SIGMA: SPATIOTEMPORAL INTERVENTIONAL GENERATION VIA MECHANISM-AWARE UNIFIED MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical management of blinding fundus diseases is a long-term process involving continuous follow-up and dynamic intervention. Beyond assessing the current pathological status, clinicians also need to evaluate how lesions may evolve over a future time horizon under a given treatment strategy. Existing medical vision AI is mainly organized around two task formulations: *static perception*, which focuses on the current pathological state in single-time-point images, and *medical image editing*, which focuses on controlled image modification under given target attributes or editing constraints. These two paradigms address current-state analysis and target-appearance control, respectively, but neither explicitly models the relationship among treatment strategy, time horizon, and pathological evolution, making them unsuitable for future image forecasting under intervention constraints. To address this gap, we introduce **Spatiotemporal Interventional Generation (SIG)** Task. Given a baseline image, intervention variables, and a time horizon, SIG requires the model to jointly predict the future image and its structured clinical rationale. For this task, we propose **SIGMA**, **a tool-augmented unified generation-understanding model**. SIGMA first extracts discrete biomarker representations from the baseline image and uses them to filter an external memory, where candidates are ranked by image similarity. The retrieved trajectories, intervention, and time horizon then condition unified forecasting. Under such priors, the unified model jointly generates the clinical rationale and the future image. Experiments reveal the challenges that SIG poses to existing models and demonstrate the effectiveness of SIGMA after task-specific fine-tuning.

1 INTRODUCTION

Clinical management of vision-threatening fundus diseases is inherently a longitudinal process that requires continuous monitoring and adaptive intervention van Timmeren et al. (2025). Beyond assessing the current pathological state, clinicians must anticipate the future progression of lesions under specified treatment strategies. Therefore, enabling reliable forecasting of future states from a baseline image, an intervention plan, and a time horizon is an important problem for bringing medical vision intelligence closer to real-world clinical applications.

Medical vision AI has achieved substantial progress in ophthalmic image analysis Lan et al. (2025); Abdi & Abdulazeez (2025); Morano et al. (2025); Wang et al. (2024); Tong et al. (2025). Existing task formulations can be broadly grouped into two categories. The first category is **static perception**, as is shown in Fig. 1 B.1, including disease classification Kulyabin et al. (2024); Nazlı et al. (2024), medical segmentation Li et al. (2024); Qiu et al. (2023), and VLM-based QA Xu et al. (2025a); Peng et al. (2025), whose core objective is to characterize the pathological information in the current image. The second category is **medical image editing**, as is shown in Fig. 1 B.2, whose core focus is to generate a target image under given control signals, target attributes, or editing constraints Huang et al. (2025); Konz et al. (2024); Zhan et al. (2024). From the perspective of task formulation, the former does not model how disease states evolve over time, while the latter usually relies on explicit appearance targets or editing instructions rather than requiring the model to infer how pathology should change under treatment and time conditions. Consequently, neither category directly supports future image forecasting under intervention constraints.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

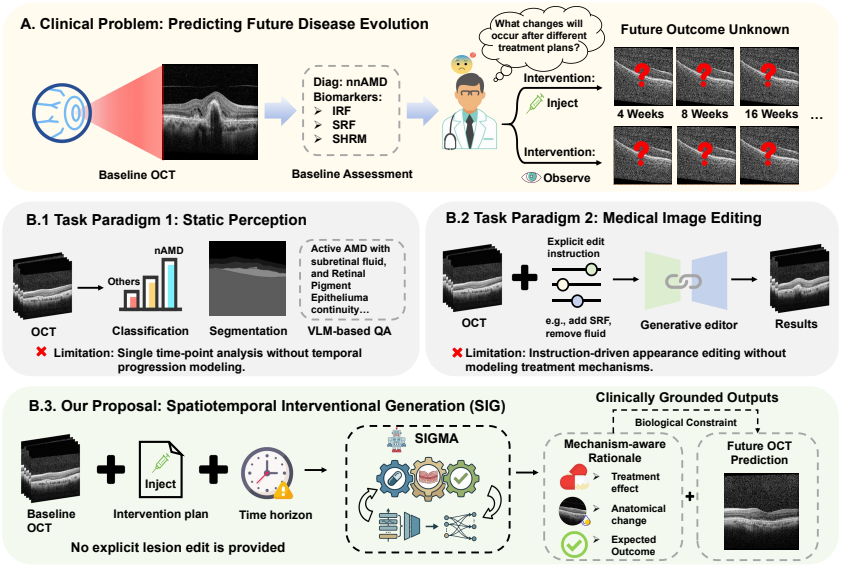


Figure 1: Clinical motivation and task formulation of Spatiotemporal Interventional Generation (SIG). **(A)** In clinical practice, ophthalmic decision-making requires not only accurate assessment of the current OCT image, but also forecasting future disease evolution under different intervention plans and time horizons. **(B.1)** Existing *static perception* tasks focus on single-time-point analysis, such as classification, segmentation, and VLM-based QA, without modeling longitudinal progression. **(B.2)** Existing *medical image editing* tasks typically rely on explicit edit instructions or target appearance constraints, and are therefore instruction-driven rather than treatment-aware. **(B.3)** We propose *Spatiotemporal Interventional Generation*, which takes a baseline OCT image, an intervention plan, and a time condition as input, and predicts a structured clinical rationale and the future OCT image.

As illustrated in Fig. 1 A, real clinical decision-making relies not only on accurate interpretation of the current OCT image, but also on answering a further question: given an intervention plan and a time horizon Qiao et al. (2025), how will the pathological state evolve in the future? Motivated by this observation, we propose a new task, **Spatiotemporal Interventional Generation (SIG)**. Given a baseline image X_t, an intervention condition I, and a time horizon ΔT, SIG requires the model to predict both the future image $\hat{X}_{t+\Delta T}$ and its structured clinical rationale. Unlike static perception, SIG requires modeling the conditional transition of pathology over time under intervention. Unlike medical image editing, the conditioning signal in SIG is not an appearance-level editing constraint, but a treatment strategy and temporal information. The model infers a plausible pathological progression from the current state and generates a consistent future image.

To support supervised learning for SIG, we construct **Ophtha-SIG**, which provides biomarker-consistent longitudinal supervision. We further propose **SIGMA**, a mechanism-aware unified generation-understanding model for intervention-conditioned future forecasting. SIGMA combines biomarker perception, retrieval-augmented priors, and unified generation to jointly produce the clinical rationale and future prediction Xu et al. (2025c). Specifically, SIGMA employs a domain-adapted **Perception Probe** to extract biomarker anchors from the baseline OCT, filters historical trajectories accordingly, and ranks the remaining candidates by image similarity. The retrieved evidence, intervention, and time horizon then condition forecasting. SIGMA adopts a unified generation–understanding formulation to jointly predict the future OCT image and the structured clinical rationale under shared state representations and retrieved evolution evidence, thereby improving the consistency between image prediction and clinical explanation, as shown in Fig. 1 B.3.

Experiments reveal the challenges that SIG poses to existing models and demonstrate the effectiveness of SIGMA after task-specific fine-tuning. Moreover, SIGMA exhibits strong disease diagnosis capability, indicating unified future forecasting does not come at the expense of medical understanding.

Our main contributions are summarized as follows:

- We introduce a new task formulation, **Spatiotemporal Interventional Generation (SIG)**, extending medical image modeling from single-time-point recognition to intervention-conditioned future forecasting.
- We propose **SIGMA**, a mechanism-aware unified generation-understanding model for SIG task, which jointly predicts the future OCT image and its structured clinical rationale under shared state representations and retrieved evolution evidence.
- We establish an initial benchmark for SIG, evaluating existing models and SIGMA across image fidelity, rationale quality, cross-modal consistency, and disease diagnosis.

2 Related Work

2.1 Static Medical Perception Tasks

Static medical perception aims to characterize current pathological states from single-time-point images. Early structure-oriented architectures, such as U-Net variants Ronneberger et al. (2015); Zhou et al. (2018), established strong foundations for precise lesion localization and medical segmentation. Subsequently, task-specific multimodal frameworks advanced capabilities in complex semantic tasks, including medical report generation Tang et al. (2025) and visual question answering (VQA) Yu et al. (2025); Liu et al. (2025); Gao et al. (2026). Most recently, the advent of multimodal medical large language models Fang et al. (2025), such as Hulu-Med Jiang et al. (2025) and Baichuan Team et al. (2025), has further pushed the boundaries of cross-modal alignment, enabling comprehensive semantic understanding of static clinical presentations.

Despite these substantial advances in current-state representation, existing paradigms remain fundamentally constrained to static, single-time-point prediction. They focus on identifying the current pathological status, rather than forecasting how the state will evolve under specific medical interventions and temporal dynamics. While these static perception models provide robust baseline representations for our problem, their inability to model intervention-conditioned temporal dynamics and continuous temporal trajectories prevents them from directly supporting future image forecasting.

2.2 Medical Image Editing Tasks

The task of medical image editing defines image modification as a controllable synthesis problem. Representative models range from early generative adversarial networks (e.g., MedGAN) for attribute manipulation to high-fidelity diffusion models (e.g., RoentGen, MedEdit) that enable precise local edits via explicit text prompts or masks. In this formulation, the input typically consists of a source image paired with these explicit control signals—such as editing intents or target appearances—with the objective of generating an image that strictly satisfies the constraints. Recent extensions of this task paradigm include counterfactual visualization and unified multimodal generation, as seen in models like ProgEmu Ma et al. (2025), UniMedVL Ning et al. (2025), and HealthGPT Lin et al. (2025). Although editing tasks inherently model visual changes, they are fundamentally instruction-following problems driven by predefined visual cues rather than longitudinal disease dynamics. In contrast, our proposed SIG task formulates a temporal, treatment-conditioned prediction problem. Instead of accepting explicit instructions on how lesions should change, the SIG task requires utilizing only the baseline state, intervention condition, and time horizon to predict the future imaging state alongside a structured clinical rationale.

