Toward an AI Co-Annotator for Interactive Clinical Diagnosis: Expert-Attention-Aligned AOI Guidance and VLM Biomarker Drafting in Age-Related Macular Degeneration

ANONYMOUS AUTHOR(S)

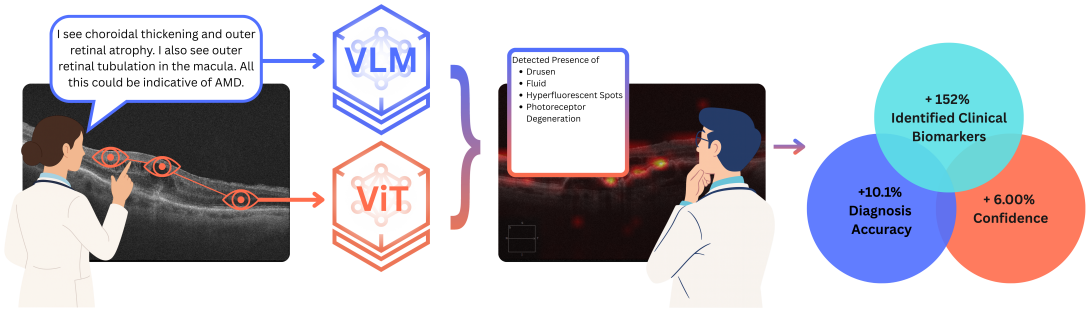


Fig. 1. The *Co-Annotator* system architecture. The system collects expert behavioral signals (gaze and dictation) to train two backend models: a Vision Transformer (ViT) and a Vision-Language Model (VLM). These models power two complementary clinician-facing interface components: (1) Fixation-aligned Areas of Interest (AOIs) that visually guide visual search, and (2) Biomarker-bounded VLM guidance that scaffolds documentation. Far right: relative change with VLM guidance vs. control in diagnosis accuracy, confidence, and identified biomarkers.

Clinical AI often optimizes prediction accuracy without directly interfacing with how clinicians make decisions, i.e., where they look and what they write. We present *Co-Annotator*, a human–AI co-annotation method that (i) trains a gaze-aligned Vision Transformer (ViT) to produce fixation-aligned areas of interest (AOIs) and (ii) fine-tunes an ontology-bounded vision-language model (VLM) to draft editable biomarker summaries. The method learns directly from expert gaze and dictations, anchoring model attention and language in clinical expertise. In a case study on retinal OCT for wet age-related macular degeneration, aligning the ViT to expert fixations improved micro-area-under-the-curve (micro-AUC) from 0.95 to 0.98; and the VLM-generated biomarker text matched experts' dictations (MedBERTScore 0.867). In a deployment with ophthalmology residents, the residents' gaze aligned with the AOI guidance. The users had significantly greater biomarker reporting with the VLM, and high rates of biomarker retention (83.1%) with preserved diagnostic accuracy.

CCS Concepts: • **Human-centered computing** → **Collaborative interaction**; **Collaborative interaction**; • **Computing methodologies** → *Artificial intelligence*; • **Applied computing** → *Imaging*; *Health informatics*.

Additional Key Words and Phrases: Human–AI collaboration, Clinical decision support, Explainable AI, Eye tracking (gaze-aligned attention), Vision–language models, Ontology-bounded generation, Co-annotation, Workflow-integrated documentation, Optical coherence tomography (OCT), Wet age-related macular degeneration

Manuscript submitted to ACM

ACM Reference Format:
Anonymous Author(s). 2025. Toward an AI Co-Annotator for Interactive Clinical Diagnosis: Expert-Attention-Aligned AOI Guidance and VLM Biomarker Drafting in Age-Related Macular Degeneration. In *Proceedings of CHI (Computer Human-Computer Interaction).* ACM, New York, NY, USA, 34 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

We build and study an optical coherence tomography (OCT)-specialized AI co-annotator for wet age-related macular degeneration (wAMD) consisting of two physician-facing, visual-language-model–backed user-interface components that highlight relevant imaging evidence to support trainees' diagnostic workflow. wAMD affects millions and carries a disproportionate share of vision loss despite being the less common AMD subtype [17]. Timely treatment can preserve vision when wAMD is identified before it causes permanent damage, yet early diagnosis is difficult: wAMD arises from growth of new abnormal blood vessels causing fluid accumulation and bleeding, which must be resolved and distinguished from other AMD subtypes on OCT—the primary non-invasive imaging modality for diagnosing and monitoring neovascular AMD in routine practice [34]. The relevant OCT biomarkers (e.g., intraretinal fluid (IRF) vs. subretinal fluid (SRF), pigment epithelial detachment (PED), etc.) can be subtle and fine-grained and are challenging even for experienced readers to distinguish [35, 41].

Retina specialty clinics operate with high imaging volume and under significant time pressure. Each imaging study ordered by the ophthalmologist demands directed visual search, structured description, and auditable documentation. Trainees are a critical user group because they show higher variance in accuracy and speed, are still forming mental models for biomarker patterns, and are calibrating trust in any computer-aided tools they use. Concretely, our system targets three workflow outcomes: (i) **faster, more focused reading** via area-of-interest guidance, (ii) **lower documentation burden** via structured biomarker drafts residents can edit, and (iii) **greater transparency** by exposing where the model "looked" and constraining language to a clinical ontology.

Despite recent advances in vision and vision–language foundation models, faithful OCT biomarker grounding remains hard due to (i) domain shift across scanners, protocols, and patient populations; (ii) the need for *spatially localized* distinctions (e.g., IRF vs. SRF) rather than global labels; and (iii) the risk of clinically unsafe hallucinations or vague free-text rationales [29, 30]. For safety-critical use, clinicians need transparency (to know where the model is attending visually), controllability (to be able to accept/edit findings), and outputs bounded by a shared ontology. These requirements motivate *expert-aligned supervision* and interaction designs that present evidence to *support* clinicians' decisions rather than replace them.

We therefore built *Co-Annotator*, a system that unifies expert-distilled modeling with workflow-integrated design. The system consists of two backend models—a gaze-aligned ViT and an ontology-bounded VLM—that drive two complementary clinician-facing interface components:

(1) **Fixation-aligned Areas of Interest (AOI):** Attention overlays trained to *visually guide* readers toward regions with higher diagnostic relevance on OCT images, aligned to expert fixation density rather than post hoc saliency.
(2) **Biomarker-bounded Vision-Language Model (VLM) guidance:** A VLM fine-tuned on OCT images paired with expert dictations and curated biomarker labels to produce *structured biomarker summaries* and answer targeted diagnostic questions.

To create and evaluate this co-annotator, we conduct two user studies: first, we collect synchronized expert gaze and dictations and align the models to these signals; second, we deploy the interfaces to ophthalmology residents to

probe efficacy and workflow fit at a large academic medical center.[1] We aim to study whether expert behavior can be translated into faithful model attention and bounded language, both of which are preconditions for trustworthy assistance. Then, we assess whether these results in favorable expert behavior transfer to trainees, as mediated by an AI Co-annotator. Specifically, we make the following contributions:

- Introducing an *OCT AI co-annotator* with two interfaces—fixation-aligned AOIs and biomarker-bounded VLM guidance—designed for specificity, transparency, and editability.
- Demonstrating that the ViT and VLM can faithfully align to experts' visual attention and language, thereby improving the co-annotator's automated diagnostic accuracy (micro-AUC: 0.98), yielding AOIs concordant with expert gaze, and identifying ontology-bounded biomarker text (MedBERTScore: 0.87).
- Reporting a deployment study with ophthalmology residents measuring diagnostic accuracy, throughput correct diagnoses per minute, time-to-final diagnosis, documentation times), viewing pattern behavior, biomarker editing behavior, and qualitative usability feedback.
- Deriving design implications for *evidence-forward* clinical AI: align to expert attention, bound language to a shared ontology, expose controls (toggle/edits), and evaluate beyond aggregate accuracy.

