GenSRL: Generative Spatiotemporal Representation Learning for Ophthalmic Prognosis Prediction

Supplementary Material

1. Public OCT4DME Experiments

1.1. OCT4DME Dataset

To evaluate the cross-dataset generalization ability of GenSRL, we further test our method on the public OCT4DME dataset [1], which focuses on prognosis prediction for diabetic macular edema. The dataset contains longitudinal OCT scans collected at baseline and follow-up timepoints, together with structured clinical annotations describing treatment response and fluid evolution. In contrast to our in-house cohort, OCT4DME only provides binary labels indicating whether subretinal fluid (SRF) has been absorbed after treatment, and does not contain annotations for visual acuity improvement. Therefore, on this public benchmark, we only evaluate the *subretinal fluid change (SFC)* classification task.

1.2. OCT4DME Results

We benchmark GenSRL against representative CNN-, transformer-based, and multimodal prognosis models on OCT4DME, and report accuracy, precision, recall, and F1 for the SFC classification task. As summarized in Table 1, GenSRL achieves the best performance across all four metrics. Compared with the strongest baseline *cd-chat*, GenSRL improves accuracy from 69.2% to 70.1% (+0.9%) and F1 from 46.8% to 49.8% (+3%, roughly +6.4% relative). The gains over earlier CNN and ViT-based methods, such as VGG-16, GoogLeNet, ResNet-152, and FDDM, are even larger, highlighting that generative spatiotemporal modeling yields more discriminative SRF-absorption representations and better cross-dataset generalization.

Table 1. Performance comparison on the public OCT4DME dataset for the binary SFC classification task. GenSRL surpasses all CNN-based, transformer-based, and multimodal prognosis baselines.

Method	ACC	Precision	Recall	F1
VGG-16 [2]	0.563	0.265	0.303	0.280
GoogleNet [3]	0.619	0.438	0.381	0.404
ResNet-152 [4]	0.629	0.428	0.441	0.432
MM-MIL [5]	0.619	0.485	0.342	0.351
MSAN [6]	0.616	0.379	0.373	0.374
FDDM [7]	0.670	0.450	0.406	0.423
CD-Chat [8]	0.692	0.522	0.445	0.468
GenSRL (Ours)	**0.701**	**0.551**	**0.454**	**0.498**

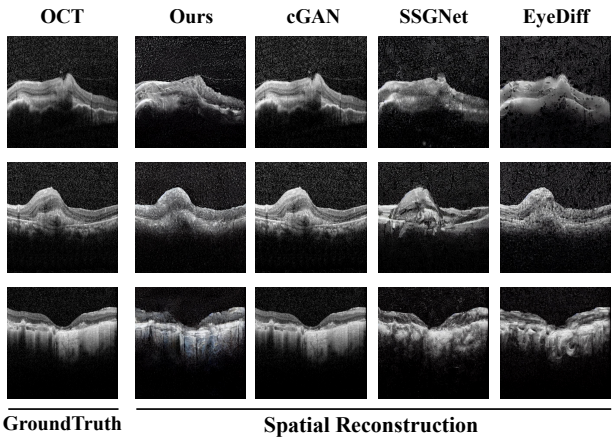

Figure 1. Qualitative visualization of Stage 1 spatial reconstruction on representative macular OCT images. Columns correspond to the input OCT, GenSRL (Ours), cGAN [9] , SSGNet [10] , and EyeDiff [11] , respectively, while rows show three different cases. All images are displayed using the same grayscale window.

2. Qualitative Analysis in Stage 1

We have illustrated the qualitative results of Stage 2 and demonstrated its effectiveness in the main paper. To investigate how the Stage 1 reconstruction model behaves on challenging macular scans, we compare GenSRL with cGAN, SSGNet, and EyeDiff in Fig. 1. Given a single pre-treatment OCT image as input, all methods are trained to reconstruct the spatial appearance of the macula. Across the three representative cases, GenSRL better preserves the global foveal contour and the relative thickness of inner retinal layers, while also recovering sharper outer retinal bands and more stable retinal pigment epithelium (RPE) profiles. The reconstructed fluid cavities remain close to the ground-truth in both shape and location, avoiding the excessive smoothing or local distortions observed in the baseline models.

In contrast, cGAN often over-smooths high-curvature regions and partially erases thin layers near the outer retina; SSGNet tends to introduce fragmented or wavy layer interfaces; and EyeDiff occasionally produces low-frequency shading artifacts that blur subtle anatomical details.

Overall, the Stage 1 reconstructions produced by GenSRL remain more faithful to the ground-truth targets, indicating that the model has learned a structure-aware prior over macular anatomy that can be exploited by the subsequent temporal generation and prognosis stages.

3. OCT Data Processing

3.1. Image Preparation

All OCT images are converted to single channel grayscale, and we normalize them to the range $[0, 1]$ using per scan z score standardization followed by rescaling. Each image is then resized to 224×224 for the ViT and LLaMA backbone and to 512×512 for the Stable Diffusion UNet decoder. For paired pre-treatment and post-treatment scans, pixel wise alignment is preserved and only synchronized geometric and intensity transformations are applied.

