*t*HPM-LDM: Integrating Individual Historical Record with Population Memory in Latent Diffusion-based Glaucoma Forecasting

Yuheng Fan[1], Jianyang Xie[1], Yimin Luo[2], Yanda Meng[3], Savita Madhusudhan[4], Gregory Y.H. Lip[5], Li Cheng[6], Yalin Zheng[1], and He Zhao[1]

[1] Eye and Vision Sciences Department, University of Liverpool, Liverpool, UK
[2] Department of Radiology, Weill Cornell Medicine, New York City, US
[3] Computer Science Department, University of Exeter, Exeter, UK
[4] St Paul's Eye Unit, Liverpool University Hospitals NHS Foundation Trust, Liverpool, UK
[5] Liverpool Centre for Cardiovascular Science, University of Liverpool, Liverpool, UK
[6] Department of Electrical and Computer Engineering, University of Alberta, Canada
He.Zhao@liverpool.ac.uk

Abstract. Longitudinal medical records offer crucial insights into disease progression, including structural changes and dynamic evolution, essential for clinicians in treatment planning. However, existing disease forecasting methods are hindered by irregular data collection intervals, negligence in inter-patient relationships, and a lack of case-reference capabilities. We introduce *t*HPM-LDM, a glaucoma forecasting framework leveraging continuous-time attention within a historical condition module to capture disease progression from irregularly acquired records. Notably, our approach integrates population memory, enabling personalized forecasting through relevant population patterns. Empirical evaluations on the SIGF glaucoma longitudinal dataset demonstrate the significant improvements of our approach in image prediction and category consistency compared to state-of-the-art methods. Furthermore, our approach provides interpretable individual-population patterns and showcases robust performance despite missing visits.

Keywords: Longitudinal Record · Irregular-time Forecasting · Latent Diffusion Model · Population Memory

1 Introduction

Glaucoma, an irreversible and leading cause of blindness, presents significant diagnostic challenges due to its chronic, often asymptomatic nature, requiring extensive long-term monitoring [13]. Given its relatively low global prevalence (estimated 3.54% among individuals aged 40-80 [20]), developing models for accurate glaucoma forecasting from longitudinal data is essential for improving early detection and reducing healthcare costs. However, the irregular acquisition patterns inherent in longitudinal medical data present a substantial obstacle

for deep learning models in capturing the dynamic nature of disease progression. While traditional methods address this by aggregating data at specific time scales [19] or estimating missing data [23], these methods may underutilize longitudinal data and weaken temporal dependencies.

Some methods focusing on architecture improvement [1], such as previous glaucoma forecasting models: DeepGF [14] employs an LSTM-based method to process fundus images for glaucoma forecasting for next visit with an indefinite time span; GLIM-Net [9] introduces a transformer that integrates image and label sequences with time intervals, enabling label predication at a given time; MSTF [21] further extends into a multi-scale framework for improved spatial dependence modeling. However, these methods primarily focus on category forecasting and lack visual prediction capabilities.

More recently, several works have been proposed for longitudinal image forecasting and completion [22, 26]. C2FLDM [25], a coarse-to-fine latent diffusion model (LDM), employs a two-stage framework for glaucoma image and category forecasting. BrLP [16] incorporates prior knowledge and progression patterns through Disease Course Mapping [12] in an LDM-based Alzheimer's disease forecasting model, which are constrained by its reliance on accurate volumetric brain segmentation. These methods model the temporal relationships in a simplistic manner and ignore related disease progression pattens across the population.

In this paper, we propose a glaucoma forecasting framework called continuous-**t**ime **H**istorical and **P**opulation **M**emory-conditioned **L**atent **D**iffusion **M**odel (*t*HPM-LDM). We employ continuous-time modeling to address the challenge of irregularly spaced historical records over time, while incorporating population evolution to enhance personalized image and category forecasting. The main contributions are: 1) to our best knowledge, this is the first disease forecasting approach that integrates disease population information with image generation; 2) a multiscale continuous-time transformer is introduced to capture dynamic evolution within the irregularly acquired records, improves visual forecasting quality and robustness toward missing visits; 3) a retrievable memory module is proposed for individual interaction with the disease population, mimicking how clinicians reference similar cases during diagnosis.