We evaluate these two components separately in controlled experiments with ophthalmology residents (User Study 2) to isolate their individual effects on diagnostic accuracy, workflow efficiency, and trust calibration before combining them in future integrated deployment. This separation allows us to establish baseline efficacy and identify potential issues (e.g., anchoring bias, attention interference) for each modality independently.

2 Background

2.1 Clinical Context and Challenges in Medical Imaging Diagnosis

Ophthalmic diagnosis for wAMD is heavily OCT-reliant and high-volume; subtle, fine-grained biomarkers (e.g., drusen vs. PED vs. CNV) drive treatment decisions and documentation, yet are difficult to perceive reliably and consistently at speed [10, 21, 34].

Decades of medical image-perception research show that expertise shapes visual search, with experts employing distinct, often idiosyncratic heuristics that novices struggle to emulate [27, 52]. While sharing an expert's gaze can sometimes aid detection [31], direct gaze transfer is operationally brittle: expert scanpaths are highly variable, often unavailable for new cases, and difficult to standardize [56]. This motivates shifting from manual gaze replay to *automated AOI augmentation*. By training a model to predict fixation density from image features, we can generate generalized attention maps for unseen patient images, scaling expert-like guidance without requiring an expert to view every scan.

Meanwhile, AI for imaging shows strong technical performance but uneven clinical fit. Structured reporting can improve clarity yet remains variably adopted [14], and qualitative studies document workflow frictions, trust calibration issues, and unclear utility when AI outputs are not designed for clinicians' intermediate reasoning steps [15, 16, 26, 55, 58]. Human-centered guidance argues that explanation alone is insufficient without integrability and control added third citation [8, 18, 48]. Together, these gaps—perceptual complexity, impracticality of gaze transfer, and human–AI misfit—motivate *evidence-forward* tools that (i) algorithmically highlight areas experts attend to and (ii) scaffold documentation with ontology-bounded language that clinicians can edit.

[1] Anonymized for review.

2.2 Human–AI Collaboration in Clinical Decision Making

There is growing consensus that clinical AI should collaborate with experts rather than replace them, supporting intermediate reasoning (e.g., where to look and what evidence to consider) and leaving judgment to clinicians [2, 7, 20, 59]. In a case study on sepsis detection [59], a binary risk score was abandoned in favor of extending the tool with trend projections, uncertainty views, and actionable next steps. This modification enabled more effective collaboration because clinicians could incorporate the AI into their reasoning rather than being forced to trust its output. This echoes guidance that interaction should be iterative, timely, and controllable, not one-shot [2, 19, 20].

Critically, simply mixing human and AI decisions does not guarantee performance gains. A large meta-analysis showed that human–AI teams often fail to outperform the stronger agent alone due to miscalibrated trust and poor interaction design [51]. Explanations and transparency can also backfire if not aligned to tasks and timing. Participants may over-accept AI advice or be distracted without improving accuracy [4, 6, 36, 49]. Recent clinical studies underscore the need for trust calibration: clinicians should rely on the AI more when it is right and discount it when it is wrong [37, 38, 40].

In addition to accuracy, workflow fit is paramount to effective implementation. Empirical HCI and imaging studies report friction when AI outputs do not match clinicians' reading stages or documentation needs, and benefits when integration is thoughtful [7, 12, 44, 47, 55]. For instance, dentists activated inspection-assistance AI only in a quarter of cases and mostly after forming an initial impression, using AI as a mid-stream second opinion. When the AI guidance was used, gaze shifted off the image, increasing time without accuracy gains [7]. Conversely, when designed for complementary roles, human–AI teams can exceed either alone in dermatology, pathology, and endoscopy [25, 50].

Taken together, HCI and clinical evidence motivate co-annotation interfaces that scaffold the diagnostic process. In particular, timely, editable user-interface components can help clinicians decide *where to look* and *what to write*, communicate uncertainty, and preserve control [2, 19]. Our approach operationalizes this via fixation-aligned AOIs and ontology-bounded, editable biomarker drafts (Section 1).

2.3 Augmenting AI with Expert Knowledge: Gaze-Guided Models and Ontology-Bounded LLMs

Bridging black-box AI and clinical transparency increasingly relies on injecting expert signals and domain knowledge into model training and interaction. Eye tracking captures where readers focus during interpretation and can serve as weak supervision with low annotation burden [13]. Empirically, expert fixations align with subsequently reported findings, motivating models that learn to attend where experts look [32]. Recent studies show that gaze-supervised models can improve lesion detection and align saliency with clinically meaningful regions [23, 53, 54]. Beyond single-modality convolutional neural networks (CNNs), multi-modal methods use radiologists' gaze to better align image features with report text, improving text–image retrieval [32]. ViT variants that explicitly align attention maps to gaze report concurrent gains in accuracy and interpretability [9]. Performance depends on careful integration, however. Some pipelines find mixed effects without careful preprocessing or task fit, underscoring the need to model gaze as supervision rather than post hoc decoration [22].

In parallel, large language models (LLMs) in medicine require knowledge grounding to curb hallucinations and standardize outputs. Augmenting LLMs with the United Medical Language System (UMLS) [5] or knowledge graphs constrains generated output to clinically valid concepts and relationships [57]. For imaging, structured report ontologies (e.g., RadGraph/RadGraph-XL) support entity–relation grounding and fact-aware report generation/rewarding [24]. Ontology-constrained decoding and retrieval-augmented prompting further reduce off-ontology drift and make outputs

easier to edit [33, 45]. Together, gaze-supervised vision and ontology-bounded language point to *co-annotation* designs: models that point out where experts would look (AOIs) and draft what experts would write (biomarkers) within a shared ontology that the clinician can accept or edit.

Despite significant promise, most prior work optimizes either gaze alignment or knowledge grounding in isolation, often outside realistic reading and documentation loops [39]. Our approach unifies both: we learn fixation-aligned AOIs from expert gaze and fine-tune an ontology-bounded VLM on expert dictations, yielding editable, evidence-forward assistance that fits the OCT workflow (Section 1).

3 Methods

Our system development follows a three-stage pipeline: (1) **Data Collection** (Section 3.1): capturing expert gaze and dictation to serve as ground truth; (2) **Model Training** (Section 3.2-3.3): training the ViT and VLM to mimic these expert behaviors; and (3) **Interface Integration** (Section 3.4): embedding these models into a clinician-facing tool (Figure 3).

Due to the limits of generalist VLMs in providing diagnostic reasoning, as well as the relative scarcity of rich, high-quality datasets for this task, we build our own corpus through **User Study 1** (US1). We then investigate whether the two-part system yields the benefits outlined in Section 1 through **User Study 2** (US2). Because of the limited dataset size due to the small number of available experts in US1, we employ a two-stage hierarchical training curriculum to ensure optimal transfer efficiency.

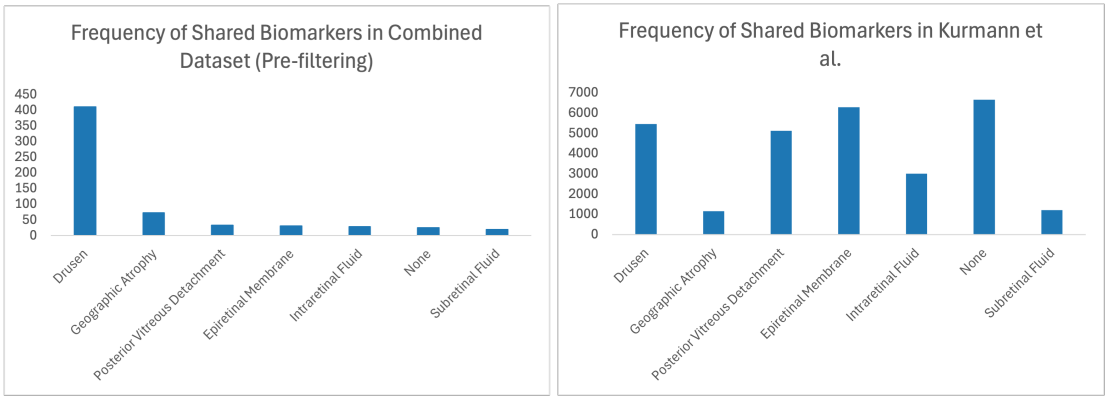
3.1 From Expert Behavior to Distillation Targets

Experts' eye movements and spoken dictations were transformed into *supervision targets* used to (i) train our gaze-aligned vision models and (ii) fine-tune and evaluate the biomarker-bounded VLM in US1. These expert-derived artifacts form the foundation of our system and serve as the substrate for validating our hypotheses in US1 (Section 5.1).