3.2. Data Augmentation Strategy

To improve robustness while preserving retinal anatomical integrity, we adopt augmentation pipelines that are specific to each stage:

- **Stage 1: Spatial Reconstruction** We apply intensity normalization followed by contrast and geometric augmentations (CLAHE, speckle style noise, anisotropic blur, horizontal flip, and mild horizontal elastic deformation [12]) on the pre-treatment scan.
- **Stage 2: Temporal Generation** We apply synchronized geometric transforms and mild intensity augmentations to both baseline and follow up scans, including occasional speckle style noise and anisotropic blur [13], shared brightness/contrast/gamma adjustments [14], and a slight contrast tweak on the post-treatment scan to simulate appearance shifts.
- **Stage 3: Clinical Prognosis Classification** We reuse synchronized geometric augmentations for paired OCT scans, keep only infrequent blur with low intensity, and disable perturbations based on noise to preserve label consistency.

For clarity, the augmentation probabilities and parameter ranges for each stage are listed in Table 2.

4. Stagewise Prompt Templates

GenSRL uses prompt templates that are specific to each stage, in an instruction format implemented in the dataset loader to interface multimodal features with the LLaMA backbone. For each sample, we construct

$$\begin{aligned}\text{Prompt} = {} & \text{[Img] \{visual tokens\} [/Img]} \\ & \text{text body \{learnable prompt tokens\},}\end{aligned} \tag{1}$$

where the visual tokens encode one or two OCT image streams depending on the training stage, and the learnable prompt tokens provide additional, trainable conditioning in the text space. We allocate 77 learnable prompt tokens per

Table 2. Intensity and geometric augmentations for longitudinal OCT images training. For each operator we list the core parameter settings and, when applicable, the application probabilities for each stage.

Operator	Description	Stage(s)
Intensity normalization	Per scan z score normalization followed by rescaling to the range $[0, 1]$.	S1 / S2 / S3
CLAHE	Contrast limited adaptive histogram equalization (clip limit = 2.0, tile grid = 8×8); applied with probability $p \approx 0.7$ on training scans.	S1 / S2 / S3
Speckle noise	Multiplicative speckle noise with standard deviation 0.02; probability $p = 0.3$ in Stage 1 and $p = 0.1$ in Stage 2.	S1 / S2
Anisotropic blur	Gaussian blur with 3×1 or 1×3 kernel, $\sigma = 0.5$; applied with probability $p = 0.1$ during training (slightly lower in Stage 3 to avoid oversmoothing fine prognostic cues).	S1 / S2 / S3
Horizontal flip	Synchronized horizontal flipping of paired scans with probability $p = 0.5$.	S1 / S2 / S3
Elastic deformation	Horizontal elastic deformation with amplitude 1.5 px and frequency 0.02; probability $p = 0.25$ in Stage 1 and $p = 0.3$ in Stage 2/Stage 3.	S1 / S2 / S3
Brightness/Contrast/Gamma	Shared brightness, contrast, and gamma scaling factors sampled from $[0.95, 1.05]$; probability $p = 0.3$.	S2 only
Post-treatment contrast tweak	Additional contrast scaling factor sampled from $[0.975, 1.025]$ on the post-treatment scan; probability $p = 0.2$.	S2 only

Table 3. Instruction style prompt templates for the three training stages of GenSRL, covering baseline reconstruction (Stage 1), temporal generation (Stage 2), and joint SFC and VR classification (Stage 3). image_feature_pre and image_feature_post are visual tokens from pre-treatment and post-treatment OCT scans. {learnable prompt} denotes a group of 77 learnable prompt tokens; multiple instances (for example, two in Stage 3) correspond to 2×77 tokens in total.

Training stage	Prompt template
Stage 1	[Img] image_feature_pre [/Img][MRG] Retinal OCT scan before treatment, baseline image, predominantly showing subretinal fluid (SRF) in the macular region, medical grayscale of high quality. {learnable prompt}
Stage 2	[Img] image_feature_post [/Img][MRG] Retinal OCT scan after treatment, medical grayscale, predicted or conditioned on the baseline OCT; the description should reflect SRF absorption or persistence and the patient's visual outcome. SRF status: . . . ; visual acuity outcome: . . . {learnable prompt} {learnable prompt}
Stage 3	[Img] image_feature_pre image_feature_post [/Img][MRG] Given paired retinal OCT scans before and after treatment, classify subretinal fluid (SRF) absorption and visual acuity improvement, and generate a structured medical description. {learnable prompt} {learnable prompt}

group, matching the context length of 77 tokens in the CLIP text encoder used in Stable Diffusion v1.5.

Concretely, Stage 1 focuses on spatial reconstruction using a single pre-treatment visual token and an OCT description that is oriented to the baseline scan. Stage 2 conditions on a post-treatment visual token and appends a label aware suffix that explicitly encodes the SFC and VR outcomes in natural language. Stage 3 takes paired pre-treatment and post-treatment visual tokens and uses an instruction that formulates prognosis prediction as a classification task. The corresponding prompt templates for each stage are summarized in Table 3. The learnable prompt tokens are optimized jointly with the LLaMA backbone under a training schedule with multiple stages.

