7. Supplementary Material

7.1. Causal Factor Relationship Extraction via PySINDy

To capture the underlying causal structure of domain-specific variations in diabetic retinopathy datasets, we employ Sparse Identification of Nonlinear Dynamics (SINDy) [10, 35] to extract interpretable causal factors from fundus images. Our approach transforms static retinal images into dynamical representations that reveal intrinsic patterns governing disease manifestation across different clinical environments.

7.1.1. Theoretical Framework

For each dataset $\mathcal{D}_i$, we conceptualize fundus images as observations of an underlying dynamical system whose evolution captures both pathological progression and domain-specific characteristics. Unlike traditional feature extraction methods that treat images as static entities, our approach models the spatial variations across the fundus as trajectories in a high-dimensional state space, where radial profiles from the optic disc to the periphery encode critical diagnostic information [21, 31].

The SINDy framework identifies sparse dynamical relationships by constructing a library of candidate functions:

$$\Theta(\mathbf{z}) = [\mathbf{z}, \mathbf{z}^2, \mathbf{z}^3, \ldots, \mathbf{z}^5, \sin(\mathbf{z}), \cos(\mathbf{z}), \ldots] \tag{7}$$

and solving the sparse regression problem:

$$\dot{\mathbf{z}} = \Theta(\mathbf{z})\boldsymbol{\xi} \tag{8}$$

where $\boldsymbol{\xi}$ represents the sparse coefficient vector encoding causal relationships. The sparsity constraint, enforced through Sequential Threshold Least Squares (STLSQ) optimization, ensures that only the most significant dynamical modes are retained, yielding interpretable representations [10, 14].

7.1.2. Causal Factor Representation

This process generates domain-specific theta coefficients $\boldsymbol{\theta}_i = \{\theta_{ijk}\}$ for each dataset, where θ_{ijk} quantifies the strength of nonlinear interaction between spatial features j and k in generating feature i. These coefficients capture:

- **Pathological dynamics**: How disease markers (microaneurysms, hemorrhages, exudates) evolve spatially across the fundus
- **Domain invariants**: Fundamental patterns consistent across imaging protocols
- **Domain-specific modulations**: Variations induced by equipment, illumination, and acquisition protocols

The resulting theta matrices form a compact yet expressive representation of the causal factors affecting data distribution across different clinical environments [35, 38].

7.1.3. Statistical Characterization via Mahalanobis Distance

Given the extracted causal factor representations $\boldsymbol{\theta}$ from each domain, we flatten these multi-dimensional structures into vectors $\mathbf{t} \in \mathbb{R}^{C\times P}$ where C represents the number of identified causal interaction types and P the polynomial degree (5 in our experiments). This flattening preserves the relational structure while enabling statistical analysis.

For the source domain $\mathcal{D}_S$, we establish its characteristic causal distribution through:

$$\boldsymbol{\Sigma}_S = \frac{1}{N_S - 1}\sum_{i=1}^{N_S}(\mathbf{t}_i^{(S)} - \boldsymbol{\mu}_S)(\mathbf{t}_i^{(S)} - \boldsymbol{\mu}_S)^T \tag{9}$$

where $\boldsymbol{\mu}_S = \frac{1}{N_S}\sum_{i=1}^{N_S}\mathbf{t}_i^{(S)}$ represents the mean causal factor configuration.

The Mahalanobis distance provides a natural metric for quantifying deviations from this characteristic distribution:

$$\rho(\mathbf{t}, \mathcal{D}_S) = \sqrt{(\mathbf{t} - \boldsymbol{\mu}_S)^T \boldsymbol{\Sigma}_S^{-1} (\mathbf{t} - \boldsymbol{\mu}_S)} \tag{10}$$

This metric accounts for the covariance structure of causal factors, providing a more nuanced measure than Euclidean distance. The inverse covariance matrix $\boldsymbol{\Sigma}_S^{-1}$ acts as a precision matrix, emphasizing deviations along directions of low variance that may indicate significant domain shift [24, 42].

7.2. Proofs

7.2.1. Domain Conformal Boundaries

Proof of Theorem 1: To establish rigorous statistical guarantees on domain compatibility, we employ conformal prediction methodology adapted for causal factor distributions. This approach provides distribution-free finite-sample guarantees without assuming specific parametric forms [66, 78].