3 Method

3.1 Problem Formulation

We study future ophthalmic image forecasting under intervention and temporal constraints, and accordingly construct **Ophtha-SIG**. Formally, the dataset is defined as:

$$\mathcal{D} = \{(X_t^{(n)}, B_t^{(n)}, I^{(n)}, \Delta T^{(n)}, M_{\text{patient}}^{(n)}, X_{t+\Delta T}^{(n)}, B_{t+\Delta T}^{(n)}, r^{(n)})\}_{n=1}^{N} \tag{1}$$

where X_t denotes the baseline OCT at time t, $X_{t+\Delta T}$ denotes the follow-up OCT, B_t and $B_{t+\Delta T}$ denote the biomarker states, I denotes the intervention condition, M_{patient} denotes patient metadata available at baseline, and r denotes the structured clinical rationale.

Based on this dataset, we define Spatiotemporal Interventional Generation (SIG). Given a baseline OCT X_t, an intervention condition I, and a time horizon ΔT, SIG predicts the future OCT and its clinical rationale, with M_{patient} provided as auxiliary context:

$$(\hat{r}, \hat{X}_{t+\Delta T}) = F_{\text{sig}}(X_t, I, \Delta T; M_{\text{patient}}). \tag{2}$$

Unlike static medical perception tasks, SIG models a *conditioned disease evolution trajectory* based on intervention and time. As shown in Fig. 2, the task requires the model to jointly predict the future OCT image and the clinical rationale behind its evolution.

3.2 Ophtha-SIG Data Construction

To facilitate supervised learning for SIG, we construct the **Ophtha-SIG dataset**. Given a longitudinal sample $(X_t, X_{t+\Delta T}, I, \Delta T)$ and associated baseline patient metadata M_{patient}, we obtain physician-confirmed biomarker states at both time points, the recorded intervention, and the overall progression trend. A Vision-Language Model then generates a candidate clinical rationale and a candidate trend label. We apply two consistency checks: the local changes described in the rationale must agree with the physician-confirmed biomarker transition, and the candidate trend must match the physician-confirmed overall trend. Experienced ophthalmologists additionally inspect a random subset of the retained clinical rationales. This process combines clinically grounded structured labels with consistency-filtered textual supervision.

The final curated tuples (baseline image, baseline patient metadata, future image, intervention condition, time horizon, biomarker transition, and rationale) serve as supervisory signals for future image prediction and rationale generation. Detailed mathematical formulations and alignment rules are provided in Appendix A. All procedures in this study adhere to the principles of the Declaration of Helsinki and were approved by the relevant Institutional Review Board (IRB).

3.3 SIGMA

To solve SIG, we propose **SIGMA**, a mechanism-aware unified generation-understanding framework. Given a baseline OCT, an intervention condition, and a time horizon, SIGMA enforces mechanism-awareness by equipping the unified network with two specialized tools: (i) a perception probe for state abstraction, and (ii) a trajectory retrieval module for evolution-prior retrieval. Conditioned on the extracted structured states and retrieved historical priors, the model generates the future OCT together with its structured clinical rationale. The overall framework is shown in Fig. 2 B.

3.3.1 Agentic Tool 1: Domain-Adapted Perception Probe

As is shown in Fig. 2 C.1, the first tool in SIGMA is a **domain-adapted perception probe** Φ to extract discrete biomarker anchors from the baseline OCT:

$$\hat{B}_t = \Phi(X_t) = \{b_1, b_2, b_3, b_4\}. \tag{3}$$

where $b_i \in \{0, 1\}$ serves as a binary indicator for the presence or absence of a specific biomarker. The perception probe uses DINOv3 Siméoni et al. (2025) as a visual backbone, continually self-supervised pretrained on 50,000 unlabeled OCT images excluding Ophtha-SIG test images and forecasting targets. We independently train four biomarker-specific binary classification heads, one for each biomarker, to obtain discrete state variables that characterize the current pathological state and guide subsequent retrieval and prediction.

3.3.2 Agentic Tool 2: Biomarker-Guided Trajectory Retrieval

Following the perception probe, SIGMA leverages the extracted anchors $\hat{B}_t$ to retrieve relevant disease-evolution trajectories from a historical memory bank Guu et al. (2020); Lewis et al. (2020). To reduce dependence on discrete predictions, we implement a coarse-to-fine retrieval mechanism. The model first performs coarse-grained pruning within a Global Case Database H_{global}, constructed exclusively from training patients. During training, leave-one-patient-out retrieval excludes all visits from both eyes of the current patient to prevent data leakage. By using $\hat{B}_t$ as a soft semantic prior, this stage identifies a narrowed Biomarker-Anchored Database H_{pruned}.

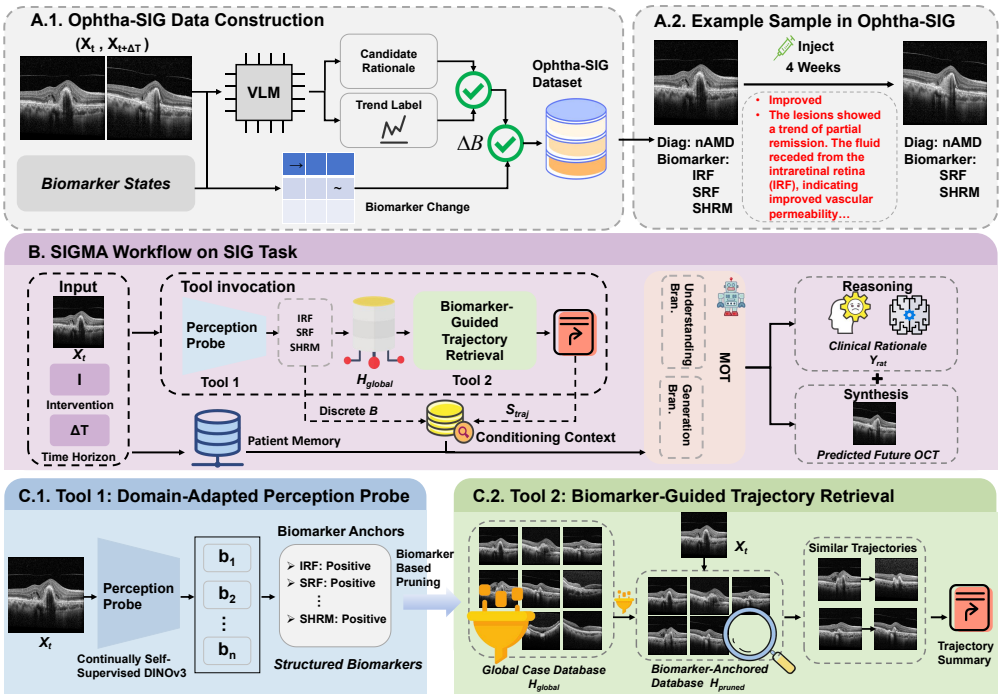


Figure 2: Overview of the Ophtha-SIG construction and the SIGMA framework. **(A)** Construction pipeline. Longitudinal OCT pairs are processed by a Vision-Language Model to generate candidate rationales, which undergo a strict consistency check against objective biomarker changes (ΔB) to yield high-quality training tuples (A.1). A representative sample is shown in A.2. **(B)** The inference workflow of SIGMA on the Spatiotemporal Interventional Generation (SIG) task. Given a baseline OCT (X_t), an intervention condition, and a time horizon (ΔT), the model invokes two specialized tools to build a conditioning context. Guided by this context, the dual-branch network (Understanding and Generation) jointly synthesizes the structured clinical rationale and the predicted future OCT. **(C.1)** Tool 1: A Domain-Adapted Perception Probe, built upon continually self-supervised DINOv3, extracts discrete biomarker anchors to characterize the baseline pathological state. **(C.2)** Tool 2: A Biomarker-Guided Trajectory Retrieval module. It first prunes the historical case database based on the extracted biomarkers, and subsequently retrieves visually similar trajectories to provide empirical evolution priors. Here, "Patient Memory" denotes the baseline patient metadata M_{patient}.

Subsequently, fine-grained retrieval is conducted within $\mathcal{H}_{\text{pruned}}$ by calculating the cosine similarity between the current image X_t and historical candidates in the dense feature space of a domain-adapted DINOv3*. This dual-stage design helps mitigate potential cascading errors originating from initial classification noise. Finally, the evolution priors of the retrieved matches are aggregated into a textual summary S_{traj} to provide longitudinal context. The details are shown in Appendix C.

3.3.3 Unified Forecasting

Prior to the generation phase, SIGMA aggregates the structured state representations and the retrieved historical priors. By incorporating the baseline patient metadata $M_{patient}$, the complete conditioning context is formulated as:

$$C = \text{Concat}(\hat{B}_t, M_{patient}, S_{traj}) \tag{4}$$

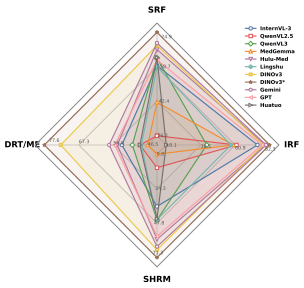
Given the baseline image X_t and the conditioning context C, SIGMA models future forecasting as a joint conditional generation process over rationale and future OCT:

$$p_\theta(r, X_{t+\Delta T} \mid X_t, I, \Delta T, C), \tag{5}$$

where r denotes the clinical rationale for the future pathological transition, and $X_{t+\Delta T}$ denotes the future OCT.