3.1.1 In-house Dataset. We collected **1,155** high-definition 5-line raster OCT scans of the macula from **231** eyes in **203** patients (five scans per eye; **104** normal, **127** wAMD). Scans were captured on the Zeiss Cirrus OCT imaging platform, which is likely the most commonly used for retinal imaging [1]. All data were collected in accordance with the principles laid out in the Declaration of Helsinki under a protocol approved by our Institutional Review Board. The data was de-identified according to national law. We will refer to this dataset as the *in-house dataset*. We collected visual attention from eight experts on a subset of 138 images, and dictations on a subset of 113 (see section 5).

3.1.2 Visual Attention. For each viewed image, fixation events were detected using velocity-thresholding [42] mapped to screen coordinates and converted to a fixation-density heatmap via kernel density estimation. Heatmaps were normalized to unit mass (probability maps), then downsampled by area interpolation to the ViT patch grid (32×32), yielding a rasterized attention target A^* for that image. We note that A^* represents the cumulative expert attention over the entire diagnostic review of the image; it links gaze to the holistic diagnosis rather than to isolated diagnostic steps or individual biomarker tokens. The resulting A^* maps directly compare with attention rollout on the same patch lattice (Section 3.2).

3.1.3 Expert Dictation and Clinical Biomarkers. Audio was recorded continuously while experts examined images in the *Single Image Input* setting and was time-synced to the active image. Recordings were transcribed to text and lightly cleaned by experts to remove noise while preserving clinical content.



(a) Transcripts + OCT-5K distribution

(b) Kurmann et al.

Fig. 2. Comparison of biomarker distributions across Transcripts+OCT-5K and other literature.

First, we converted the free-text dictations into multi-label biomarker targets. These biomarkers were curated from the diagnostic categories in OCT5k, a public OCT corpus [3]. To convert free-text dictations to multi-label biomarker targets, we binarized the words within each dictation using predefined prompt rules. The model was given only the clinician transcript (and, if available, a brief caption **see commented out section below, maybe remove** typed by the expert at the time of data collection to ensure the accuracy of the data produced) and was required to output biomarkers from an exclusive list of pathological symptoms. Only presence cues tied to explicit nouns/phrases were considered, and modifiers such as severity adjectives (e.g., "mild", "significant") were disregarded to ensure uniform labels. A biomarker was included only if it was explicitly supported by the transcript or caption. The model was instructed to avoid speculation/inferred findings, and if no biomarker was supported, it returned 'None'. Outputs were subsequently reviewed by a retina specialist and no changes were made to the final dataset constructed. Finally, we enhanced the in-house dataset by combining the 113 image-biomarker pairs from US1 with a subset of OCT5k to form a total of 573 pairs. The full distribution of the biomarkers is provided in Appendix A.3

To manage label complexity, biomarkers with fewer than 10 occurrences were excluded, resulting in a final set of 12 predefined biomarkers: "Drusen", "Photoreceptor Degeneration", "Pigment Epithelial Detachment", "Geographic Atrophy", "Choroidal Fold", "Epiretinal Membrane", "Hyperfluorescent Spots", "Intraretinal Fluid", "Posterior Vitreous Detachment", "Fluid", "Subretinal Fluid", and "Choroidal Neovascularization". As a benchmark, we compare our biomarker distribution to an external population-level baseline to contextualize the diversity and representativeness of our dataset [28]. Our dataset contains additional complex OCT findings (e.g., CNV, PED), but because these are heterogeneous and not represented in the atomic biomarker ontology of Kurmann et al., we compare distributions only across the shared atomic biomarkers to ensure meaningful cross-dataset comparison. The two tables are below, and their full distributions for both the combined dataset and Kurmann et al. can be found in Appendix A.3 **Maybe don't use this, have Benji look if this is a valid reason** Because our dataset is derived from real-world retina visits and includes transcript-derived labels, its biomarker prevalence differs from the population-level distribution reported in Kurmann et al. These differences reflect underlying clinical case mix rather than inconsistencies in extraction, and our comparison is intended to contextualize diversity rather than assert distributional equivalence. This phase of US1 ultimately produced the Expert Distillation Corpus: (1) patch-aligned attention targets A^* per image (for gaze-aligned

supervision), (2) ontology-bounded biomarker multi-labels, and (3) eight caption variants per image grounded in those biomarkers. We use this corpus to train the expert-aligned ViT, fine-tune the biomarker-bounded VLM, and quantify model–expert correspondence in US1.

3.2 Distilling expert knowledge via expert-alignment loss

To ensure transparent and clinically grounded visual saliency, we trained a ViT to align its attention to expert fixations using a multi-task learning objective that combines classification with gaze alignment. The vision backbone is a vanilla ViT-Base/16 [11] with 12 transformer layers, 768-dimensional embeddings, and 16×16 pixel patches, yielding a 32×32 patch grid for 512×512 input images. We initialize from ImageNet-21k pretrained weights and fine-tune end-to-end for 100 epochs. Complete architecture specifications, hyperparameters, and training configuration are provided in Supplement Section 3.2.. At each training step, the ViT produces two outputs: (i) a classification prediction $\hat{y}$ for the binary diagnosis task (normal vs. wAMD), and (ii) an attention map $\hat{A}^{(L)}(x)$ extracted from the last transformer layer via attention rollout, a method that recursively aggregates attention weights across all 12 layers to compute the effective attention from input patches to the output (detailed algorithm in Supplement Section 3.1).

Let A^* denote the expert fixation-density target on the 32×32 ViT patch grid (Section 3.1). We penalize divergence between the model's predicted attention distribution $\hat{A}^{(L)}(x)$ and the expert target A^* using a cross-entropy loss:

$$\mathcal{L}_{\text{align}} = \text{CrossEntropy}\big(\hat{A}^{(L)}(x),\, A^*\big) \tag{1}$$

The total training loss is a weighted combination of the binary cross-entropy classification loss $\mathcal{L}_{\text{cls}}$ and the attention-alignment objective:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \alpha \cdot \mathcal{L}_{\text{align}} = \text{BCE}(\hat{y}, y) + \alpha \cdot \text{CrossEntropy}\big(\hat{A}^{(L)}(x),\, A^*\big) \tag{2}$$

where $\alpha \in \{0, 0.01, 0.05, 0.1, 0.2, 0.3\}$ controls the strength of gaze supervision. Intuitively, this encourages the network to concentrate probability mass on regions that experts fixated, while retaining strong diagnostic performance.

Crucially, this objective allows the ViT to learn a generalized mapping between visual features (e.g., fluid pockets, hyperreflective spots) and expert attention. At inference time, the trained model predicts fixation-aligned AOIs for *unseen images*—generating guidance for new patients without requiring concurrent expert eye-tracking.

3.3 Two-Stage VLM Training for Ophthalmic Diagnosis

Having aligned the ViT's visual attention to expert gaze, we now describe the complementary language component. We finetune MedGemma [43], a large-scale VLM pretrained on diverse medical datasets, to specialize it at interpreting ophthalmic OCT scans. Our framework produces a single, unified model trained to perform three tasks: (1) diagnosis, (2) biomarker discrimination, and (3) biomarker identification (Table 1). To ensure clinical utility, the model interaction is governed by a structured prompting methodology to minimize ambiguity and a strict overarching system prompt (Table 1). We designed a two-staged curriculum to help the model to first build a robust and domain-specific visual foundation (for tasks 1 and 2), before attempting the more challenging task of generating biomarkers (task 3).

3.3.1 Stage One: Foundation Training. To first encourage specialization on OCT data, we first trained the MedGemma model exclusively on two "easier", albeit large-corpus tasks: (1) detecting the presence of wAMD, and (2) deciding

Table 1. VLM tasks, objectives, outputs, and prompt placeholders.