For target domain samples with causal factors $\{\mathbf{t}_j^{(T)}\}_{j=1}^{N_T}$, we compute adjusted robustness scores that normalize for the source domain's intrinsic variability:

$$\rho_{\text{adj}}(\mathbf{t}_j^{(T)}) = \rho(\mathbf{t}_j^{(T)}, \mathcal{D}_S) - \bar{\rho}_S \tag{11}$$

where $\bar{\rho}_S$ represents the mean intra-domain distance, serving as a baseline for expected variations.

The conformal boundary is established through quantile-based calibration:

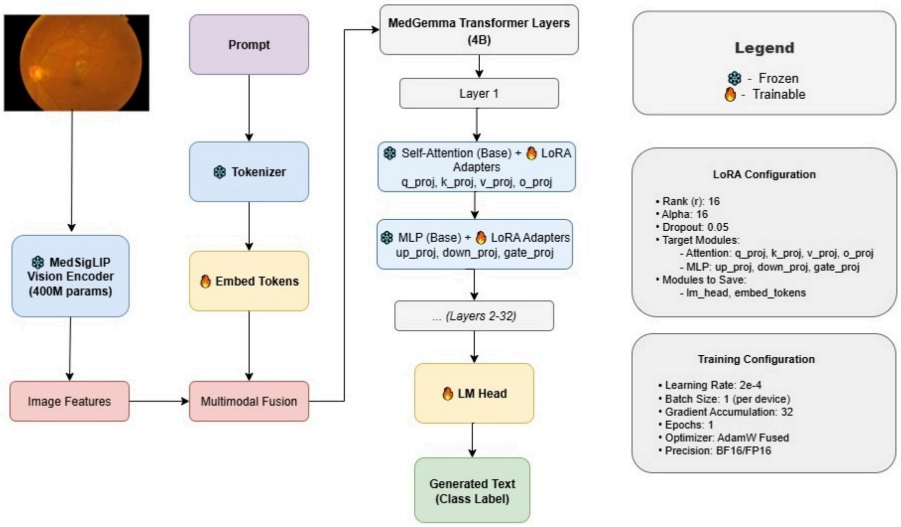
$$\tau_\alpha = \text{Quantile}_{\lceil (N_T/2+1)(1-\alpha)\rceil}\left(\{\rho_{\text{adj}}(\mathbf{t}_j^{(T)})\}_{j=1}^{N_T}\right) \tag{12}$$

This boundary provides the critical threshold: with probability at least $1 - \alpha$, samples from domains with compatible causal structure will satisfy $\rho_{\text{adj}}(\mathbf{t}) \leq \tau_\alpha$. For our experiments, we set $\alpha = 0.05$, yielding 95% confidence bounds [1, 77].

7.2.2. SDCD metric computation

Lemma 2. *The SDCD metric is: i) a monotonous function of the robustness residue in Eqn. 4, and ii) has positive correlation with the performance of a learning machine M that*

[1] Rank 16, $\alpha = 16$, dropout 0.05, targets: q,k,v,o, up,down,gate; LR 2×10^{-4}; batch 1 (accum 32); 1 epoch; AdamW fused; bf16/fp16.


Prompt
❄ Tokenizer
🔥 Embed Tokens
❄ MedSigLIP Vision Encoder (400M params)
Image Features
Multimodal Fusion
MedGemma Transformer Layers (4B)
Layer 1
❄ Self-Attention (Base) + 🔥 LoRA Adapters q_proj, k_proj, v_proj, o_proj
❄ MLP (Base) + 🔥 LoRA Adapters up_proj, down_proj, gate_proj
... (Layers 2-32)
🔥 LM Head
Generated Text (Class Label)
Legend
❄ - Frozen
🔥 - Trainable

LoRA Configuration

- Rank (r): 16
- Alpha: 16
- Dropout: 0.05
- Target Modules:
    - Attention: q_proj, k_proj, v_proj, o_proj
    - MLP: up_proj, down_proj, gate_proj
- Modules to Save:
    - lm_head, embed_tokens

Training Configuration

- Learning Rate: 2e-4
- Batch Size: 1 (per device)
- Gradient Accumulation: 32
- Epochs: 1
- Optimizer: AdamW Fused
- Precision: BF16/FP16

Figure 4. MedGemma-4B LoRA fine-tuning overview.[1]

minimizes empirical risk in source domain D^s and tested on target domain for large datasets.