Table 1: Evaluation of fine-grained biomarker perception capabilities across general VLMs, medical VLMs, closed-source commercial VLMs, and vision foundation models. This table reports the average performance across the four selected biomarkers (IRF, SRF, SHRM, and DRT/ME), while the adjacent radar chart details the individual performance for each specific biomarker. Evaluation metrics include Average Accuracy, Average Recall, Average Jaccard Index, Hamming Loss, and Hit Rate. Arrows indicate whether higher or lower values are better. Bold values indicate the best performance in each metric. A dash indicates undisclosed parameter counts. The symbol * denotes the continually self-supervised domain-adapted version of the baseline model.

Models	Params	Avg. Acc ↑	Avg. Rec ↑	Avg. Jacc ↑	Hamm ↓	Hit Rate ↑
Closed-source Commercial VLMs						
GPT-5.5	—	64.47	65.74	45.72	35.53	80.57
Gemini-3.5-Flash	—	68.53	74.90	52.30	31.47	83.53
General VLMs						
InternVL3 Chen et al. (2024b)	78B	62.22	54.43	42.99	37.78	77.35
Qwen2.5-VL Bai et al. (2025b)	72B	57.80	34.80	29.70	42.20	53.15
Qwen3-VL Bai et al. (2025a)	32B	53.82	47.76	33.35	46.18	76.32
Medical VLMs						
MedGemma Sellergren et al. (2025)	27B	59.11	32.99	22.11	40.69	42.21
Hulu-Med Jiang et al. (2025)	32B	65.01	75.40	51.01	34.99	88.55
Lingshu Xu et al. (2025b)	32B	58.49	44.30	29.37	41.51	57.92
HuatuoGPT-Vision Chen et al. (2024a)	34B	49.76	42.59	29.94	50.24	73.49
Vision Foundation Models						
DinoV3 Siméoni et al. (2025)	1.13B	72.85	78.71	58.69	27.15	90.55
DinoV3* Siméoni et al. (2025)	1.13B	**77.98**	**83.32**	**65.14**	**22.02**	**94.00**



3.3.4 Training and Inference

Our model employs a Mixture-of-Transformer-Experts (MoT) architecture, initialized with pre-trained BAGEL weights Deng et al. (2025) and fine-tuned on the Ophtha-SIG dataset. During training, the model takes the baseline OCT X_t, intervention I, time horizon ΔT, and context C as input, optimized via the joint conditional likelihood:

$$\mathcal{L}_{SIG} = -\mathbb{E}_{(X_t, I, \Delta T, C, r, X_{t+\Delta T}) \sim \mathcal{D}} \left[\log p_\theta(r, X_{t+\Delta T} \mid X_t, I, \Delta T, C) \right] \tag{6}$$

For inference, the context C is constructed by concatenating extracted baseline features $\hat{B}_t$, baseline patient metadata M_{patient}, and similar historical trajectories S_{traj} retrieved from a global repository. The model then conditionally predicts the response and future state: $(\hat{r}, \hat{X}_{t+\Delta T}) \sim p_\theta(\cdot \mid X_t, I, \Delta T, C)$. The complete inference procedure is detailed in Appendix D.

4 Experiments

We evaluate the SIGMA framework to answer three questions: (1) Does the domain-adapted perception probe mitigate feature space misalignment in structured state abstraction? (2) How does SIGMA perform on the proposed Spatiotemporal Interventional Generation (SIG) task regarding rationale reasoning, visual fidelity, and cross-modal semantic grounding? (3) Can the unified generation-understanding paradigm maintain discriminative capabilities without significant performance trade-offs? Detailed metric formulations and implementations are in Appendix B.

4.1 Pre-experiment: Biomarker Perception

Fine-grained biomarker extraction is fundamental for ophthalmic image understanding. Although Large Multimodal Models (LMMs) exhibit strong general capabilities, their limited fine-grained perception in specialized medical domains motivates the design of our domain-adapted perception probe. Table 1 evaluates its biomarker recognition performance against various foundational models. General VLMs like InternVL3 (78B) struggle with detailed ophthalmic features (42.99% Average Jaccard Index), indicating that parameter scaling alone inadequately addresses domain shifts. Medical VLMs such as Hulu-Med reduce this gap but still face anatomical challenges (51.01% Jaccard Index). Conversely, our continually self-supervised DINOv3* achieves an Average Recall of 83.32%. This improvement over vanilla DINOv3 demonstrates that domain-specific adaptation effectively aligns visual features with clinical biomarkers, empirically validating our module design.

Table 2: Main quantitative results on the Spatiotemporal Interventional Generation (SIG) task. We evaluate textual rationale quality (ROUGE-L: lexical overlap; BERTScore: token-level similarity; SBERT-Cos: sentence-level similarity), visual fidelity (LPIPS: perceptual distance; DINOv3-FID/KID: feature distribution discrepancy), and cross-modal consistency (CRVA: alignment between the future OCT $\hat{X}_{t+\Delta T}$ and its clinical rationale $\hat{r}$). ↓ indicates lower is better and ↑ indicates higher is better. Best results are shown in bold, and "–" denotes unavailable results.

Model	Text			Visual Fidelity			Cross-modal Consistency
	ROUGE-L ↑	BERTScore ↑	SBERT-Cos ↑	LPIPS ↓	DINOv3-FID ↓	DINOv3-KID ↓	CRVA ↑
Closed-source Commercial Models							
GPT-Image-2	–	–	–	0.4308	1.98	2.10×10^{-3}	–
Gemini-3.1-Flash-Image	0.1431	0.6856	0.6951	**0.4255**	2.42	2.57×10^{-3}	3.06
w/o reasoning							
Lumina-DiMOO Xin et al. (2025)	–	–	–	0.5038	8.92	13.66×10^{-3}	–
Qwen-Image-Edit Wu et al. (2025a)	–	–	–	0.7823	28.12	60.93×10^{-3}	–
OmniGen2 Wu et al. (2025b)	–	–	–	0.9502	32.08	61.19×10^{-3}	–
w/ reasoning							
InternVL-U Tian et al. (2026)	0.1745	0.6905	0.6882	0.8200	28.61	46.06×10^{-3}	2.19
UniMedVL Ning et al. (2025)	0.0850	0.5684	0.5190	0.7458	12.07	18.38×10^{-3}	2.29
UniCoT Qin et al. (2025)	0.1482	0.6712	0.5044	0.6567	6.20	6.13×10^{-3}	2.16
Bagel Deng et al. (2025)	0.1333	0.6598	0.4116	0.5652	4.25	3.86×10^{-3}	1.99
SIGMA (Ours)	**0.7942**	**0.7978**	**0.8948**	0.4603	**1.48**	$\mathbf{1.07 \times 10^{-3}}$	**3.82**

4.2 Evaluation on the SIG Task

We evaluate the proposed SIG task across textual reasoning, structural visual fidelity, and cross-modal semantic grounding. Table 2 summarizes the quantitative results.

Existing instruction-driven editors and untuned unified multimodal models struggle with SIG, particularly in generating clinically grounded rationales and intervention-consistent future images. After task-specific fine-tuning on Ophtha-SIG, SIGMA adapts effectively to the task, achieving strong rationale quality, competitive visual fidelity, and improved cross-modal consistency, with a BERTScore of 0.7978 and a CRVA score of 3.82.

Cross-modal semantic grounding, which measures whether anatomical shifts align with predicted clinical rationales, is evaluated using the CRVA metric. SIGMA achieves a CRVA of 3.82, outperforming existing baselines. This indicates stronger consistency between the generated future OCT images and their accompanying clinical rationales.

We further assess the visual plausibility of generated OCT follow-up images through a blinded Turing test conducted by an ophthalmologist (Appendix E.1) and evaluate agreement between automated CRVA scores and expert ratings in an independent study involving two ophthalmologists (Appendix E.2).

Baseline Selection Justification: We evaluate unified MLLMs and instruction-driven editors to assess how existing models transfer to SIG. SIGMA is fine-tuned on Ophtha-SIG to establish a task-adapted reference.

4.3 Disease Diagnosis: Assessing Understanding Capabilities

Having established SIGMA's performance on the generative SIG task, we further investigate its fundamental discriminative reasoning capabilities. A common consideration in unified multimodal models is whether the optimization for generative tasks inadvertently degrades discriminative performance. To evaluate SIGMA's diagnostic integrity, we assess its current-state disease classification accuracy at both the slice and patient levels. In clinical practice, longitudinal treatment can introduce a diagnostic discrepancy: while a specific 2D cross-sectional image (slice-level) might exhibit morphological recovery following intervention, the patient's global medical record (patient-level) often persistently logs the underlying disease. Evaluating at both levels thus provides a more comprehensive assessment.

As detailed in Table 3, we evaluate the models under identical diagnostic supervision. The results indicate that SIGMA exhibits disease understanding capabilities after spatiotemporal generative fine-tuning. Specifically, SIGMA achieves an F1-score of 70.80% at the slice-level and 66.45% at the patient-level. Standard unified models lacking specialized visual priors tend to exhibit limited diagnostic capabilities on this dataset. In contrast, the integration of the domain-adapted perception

Table 3: Quantitative comparison of disease diagnosis performance at slice-level and patient-level. Performance is measured using Accuracy (Acc), Recall, and F1-score. Results with and without the domain-aware expert are shown. Best results are in **bold**.