2 Methods

Given a longitudinal dataset $D = \{R_{1:T}^{(m)} = [(x_i, l_i, t_i)]_1^T\}_{m=1}^M$ containing T visits from M participants, where $x_i \in \mathbb{R}^{W\times H\times 3}$ represents the image from the i-th visit, $l_i \in \{0, 1\}$ denotes the disease label (0 for healthy, 1 for glaucoma), and $t_i \in \mathbb{R}$ indicates the time interval between the i-th visit and the first visit. Our goal is to train a model to leverage the historical information from time steps $1 : T-1$ to forecast the participant's future image $\hat{x}_T$ and its category $\hat{l}_T$. To achieve this, we propose a LDM that generates future images, conditioned on both individual history c_h and population reference c_p. The generated images are subsequently fed into a pretrained classifier for category prediction. Fig. 1

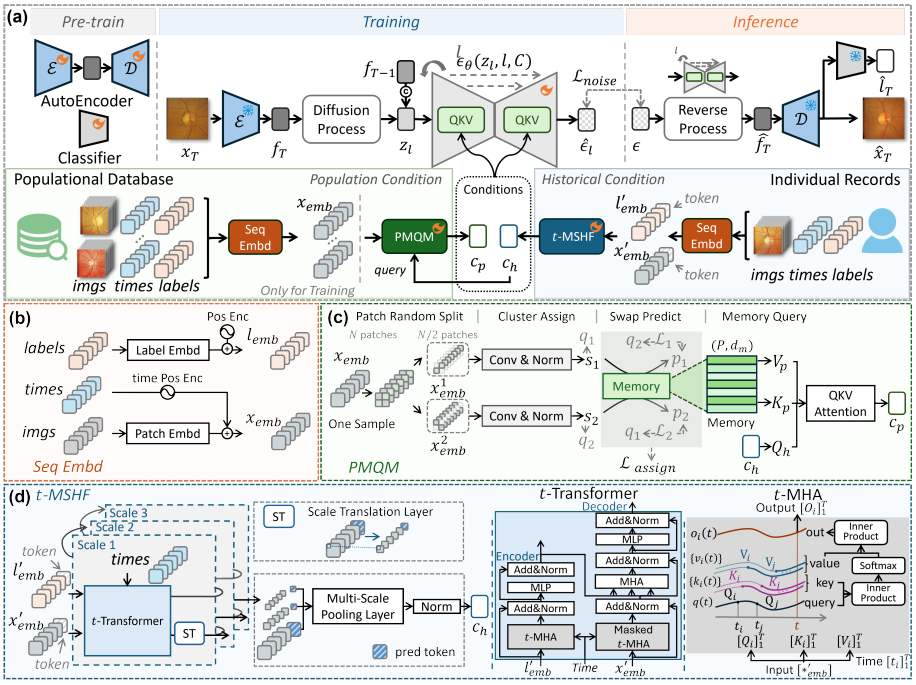


Fig. 1. Overview of our framework. (a) depicts the training and inference processes. (b) presents the embedding module, which generates embedding from longitudinal records. (c) illustrates the pipeline of PMQM in generating the population condition c_p. (d) outlines the model structure of t-MSHF in extracting the historical condition c_h.

provides an overview of our approach. In the follow subsections, we introduce each of the modules in detail.

2.1 Disease forecasting with Conditional Latent Diffusion

Our forecast framework is presented in Fig. 1(a), which contains a pre-trained auto-encoder ($\mathcal{E}$-$\mathcal{D}$) for image x_T and latent feature f_T translation, and a diffusion model for future latent prediction. In the training stage, the latent feature f_T is decomposed into noisy status $\{z_l\}_0^L$ with L-steps diffusion as:

$$z_l = \sqrt{\bar{\alpha}_l} z_0 + \sqrt{1-\bar{\alpha}_l}\epsilon, \text{ with } \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\ \bar{\alpha}_l = \prod_{i=1}^{l} \alpha_l,\ \alpha_l = 1 - \beta_l, \quad (1)$$

where $z_0 = f_T$, $l \in [1, L]$, and $\{\beta_i\}_0^L$ is the linear noise schedule. During inference, the latent feature f_T is restored from noise z_L with a denoising process:

$$z_{l-1} = \frac{1}{\sqrt{\alpha_l}}\left(z_l - \frac{1-\alpha_l}{\sqrt{1-\bar{\alpha}_l}}\epsilon_\theta(z_l, l, C)\right) + \sigma_l \epsilon \text{ with } \sigma_l = \sqrt{\frac{1-\bar{\alpha}_{l-1}}{1-\bar{\alpha}_l}\beta_l}, \quad (2)$$