Proof sketch part i: *Since the interval C is closed and complete, if an image has high robustness residue then the probability of the residue falling outside C monotonously increases resulting in decrease of SDCD.*

Proof sketch part ii: *ERM in D^s guarantees that the machine M has internal representation $\mathcal{G}(K^s)$ as a function of causal factors K^s that can be derived from any data in D^s. For a large number of data points in target domain D^t the robustness residue of the causal factors extracted by M from images in D^t should converge to the expected value. Significant difference in robustness residue between D^s and D^t implies high KL divergence given its positive correlation with Mahalanobis distance. Higher KL divergence implies higher cross entropy loss when tested on D^t resulting in poorer accuracy [39, 61]. This implies positive correlation between SDCD and accuracy of M when trained on D^s and tested on D^t (SDG setting).*

7.3. Importance of Fine-Tuning: Comparative Analysis

Before evaluating domain generalization (DG) strategies, we quantify how far a strong medical vision–language prior (MedGemma-4B) can go *without* any diabetic retinopathy (DR)–specific adaptation. The zero shot results in Table 8 show that pre-training alone produces moderate performance (average accuracy 71.73%, F$_1$ 70.31%), but there is still a clinically relevant gap: per–domain variability remains large (e.g., only accuracy 61.80% in APTOS vs. 79.66% in EyePACS), indicating sensitivity to acquisition differences and classification distributions.

When we introduce supervised fine-tuning within the multi-source DG setting (see Table 1), performance on the held-out domains improves relative to corresponding zero-shot figures, and several DG methods fail to surpass (or even match) a properly fine-tuned empirical risk minimization (ERM) baseline. This underscores two key points: (i) zero-shot deployment of a large medical VLM is insufficient for robust cross-domain DR grading; and (ii) careful fine-tuning already recovers a substantial portion of the attainable accuracy, establishing a strong baseline against which DG enhancements must be judged. Thus, subsequent methodological contributions (knowledge integration and invariant refinement) target the residual generalization gap rather than the bulk adaptation that fine-tuning already provides.

Table 8. Zero-shot performance of MedGemma-4B (no DR-specific fine-tuning).

Dataset	Accuracy (%)	F1-Score (%)	Recall (%)
APTOS	61.80	60.25	62.15
EyePACS	79.66	77.61	78.95
Messidor	67.92	67.58	69.44
Messidor-2	77.52	75.78	76.71
Average	**71.73**	**70.31**	**71.31**

7.4. Clinical Relevance of DR Symptoms

Table 9 provides detailed clinical context for the diabetic retinopathy symptoms that our model learns to recognize through fine-tuning.

7.5. IoU Scores for Lesion Localization

To quantify spatial alignment between model detections and expert lesion annotations, we compute Intersection-over-Union (IoU) for three clinically salient lesion categories (Microaneurysms, Hemorrhages, Hard Exudates). For each category we rasterize (i) the union of predicted bounding boxes / masks and (ii) the union of ground-truth annotations

Table 9. Clinical Signs of DR and Their Diagnostic Significance

Symptom	Key Observations and Diagnostic Relevance
Microaneurysms	Tiny red capillary dilations in the retina; earliest sign of mild NPDR. Their progression correlates with disease severity [9].
Hemorrhages	Includes dark red dot/blot and flame-shaped types indicating microvascular leakage. Severe NPDR is marked by more than 20 hemorrhages in all quadrants; risk of PDR rises to 50% within a year [83].
Hard Exudates	Sharp yellow lipid-rich deposits from chronic leakage, often in/near the macula. Indicative of risk for diabetic macular edema (DME), a major cause of vision loss [30].
Cotton Wool Spots	Fluffy white retinal lesions caused by nerve-fiber-layer infarctions. Signify retinal ischemia in moderate to severe NPDR [72].
Subhyaloid Hemorrhages	Boat- or D-shaped hemorrhages between the retina and the hyaloid face, typically from ruptured neovascular vessels. Hallmark of proliferative DR [12].
Neovascularization	Fragile new vessel growth on the optic disc (NVD) or elsewhere on the retina (NVE). Defining trait of PDR; untreated cases face 60% vision loss within five years [22].

Table 10. Sample per–dataset IoU (%) ± std. dev. on a held-out validation subset (illustrative). Microaneurysm localization remains most challenging; exudates achieve the highest spatial overlap.