Model	Params	Slice-level						Patient-level					
		w/o Domain Expert			w/ Domain Expert			w/o Domain Expert			w/ Domain Expert		
		Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1
Visual foundation models													
DINOv3 Siméoni et al. (2025)	1.13B	69.41	67.64	68.22	–	–	–	68.57	66.67	64.82	–	–	–
DINOv3* Siméoni et al. (2025)	1.13B	**73.25**	**71.48**	**72.06**	–	–	–	74.04	71.14	**70.02**	–	–	–
General-purpose MLLMs													
Qwen2.5-VL Bai et al. (2025b)	7B	32.86	32.19	31.59	51.88	50.92	37.98	40.22	38.91	31.41	54.13	52.14	31.92
Qwen3-VL Bai et al. (2025a)	8B	40.09	39.69	38.44	52.41	51.31	49.34	44.35	43.37	36.03	53.26	51.34	45.57
InternVL3.5 Wang et al. (2025)	8B	25.98	24.70	24.59	51.69	50.32	48.44	33.04	31.12	24.88	52.17	50.04	44.45
MiMo-VL-RL Xiaomi (2025)	7B	41.07	40.52	40.67	49.73	48.23	47.71	44.35	42.90	39.00	48.91	46.56	44.02
Medical MLLMs													
QoQ-Med-VL Dai et al. (2025)	7B	24.64	24.33	22.95	49.20	48.57	34.94	33.26	32.61	25.39	50.43	48.80	32.74
Lingshu Xu et al. (2025b)	7B	37.59	37.05	35.01	54.82	53.50	40.17	42.61	41.74	34.36	55.43	53.41	37.55
Hulu-Med Jiang et al. (2025)	7B	48.75	47.84	41.52	53.75	52.43	42.08	52.17	50.83	37.38	55.87	53.62	36.16
Hulu-Med Jiang et al. (2025)	14B	53.04	52.10	47.33	59.64	58.39	54.98	56.52	54.78	42.90	63.91	61.52	52.22
Hulu-Med Jiang et al. (2025)	32B	57.23	56.10	47.57	60.18	58.91	51.21	64.78	62.79	44.90	63.70	61.59	46.58
Unified generation-understanding models													
Bagel Deng et al. (2025)	14B	7.77	6.49	6.90	13.75	12.38	11.86	10.65	9.06	8.47	20.22	18.41	14.59
SIGMA (Ours)	14B	70.54	69.55	68.96	**73.93**	**72.78**	**70.80**	73.26	71.20	65.50	**75.43**	**73.08**	66.45

Table 4: Ablation study of SIGMA on the proposed SIG task. Tool 1 denotes the domain-adapted perception probe, and Tool 2 denotes the biomarker-guided trajectory retrieval module. We follow the same evaluation protocol as in the main results. Higher is better for ROUGE-L, BERTScore, and SBERT-Cos, while lower is better for LPIPS, DINOv3-FID, and DINOv3-KID.

Model Variant	**Configuration**		**Text**			**Visual Fidelity**		
	Tool 1	**Tool 2**	**ROUGE-L ↑**	**BERTScore ↑**	**SBERT-Cos ↑**	**LPIPS ↓**	**DINOv3-FID ↓**	**DINOv3-KID ↓**
Base SIGMA			0.4211	0.7917	0.8887	0.4618	1.59	1.20×10^{-3}
+ Tool 1 only	✓		0.8023	0.8019	0.8953	0.4611	1.53	1.14×10^{-3}
+ Tool 2 only		✓	0.3817	0.7635	0.8562	0.4584	1.42	1.02×10^{-3}
Full SIGMA	✓	✓	0.7942	0.7978	0.8948	0.4603	1.48	1.07×10^{-3}

expert into SIGMA provides beneficial structural priors that complement the unified architecture, aiding in balancing perception and generation tasks.

4.4 Ablation Study

To evaluate SIGMA's core components, we ablate **Tool 1** (domain-adapted perception probe for baseline state abstraction) and **Tool 2** (biomarker-guided trajectory retrieval for historical priors). Table 4 compares four variants following the main experimental protocol.

Results indicate complementary contributions: **Tool 1** primarily enhances textual quality, suggesting that explicit biomarker anchors help stabilize baseline modeling, whereas **Tool 2** improves visual distribution alignment, demonstrating that historical trajectories offer valuable forecasting priors. Notably, introducing **Tool 2** occasionally lowers automated text similarity scores compared to ground truths. We attribute this to the retrieved trajectories introducing diverse historical evidence, which prompts clinically plausible rationales that diverge from reference phrasing rather than indicating reduced medical validity. Ultimately, integrating both tools achieves balanced performance across visual fidelity, textual quality, and cross-modal consistency. Beyond optimizing automated metrics, this integration introduces traceable historical evidence, thereby enhancing the transparency and interpretability of the forecasting process.

4.5 Qualitative and Visual Analysis

Intervention and Temporal Dynamics. Fig. 3 presents the visual results of SIGMA under different intervention conditions and forecasting horizons. Given the same baseline OCT, the model directly

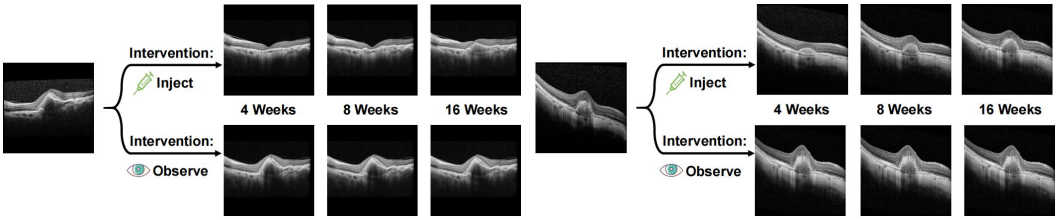


Figure 3: Time- and intervention-aware forecasting. For each baseline OCT image, SIGMA predicts future scans at 4, 8, and 16 weeks under two intervention conditions: Inject and Observe. All predictions are generated directly from the baseline image conditioned on the target time horizon, rather than in a cascaded or step-wise manner.

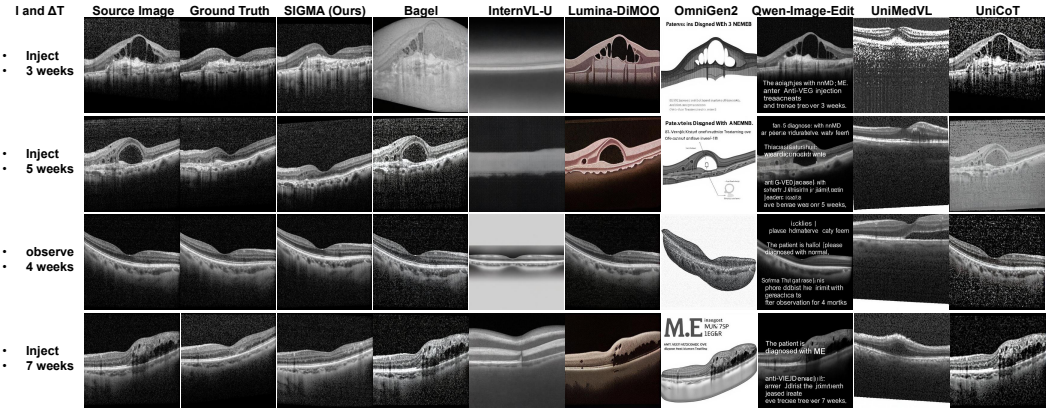


Figure 4: Qualitative comparison on the SIG task. We visualize the generated future OCTs alongside their corresponding Ground Truth to evaluate clinical fidelity. Compared to unified multimodal models (e.g., Bagel, InternVL-U) and instruction-driven image editors, SIGMA more accurately synthesizes specific pathological evolutions while preserving overall structural integrity.

generates future images at 4, 8, and 16 weeks under the two intervention settings. Predictions at different time points are all obtained directly from the baseline image conditioned on the corresponding time horizon, rather than being generated progressively from earlier predictions. These results show that SIGMA can incorporate both intervention cues and temporal conditions within a unified framework to produce future retinal structure predictions under different settings.

Comparison with Baselines. To comparatively evaluate generation quality, we select specific follow-up states with available Ground Truth for visualization in Fig. 4. Pure instruction-driven editors and unified multimodal models often produce structural distortions or superficial stylistic changes without accurately reflecting the pathological mechanism. SIGMA produces future OCT images that more closely match the observed follow-up scans in these examples. It accurately captures targeted pathological evolutions, such as the absorption of retinal fluid, demonstrating robust cross-modal semantic grounding.

5 Conclusion

This paper introduces the Spatiotemporal Interventional Generation (SIG) task, advancing medical vision AI from static perception toward proactive, intervention-conditioned forecasting. To address this, we propose SIGMA, a unified framework that utilizes discrete biomarker abstraction and trajectory retrieval to model pathological evolution under specified treatments and time horizons. Our results demonstrate that SIGMA effectively couples future image generation with structured clinical rationales. Furthermore, it successfully integrates predictive generative capabilities without compromising foundational diagnostic accuracy. Ultimately, this work provides a scalable approach for intervention-aware clinical AI and long-term patient management.

AI USE STATEMENT

Generative AI tools were used for limited literature-search assistance and language editing during manuscript preparation. Gemini-3.1 Pro was used to generate candidate longitudinal rationales and trend labels and to aggregate retrieved trajectories into textual summaries. A candidate rationale was included only when its described biomarker changes agreed with the physician-confirmed biomarker transition and its trend label matched the physician-confirmed overall trend. GPT-5.4 was used as an automated evaluator for the CRVA metric, and its scores were compared with independent ophthalmologist assessments on a separate sample. The authors reviewed the annotation and evaluation protocols, verified the reported results, and take full responsibility for the final manuscript.