where the estimator $\epsilon_\theta(z_l, l, C)$ predicts the noise ϵ_l at each l-step under conditions $C = \{c_h, c_p\}$. Here, c_h is the historical feature and c_p is associated population condition, and all combined into intermediate layers of ϵ_θ via cross-attention. We also uses latent alignment by concatenating f_{T-1} with z_l for better anatomical structure [11]. The loss function of the diffusion model is shown as:

$$\mathcal{L}_{\text{noise}} := \mathbb{E}_{\mathcal{E}(x_T),\epsilon\sim\mathcal{N}(0,1),l\sim\mathcal{U}(0,L)}\left[\|\epsilon - \epsilon_\theta(z_l, l, C)\|_2^2\right] \tag{3}$$

2.2 Conditional modules

Two modules, namely t-MSHF and PMQM, are proposed to generate historical and population conditions, respectively, as shown in Fig. 1(d) and (c). A sequence embedding module `Seq Embd` is utilized to produce image embeddings $x_{\text{emb}} \in \mathbb{R}^{T\times N\times d_m}$ and label embeddings $l_{\text{emb}} \in \mathbb{R}^{T\times d_m}$ from the images and labels of all visits with the process depicted in Fig. 1(b). Here, N represents the patch number and d_m denotes the embedding dimension. These embeddings serve as inputs for t-MSHF and are used to learn the memory in PMQM. Notably, to prevent future information leakage, the last visit embeddings of x_{emb} and l_{emb} are replaced with learnable tokens for t-MSHF.

Continuous-time Multi-scale Historical Feature Module (t-MSHF). Our t-MSHF consists of t-Transformers with continuous-time attention module (t-MHA) at different scales, which captures multiscale dynamic information from individual historical records. Inspired by [4], t-MHA uses continuous functions in modeling the attention of dynamic evolution based on discrete observations. More precisely, the input embeddings at all visit times are first copied to observed sequences $[Q_i]_1^T$, $[K_i]_1^T$ and $[V_i]_1^T$. The query $q(t)$ is an interpolation function (e.g. bilinear interpolation) with observed point $q(t_i) = Q_i$. And the continuous key and value are sets of functions $\{k_i(t)\}$ and $\{v_i(t)\}$ based on each observation i. Neural networks f_{θ_k} and f_{θ_v} estimate the dynamics of key and value, where the changes of $\{k_i(t)\}$ and $\{v_i(t)\}$ described as linear ODEs [3] as follow:

$$\frac{\mathrm{d}k_i}{\mathrm{d}t} = f_{\theta_k}(k_i(t), t),\ k_i(t_i) = K_i \qquad \frac{\mathrm{d}v_i}{\mathrm{d}t} = f_{\theta_v}(v_i(t), t),\ v_i(t_i) = V_i. \tag{4}$$

Observed Q_i, K_i at time t_i are initial conditions of ODEs, and we solve these ODEs by Runge-Kutta-4 algorithm. For any visit time t, the i-th continuous correlation of query and key functions is defined as the scaled inner product of $q(t)$ and $k_i(t)$, i.e. $\alpha_i(t) := (\int_{t_i}^t q(\tau)\cdot k_i(\tau)^\mathsf{T} d\tau)/(t - t_i)$ and $\alpha_i(t_i) = q(t_i)\cdot k_i(t_i)^\mathsf{T}$ for each t_i to ensure continuity. The expected value function based on i-th observation is $\hat{v}_i(t) = (\int_{t_i}^t v_i(\tau)d\tau)/(t - t_i)$. We approximate these integrations by Gauss-Legendre Quadrature. Therefore, the output function is computed as:

$$O(t) = \sum_{i=1}^{T} \text{softmax}(\alpha_i(t)/\sqrt{d_k})\cdot \hat{v}_i(t) \tag{5}$$

where d_k is the dimension of inputs. Discrete output sequence $[O_i]_{i=1}^T$ is computed based on Eq. 5 according to time $[t_i]_1^T$, which will be used for next layers.

Our t-Transformer, equipped the t-MHA, will handle label embeddings in the Transformer encoder and process image embeddings with its decoder. Additionally, the t-Transformer will be applied iteratively in a multi-stage manner with a scale translation layer to enable multi-scale learning. The final historical condition $c_h \in \mathbb{R}^{1\times d_m}$ is obtained by the average of different scales.