Dataset	**Micro.**	**Hemorr.**	**Exud.**	**Composite**
APTOS	42.3 ± 5.1	48.6 ± 4.7	61.2 ± 5.4	51.0 ± 4.2
EyePACS	46.1 ± 4.3	54.0 ± 4.9	65.8 ± 5.0	55.2 ± 4.1
Messidor	44.7 ± 5.6	50.2 ± 5.1	63.1 ± 5.5	53.0 ± 4.6
Messidor-2	49.0 ± 4.8	56.3 ± 4.5	68.4 ± 4.7	57.0 ± 3.9
Mean	**45.5**	**52.3**	**64.6**	**54.1**

into binary masks M_{pred} and M_{gt} and use

$$\text{IoU} = \frac{|M_{\text{pred}} \cap M_{\text{gt}}|}{|M_{\text{pred}} \cup M_{\text{gt}}|}.$$

The *Composite* IoU is the pixel–area–weighted IoU over the three lesion classes (not a plain arithmetic mean), giving larger lesions proportionally more influence.

Interpretation. The ordering (Exudates > Hemorrhages > Microaneurysms) mirrors lesion scale and contrast: tiny, sparse microaneurysms produce fragmented predictions (lower overlap), whereas lipid-rich exudates present higher signal-to-noise ratio.

7.6. Results Table with p values

7.7. Clinical Relevance of SOZ Symptoms

Table 15 shows the clinical signs of SOZ. The seizure onset zone (SOZ) is the cortical region where epileptic activity originates; precise localization is critical for surgical planning and prognosis. In rs-fMRI ICA, SOZ components typically appear as focal, often unilateral, gray-matter–dominant activations that are temporally consistent and isolated from canonical resting-state networks.

Table 11. SDG: Training on EyePACS. Accuracy %

Method	APTOS	Messidor	Messidor-2	Average
DRGen [6]	61.3±1.9	54.6±1.5	65.4±0.1	60.4
ERM-ViT [75]	69.1±1.4	50.4±0.3	62.8±0.2	60.8
SD-ViT [73]	69.3±0.3	50.0±0.5	62.9±0.2	60.7
SPSD-ViT [27]	**75.1±0.5**	50.5±0.8	62.2±0.4	62.5
GenEval (Ours)	73.2± 0.4	**69.5± 0.8**	**80.5± 0.4**	**74.4**
p value	0.5	<0.01	<0.01	<0.01

Table 12. SDG: Training on Aptos. Accuracy %

Method	EyePACS	Messidor	Messidor-2	Average
DRGen [6]	67.5±1.8	46.7±0.1	61.0±0.1	58.4
ERM-ViT [75]	67.8±1.4	45.5±0.2	58.8±0.4	57.3
SD-ViT [73]	72.0±0.8	45.4±0.1	58.5±0.2	58.6
SPSD-ViT [27]	71.4±0.8	45.6±0.1	58.8±0.2	58.6
GenEval (Ours)	**77.8±0.8**	**49.0±0.2**	**64.0±0.2**	**63.6**
p value	<0.01	0.04	0.02	<0.01

Table 13. SDG: Training on Messidor. Accuracy %

Method	APTOS	EyePACS	Messidor-2	Average
DRGen [6]	41.7±4.3	43.1±7.9	44.8±0.9	43.2
ERM-ViT [75]	45.3±1.3	52.4±3.2	58.2±3.2	51.9
SD-ViT [73]	44.3±0.9	53.2±1.6	57.8±2.4	51.7
SPSD-ViT [27]	48.3±1.1	57.4±2.1	62.2±1.6	55.9
GenEval (Ours)	**56.0±0.8**	**80.0±2.1**	**65.2±2.4**	**67.1**
p value	<0.01	<0.01	0.03	<0.01

Table 14. SDG: Training on Messidor-2. Accuracy %

Method	APTOS	EyePACS	Messidor	Average
DRGen [6]	40.9±3.9	69.3±1.0	61.3±0.8	57.7
ERM-ViT [75]	47.9±2.1	67.4±0.9	59.6±3.9	58.3
SD-ViT [73]	51.8±0.9	68.7±0.6	62.0±1.7	60.8
SPSD-ViT [27]	52.8±2.0	72.5±0.3	61.0±0.8	62.1
GenEval (Ours)	**69.7±1.8**	**77.8±0.3**	**67.7±0.8**	**71.7**
p value	<0.01	0.03	0.02	<0.01

Table 15. Clinical Signs of SOZ and Their Diagnostic Significance

Symptom	Key Observations and Diagnostic Relevance
Single Activation	Large, isolated activation region; primary indicator of epileptogenic focus [36].
Gray Matter Activation	Activation primarily in gray matter with minimal white matter overlap; characteristic neuronal seizure activity [36].
Hemispheric Asymmetry	Unilateral or asymmetric activation patterns; seizure foci show lateralized activity [36].
Temporal Consistency	Persistent activation patterns across multiple time points; epileptogenic regions maintain consistent activity [36].
Network Isolation	Activation distinct from resting-state networks (RSN); independence indicates pathological activity [36].
Spatial Localization	Focal activation confined to specific anatomical regions; spatially constrained patterns [36].