ETHICS STATEMENT

Ophtha-SIG was constructed from retrospective ophthalmic OCT records under approval from the relevant institutional review board. All data were de-identified before analysis, and patient identifiers were removed or protected according to the approved protocol. Physician annotation and ophthalmologist evaluation were limited to image-content assessment and longitudinal disease interpretation. The generated images were not used for clinical decision-making or patient care. Because the source data contain clinical information, the raw OCT images cannot be publicly released in full. Data access and any release of derived materials will follow the applicable institutional and data-use requirements.

REPRODUCIBILITY STATEMENT

The paper and supplementary materials specify the patient-disjoint data split, dataset statistics, biomarker definitions, annotation and quality-control procedures, trajectory-retrieval algorithm, model architecture and initialization, inference pipeline, evaluation metrics, and the prompt used for CRVA scoring. The definitions of the reported metrics and the procedures used for human evaluation are also provided. Because the raw clinical OCT images and some third-party model services are subject to privacy, licensing, and access restrictions, complete reproduction may require access to an authorized copy of the data and the corresponding external services.

REFERENCES

Abdo Sulaiman Abdi and Adnan Mohsin Abdulazeez. A comprehensive review of deep learning in oct image segmentation and classification. *Medicine in Novel Technology and Devices*, pp. 100396, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans, 2021. URL https://arxiv.org/abs/1801.01401.

Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale, 2024a. URL https://arxiv.org/abs/2406.19280.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Wei Dai, Peilin Chen, Chanakya Ekbote, and Paul Pu Liang. Qoq-med: Building multimodal clinical foundation models with domain-aware grpo training. *arXiv preprint arXiv:2506.00711*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Mengjie Fang, Zipei Wang, Sitian Pan, Xin Feng, Yunpeng Zhao, Dongzhi Hou, Ling Wu, Xuebin Xie, Xu-Yao Zhang, Jie Tian, et al. Large models in medical imaging: Advances and prospects. *Chinese Medical Journal*, 138(14):1647–1664, 2025.

Shujian Gao, Yuan Wang, Jiangtao Yan, Zuxuan Wu, and Yu-Gang Jiang. Thinking with deltas: Incentivizing reinforcement learning via differential visual reasoning policy, 2026. URL https://arxiv.org/abs/2601.06801.

Peyman Gholami, Priyanka Roy, Mohana Kuppuswamy Parthasarathy, and Vasudevan Lakshminarayanan. Octid: Optical coherence tomography image database, 2019. URL https://arxiv.org/abs/1812.07056.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen. Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, Jin Hao, Zijian Chen, Ruijia Wu, Tao Tang, Junhui Lv, Hongxia Xu, Hongwei Wang, Jun Xiao, Bin Feng, Fudong Zhu, Kenli Li, Weidi Xie, Jimeng Sun, Jian Wu, and Zuozhu Liu. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding, 2025. URL https://arxiv.org/abs/2510.08668.

Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 88–98. Springer, 2024.

Mikhail Kulyabin, Aleksei Zhdanov, Anastasia Nikiforova, Andrey Stepichev, Anna Kuznetsova, Mikhail Ronkin, Vasilii Borisov, Alexander Bogachev, Sergey Korotkich, Paul A Constable, et al. Octdl: Optical coherence tomography dataset for image-based deep learning methods. *Scientific data*, 11(1):365, 2024.

Zibo Lan, Ying Hu, Shuang Yang, Meile Wu, Yuheng Huang, and He Zhang. Deep learning in ophthalmic image analysis: A task-driven review of segmentation, diagnosis, and progress prediction. *Neurocomputing*, pp. 132282, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Xiang Li, Huihui Fang, Mingsi Liu, Yanwu Xu, and Lixin Duan. Diffusion-enhanced transformation consistency learning for retinal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 221–231. Springer, 2024.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, Siliang Tang, Jun Xiao, Hui Lin, Yueting Zhuang, and Beng Chin Ooi. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation, 2025. URL https://arxiv.org/abs/2502.09838.

Bo Liu, Ke Zou, Li-Ming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21310–21320, 2025.

Chenglong Ma, Yuanfeng Ji, Jin Ye, Lu Zhang, Ying Chen, Tianbin Li, Mingjie Li, Junjun He, and Hongming Shan. Towards interpretable counterfactual generation via multimodal autoregression. *arXiv preprint arXiv:2503.23149*, 2025.

José Morano, Botond Fazekas, Emese Sükei, Ronald Fecso, Taha Emre, Markus Gumpinger, Georg Faustmann, Marzieh Oghbaie, Ursula Schmidt-Erfurth, and Hrvoje Bogunović. Multimodal foundation model and benchmark for comprehensive retinal oct image analysis. *NPJ Digital Medicine*, 8(1):576, 2025.

M Serdar Nazlı, Yasemin Turkan, Faik Boray Tek, Devrim Toslak, Mehmet Bulut, Fatih Arpacı, and Mevlüt Celal Öcal. Retinal disease diagnosis in oct scans using a foundational model. In *International Conference on Pattern Recognition*, pp. 208–220. Springer, 2024.

Junzhi Ning, Wei Li, Cheng Tang, Jiashi Lin, Chenglong Ma, Chaoyang Zhang, Jiyao Liu, Ying Chen, Shujian Gao, Lihao Liu, Yuandong Pu, Huihui Xu, Chenhui Gou, Ziyan Huang, Yi Xin, Qi Qin, Zhongying Deng, Diping Song, Bin Fu, Guang Yang, Yuanfeng Ji, Tianbin Li, Yanzhou Su, Jin Ye, Shixiang Tang, Ming Hu, and Junjun He. Unimedvl: Unifying medical multimodal understanding and generation through observation-knowledge-analysis, 2025. URL https://arxiv.org/abs/2510.15710.

Eric J. Nunn, Pejman Khadivi, and Shadrokh Samavi. Compound frechet inception distance for quality assessment of gan created images, 2021. URL https://arxiv.org/abs/2106.08575.

Qi Peng, Jialin Cui, Jiayuan Xie, Yi Cai, and Qing Li. Tree-of-reasoning: Towards complex medical diagnosis via multi-agent reasoning with evidence tree. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1744–1753, 2025.

Mohit Prabhushankar, Kiran Kokilepersaud, Yash-yee Logan, Stephanie Trejo Corona, Ghassan AlRegib, and Charles Wykoff. Olives dataset: Ophthalmic labels for investigating visual eye semantics. *Advances in Neural Information Processing Systems*, 35:9201–9216, 2022.

Huiyu Qiao, Feilong Tang, Huanfen Zhou, Yun Cai, Kairou Guo, Jin Wang, Tong Ma, Lie Ju, Wei Feng, Zhiqiang Ma, et al. Forecasting the diabetic retinopathy progression using generative adversarial networks. *Communications Medicine*, 5(1):368, 2025.

Luozheng Qin, Jia Gong, Yuqing Sun, Tianjiao Li, Mengping Yang, Xiaomeng Yang, Chao Qu, Zhiyu Tan, and Hao Li. Uni-cot: Towards unified chain-of-thought reasoning across text and vision. *arXiv preprint arXiv:2508.05606*, 2025.

Yansheng Qiu, Ziyuan Zhao, Hongdou Yao, Delin Chen, and Zheng Wang. Modal-aware visual prompting for incomplete multi-modal brain tumor segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3228–3239, 2023.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2025. URL https://arxiv.org/abs/2507.05201.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Yuhao Tang, Ye Yuan, Fei Tao, and Minghao Tang. Cross-modal augmented transformer for automated medical report generation. *IEEE Journal of Translational Engineering in Health and Medicine*, 13: 33–48, 2025.

M2 Team, Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, Xiangrong Zeng, and et al. Yijie Zhou. Baichuan-m2: Scaling medical capability with large verifier system, 2025. URL https://arxiv.org/abs/2509.02208.

Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, et al. Internvl-u: Democratizing unified multimodal models for understanding, reasoning, generation and editing. *arXiv preprint arXiv:2603.09877*, 2026.

Qinyue Tong, Ziqian Lu, Jun Liu, Yangming Zheng, and Zhe-Ming Lu. Medisee: Reasoning-based pixel-level perception in medical images. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2742–2751, 2025.

Janita E van Timmeren, Johan Bussink, Peter Koopmans, Robert J Smeenk, and Rene Monshouwer. Longitudinal image data for outcome modeling. *Clinical Oncology*, 38:103610, 2025.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Yuran Wang, Zhijing Wan, Yansheng Qiu, and Zheng Wang. Devil is in details: Locality-aware 3d abdominal ct volume generation for self-supervised organ segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10640–10648, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo,

Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

LLM-Core-Team Xiaomi. Mimo-vl technical report, 2025. URL https://arxiv.org/abs/2506.03569.

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025.

Dexuan Xu, Yanyuan Chen, Yu Huang, Shihao E, Yiwei Lou, Yongzhi Cao, Hanpin Wang, and Meikang Qiu. Medical vision-language pre-training with multimodal variational masked autoencoder for robust medical vqa. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3991–4000, 2025a.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025b.

Yifeng Xu, Zhenliang He, Meina Kan, Shiguang Shan, and Xilin Chen. Jodi: Unification of visual generation and understanding via joint modeling. *arXiv preprint arXiv:2505.19084*, 2025c.

Suhao Yu, Haojin Wang, Juncheng Wu, Luyang Luo, Jingshen Wang, Cihang Xie, Pranav Rajpurkar, Carl Yang, Yang Yang, Kang Wang, et al. Medframeqa: A multi-image medical vqa benchmark for clinical reasoning. *arXiv preprint arXiv:2505.16964*, 2025.

Chenlu Zhan, Yu Lin, Gaoang Wang, Hongwei Wang, and Jian Wu. Medm2g: Unifying medical multi-modal generation via cross-guided diffusion with visual invariant. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11502–11512, 2024.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation, 2018. URL https://arxiv.org/abs/1807.10165.