Population Memory Query Module (PMQM). In practice, clinicians will evaluate the patient's current status by considering both individual medical histories and the knowledge of disease progression from other cases. Motivated by this practice, we propose PMQM, which initializes a population memory $\mathcal{M}_p \in \mathbb{R}^{P\times d_m}$ to capture evolution patterns of P clusters of representative patients. Following the SwAV [2], we apply online contrastive learning with swapping strategy to update the memory from the disease population. As shown in Fig.1(c), one image embedding x_{emb} is first divided into two sub-embeddings randomly, resulting in $x^1_{\text{emb}}, x^2_{\text{emb}} \in \mathbb{R}^{T\times N/2\times d_m}$ and further projected to $s_1, s_2 \in \mathbb{R}^{1\times d_m}$. The loss function to update the memory is shown as:

$$\mathcal{L}_{\text{assign}}(s_1, s_2) := -\frac{1}{2}\sum(q_1 \log p_2 + q_2 \log p_1), \tag{6}$$

where $p_* = \text{softmax}(s_*^{\mathsf{T}} \cdot \mathcal{M}_p)$ represents the predicted assign probability, and q_* is the optimal assign probability computed by the Sinkhorn-Knopp algorithm [5]. By minimizing the loss over the training samples, the memory $\mathcal{M}_p$ learns to represent the training participants into P clusters with centers stored in the memory. Moreover, we can build an offline cache using predicted assignment p that stores the probability of the patient from each cluster, providing a retrievable strategy to determine the top-k relevant clusters and their corresponding participants in certain patient forecasting. Finally, the condition c_p is obtained by the cross attention of historical condition c_h (Query) and the learned memory $\mathcal{M}_p$ (Key, Value) by the following equation:

$$c_p = \text{softmax}\left((Q_h \cdot K_p^{\mathsf{T}})/\sqrt{d_m}\right) V_p \tag{7}$$

2.3 Learning conditional LDM and Memory all-in-once.

After pre-training the autoencoder ($\mathcal{E}$-$\mathcal{D}$) and classifier $\mathcal{P}$, our methods can learn the LDM with condition modules all-in-once. We first perform $W = 1500$ iterations of warm-up training on the noise estimator ϵ_θ and t-MSHF with loss in Eq. 3 to prevent the random query for PMQM. After that, joint training is conducted using the total loss:

$$\mathcal{L}_{total} := \mathcal{L}_{\text{noise}} + \lambda_m \mathcal{L}_{assign}, \tag{8}$$

where λ_m is the balanced hyper-parameter, empirically setting to 0.1.

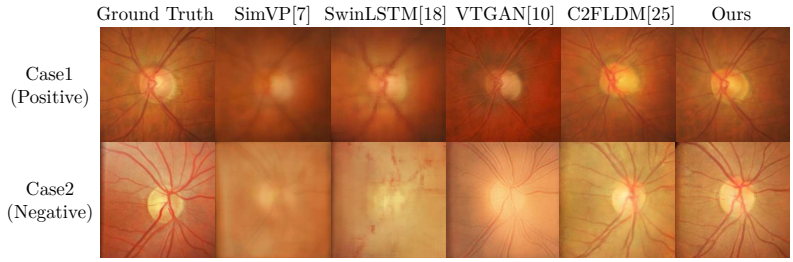


Fig. 2. Two image prediction cases using different methods.

3 Experiments

3.1 Setup Details

Dataset. We conducted experiments on the glaucoma longitudinal dataset named SIGF [14]. Following the conventional process [9], we resize fundus images to 256×256, compute the time interval in years, and collect $T = 6$ consecutive visits as one clip, resulting 1146 training, 16 validation and 351 testing clips.

Settings. The autoencoder ($\mathcal{E}$-$\mathcal{D}$) is a U-Net structure [17] with latent dimension $32 \times 32 \times 4$ and pre-trained on two external glaucoma dataset named LAG [15] and ACRIMA [6]. The classifier $\mathcal{P}$ is a ResNet50 with two hidden layers and a binary head per-trained on SIGF training set, LAG and ACRIMA. The embedding dimension d_m is 128. More details can refer to our code[7].