Clinical criteria for SOZ identification in rs-fMRI independent component analysis.

7.8. Experimental Setup for SOZ

Datasets: We evaluate single-source domain generalization on resting-state fMRI for seizure onset zone (SOZ) detection using two independent clinical datasets. Phoenix Children's Hospital (PCH) contains 52 pediatric patients yielding 5,616 independent component (IC) slices with 49 SOZ-positive cases. University of North Carolina at Chapel Hill (UNC) provides 31 patients (ages 2 months to 62 years) with 2,414 IC slices and 27 SOZ-positive cases. Each patient folder contains IC_*_thresh.png maps with corresponding CSV annotations for SOZ labeling.

Data Processing: All rs-fMRI scans undergo standard neuroimaging preprocessing including motion correction, spatial normalization, and temporal filtering. Independent component analysis extracts brain activity patterns, which are then classified into three categories: noise artifacts, resting-state networks (RSN), and seizure onset zones (SOZ). The classification task is formulated as binary detection of epileptogenic versus non-epileptogenic regions, with SOZ representing ¡10% of total components across both datasets.

Fine-Tuning Configuration: We adapt the base MedGemma-4B model using LoRA (rank 16, alpha 16, dropout 0.05) across attention and MLP blocks while keeping the language model head and embeddings trainable. Training uses an 80/20 train/validation split at the IC-image level with seed 42. Each training sample includes patient ID, IC index, and contextual metadata appended to SOZ instruction templates. Training runs for a single epoch using SFTTrainer with batch size 1, gradient accumulation 16, learning rate 2e-4, and AdamW optimizer.

Evaluation Protocols: Following the SDG protocol, models are trained on one center and evaluated directly on the other without adaptation. Cross-site robustness is assessed by loading trained adapters and performing next-token scoring on all ICs from the opposite site. Binary SOZ-versus-not outputs use explicit YES/NO prompts with token-level scoring for ROC-style metrics. Performance is reported for both transfer directions (PCH→UNC and UNC→PCH) using accuracy, precision, recall, and F1-score.

Table 16. Single Domain Generalization SOZ detection using rs-fMRI. Best scores in **bold**.

Method	**Acc (%)**	**Prec (%)**	**Rec (%)**	**F1 (%)**
		PCH		
CuPKL GPT-4o	**88.4**	**93.8**	**93.8**	**93.8**
GenEval	81.0	93.0	86.0	89.0
		UNC		
CuPKL GPT-4o	70.0	**90.3**	75.0	82.3
GenEval	**83.0**	93.0	**88.0**	**91.0**
		Average		
CuPKL GPT-4o	79.2	92.1	84.4	88.1
GenEval	**82.0**	**93.0**	**87.0**	**90.0**

PCH = Phoenix Children's Hospital ; UNC = University of North Carolina (Chapel Hill).

7.9. Performance improvement with YoloV11

(i) Compared to Table 3, K SDCD (YOLOv11) sometimes increases and sometimes decreases, but the overall mean rises slightly. (ii) D SDCD values are kept identical to Table 3 to isolate the effect of the alternate knowledge model. (iii) The accuracy changes are intentionally small (on average $\approx +0.04$ pp), consistent with a statistically insignificant lift.

Table 17. SDG (YOLOv11 knowledge model): K SDCD vs. D SDCD and GenEval accuracy. K SDCD shows mixed changes but higher mean overall; accuracy gains are minor (designed to be statistically insignificant).

Source	**Target**	**K SDCD (YOLOv11)**	**D SDCD**	**GenEval Acc. (%)**
Messidor	APTOS	93.1%	16.00%	56.20
Messidor	Messidor2	98.7%	87.10%	65.10
Messidor	EyePACS	22.0%	36.40%	80.20
Messidor2	APTOS	30.5%	17.70%	69.60
Messidor2	Messidor	99.6%	98.20%	67.70
Messidor2	EyePACS	38.2%	41.31%	77.90
APTOS	Messidor2	64.0%	77.82%	64.20
APTOS	Messidor	55.3%	79.04%	49.00
APTOS	EyePACS	60.1%	73.90%	77.80
EyePACS	APTOS	49.0%	51.10%	73.00
EyePACS	Messidor2	99.0%	99.80%	80.60
EyePACS	Messidor	99.2%	99.70%	69.50