5. Training Configuration

The unified training and optimization hyperparameters used across all three stages, together with the remaining data loading, optimizer, and model adaptation settings such as LLaMA quantization and SD-UNet LoRA configuration, are summarized in Table 4. Across all stages, we adopt a per-GPU batch size of 3 with 4 gradient accumulation steps, yielding an effective batch size of 12, and train each stage for 30 epochs with a base learning rate of 1×10^{-4} and a weight decay of 0.01. We follow the three-stage training schedule described in the main paper.

Table 4. Unified training and optimization hyperparameters shared across all three stages of GenSRL.

Config	**Value**
Training / val / test batch size	3 / 3 / 3
Gradient accumulation steps	4
Epochs per stage	30
Context length	1024 tokens
Learnable prompt tokens	77 per image stream
Base learning rate	1×10^{-4}
Weight decay	0.01
LR decay factor γ	0.95
Warmup epochs	2
Gradient clipping threshold	0.25
LLaMA quantization	4-bit
Mixed precision	enabled
UNet LoRA rank / α / dropout	32 / 64 / 0.05
UNet LoRA learning rate	2e-5
LLaMA LoRA learning rate	1e-5
DDIM sampling steps (stage 2)	50
CFG scale	1.0
Random seed	42

References

[1] W. Zhang, P. Chotcomwongse, Y. Li, P. Xu, R. Yao, L. Zhou, Y. Zhou, H. Feng, Q. Zhou, X. Wang, *et al.*, "Predicting diabetic macular edema treatment responses using oct: Dataset and methods of aptos competition," *arXiv preprint arXiv:2505.05768*, 2025. 1

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 1

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015. 1

[4] S. Wu, S. Zhong, and Y. Liu, "Deep residual learning for image recognition," *Multimed. Tools Appl*, pp. 1–17, 2017. 1

[5] X. Li, Y. Zhou, J. Wang, H. Lin, J. Zhao, D. Ding, W. Yu, and Y. Chen, "Multi-modal multi-instance learning for retinal disease recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2474–2482, 2021. 1

[6] X. He, Y. Deng, L. Fang, and Q. Peng, "Multi-modal retinal image classification with modality-specific attention network," *IEEE transactions on medical imaging*, vol. 40, no. 6, pp. 1591–1602, 2021. 1

[7] L. Wang, W. Dai, M. Jin, C. Ou, and X. Li, "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. 1

[8] M. Noman, N. Ahsan, M. Naseer, H. Cholakkal, R. M. Anwer, S. Khan, and F. S. Khan, "Cdchat: A large multimodal model for remote sensing change description," *arXiv preprint arXiv:2409.16261*, 2024. 1

[9] O. N. Hassan, S. Sahin, V. Mohammadzadeh, X. Yang, N. Amini, A. Mylavarapu, J. Martinyan, T. Hong, G. Mahmoudinezhad, D. Rueckert, *et al.*, "Conditional gan for prediction of glaucoma progression with macular optical coherence tomography," in *International Symposium on Visual Computing*, pp. 761–772, Springer, 2020. 1

[10] X. Zhao, X. Zhang, B. Lv, L. Meng, C. Zhang, Y. Liu, C. Lv, G. Xie, and Y. Chen, "Optical coherence tomography-based short-term effect prediction of anti-vascular endothelial growth factor treatment in neovascular age-related macular degeneration using sensitive structure guided network," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 259, no. 11, pp. 3261–3269, 2021. 1

[11] R. Chen, W. Zhang, B. Liu, X. Chen, P. Xu, S. Liu, M. He, and D. Shi, "Eyediff: text-to-image diffusion model improves rare eye disease diagnosis," *arXiv preprint arXiv:2411.10004*, 2024. 1

[12] D. Bar-David, L. Bar-David, Y. Shapira, R. Leibu, D. Dori, A. Gebara, R. Schneor, A. Fischer, and S. Soudry, "Elastic deformation of optical coherence tomography images of diabetic macular edema for deep-learning models training: How far to go?," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 11, pp. 487–494, 2023. 2

[13] K. J. Halupka, B. J. Antony, M. H. Lee, K. A. Lucy, R. S. Rai, H. Ishikawa, G. Wollstein, J. S. Schuman, and R. Garnavi, "Retinal optical coherence tomography image enhancement via deep learning," *Biomedical optics express*, vol. 9, no. 12, pp. 6205–6221, 2018. 2

[14] S. Apostolopoulos, J. Salas, J. L. Ordóñez, S. S. Tan, C. Ciller, A. Ebneter, M. Zinkernagel, R. Sznitman, S. Wolf, S. De Zanet, *et al.*, "Automatically enhanced oct scans of the retina: a proof of concept study," *Scientific reports*, vol. 10, no. 1, p. 7819, 2020. 2