Metrics and Baseline Methods. We take 6th image in test clip as ground truth and evaluate the visual quality and category consistency of all methods. For future image generation, we use PSNE, SSIM and MSE to measure image quality, along with semantic similarity metrics such as FID [8] and LPIPS [24]. For category forecasting, we report the performance of pre-trained classifier using accuracy, sensitivity, specificity, and AUC. Cohen's Kappa (Kappa) is utilized to compare the forecasting consistency between ground truth and generative results. We compare our approach with four baseline methods: two video models for future image generation from image sequence (CNN-based SimVP [7] and Transformer-LSTM based SwinLSTM [18]); one GAN-based image-to-image translation method VTGAN [10]; and one LDM-based sequence-to-image method C2FLDM [25]. All methods are trained using their recommended settings with identical batch sizes for a maximum of 700 epochs until convergence. Quantitative results are reported as the mean and std over five runs.

3.2 Experimental Results

Results on image prediction. Qualitative and quantitative results are shown in Fig. 2 and Table 1, respectively. Two cases generated by different

[7] The implementation is available at https://github.com/yhf42/tHPM-LDM

Table 1. Quantitative results of different methods for image prediction.

Model	PSNR ↑	SSIM ↑	MSE (10^2) ↓	FID ↓	LPIPS ↓
SimVP [7]	19.02±0.02	0.70±0.00	11.64±0.09	298.75±0.75	0.50±0.00
SwinLSTM [18]	19.01±0.04	0.67±0.00	11.35±0.06	300.23±0.86	0.51±0.00
VTGAN [10]	18.42±0.01	0.65±0.00	13.67±0.03	189.33±1.15	0.47±0.00
C2FLDM [25]	18.32±0.06	0.65±0.00	13.91±0.16	210.67±0.71	0.50±0.00
Ours	**19.30**±0.03	**0.71**±0.00	**10.40**±0.05	**141.19**±1.32	**0.44**±0.00

Table 2. Quantitative results of different methods for category forecasting.

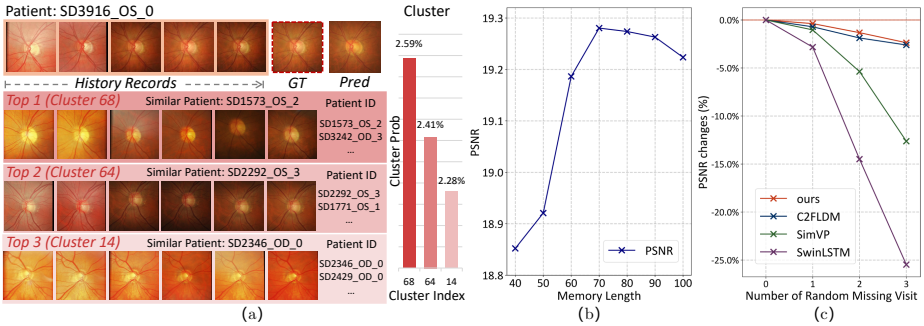
Model	Accuracy(%)	Specificity(%)	Sensitivity(%)	AUC(%)	Kappa ↑
Ground Truth	77.09±1.48	77.31±1.79	72.50±5.00	83.03±0.35	-
SimVP [7]	29.12±1.25	26.03±1.31	93.75±0.00	59.48±1.57	0.14±0.02
SwinLSTM [18]	12.65±0.39	9.25±0.46	83.75±3.06	42.60±1.28	-0.01±0.01
VTGAN [10]	70.14±1.32	70.39±1.15	61.00±6.09	74.07±1.99	0.35±0.02
C2FLDM [25]	49.86±0.51	47.70±0.40	**95.00**±4.68	75.22±0.82	0.29±0.02
Ours	**74.02**±1.16	**74.27**±1.02	68.75±5.59	**82.06**±1.23	**0.45**±0.02

methods are displayed in Fig. 2, where Cases 1 has positive findings in real diagnosis while Cases 2 is negative. Compared to the ground truth images shown in the leftmost column, our predicted images demonstrate the best semantic similarity and clearer OC/OD and vessel structures among other methods. Although SimVP and SwimLSTM utilize historical images for image prediction, their results fail to show a clear fundus structure. VTGAN and C2FLDM generates images with better semantic information, but unstable contrast and structural variation poses challenges in identifying glaucoma. Table 1 demonstrates the superior performance of our model in both image quality and semantic consistency. Notably, while SimVP and SwinLSTM show relatively better PSNR, SSIM, and MSE since the low clarity may reduce prediction errors, their blurry structure leads to semantic dissimilarity, such as higher FID and LPIPS, making them inadequate for further analysis.