A Dataset Details

The general informed consent, approved by the relevant Institutional Review Board (IRB), explicitly authorizes the use of clinical data for scientific research and academic publication. Given the retrospective nature of this study, the requirement to obtain explicit informed consent was waived. Prior to data acquisition, all data were fully de-identified and anonymized to safeguard patient privacy.

A.1 Dataset Statistics

To support the proposed **Spatiotemporal Interventional Generation (SIG)** task, we construct **Ophtha-SIG**. Patients are first assigned to the training or test split. Both eyes and all visits of each patient remain in the same split. Within each split, OCT images are then grouped by patient and eye into longitudinal sequences, consistent with the intervention-conditioned and time-aware forecasting setting. The resulting training and test sets are strictly patient-disjoint.

Ophtha-SIG is divided into training and test sets. The training split contains 1,896 longitudinal sequences and 7,304 image references, while the test split contains 381 sequences and 1,264 image references. The average number of stages per sequence is 2.93 for training and 2.66 for testing, and the average number of image references per sequence is 3.85 and 3.32, respectively. Overall, the dataset is dominated by short sequences while still preserving longer disease trajectories. The temporal interval distribution is concentrated in clinically common short- to mid-term follow-up windows, particularly the 4–8 week range, while also covering longer horizons. Table 5 summarizes the overall statistics, and Fig. 5 presents the detailed distributions of stage counts and prompt intervals.

Table 5: Summary statistics of the Ophtha-SIG dataset. This table reports the number of sequences and image references in the training and test splits, as well as the average number of stages per sequence, average number of image references per sequence, and the minimum and maximum number of stages observed.

Split	#Sequences	#Image Refs	Avg. Stages / Seq.	Avg. Image Refs / Seq.	Min Stages	Max Stages
Train	1,896	7,304	2.93	3.85	2	19
Test	381	1,264	2.66	3.32	2	15

A.2 Comparison with Existing OCT Datasets

As shown in Table 6, most existing OCT datasets focus on static tasks such as disease classification or lesion segmentation. For instance, OCTID Gholami et al. (2019) and OCTDL Kulyabin et al. (2024) provide single-timepoint cross-sectional scans explicitly for multi-class retinal pathology categorization. Consequently, they consist entirely of independent images without longitudinal tracking.

While OLIVES Prabhushankar et al. (2022) introduces longitudinal and treatment-related information, it is primarily curated for biomarker learning with discrete clinical labels. It lacks standardized intervention-time conditioning and explicit dynamic rationales (i.e., mechanistic explanations for temporal changes). Furthermore, its limited scale (96 eyes) constrains its utility for training data-intensive generative models.

In contrast, **Ophtha-SIG** operates at a significantly broader scale and jointly incorporates longitudinal trajectories, clinical interventions, precise time horizons, and explicit mechanistic context into a unified formulation, directly facilitating intervention-conditioned future prediction and multimodal reasoning.

A.3 Detailed Data Construction Pipeline

For each real longitudinal sample $(X_t, X_{t+\Delta T}, I, \Delta T)$, physician-confirmed annotations provide the baseline and follow-up biomarker states B_t and $B_{t+\Delta T}$, the recorded intervention (I), and the overall trend $s^* \in \{\text{improvement}, \text{stability}, \text{worsening}\}$. Gemini-3.1 Pro generates a candidate evolution description $\tilde{r}$ and a candidate trend $\tilde{s}$. The observed biomarker transition is computed as:

$$\Delta B = B_{t+\Delta T} - B_t. \tag{7}$$

Table 6: Comparison with representative OCT datasets. *Scale* reports the primary unit used by each dataset. *Dynamic Rationale* indicates whether the dataset provides explicit mechanistic explanations for longitudinal changes.

Dataset	Scale	Longitudinal	Intervention	Dynamic Rationale	Task Formulation
OCTID Gholami et al. (2019)	∼500 images	×	×	×	Classification
OCTDL Kulyabin et al. (2024)	∼1,600 images	×	×	×	Classification
OLIVES Prabhushankar et al. (2022)	96 eyes	✓	✓	×	Biomarker Learning
Ophtha-SIG (ours)	2,277 seqs / 8,568 refs	✓	✓	✓	Forecasting & Reasoning

A candidate is retained when the local biomarker changes described in $\tilde{r}$ agree with ΔB and the candidate trend satisfies $\tilde{s} = s^*$. The retained description becomes the reference rationale (r) used for textual supervision and evaluation. For the test split, these checks retain 634 of 689 candidate tuples. Ophthalmologists subsequently review a random subset of the retained rationales and exclude two inaccurate samples, yielding 632 final test tuples.

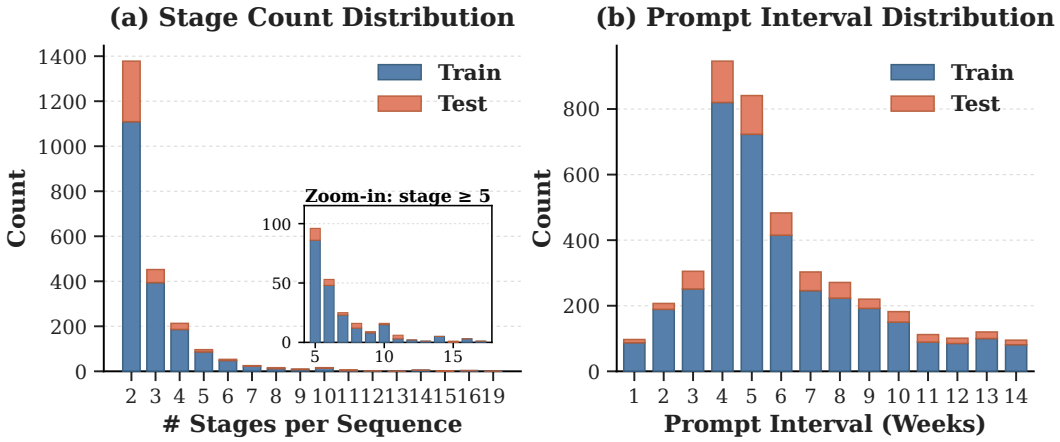


Figure 5: Detailed statistics of Ophtha-SIG. Left: distribution of the number of stages in each longitudinal sequence for the training and test splits. Right: distribution of sample pairs across different prompt intervals. The dataset is dominated by short sequences and short- to mid-term follow-up windows, while still retaining a small number of longer trajectories and longer temporal horizons.

B Detailed Evaluation Metrics

To comprehensively evaluate the SIGMA framework on the Spatiotemporal Interventional Generation (SIG) task, and to verify both the efficacy of the domain-adapted expert and the preservation of foundational discriminative capabilities within the unified architecture, we construct the following multi-dimensional evaluation metrics.

B.1 Core Metrics for the SIG Task

For the proposed SIG task, we evaluate the models along three automatic dimensions: textual rationale quality, spatiotemporal visual fidelity, and cross-modal semantic consistency. In addition, we further assess clinical plausibility via a human Turing test.

- **Textual Rationale Quality:** We evaluate generated clinical evolution rationales using ROUGE-L, BERTScore, and SBERT-Cos against the consistency-filtered reference rationales described in Appendix A.3. ROUGE-L measures structural coherence based on the longest common subsequence; BERTScore calculates word-level semantic similarity using deep contextual embeddings; and SBERT-Cos captures higher-level macro-semantic consistency by computing the cosine similarity between sentence-level feature embeddings of the generated and ground-truth rationales.

- **Spatiotemporal Visual Fidelity:** In addition to employing LPIPS to calculate the microscopic perceptual similarity between the generated predictions and the real follow-up OCTs at the feature level, we introduce domain-specific DINOv3-FID and DINOv3-KID metrics. By extracting feature distributions using our continually self-supervised domain-adapted expert model, these metrics compute the Fréchet Inception Distance (FID) Nunn et al. (2021) and Kernel Inception Distance (KID) Bińkowski et al. (2021) to more acutely capture the distribution fidelity of fine-grained pathological evolutions. These metrics measure distributional similarity in the same feature space used for retrieval and are therefore interpreted alongside LPIPS and expert assessment.

- **Cross-modal Rationale-Visual Alignment (CRVA):** To quantitatively assess the semantic consistency between the generated future OCT image and the predicted clinical rationale, we introduce the Cross-modal Rationale-Visual Alignment (CRVA) metric. Evaluating cross-modal consistency in medical imaging is highly challenging, as it requires clinical reasoning to determine whether the textual description matches the fine-grained anatomical changes in the images. Inspired by recent advances in LLM-as-a-judge paradigms Gu et al. (2025), we use GPT-5.4 as the automated CRVA evaluator. Its inputs are the baseline OCT, the generated future OCT, the extracted baseline biomarkers, and the model-generated rationale; the reference rationale is not included. CRVA measures consistency between the generated image and its explanation, rather than agreement with the observed follow-up outcome. It complements LPIPS and DINOv3-FID/KID comparisons with observed follow-up images.

  The evaluator assigns a CRVA score ranging from 1 to 5 based on a predefined clinical rubric. The score distribution across different baselines is visualized in Fig. 6, and the exact evaluation prompt is provided in Fig. 8. We present two representative examples from SIGMA to illustrate the scoring criteria. Fig. 9 shows a high-scoring instance (5/5), where the generated rationale accurately describes the pathological changes between the baseline and generated future OCT images. Conversely, Fig. 10 shows a low-scoring case (1/5), where the rationale predicts improvement but the generated future OCT exhibits morphological deterioration, resulting in a low alignment score.

