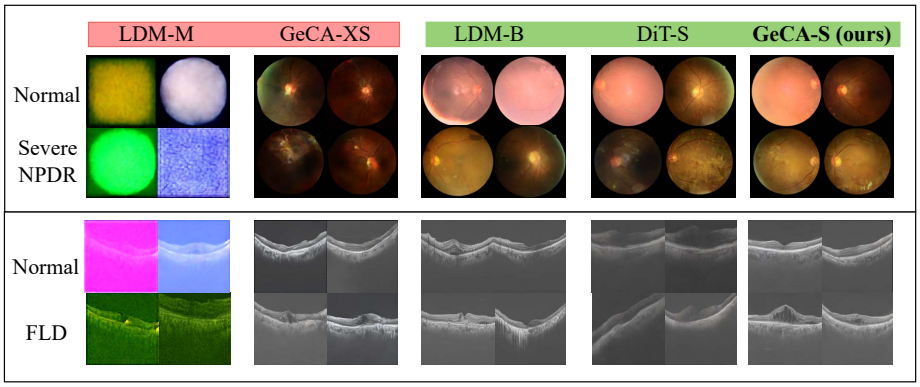



Fig. 4: Qualitative examples for the Fundus and OCT datasets are provided; images are downsampled for visualization purposes, with high-resolution versions available in the supplementary material.

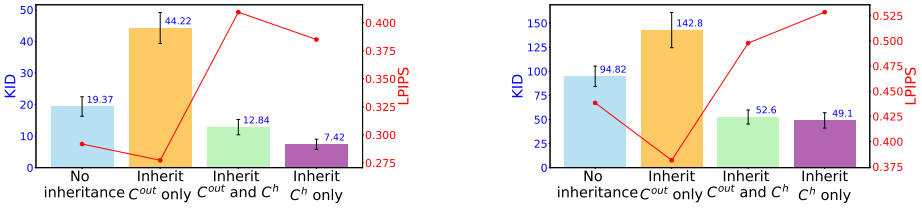
For the DiT, we report DiT-S, with an optimal patch size of 2. Given our GeCA trains *a single DiT layer*, we take $M = 12$ equivalent to the number of layers in DiT-S; See Appendix for details.

**Implementation Details.** For all methods, generation is conducted in the latent space akin to LDM [24] with an output resolution of 256x256. Training acceleration for all methods is done with mixed-precision. We utilize a batch size of 128 and train all models for 14,000 epochs until convergence. For the downstream classification task, ResNet-34 is utilized with Adam optimizer.

**Generative Modeling Evaluation.** Tab. 1 presents the quantitative results to assess the quality of the generated samples. Noting the limitations of the Fréchet Inception Distance (FID) score observed in prior works [19,10], we employ the Kernel Inception Distance (KID) for *fidelity* due to its sensitivity to dataset size [10,1]. Additionally, we report the the perceptual LPIPS *diversity* [34] to measure the image variability. Finally, we present the generalization gap (GG) as quantified by the Feature Likelihood Divergence (FLD) [10], encapsulating the triplet *novelty* (different from the training samples), *fidelity*, and *diversity* of the synthetic samples. Overall, our *GeCA demonstrates superior image quality*, both quantitatively and qualitatively, as depicted in Fig. 4. We show samples from the high-resolution GeCA model in Fig. 1 and the appendix.

Table 2: Performance results on our in-house multi-label OCT dataset, employing a five-fold cross-validation approach at the patient level. Each fold trains a separate diffusion model to generate synthetic data for the downstream classification task. All downstream experiments use ResNet34 as the backbone. We follow prior works [30,14] for eye-level performance evaluation, considering the multiple scans per eye in our dataset. The $\mathrm{F1}_{sen/pe}$ quantifies the harmonic mean of Sensitivity (Sen.) and Specificity (Spe.). (****) denote statistical significance with a p-value less than 0.0001. All reported metrics are macro-averaged.