Results on category forecasting. Table 2 first summarizes the performance of pre-trained classifier $\mathcal{P}$ on test set ground truth images, which regard as the upper bound of the category forecast. Compared to other generative methods, our approach achieves a more balanced performance in glaucoma forecasting and has the highest consistency with the prediction on ground truth. The predicted images from SimVP and SwinLSTM confused the classifier's judgment due to the lack of fine-grain structure. VTGAN's results lead to a higher false-negative rate since the higher brightness increases the area of the optic cup. And the classifier tends to predict C2FLDM's results as positive due to their blurred optic disc boundaries.

Table 3. Results of Ablation Study. LA: Latent Alignment Strategy, MS: Multi-Scale Translation, tM: t-MHA, QM: Population Memory Query Module.

LA	MS	tM	QM	PSNR ↑	SSIM ↑	MSE (10^2) ↓	FID ↓	LPIPS ↓	Kappa ↑
✓	✗	✗	✗	18.16±0.10	0.71±0.00	14.78±0.31	176.30±2.19	0.48±0.00	0.35±0.02
✓	✓	✗	✗	18.44±0.06	0.71±0.00	14.26±0.22	155.09±1.22	0.46±0.00	0.44±0.02
✓	✓	✓	✗	18.81±0.10	**0.72**±0.00	12.17±0.29	149.32±1.10	0.45±0.00	**0.45**±0.01
✓	✓	✓	✓	**19.30**±0.03	0.71±0.00	**10.40**±0.05	**141.19**±1.32	**0.44**±0.00	0.45±0.02



(a) (b) (c)

Fig. 3. (a) Memory retrieval results for forecasting of one glaucoma patient. (b) The ablation study of the memory length P in PMQM. (c) The relative changes (%) of PSNR under different number of missing visit.

3.3 Ablation Study

Our ablation study results are shown in Table 3. Latent alignment preserves basic anatomical structures, serving as a default setting. Multi-scale translation enhances performance by capturing spatial features at different scales, and t-MHA enables the LDM to better capture the dynamic progression within records. PMQM results in the best image quality by leveraging population patterns for individual prediction, though mirroring real diagnosis that considering population trends may influence the judgment. Figure 3(b) illustrates the influence of memory length P in PMQM, where $P = 70$ achieves the best result.

Population Memory Retrieve. Another benefit of our proposed population memory is that its ability to retrieve the high relevant cases. This enables our method with stronger interpretability in patient-level disease progression forecasting. Fig.3 (a) displays one example of the memory retrieval, which demonstrates that our method can successfully query groups of participants with similar structural characteristics and evolutionary trajectories.

Random Missing Visit Analysis. To evaluate the performance of sequence-based models under varying levels of missing visits, we randomly mask out different numbers of visits (2nd to 4th) to simulate records missing scenario. Fig. 3(c) shows the relative changes in PSNR of SimVP, SwinLSTM, C2FLDM and our method under different levels of missing visits. The results demonstrate that conditional LDM-based methods (C2FLDM and ours) achieve better pre-

diction performance with missing data. Furthermore, our model exhibits better robustness toward missing visits thanks to the t-MHA's capacity in modeling the irregular series and the sematic reference in population memory.

4 Conclusion

In this paper, we propose a conditional latent diffusion-based glaucoma forecasting framework. Specifically, we design t-MSHF, which incorporates continuous-time attention into multi-scale transformers to better capture the spatiotemporal dynamics of irregularly acquired records. We further enable population-assisted forecasting by querying individual-related features in PMQM. These conditions are jointly learned with the LDM to improve glaucoma image and category forecasting. Furthermore, our method enables the retrieval of participants with similar progression patterns to a given individual. This individual-population interaction does not rely on expertise in a specific disease and has potential for analyzing other diseases with longitudinal follow-up.

Acknowledgments. This work is funded by the UK Engineering and Physical Sciences Research Council (grant number EP/X01441X/1) and Medical Research Council (grant number MR/Z506175/1).