- **Clinical Plausibility:** To quantify the real-world clinical plausibility of the synthetic follow-up forecasts, one ophthalmologist with more than ten years of clinical experience conducted a blinded Turing test comparing real clinical progressions with model-generated predictions. The evaluation metrics include the AI Fooling Rate, Real Recognition Rate, and a Balanced Turing Score, which assess how closely generated OCT images resemble real scans.

B.2 Biomarker Perception Metrics

This set of metrics is designed to validate the efficacy of the Domain-Adapted Perception Probe in resolving feature space misalignment and extracting precise structured state representations. We employ Average Accuracy, Average Recall, Average Jaccard Index, Hamming Loss, and Hit Rate. These metrics quantify the model's proficiency in extracting discrete state anchors from baseline OCT scans, thereby validating its reliability as a foundational input condition for subsequent trajectory retrieval and unified generation.

B.3 Preservation of Disease Diagnosis Capabilities

To verify whether the unified generation-understanding paradigm degrades foundational discriminative reasoning during the joint conditional generation process, we evaluate the model's disease classification performance on the current state. The quantitative benchmarks include Accuracy, Recall, and F1-score. We comprehensively evaluate these metrics across two dimensions: slice-level and patient-level. The slice-level metrics reflect the immediate morphological diagnostic performance based on a single 2D cross-sectional image post-intervention. The patient-level metrics integrate the patient's global medical record. Given the diagnostic "inertia" in clinical treatments (where morphological recovery is evident on a 2D slice, but the overall patient diagnostic label persists), combining both levels objectively demonstrates the model's ability to maintain diagnostic integrity over long-term follow-up trajectories.

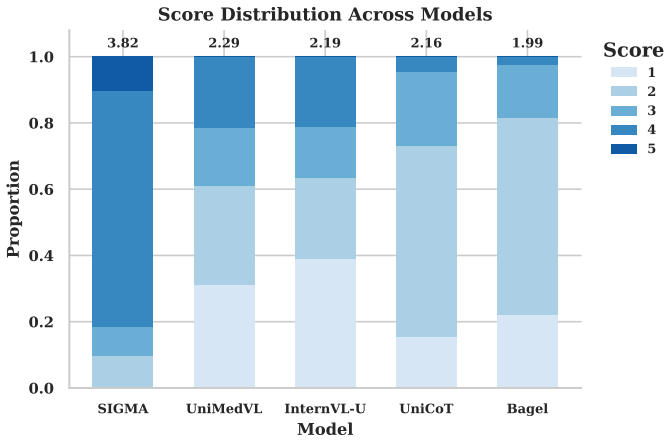


Figure 6: Score distribution of the Cross-modal Rationale-Visual Alignment (CRVA) metric across different baselines, evaluated by the GPT-5.4 MLLM judge. The bar chart visualizes the proportion of scores ranging from 1 (Severe Mismatch) to 5 (Perfect Support). Compared to unified multimodal models, SIGMA demonstrates a higher concentration in the high-score brackets (scores 4 and 5), suggesting improved cross-modal semantic consistency in jointly forecasting future OCT images and their corresponding trajectory changes.

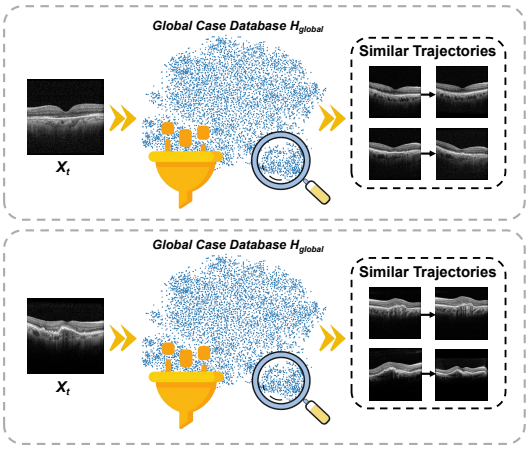


Figure 7: Qualitative examples of the multimodal trajectory retrieval process. Given a baseline OCT scan, the retrieval module identifies visually similar historical source images from the biomarker-pruned database. The figure further presents the corresponding real-world follow-up states (target images) of these matched cases under specific interventions, intuitively illustrating the trajectory variation information from the source to the target state.

C Biomarker-Guided Trajectory Retrieval

As illustrated in Fig. 7, the Similar Trajectory module first performs coarse-grained pruning over the global case database $\mathcal{H}_{\text{global}}$ using the biomarker anchors $\hat{B}_t$ predicted by Tool 1, producing a candidate subset $\mathcal{H}_{\text{pruned}}$. Specifically, biomarker overlap is defined as the Jaccard similarity between the predicted binary biomarker set of the input sample and that of each historical case. We retain historical cases with a Jaccard similarity at least 0.5 for coarse-grained pruning. Fine-grained retrieval is then performed within $\mathcal{H}_{\text{pruned}}$ by computing cosine similarity in the feature space of the domain-adapted DINOv3*.

After obtaining the most similar historical images, the module directly extracts the trajectory information utilizing them as source images (including their real follow-up outcomes under specific intervention conditions and time horizons). As shown in Fig. 7, the left panel presents the input

Prompt template: MLLM-as-a-judge for CRVA

 # System role. You are an ophthalmic clinical expert and medical imaging reviewer. Your objective is to evaluate whether a generated "Textual Rationale" faithfully supports the pathological evolution from a "Source Image" to a "Target Image".

 # Input data. Please read the following standardized fields:

- **Source Image:** The patient's baseline OCT scan prior to intervention.
- **Target Image:** The AI-generated predicted OCT scan following the intervention.
- **Baseline Biomarkers:** `{baseline_biomarkers}`.
- **Rationale:** "`{rationale}`".

 # Evaluation task. Strictly follow the three-step evaluation logic below:

- **Step 1: Visual delta.** Compare the Source Image and the Target Image to identify micro-level pathological changes, such as subretinal fluid absorption, variations in hyperreflective foci, or changes in retinal topology.
- **Step 2: Prior grounding.** Use the provided Baseline Biomarkers as clinical prior information to ground the interpretation of the baseline lesion state.
- **Step 3: Logical deduction.** Evaluate whether the **Rationale** provides a logically sound, accurate, and sufficiently complete justification for the observed visual changes. Check for omissions, exaggerations, medical hallucinations, and inconsistency with the baseline biomarker evidence.

 # Scoring rubric. Assign a `CRVA_Score` from 1 to 5:

- **5: Perfect support.** Fully consistent with visual evolution and baseline biomarkers, without omissions or hallucinations.
- **4: Good support.** Correct primary trajectory with only minor deviations in severity, fine-grained scale, or biomarker grounding.
- **3: Moderate support.** Captures macroscopic trends but misses critical details, includes weakly grounded reasoning, or incompletely uses biomarker context.
- **2: Weak support.** Significant conflict with the actual visual changes or substantial misinterpretation of biomarker evidence.
- **1: Severe mismatch.** Contradicts the image evolution or fundamentally conflicts with biomarker evidence.

 # Output format. Return only the following JSON object:

```
{
  "Reasoning":  "<Concise analysis of the visual delta,
baseline context, and rationale validity.>",
  "CRVA_Score":  <Integer from 1 to 5>
}
```

Figure 8: Prompt template used by the MLLM judge to evaluate Cross-modal Rationale–Visual Alignment (CRVA). The judge compares the source and target OCT images, incorporates baseline biomarker evidence, and assigns a 1–5 alignment score.

Algorithm 1 SIGMA Inference Pipeline

Require: Baseline OCT image X_t, patient metadata M_{patient}, intervention condition I, time horizon ΔT, global case database $\mathcal{H}_{\text{global}}$

Ensure: Future rationale $\hat{r}$ and predicted future OCT image $\hat{X}_{t+\Delta T}$

Perception Probe:
  $\hat{B}_t \leftarrow \Phi(X_t)$
Trajectory Retrieval:
  $\mathcal{H}_{\text{pruned}} \leftarrow \text{Prune}(\mathcal{H}_{\text{global}}, \hat{B}_t)$
  $S_{\text{traj}} \leftarrow \text{Retrieve}_{\mathcal{H}_{\text{pruned}}}(X_t)$ ▷ ranked by image similarity
Context Construction:
  $C \leftarrow \text{Concat}(\hat{B}_t, M_{\text{patient}}, S_{\text{traj}})$
Future Generation:
  $(\hat{r}, \hat{X}_{t+\Delta T}) \sim p_\theta(r, X_{t+\Delta T} \mid X_t, I, \Delta T, C)$
return $(\hat{r}, \hat{X}_{t+\Delta T})$

baseline OCT, the middle panel indicates the biomarker-pruned candidate space, and the right panel shows the retrieved similar historical source images together with their corresponding real follow-up target images, which collectively provide trajectory-level references from the source state to the target state.

Finally, the source-target changes, intervention conditions, and time horizons of the matched cases are aggregated by Gemini-3.1 Pro into a textual trajectory summary S_{traj}, which is used as part of the conditioning context for subsequent unified prediction.

D Detailed Inference Algorithm

Algorithm 1 formalizes the inference pipeline of the SIGMA framework on the SIG task. It details the sequential execution of the domain-adapted perception probe (Tool 1), biomarker-guided trajectory retrieval (Tool 2), context construction, and the final unified forecasting pass.

E Additional Experimental Results

E.1 Human Evaluation via Turing Test

To assess the visual plausibility of generated OCTs, one ophthalmologist with more than ten years of clinical experience conducted a blinded Turing test. We randomly sampled 63 test tuples and evaluated eight models in separate sessions. Each session contained 63 real follow-up OCTs and 63 generated OCTs from the corresponding model, presented individually in random order. The reader was blinded to model identity and was not informed of the 1:1 real/generated ratio. Breaks were arranged between sessions to reduce fatigue. AI Fooling Rate and Real Recognition Rate were computed from the 63 generated and 63 real images, respectively.