Task	Objective	Output Format	Prompt
Diagnosis	Binary classification of the OCT scan's primary pathology.	`Normal` or `wet-AMD`	What is the diagnosis for this OCT scan, Normal or wet-AMD?
Biomarker Discrimination	Verify presence of a queried biomarker (VQA-style).	`Yes` or `No`	Is <BIOMARKER> present in this OCT scan?
Biomarker Identification/ Generation	List all present biomarkers from the ontology, or none.	Comma-separated ontology terms or `None`	What biomarkers are visible in this OCT scan?

System prompt (shared across tasks): You are an expert ophthalmic AI assistant. Analyze the provided retinal OCT scan. Respond concisely and accurately, sticking strictly to the requested format without additional explanation or conversational text.
You are a specialized ophthalmic AI assistant. Your function is to analyze retinal OCT scans with high precision.
For diagnostic queries, provide only the diagnosis (e.g., 'wet-AMD' or 'Normal').
For biomarker identification, provide only a comma-separated list of findings (e.g., 'Drusen, Subretinal fluid') or 'None' if no biomarkers are present.
For direct questions, answer only with 'Yes' or 'No'.
Do not add any introductory phrases, explanations, or disclaimers.

whether a specific biomarker was present in the image. The goal was to transfer the pretrained model's domain to the OCT domain, retaining its diagnostic accuracy while learning salient features of ophthalmic pathology.

For the diagnosis task (1), we combined our in-house datasets as discussed in Section 3.1 with OCT-C8 [46], a multi-class dataset of 24000 high-quality retinal OCT images categorized into eight retinal conditions. We used a subset of OCT-C8, comprising a total of 4,600 OCT images, balanced with 2,300 samples each from the 'Normal' and'wet-AMD' classes.

The biomarker discrimination task (2) employs the Expert Distillation Corpus from US1. Compared to the diagnosis task, the corpus provides semantically richer training signals for improved visual grounding. Our initial experiments revealed that an unbalanced dataset would be overwhelmingly populated with "No" answers to the biomarker discrimination queries. To circumvent this, we curated dataset of 1,600 samples by randomly selecting an equal number of "Yes" and "No" instances for each of the most common biomarkers identified in US1 (Section 3.1).

3.3.2 Stage Two: Unified Training. After stage one, we included the biomarker generation task (3): identifying all relevant pathological OCT imaging signs that may assist an ophthalmologist in making a diagnosis. The Expert Distillation Corpus from US1 was again employed for this task (Section 3.1). A balanced sampler was configured to ensure a uniform mixture of all tasks within each training batch, resulting in heavy oversampling of the smaller biomarker dataset. This prevents forgetting of previously-learned tasks while forcing the model to map previously learned visual features to the generated semantically rich biomarker labels.

3.4 User Interface Design Rationale

The Co-Annotator interface (Figure 3) translates the model outputs into interactive clinical decision support. It consists of two primary components corresponding to our training pipeline:

- **Visual Guidance (from ViT, Sec 3.2):** A toggleable heatmap overlay derived from the gaze-aligned ViT attention map. The AOI blend mode and color were chosen...

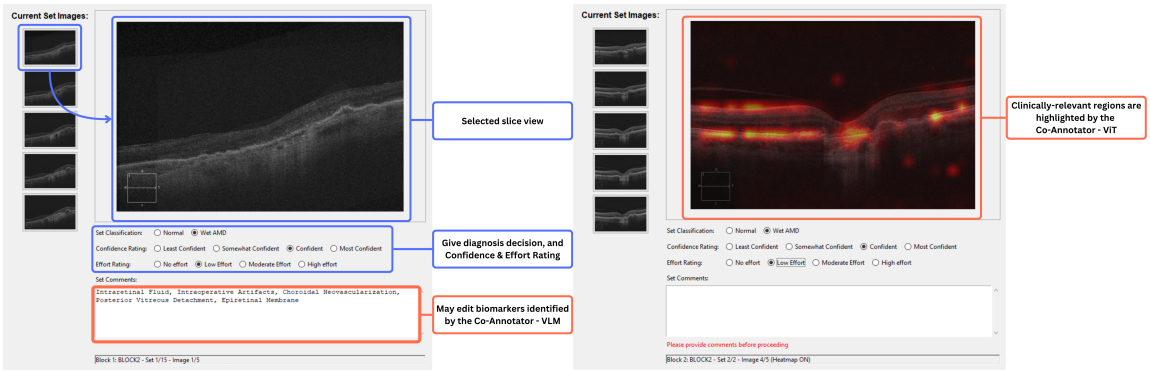


Fig. 3. The Co-Annotator Interfaces. Left: VLM Guidance (documentation support) with editable biomarker text. Right: ViT Guidance (visual search support) with gaze-aligned attention heatmaps.

- **Documentation Guidance (from VLM, Sec 3.3):** An editable text box pre-filled with the VLM's biomarker draft...

The interface employed in US2 was designed to closely mimic the viewing conditions that residents are already accustomed to, to strive for maximally realistic clinical application and to isolate effects of the guidance itself. Thumbnails were also displayed to provide experienced readers with a view of all five images in a set at once. Radio buttons were employed for rapid classification and scoring of image sets, and the free-text comment box allowed for biomarker summary collection. Participants had to fill every field before advancing to the next image set for data completeness. They also could not return to previously-rated set, to maximize independence of reviews. The interface also collected a granular record of timestamped interactions for measurement of outcome variables. The AOI blend mode and color were chosen to be quickly identifiable while avoiding entirely masking the underlying features. AOI overlays were upscaled from ViT rollout using linear interpolation.

4 Data and model availability

The anonymized Expert Distillation Corpus collected from US1 will be released at a later date (pursuant to CHI submission anonymization requirements). For reproducibility, we have released the fine-tuned model weights (**HuggingFace link anonymized**.) The source code for training and evaluation is available at **GitHub link anonymized**.

5 User Study 1: Distilling Expert Knowledge via Gaze and Dictation

Capturing expert data on this problem domain is necessary due to the lack of open-source, high-quality, grouped dictation and eye tracking data. Generalist models, even those pre-trained on multi-modal biomedical data such as MedGemma, do not translate well to this domain. In general, we seek to evaluate whether expert behavioral signals—visual fixations and concurrent dictations—can be faithfully captured and used to (i) align model attention with expert gaze and (ii) improve OCT diagnosis while laying the foundation for biomarker-bounded language guidance.

5.1 Hypotheses

For this user study, we formulated the following two hypotheses:

H1.1 (Attention faithfulness): A ViT trained with fixation-alignment loss performs better in terms of diagnostic accuracy when compared to a baseline model.

H1.2 (Ontology-grounded language): A biomarker ontology derived from expert dictations provides effective supervision and assessment targets for the VLM's biomarker text generation as measured by semantic fidelity to reference text (e.g., BERTScore/MedBERTScore).

5.2 Participants, Study Design, and Apparatus

Eight retina specialists at our organization provided their clinical reads of the images in the in-house dataset (Section 3.1.1). To best gauge downstream knowledge distillation, we employed two complementary setups:

- **Setup A: Five-image bundles, without dictation.** Experts reviewed the per-eye five-image sets without dictation, while their eye movements were recorded. They advanced through the five images at their own pace using the keyboard and mouse. This approach aimed to maximize the volume and diversity of fixation-derived Areas of Interest (AOIs) to supervise ViT attention alignment at scale.
- **Setup B: Single image, with concurrent dictation.** Experts reviewed one image at a time and provided free-form dictation describing the presence or absence of biomarkers, as well as any diagnostic impressions. The goal was to collect high-quality, fine-grained language aligned to image evidence to train the biomarker-bounded VLM. For this setup, experiment, we encouraged the participants to dwell longer and conduct more deliberate examination of each image than in Setup A.

In both setups, a brief calibration and practice block preceded the recorded block. No hard time limits were imposed; participants could proceed at a self-selected pace.