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

References

1. Bridge, J., Harding, S., Zheng, Y.: End-to-end deep learning vector autoregressive prognostic models to predict disease progression with uneven time intervals. In: Medical Image Understanding and Analysis: 25th Annual Conference, MIUA 2021, Oxford, United Kingdom, July 12–14, 2021, Proceedings 25. pp. 517–531. Springer (2021)
2. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020)
3. Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. Advances in neural information processing systems **31** (2018)
4. Chen, Y., Ren, K., Wang, Y., Fang, Y., Sun, W., Li, D.: Contiformer: Continuous-time transformer for irregular time series modeling. Advances in Neural Information Processing Systems **36**, 47143–47175 (2023)
5. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
6. Diaz-Pinto, A., Morales, S., Naranjo, V., Köhler, T., Mossi, J.M., Navea, A.: Cnns for automatic glaucoma assessment using fundus images: an extensive validation. Biomedical engineering online **18**, 1–19 (2019)

7. Gao, Z., Tan, C., Wu, L., Li, S.Z.: Simvp: Simpler yet better video prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3170–3180 (2022)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
9. Hu, X., Zhang, L.X., Gao, L., Dai, W., Han, X., Lai, Y.K., Chen, Y.: Glim-net: chronic glaucoma forecast transformer for irregularly sampled sequential fundus images. IEEE Transactions on Medical Imaging **42**(6), 1875–1884 (2023)
10. Kamran, S.A., Hossain, K.F., Tavakkoli, A., Zuckerbrod, S.L., Baker, S.A.: Vtgan: Semi-supervised retinal image synthesis and disease prediction using vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 3235–3245 (2021)
11. Kim, M., Liu, F., Jain, A., Liu, X.: Dcface: Synthetic face generation with dual condition diffusion model. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 12715–12725 (2023)
12. Koval, I., Bône, A., Louis, M., Lartigue, T., Bottani, S., Marcoux, A., Samper-Gonzalez, J., Burgos, N., Charlier, B., Bertrand, A., et al.: Ad course map charts alzheimer's disease progression. Scientific Reports **11**(1), 8020 (2021)
13. Leite, M.T., Sakata, L.M., Medeiros, F.A.: Managing glaucoma in developing countries (2011)
14. Li, L., Wang, X., Xu, M., Liu, H., Chen, X.: DeepGF: Glaucoma forecast using the sequential fundus images. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, vol. 12265, pp. 626–635. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-59722-1_60
15. Li, L., Xu, M., Wang, X., Jiang, L., Liu, H.: Attention based glaucoma detection: a large-scale database and cnn model. 2019 ieee. In: CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10563–10572 (2019)
16. Puglisi, L., Alexander, D.C., Ravì, D.: Enhancing spatiotemporal disease progression models via latent diffusion and prior knowledge. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 173–183. Springer (2024)
17. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
18. Tang, S., Li, C., Zhang, P., Tang, R.: Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13470–13479 (2023)
19. Tomašev, N., Glorot, X., Rae, J.W., Zielinski, M., Askham, H., Saraiva, A., Mottram, A., Meyer, C., Ravuri, S., Protsyuk, I., et al.: A clinically applicable approach to continuous prediction of future acute kidney injury. Nature **572**(7767), 116–119 (2019)
20. Weinreb, R.N., Aung, T., Medeiros, F.A.: The pathophysiology and treatment of glaucoma: a review. Jama **311**(18), 1901–1911 (2014)
21. Yang, X., Wu, J., Wang, X., Yuan, Y., Li, J., Chen, G., Wang, N.L., Heng, P.A.: Multi-scale spatio-temporal transformer-based imbalanced longitudinal learning for glaucoma forecasting from irregular time series images. IEEE Journal of Biomedical and Health Informatics (2024)

22. Yoon, J.S., Zhang, C., Suk, H.I., Guo, J., Li, X.: Sadm: Sequence-aware diffusion model for longitudinal medical image generation. In: International Conference on Information Processing in Medical Imaging. pp. 388–400. Springer (2023)
23. Yoon, J., Zame, W.R., Van Der Schaar, M.: Estimating missing data in temporal data streams using multi-directional recurrent neural networks. IEEE Transactions on Biomedical Engineering **66**(5), 1477–1490 (2018)
24. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
25. Zhang, Y., Huang, K., Yang, X., Ma, X., Wu, J., Wang, N., Wang, X., Heng, P.A.: Coarse-to-fine latent diffusion model for glaucoma forecast on sequential fundus images. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, vol. 15005, pp. 166–176. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-72086-4_16
26. Zhu, Z., Tao, T., Tao, Y., Deng, H., Cai, X., Wu, G., Wang, K., Tang, H., Zhu, L., Gu, Z., et al.: Loci-diffcom: Longitudinal consistency-informed diffusion model for 3d infant brain image completion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 249–258. Springer (2024)