Table 7 reports the AI Fooling Rate, Real Recognition Rate, and Balanced Turing Score for each model. Higher AI Fooling Rate and Balanced Turing Score indicate that the synthetic follow-ups are visually closer to real clinical scans.

E.2 Expert Agreement on CRVA Scoring

To validate automated CRVA scoring against expert judgment, two independent ophthalmologists, each with more than ten years of clinical experience, blindly evaluated 100 outputs from 20 cases and five models using the same 1–5 CRVA rubric. Agreement was assessed using quadratic-weighted Cohen's κ, Spearman's ρ, and mean absolute error (MAE).

As shown in Table 8, GPT-5.4 demonstrated substantial agreement with both experts, with weighted κ values of 0.775 and 0.743, close to the inter-reader value of 0.791. The corresponding Spearman correlations were 0.779 and 0.755, compared with 0.802 between readers. MAEs were 0.49 and

0.56 points for GPT-5.4 versus experts and 0.35 points between readers. These results support the consistency of automated CRVA scoring with expert judgments of rationale–image alignment.

F LIMITATIONS

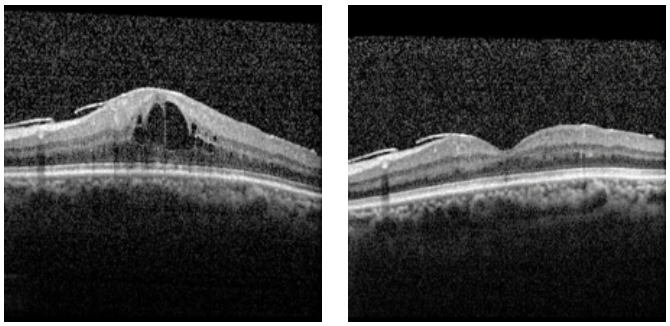
SIGMA provides a baseline for spatiotemporal interventional generation, with three directions for further investigation: (1) validation currently focuses on ophthalmic OCT; extending to other modalities would further assess generalizability; (2) evaluation considers one predicted outcome per target horizon, leaving multi-step forecasting and uncertainty quantification for future work; and (3) rationale construction and CRVA scoring rely on proprietary models, with expert assessment limited to a subset of samples. Open-model alternatives and larger reader studies would strengthen reproducibility and validation, while prospective studies would further assess the clinical utility of treatment-conditioned forecasts learned from observational data.

Table 7: Human Turing test results on the generated future OCT images. The ophthalmology specialist evaluated the synthetic follow-up scans against real clinical progressions to quantify real-world clinical plausibility. Metrics include AI Fooling Rate, Real Recognition (Recog.) Rate, and Balanced Turing Score. Best results are in **bold**.

Model	AI Fooling Rate	Real Recog. Rate	Balanced Turing Score
w/o reasoning			
Lumina-DiMOO Xin et al. (2025)	47.62	88.89	68.25
Qwen-Image-Edit Wu et al. (2025a)	0.00	**93.65**	46.83
OmniGen2 Wu et al. (2025b)	1.59	90.48	46.03
w/ reasoning			
InternVL-U Tian et al. (2026)	0.00	**93.65**	46.83
UniMedVL Ning et al. (2025)	6.35	90.48	48.41
UniCoT Qin et al. (2025)	9.52	**93.65**	51.59
Bagel Deng et al. (2025)	11.11	**93.65**	52.38
SIGMA (Ours)	**73.02**	88.89	**80.95**

Table 8: Agreement between GPT-5.4 and ophthalmologists on CRVA scoring, with inter-reader agreement as a reference. Weighted κ uses quadratic weights, and MAE denotes mean absolute error on the 1–5 scale. Parentheses indicate 95% confidence intervals. $\uparrow$ indicates higher is better; $\downarrow$ indicates lower is better.

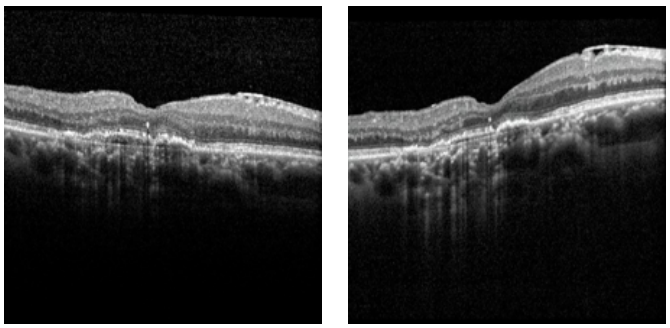
Comparison	Weighted $\kappa \uparrow$ (95% CI)	Spearman's $\rho \uparrow$ (95% CI)	MAE $\downarrow$
AI–expert agreement			
GPT-5.4 vs. Reader 1	0.775 (0.724, 0.816)	0.779 (0.711, 0.824)	0.49
GPT-5.4 vs. Reader 2	0.743 (0.667, 0.802)	0.755 (0.660, 0.817)	0.56
Inter-reader agreement (reference)			
Reader 1 vs. Reader 2	0.791 (0.679, 0.872)	0.802 (0.697, 0.880)	0.35



Baseline OCT (Input) Predicted OCT $\hat{X}_{t+\Delta T}$

Input Prompt	The patient is diagnosed with ME. Please generate the image after anti-VEGF injection treatment over 5 weeks.
Predicted Structured Rationale $\hat{r}$	At baseline, the OCT is expected to represent macular edema with marked diffuse retinal thickening and prominent intraretinal hyporeflective cystoid spaces centered in the foveal/parafoveal region, with associated sponge-like DRT/ME and mild distortion of the normal foveal contour, consistent with active vascular leakage. Following anti-VEGF treatment over 5 weeks, the edema is predicted to show a substantial but incomplete therapeutic response because suppression of abnormal permeability will likely reduce exudation without fully normalizing retinal architecture. Consequently, the large central intraretinal fluid cavities are expected to collapse and decrease markedly, and the diffuse retinal thickening will likely regress with restoration of a more physiologic foveal profile; however, mild residual IRF and persistent but reduced DRT/ME are anticipated to remain on follow-up OCT.
MLLM-as-a-judge Scoring Rationale	Baseline biomarkers (IRF, DRT/ME) are supported by the source OCT, which shows marked central retinal thickening with a large foveal/parafoveal intraretinal hyporeflective cystic space and surrounding diffuse sponge-like edema, causing loss of normal foveal depression. In the target OCT, there is clear interval improvement: the large central cystic IRF cavity has largely collapsed/resolved, overall retinal thickness is reduced, and the foveal contour is more physiologic, though not completely normal. Mild residual intraretinal hyporeflective changes/thickening persist, consistent with reduced but not fully resolved DRT/ME. No major alternative lesion evolution (such as SRF dynamics) is evident or ignored. The rationale accurately describes the baseline edema pattern and the predicted post-anti-VEGF partial response, including substantial reduction of central IRF and regression of diffuse thickening with incomplete normalization. It is well aligned with both images and the biomarker context. Minor limitation: it is somewhat generic and does not comment on finer structural detail, but there is no clear hallucination or contradiction.
CRVA Score	**5**

Figure 9: Qualitative example of a high Cross-modal Rationale-Visual Alignment (CRVA) score (5/5). The generated clinical rationale demonstrates strong cross-modal consistency with the predicted visual trajectory from the source to the target OCT image. It appropriately captures the spatiotemporal evolution, such as the substantial reduction of central intraretinal fluid (IRF) and the regression of diffuse thickening, aligning well with the visual evidence.



Baseline OCT (Input) Predicted OCT $\hat{X}_{t+\Delta T}$

Input Prompt	The patient is diagnosed with nAMD. Please generate the image after anti-VEGF injection treatment over 6 weeks.
Predicted Structured Rationale $\hat{r}$	In this eye with nAMD, the baseline OCT is expected to represent a relatively mild exudative state, with subtle irregular elevation of the RPE/outer retinal complex, slight foveal contour distortion, and shallow low-reflective subretinal or sub-RPE fluid without marked intraretinal cystic change. After anti-VEGF treatment over 6 weeks, the lesion is predicted to show a favorable but partial anatomic response, as VEGF suppression is expected to reduce abnormal choroidal neovascular leakage rather than produce complete drying in this interval. Consequently, the shallow hyporeflective fluid spaces are anticipated to decrease, the foveal contour is expected to become slightly smoother and less elevated, and residual low-grade RPE irregularity with persistent thin outer retinal separation will likely remain, consistent with incomplete but definite drying.
MLLM-as-a-judge Scoring Rationale	The source OCT shows a relatively dry to mildly exudative macula with irregular RPE/outer retinal elevation and only limited shallow hyporeflective separation; there is no prominent intraretinal cystic edema. The target OCT, however, demonstrates worsening rather than improvement: increased retinal thickening and dome-like elevation, more pronounced foveal/parafoveal contour distortion, and clearer hyporeflective subretinal and/or intraretinal fluid spaces, especially in the superior/right macula. RPE irregularity also appears more conspicuous. The rationale claims a favorable partial anti-VEGF response with reduced fluid and smoother contour, which is opposite to the visual trajectory shown in the target image. It undercalls the degree of fluid and elevation in the target and invents interval drying not supported by the scan. Therefore, the rationale is in major conflict with the actual image evolution.
CRVA Score	**1**

Figure 10: Qualitative example of a low Cross-modal Rationale-Visual Alignment (CRVA) score (1/5). This case highlights a semantic mismatch where the generated rationale diverges from the visual pathological evolution. While the text describes a favorable partial anatomic response with reduced fluid, the target OCT image exhibits a trajectory with increased retinal thickening and elevated contour distortion.