Eye movements were recorded with a *Tobii Pro Fusion* eye tracker at a 250 Hz sample rate, mounted on the bezel of a 1920×1080 monitor with a 60 Hz refresh rate. A standard 9-point calibration was performed for every participant prior to data collection (max error $< 0.5^\circ$). Participants sat comfortably in front of the display and interacted via keyboard and mouse. Dictation audio (Setup B) was captured using an *Apple AirPods* microphone and time-stamped alongside gaze and stimulus events. OCT images were presented in a custom viewing interface that logged stimulus onsets, image indices, and user interactions to enable precise temporal alignment.

5.3 Results and Discussion

As discussed in Section 3.1, we collected expert visual attention-based eye-tracking for **138** OCT images, and expert dictations (later converted to a list of biomarkers) for 113 images. The OCT scan-dictation pairs are enhanced with samples from OCT5k [3], from which we derived the final 573 samples.

5.3.1 Expert visual attention improves model diagnostic accuracy. To evaluate **H1.1**, we computed the diagnostic accuracy of the ViT with its attention biased toward clinically meaningful regions, using the auxiliary *expert-alignment loss* (Section 3.2). We did so across a range of alignment weights, α, to vary the extent to which expert fixation factors in the ViT training. Because each eye's scan has five corresponding layers, it is important to test **H1.1** in either single or multi-layer cases, corresponding to two input regimes:

(1) **Patient-level (multi-image):** All 5 OCT images per eye concatenated and jointly processed to mirror what a clinician sees.
(2) **Single image:** Only single images with expert visual attention data. Each image is treated as an independent sample.

Table 2. ViT performance vs. alignment weight α (averaged across 5 folds).

α (alignment weight)	Validation Accuracy	Validation F1-Score
0 (no attention-alignment)	0.8607	0.8380
0.01	0.8403	0.8073
0.05	**0.8837**	**0.8628**
0.1	0.8347	0.8140
0.2	0.8687	0.8500
0.3	0.8630	0.8429

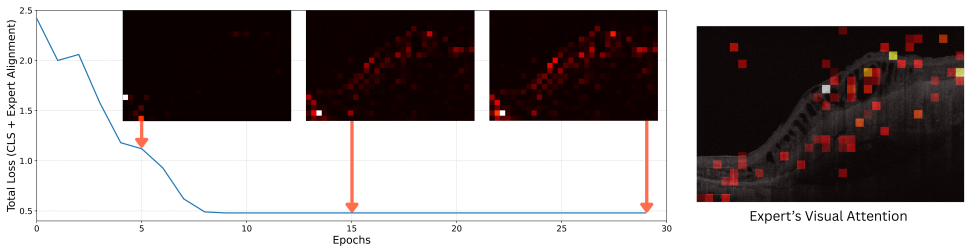


Fig. 4. Gaze-aligned attention converges during training. The curve shows total loss (classification + expert-alignment; Sec. 3.2) across epochs. Insets display the ViT's last-layer attention rollout (normalized on the 32×32 patch grid) across epochs (5, 15, and final) for the same image; The right panel displays the target expert visual attention A^* derived from a single expert's raw fixation density (including minor peripheral calibration artifacts/noise). Early attention is diffuse and off-target; as training progresses, the model's attention concentrates on the same structures emphasized by experts, while the loss improves.

For the *patient-level input*, attention alignment consistently improved diagnostic accuracy. Adding gaze alignment raised **micro-AUC from 0.95 to 0.98** with five-fold cross-validation. The amount of data is limited under this setting, as not all images in each set of 5 had expert visual attention data. Thus, we focus on the results in the **single–image** setting. The best trade-off occurred at $\alpha = 0.05$, yielding **88.37%** validation accuracy and **F1 = 0.8628**, outperforming the baseline ViT without alignment (Table 2). The improved classification supports **H1.1**. Qualitatively, as in Figure 4, model saliency resembled expert fixation distributions, which is consistent with the intuition that experts' visual attention can cue the model to attend to clinically relevant regions.

5.3.2 Dictation-Derived Ontology Grounds Accurate Biomarker Generation. We evaluated the efficacy of the two-stage VLM training across the three tasks described in Section 3.3. The model achieved high accuracy on discriminative tasks. Diagnostic accuracy was **0.920** on the OCT-C8 held-out test set and **0.910** on a larger **US1** test set (Table 3). Biomarker discrimination had an accuracy of **0.800**. (Table 3). For generative biomarker identification, the model attained strong semantic fidelity to dictation-derived references ($\textbf{BERTScore}_{\textbf{F1}} = \textbf{0.880}$, $\textbf{MedBERTScore}_{\textbf{F1}} = \textbf{0.867}$). An ablation study revealed that integrating OCT-5K with our Expert Distilled Corpus was critical; training only on the smaller transcript dataset led to marked overfitting.

Taken together, these results support **H1.2**: the dictation-derived, ontology-bounded biomarker targets served as effective supervision and assessment signals for biomarker text generation, yielding high semantic similarity while maintaining strong discriminative performance.

Table 3. Performance of the Fine-Tuned MedGemma Model on all three tasks after the two-stage training. To gauge generalization, diagnosis performance is evaluated on OCT-C8 [46] and US1 data. Base refers to the pretrained MedGemma without finetuning.

Task Category	Task	Metric	finetuned on OCT5K + US1	finetuned on US1	Base
Binary Choice	Diagnosis (US 1)	Accuracy	0.910	0.800	0.290
	Diagnosis (OCT-C8 [46])	Accuracy	0.920	0.900	0.400
	Biomarker Discrimination	Accuracy	0.800	0.840	0.535
Generative	Biomarker Generation	BERT Score (F1)	0.880	0.818	0.790
		MedBERT Score (F1)	0.867	0.774	0.698

5.3.3 Bridge to clinical deployment. Expert attention alignment and dictation-derived ontology jointly validate our hypotheses: the attention-aligned ViT improved diagnosis (single–image F1 up to **0.8628** at α=0.05 from the baseline of 0.8380; patient-level micro-AUC **0.98** vs. 0.95), and the two-stage, ontology-bound VLM produced high-fidelity biomarker text ($\text{BERTScore}_{\text{F1}}$ **0.880**, $\text{MedBERTScore}_{\text{F1}}$ **0.867**) while maintaining strong diagnostic accuracy (**0.91–0.92**). These results motivate deployment with ophthalmology residents to answer practical questions in clinical workflow: Do fixation-aligned AOIs change gaze behavior and throughput without harming accuracy? Do editable, ontology-bounded VLM drafts reduce documentation effort, align vocabulary, and preserve (or improve) diagnostic performance and perceived trust? We examine these in US2.

6 User Study 2: Improving Resident Clinical Workflow

US1 established that expert-aligned visual attention improves diagnostic classification and that dictation-derived, ontology-bounded language yields high-fidelity biomarker text. As a pilot deployment with trainees, we intentionally evaluated the two guidance modalities in two separate experiments, one testing AOI overlays (Experiment 1) and one testing VLM drafts (Experiment 2), rather than a joint AOI + VLM condition. This separation (i) isolates causal mechanisms (perceptual guidance for where to look versus documentation guidance for what to write) and avoids interaction confounds such as text anchoring gaze or AOIs biasing language; (ii) preserves statistical power and keeps sessions clinically realistic in a small, fixed resident pool—adding a joint arm would either dilute per-condition samples or inflate the session time beyond 40 minutes; and (iii) follows a safety-first gating strategy, requiring each component to demonstrate accuracy, non-inferiority, and appropriate trust calibration before combining them in future integrated deployment. Practically, simultaneous eye-tracking during active text editing also introduces measurement interference (cursor/typing artifacts) that can distort gaze metrics. With these considerations, US2 uses two experiments to assess whether the use of different Co-Annotator components (visual and textual) affects diagnostic performance, efficiency, viewing behavior, and documentation practices.

6.1 Hypotheses

We evaluated the following hypotheses in **US2**, drawing from insights from **US1** and covering major aspects of the clinical workflow:

H2.1 (Accuracy non-inferiority): Across both guidance modalities, diagnostic accuracy is non-inferior to unguided reading within a margin of δ=0.10 at the eye-level.