| **Synthetic Data** | **Sen.** | **Spe.** | **AUC** | **F1** | $\mathbf{F1}_{sen/pe}$ | **mAP** | $p <$ |
|---|---|---|---|---|---|---|---|
| Baseline (Geometric Aug) | $54.66_{\pm 1.53}$ | $\mathbf{96.50}_{\pm 0.16}$ | $92.47_{\pm 0.85}$ | $55.47_{\pm 0.99}$ | $60.80_{\pm 1.49}$ | $68.85_{\pm 1.41}$ | - |
| Baseline w/o Aug. | $48.34_{\pm 1.45}$ | $96.39_{\pm 0.20}$ | $89.99_{\pm 0.82}$ | $54.56_{\pm 1.77}$ | $50.07_{\pm 0.89}$ | $64.58_{\pm 1.19}$ | ** |
| LDM-B [24] | $58.83_{\pm 1.90}$ | $96.12_{\pm 0.29}$ | $91.22_{\pm 0.74}$ | $59.65_{\pm 3.19}$ | $67.74_{\pm 2.97}$ | $70.49_{\pm 2.64}$ | ** |
| DiT-S [23] | $59.25_{\pm 4.54}$ | $95.87_{\pm 0.37}$ | $91.80_{\pm 1.74}$ | $59.11_{\pm 2.57}$ | $67.13_{\pm 4.87}$ | $69.89_{\pm 3.34}$ | *** |
| **GeCA-S (ours)** | $\mathbf{59.95}_{\pm 5.32}$ | $96.38_{\pm 0.40}$ | $\mathbf{92.74}_{\pm 2.21}$ | $\mathbf{61.62}_{\pm 3.93}$ | $\mathbf{68.38}_{\pm 4.61}$ | $\mathbf{73.28}_{\pm 5.58}$ | **** |




(a) Public Fundus dataset. (b) OCT-ML.

Fig. 5: Ablation of the proposed Gene Heredity Guidance (GHG).

**Retinal Disease Classification.** Tab. 2 presents the 11 multi-label classification results on *OCT-ML* expanded by synthetic data via generative modeling discussed in Sec. 2.4. All generative models remarkably improved the performance across various metrics. Notably, expanding the training dataset with our proposed GeCA achieved the highest mean average precision (mAP of 73.28%). GeCA significantly surpass the baseline with geometric augmentation by **4.43%** in mAP and **7.58%** in the harmonic mean of Sensitivity and Specificity ($\mathrm{F1}_{sen/pe}$). Furthermore, in terms of the traditional F1-score, which evaluates precision and recall, GeCA achieved a significant **6.15%** gain over the baseline. Despite being significantly more parameter-efficient, requiring only 40% of the parameters compared to the SOTA DiT-S [23], GeCA still manages to surpass it by **3.39%** in mAP. Furthermore, GeCA not only exceeds the performance of the leading baseline, LDM-B [24], by **2.79%** in mAP, but it also secures the highest degree of statistical significance (****). These results highlight GeCA's very promising performance in the realm of generative modeling.

**GHG Ablation.** Fig. 5 reveals the impact of Gene Heredity Guidance (GHG), introduced in Sec. 2.3, on two datasets. On the Fundus dataset, without inheritance, the model yields a moderate KID of 19.37, lacking the benefits of long-range dependencies. Inheriting $C^{out}$ alone drastically impairs performance,

spiking the KID to 44.22, suggesting that inheriting $C^{out}$ propagates noise. Conversely, inheriting both $C^{out}$ and hidden states $C^h$ partially mitigates this effect, reducing the KID to 12.84. Optimal performance is observed when only $C^h$ is inherited, achieving the *lowest KID of* 7.42. In contrast to $C^{out}$, whose primary function is to predict noise, inheriting $C^h$ facilitates the propagation of long-range dependencies, capturing the global context across the image.

## 4 Conclusion

We present GeCA, an innovative model outperforming current image generation benchmarks through neural cellular automata, demonstrated on challenging multi-label OCT classification. Future directions include broadening GeCA's validation across various domains and exploiting its unique capabilities, such as channel dimension selective sampling and temporal scheduling of its updates.

**Acknowledgments.** This work was supported in part by the grants from Foshan HKUST Projects, Grant Nos. FSUST21-HKUST10E and FSUST21-HKUST11E and in part by Project of Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone (HZQB-KCZYB-2020083). Marawan Elbatel is supported by the Hong Kong PhD Fellowship Scheme (HKPFS) from the Hong Kong Research Grants Council (RGC), and by the Belt and Road Initiative from the HKSAR Government.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs. In: International Conference on Learning Representations (2018)
2. Chen, C.F., Panda, R., Fan, Q.: Regionvit: Regional-to-local attention for vision transformers. In: International Conference on Learning Representations (2022)
3. Chowdary, G.J., Yin, Z.: Diffusion transformer u-net for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 622–631. Springer Nature Switzerland, Cham (2023)
4. Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., Kaiser, L.: Universal transformers. In: International Conference on Learning Representations (2019)
5. Dosovitskiy, A., Beyer, L., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
6. Frisch, Y., Fuchs, M., et al.: Synthesising rare cataract surgery samples with guided diffusion models. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 354–364. Springer Nature Switzerland, Cham (2023)
7. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. ArXiv **abs/2312.00752** (2023)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS'20, Curran Associates Inc., Red Hook, NY, USA (2020)

9. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021)
10. Jiralerspong, M., Bose, J., Gemp, I., Qin, C., Bachrach, Y., Gidel, G.: Feature likelihood score: Evaluating the generalization of generative models using samples. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
11. Kalkhof, J., González, C., Mukhopadhyay, A.: Med-nca: Robust and lightweight segmentation with neural cellular automata. In: International Conference on Information Processing in Medical Imaging. pp. 705–716. Springer (2023)
12. Kalkhof, J., Kühn, A., Frisch, Y., Mukhopadhyay, A.: Frequency-time diffusion with neural cellular automata. ArXiv **abs/2401.06291** (2024)
13. Kalkhof, J., Mukhopadhyay, A.: M3d-nca: Robust 3d segmentation with built-in quality control. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 169–178. Springer (2023)
14. Li, X., Zhou, Y., Wang, J., Lin, H., Zhao, J., Ding, D., Yu, W., Chen, Y.: Multimodal multi-instance learning for retinal disease recognition. In: Proceedings of the 29th ACM International Conference on Multimedia. p. 2474–2482. MM '21, Association for Computing Machinery, New York, NY, USA (2021)
15. Li, Y., Zhang, R., et al.: Predicting systemic diseases in fundus images: systematic review of setting, reporting, bias, and models' clinical availability in deep learning studies. Eye (Jan 2024)
16. Liu, R., Wang, X., et al.: Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. Patterns p. 100512 (2022)
17. Midena, E., Frizziero, L., et al.: Optical coherence tomography and color fundus photography in the screening of age-related macular degeneration: A comparative, population-based study. Plos one **15**(8), e0237352 (2020)
18. Mordvintsev, A., Randazzo, E., Niklasson, E., Levin, M.: Growing neural cellular automata. Distill (2020). https://doi.org/10.23915/distill.00023
19. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 7176–7185. PMLR (13–18 Jul 2020)
20. Oh, H.J., Jeong, W.K.: Diffmix: Diffusion model-based data synthesis for nuclei segmentation and classification in imbalanced pathology image datasets. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 337–345. Springer Nature Switzerland, Cham (2023)
21. Pajouheshgar, E., Xu, Y., Zhang, T., Süsstrunk, S.: Dynca: Real-time dynamic texture synthesis using neural cellular automata. CVPR pp. 20742–20751 (2022)
22. Palm, R.B., Duque, M.G., Sudhakaran, S., Risi, S.: Variational neural cellular automata. In: International Conference on Learning Representations (2022)
23. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4172–4182. IEEE Computer Society, Los Alamitos, CA, USA (oct 2023)
24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham
26. Sudhakaran, S., Najarro, E., Risi, S.: Goal-guided neural cellular automata: Learning to control self-organising systems. In: From Cells to Societies: Collective Learning across Scales (2022)

27. Tesfaldet, M., Nowrouzezahrai, D., Pal, C.: Attention-based neural cellular automata. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022)
28. Tolstikhin, I., Houlsby, N., et al.: MLP-mixer: An all-MLP architecture for vision. In: Advances in Neural Information Processing Systems (2021)
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS. p. 6000–6010. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
30. Wang, L., Dai, W., Jin, M., Ou, C., Li, X.: Fundus-enhanced disease-aware distillation model for retinal disease classification from oct images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 639–648. Springer (2023)
31. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)
32. Yang, Y., Fu, H., Aviles-Rivero, A.I., Schönlieb, C.B., Zhu, L.: Diffmic: Dual-guidance diffusion network for medical image classification. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 95–105. Springer Nature Switzerland, Cham (2023)
33. Zhang, P., Dai, X., et al.: Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. ICCV (2021)
34. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 586–595 (2018)
35. Zhang, Y., Zhou, D., Hooi, B., Wang, K., Feng, J.: Expanding small-scale datasets with guided imagination. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=82HeVCqsfh
36. Zhao, H., Li, H., Maurer-Stroh, S., Guo, Y., Deng, Q., Cheng, L.: Supervised segmentation of un-annotated retinal fundus images by synthesis. IEEE Transactions on Medical Imaging **38**, 46–56 (2019)