H2.2 (In-guidance efficiency): Exposure to either guidance condition increases diagnostic workflow efficiency and confidence during the guidance block.

H2.3 (Post-guidance efficiency): Efficiency and confidence in the post-guidance block improve with respect to baseline (i.e., a short-term efficiency carryover).

H2.4 (Inter-rater agreement): Guidance increases agreement between multiple residents reading the same images.

H2.5 (Gaze–AOI compatibility): During AOI blocks, resident gaze–AOI divergence is comparable to baseline, i.e., there is alignment between natural viewing behavior and the AOI-suggested regions.

H2.6 (VLM biomarker breadth): With VLM guidance, residents identify a larger number of distinct biomarkers per AMD eye than in the control condition.

H2.7 (VLM biomarker retention): Residents retain ≥70% of model-suggested biomarkers after edits.

H2.8 (Vocabulary alignment) VLM guidance increases overlap between resident- and model-vocabulary for biomarkers.

Efficiency as discussed in **H2.2** and **H2.3** was measured with variables described in Section 6.2. **H2.3** was not tested in the VLM-guidance condition for reasons described in Section 6.2.

6.2 Participants, Study Design, and Apparatus

Residents were recruited from a pool of 16 ophthalmology trainees. Power analyses targeted the detection of a 0.10 absolute accuracy difference at the eye level and medium effects on timing and behavioral measures (Cohen's $d = 0.5$), translating to image-set sample goals of n=74 for accuracy-focused tests and n=128 for other outcomes, achievable with 5–9 residents reading ~15 eyes per block. Ten residents completed the AOI study and four completed the VLM study (four overlapped), totaling 11 unique participants. This represents 69% of the total possible recruiting pool.

6.2.1 Experimental protocol. Experiments used a lightweight Python interface,[2] matching clinical viewing conventions (Fig. **??**,**??**). Each eye-level case comprised five images at distinct anatomical cross-sections (Fig. 5a). The block sequence varied by guidance condition. To model the standard clinical workflow ("Standard of Care"), each experiment included unguided control blocks where residents reviewed images without AI augmentation, relying solely on their visual inspection as they would in current practice. The AOI-guidance condition is therefore consisted of three blocks: one unguided control (pre), one with the AOI overlay as toggleable heatmaps, and another unguided control (post). Images were sampled without replacement in the first block to maximize coverage; subsequent blocks were sampled from the remaining pool. The VLM-guidance condition consisted of two blocks: an unguided control and a VLM-guided block. In the guidance block, the free-text box was pre-filled with the VLM's biomarker draft aggregated across the five scans for the model's majority diagnosis (defaulting to "No significant abnormality." when the model predicted a "Normal" diagnosis). Residents edited the draft for accuracy, ensuring, as in a real clinical scenario, that the final list of biomarkers aligned with their own assessment (Figure **??**). Block orders were fixed for the AOI condition to measure carryover and for the VLM condition to bound session length < 40 minutes.

All experiments began with a short practice run. Eye movements were captured with a *Tobii Pro Fusion* eye tracker sampling at 250 Hz on a 1920×1080 monitor with a 60 Hz refresh rate. A keyboard and mouse were used for navigation. Participants could navigate through images for each eye with the scroll wheel or by selecting them from their thumbnails (Figure **??**). The heatmap could be toggled with the right mouse button. For each eye, residents (i) reviewed the five images, (ii) provided a binary diagnosis (normal or wAMD) with confidence rating, and (iii) wrote or edited a brief biomarker summary.

[2]Web demo anonymized for review

6.2.2 Outcome variables. The **primary outcome** variable was accuracy at the eye level, accompanied by false positive and false negative rates, to assess **H2.1**. The other **secondary outcomes** of efficiency in **H2.2** and **H2.3** were assessed with time per eye, time to final diagnosis, number of correct diagnoses per minute, and comment edit time. Confidence and perceived effort were also measured on a four-point scale. Inter-rater agreement was computed with Cohen's κ on cases read by multiple residents for **H2.4**. In the AOI condition, gaze–AOI divergence was computed by relative entropy between normalized resident fixation density and model-generated AOI map (Eq. 3), characterizing **H2.5**. In the VLM-guidance condition, biomarker retention was defined as the fraction of VLM-suggested items kept after edits (**H2.7**). Explicit corrections ("negations") and deletions were also tracked, as well as the proportion of out-of-ontology items ("hallucination rate"). Vocabulary overlap was computed with the Jaccard index (**H2.8**).

$$\text{relative entropy}(x, y) = \begin{cases} x \log(x/y) & x > 0, y > 0 \\ 0 & x = 0, y \geq 0 \\ \infty & \text{otherwise} \end{cases}. \tag{3}$$

For each outcome, we compared within-subject distributions across blocks. Accuracy non-inferiority (H2.1) used a one-sided test with margin δ=0.10. For timing/behavioral measures, we used non-parametric tests (Wilcoxon signed-rank within-subjects; rank-sum if independent), applying Benjamini–Hochberg correction across families of related tests (α=0.05). As sensitivity analyses, we performed mixed-effects modeling, allowing for random by-participant effects on the diagnostic correctness while controlling for ground truth classification (normal vs wet AMD) and image position in the experimental block (to control for possible fatigue effects). The latter two were considered fixed effects to ensure model convergence. Likelihood ratio testing was performed to identify which variables had significant effects on model fit. We similarly modeled time spent per eye.

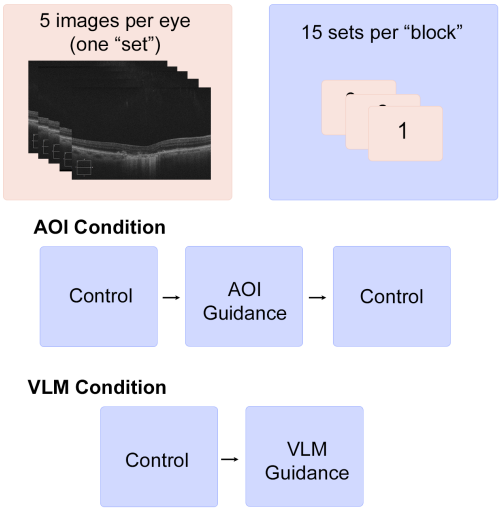
6.3 Quantitative Results & Discussion

Results from **US2** were analyzed as described above.

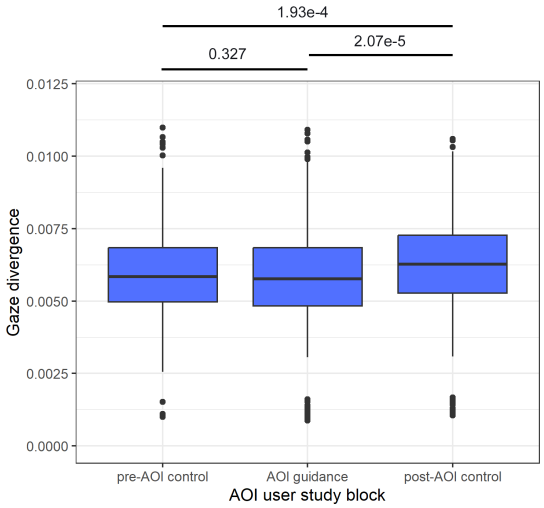
6.3.1 AOI Guidance. In the AOI-guidance experiment, participants had a median of 24 months of residency training. Diagnostic accuracy with AOI guidance was non-inferior to control (**H2.1**), as accuracy was comparable (within $\delta = 0.1$) across experiment blocks and the study was powered to detect such a difference (Table 4). There was no significant difference in confidence or time to final diagnosis. There was, however, significantly increased speed in the post-guidance block, in terms of time spent per eye and time to diagnosis (Table 4, S7; $p_{pre-post} = 1.34 \times 10^{-4}$ and $p_{pre-post} = 1.55 \times 10^{-4}$, respectively), supporting **H2.3**.

Mixed effects modeling identified wAMD images as a significant predictor of diagnostic accuracy (as these are likely the greatest source of difficulty for participants). Controlling for image order and allowing for random participant effects, AOI guidance was likewise associated with a mild ($log - odds = 0.647$), but non-significant increase in accuracy (Table S8). Though mixed effects modeling does suggest a task order effect, with later images being read faster, the post-guidance control was still on average faster that pre-guidance control by 11.3 seconds when controlling for each image's position in the overall experiment timeline (Table S9).

Participants reviewed the AOI guidance for roughly 40% of the time until arriving at their final diagnosis. Participants spent less time writing comments with each successive block of the AOI study, though comment editing time took 40–50% as long as the time to diagnosis across all blocks (Table 4). The appreciable proportion of time spent writing comments motivated the use of the VLM to generate comments automatically, to cut down on overall time per image



(a) Experimental block structure across US2.



(b) Gaze divergence by AOI study block.

Fig. 5. User Study 2 Interface and Results. (a) The two clinician-facing components of the Co-Annotator system, designed to assist with documentation (VLM) and visual search (ViT). Annotations highlight the interactive elements (toggles, editable text) that support human-in-the-loop diagnosis. (b) The study flow for the deployment with residents. (c) Analysis of gaze behavior showing residents' attention alignment with the model.

review. We also analyzed the alignment of resident gaze preferences with the generated AOIs by computing the divergence (or relative entropy) between spatial fixation times with model-generated attention maps.

Participants exhibited somewhat higher diagnostic agreement with each other in the AOI-augmentation condition (rating 24 image sets) compared to the post-AOI control ($\kappa_{AOI} = 0.66$ vs $\kappa_{control} = 0.48$), possibly supporting **H2.4**. Comparison to the pre-AOI control was not available due to the non-replacement sampling scheme mentioned above. Additionally, gaze divergence was similar between the control and guidance conditions, possibly suggesting that the guidance model learned fixation patterns comparable to the existing viewing patterns of trainees. This supports **H2.5**. Interestingly, divergence was significantly higher in the post-test block, perhaps owing to a different viewing pattern retained from the AOI guidance or the faster viewing rate due to increased comfort on return to the baseline task.

6.3.2 VLM Diagnostic Guidance. Participants in the VLM experiment had a median of 19.5 months of residency training. Accuracy was slightly higher in the VLM block than in the control, but this difference did not reach statistical significance. In any case, accuracy was non-inferior in this condition (**H2.1**), and mixed effects modeling agrees with a mild ($log - odds = 0.441$) but non-significant increase in diagnostic correctness when controlling for order and ground truth diagnosis (Table S8). There was no significant difference in confidence, participant-rated effort, or time to final diagnosis (**H2.2**, Table 4, S7). Time spent commenting on the images was overall longer in the VLM guidance condition compared to the AOI experiment, perhaps due to a priming effect when residents expected the text summary portion of the experiment to be more relevant than in the AOI condition. Within the VLM condition, comment time

was comparable between control and guidance blocks. Comments took a comparable proportion of the time per eye as in the AOI experiment, but a larger proportion of the time to diagnosis (Table 4).

Table 4. User study summary results across guidance conditions."Dx" = Diagnosis; * = significant relative to pre-guidance control and †= significant relative to guidance condition at $\alpha = 0.05$ level with Wilcoxon rank-sum test after Benjamini-Hochberg correction.

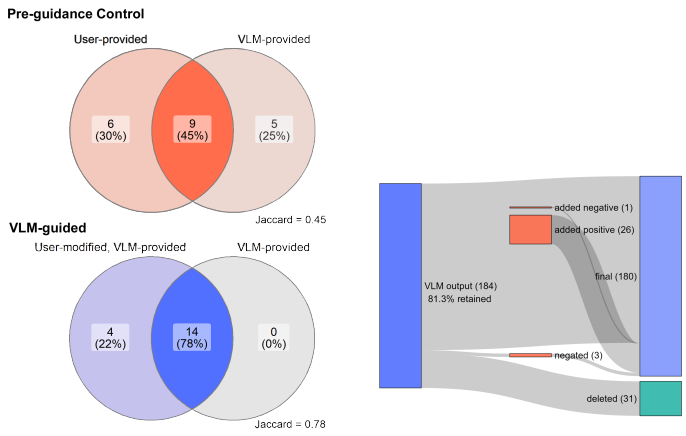
metric	pre-AOI control	AOI guidance	post-AOI control	pre-VLM control	VLM guidance
Accuracy	0.927 ± 0.097	0.887 ± 0.137	0.904 ± 0.121	0.833 ± 0.086	0.917 ± 0.084
FPR	0.000 ± 0.000	0.007 ± 0.021	0.007 ± 0.022	0.050 ± 0.100	0.033 ± 0.067
FNR	0.073 ± 0.097	0.107 ± 0.141	0.089 ± 0.125	0.117 ± 0.114	0.050 ± 0.033
Time per eye (s)	20 ± 14.4	18.9 ± 12.3	14.4 ± 11.1*,†	32.5 ± 17.8	30.1 ± 21.1
Correct Dx/min	2.9 ± 0.9	3.0 ± 1.0	4.4 ± 1.9	1.6 ± 0.6	2.0 ± 0.8
Comment edit time (s)	4.1 ± 8	3.3 ± 5.9	3.1 ± 5.7	11.3 ± 10.6	12.5 ± 15.2
Time to final Dx	10 ± 8.2	10.8 ± 8	7.4 ± 6.4*,†	14.2 ± 11.9	17.4 ± 14.9
Confidence (max 4)	3.17 ± 0.92	3.14 ± 0.9	3.12 ± 0.95	2.67 ± 0.88	2.83 ± 0.87
Heatmap on-time		4.56 ± 4.14			
Effort (max 4)				2.33 ± 0.88	2.27 ± 0.8

6.3.3 VLM Biomarker Guidance. In the control condition, users quoted nine unique biomarkers that VLM guidance would have provided, as well as six that it would not have provided (Figure 6a, Jaccard index = 0.45). The VLM quoted five unique biomarkers that the users did not mention (Figure 6a). In the VLM-guided condition, no biomarker was VLM-isolated, four were uniquely added by the user, and 14 were shared between the user and the VLM (Figure 6a, Jaccard index = 0.78). The greater degree of overlap in the VLM-guided condition suggests that the user adopted (or failed to alter) the VLM's biomarker vocabulary (**H2.8**). There seemed to be a moderately shared biomarker vocabulary at baseline, though with some disagreement. A total of 24 unique biomarkers were identified across all conditions of the experiment, with the participants providing 10 that were not in the VLM's scope. Three biomarkers provided by the VLM guidance were not in the allowed ontology, indicating that these were model hallucinations. In the 11 individual appearances of hallucinated biomarkers in the VLM guidance, participants deleted them five times (45%) and retained them for the remainder.

Across the 184 system-suggested biomarkers in the guidance block, residents actively curated the drafts: they deleted 31 items (16.8%), added 27 new items (15.0% of the final set), and negated three (1.6% of suggestions; 1.7% of the final list) (Figure 6b). Overall, residents retained 83.1% of suggested biomarkers in the summaries they submitted, meeting **H2.7** (Figure 6b). With guidance, residents documented significantly more biomarkers per AMD eye than in the control condition (mean 5.8 vs. 2.3, $p = 8.8 \times 10^{-9}$, Wilcoxon rank-sum), supporting **H2.6**. Distinct biomarkers were also listed at significantly different rates across conditions ($p = 8.9 \times 10^{-6}$, Fisher's exact; Figure 6c). Notably, residents detected epiretinal membrane (ERM) and posterior vitreous detachment (PVD) far more often with guidance—findings many had deprioritized in unguided notes as less central to a wet-AMD call. Taken together, residents used the guidance to broaden coverage and align vocabulary while keeping editorial control, preserving (and in some cases improving) diagnostic performance; when off-ontology or incorrect suggestions appeared, they typically identified and removed them (11 total appearances across three items; 5/11 deleted, 45%).

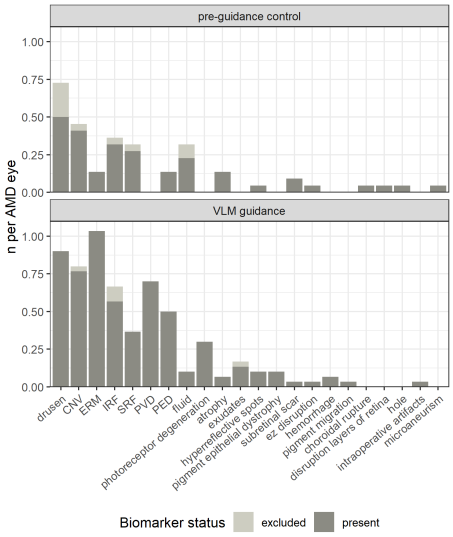
7 Discussion

Our post-experiment interviews and behavioral observations (Sections 7.1–7.3) reveal four key design principles for clinical AI guidance. Residents want assistance that is sparse and specific rather than diffuse like "condensing the



(a) Overlap in unique biomarkers across user study blocks

(b) Tracking fate of biomarkers after review by residents



(c) Distributions of identified biomarkers

Fig. 6. VLM guidance biomarker generation analysis

heat map into 2 or 3 main points of interest" (PGY-3), avoiding overlays where "too many areas were highlighted" (PGY-1). They want control and timing flexibility—many described a "read-first, check-later" rhythm, arriving at initial impressions before toggling guidance ("would answer first...then in hindsight", PGY-4). They value visible grounding in image evidence as three of four residents rated "highlighted areas of relevance from which the descriptions are drawn" as most helpful for VLM summaries. Finally, they need succinct, ontology-specific language rather than vague labels, criticizing "non-specific labels like 'fluid'..questionable for ERM and differentiating IRF vs SRF" (PGY-3). These preferences, consistently expressed across both guidance modalities, motivate the integrated visual-text design principles we propose in Section 7.6.

7.1 Resident experience and workflow fit

Residents naturally adopted a *read-first, check-later* rhythm. Several described arriving at an initial impression and only then toggling the guidance: "*would answer first (reading the image normally), then in hindsight*" (PGY-4); "*I looked at the summary and would cross check with the picture mostly*" (PGY-4). This pattern underscores the importance of **deferring assistance** until after a first unaided pass to avoid premature anchoring, and of making reveal/contrast actions fast and reversible. Where AOIs aligned with the resident's mental model, they reinforced attention allocation ("*pointed to areas of exudation and CNV… I knew which layers to focus on with more confidence*," PGY-3). Yet many turned the overlays off when they occluded structure or felt noisy: "*it's actually blocking the view because the heat map was very diffuse*" (PGY-2); "*I turned it off as it was getting in the way of my view*" (PGY-3); "*kill the noise*" (PGY-4). These responses explain our quantitative pattern—non-inferior accuracy with AOIs but limited in-block efficiency gains—and they motivate designs that respect the primacy of undistorted imagery while offering targeted cues on demand.

7.2 From saliency to suggestion: designing AOI overlays

Heatmaps were most acceptable when they behaved like *suggestions* rather than paints. Participants asked for fewer, smaller, and less opaque highlights—"*condensing the heat map into 2 or 3 main points of interest*" (PGY-3), and "*heat map was not specific enough, too many areas were highlighted*" (PGY-1). A side-by-side or thumbnail-plus-overlay compare was preferred to an overlay that covers only the view ("*overlay could be seen side-by-side with the original*," PGY-4). Together, these requests suggest an interaction contract: default to a clean image, reveal *few* AOIs with precise extents, keep their opacity adaptive to local contrast, and tie each hotspot to an interpretable rationale (e.g., "IRF candidate, 0.82 confidence") rather than a generic saliency map. Novices also asked for brief, upfront scaffolding about task criteria ("*basic guidance… what OCT features define wet AMD*," PGY-2), whereas seniors preferred minimal prompting—supporting profile-aware defaults that tune AOI density and explanations to experience.

7.3 Text summaries as editable scaffolds, not verdicts

Auto-generated summaries worked best as a *checklist and drafting aid.* Residents described using them to catch misses and to speed documentation: "*helped me double check that I had identified the key findings… if I saw something in the text box that I did not initially note, I would take a closer look*" (PGY-3); "*yes, honestly mostly helped me see stuff not wAMD related that should be noted eg ERM*" (PGY-4). The main failure mode was vagueness ("*non-specific labels like 'fluid'… questionable for ERM and differentiating IRF vs SRF*," PGY-3), pointing to the need for ontology-specific phrasing (IRF/SRF/PED/ERM) and per-item confidence rather than flat prose. Importantly, residents wanted visible grounding to the image: three of four rated "highlighted areas of relevance from which the descriptions are drawn" as most helpful, effectively asking for click-through linking between each token in the summary and its visual evidence. They preferred succinct defaults with optional expansion ("*more succinct descriptions*" helpful; verbose neutral), aligning with the time pressure of clinics and our finding that comment editing time remains a substantial slice of per-eye work.

7.4 Trust, bias, and safety: timing and accountability over explanations alone

Both overlays and draft text can anchor judgments. Our data and quotes argue for timing and accountability mechanisms over more explanation alone. A first-pass period without aids helps preserve independent assessment; after reveal, every retained item should require an explicit confirmation or negation, with the system logging these edits for auditability.

Per-item confidence badges can calibrate reliance ("*a confidence score… most helpful*," 3/4 residents), while very low-confidence items should be collapsed under "needs review" rather than emitted as facts. Because overlays may hide subtle structures, the user interface should guarantee instant "peek-through" (e.g., press-and-hold to hide AOIs) and never present AOIs at the fovea without translucency safeguards. Finally, measurement should not distort behavior: residents noted toggling costs and distraction; eye-tracking during active typing introduces artifacts, so separating viewing from editing panes reduces interference and yields cleaner gaze metrics.

7.5 Ethical and deployment considerations

Clinical documentation demands provenance and editorial control. Auto-generated content must be clearly marked, never silently committed to the record, and always require clinician approval. Logs should capture what the system proposed and what the clinician changed to support accountability and learning health-system loops. Because our models leverage expert gaze and dictations, we must treat these as biometric signals—minimize retention, de-identify early, and make their use transparent in consent and in the user interface. Generalizability remains a risk across scanners, protocols, and populations; surfacing context (e.g., scanner vendor) and confidence, and providing an easy "flag as incorrect" approach, are practical checks against silent drift. Finally, education should be a co-goal: linking each textual claim to an evidence tile and supporting progressive disclosure protects against deskilling while still offering speed when appropriate.

7.6 Toward integrated visual–text guidance

Residents' behaviors suggest a path to combining modalities without amplifying bias: a user-first first pass; then a linked reveal where each concise biomarker token is clickable to its AOI tile, both carrying confidence; presented side-by-side with the raw image to avoid occlusion; and requiring explicit keep/edit/delete decisions before insertion into the note. This design matches how participants already worked ("*either read the description first or… do my own attempt then see if the text agreed*," PGY-1) while aligning with our quantitative evidence of increased biomarker breadth and high retention after edits (Fig. 6b).

8 Limitations and Future Directions

Our participant-level sample (n = 10 AOI, n = 4 VLM, 11 unique residents) is small, reflecting clinical recruitment constraints: ophthalmology residency programs are small (typically 12˜20 residents per institution), and our protocol required 30˜40 minute eye-tracked sessions during clinical training hours. While we achieved 69% recruitment from the available pool and conducted prospective power analyses ensuring adequate image-set-level samples (n = 150 AOI, n = 120 VLM observations exceeding targets of n = 74 and n = 128), participant-level constraints preclude detection of small effects, subgroup analyses, or robust individual-differences modeling. Confidence intervals around effect estimates are wide (e.g., VLM biomarker increase: mean 3.5 per eye, 95% CI [2.8, 4.1]), and results should be interpreted as hypothesis-generating evidence of feasibility and workflow fit rather than definitive proof of clinical efficacy.

While our specific findings are limited by sample size, single-scanner dataset, and single-site scope, our methodological approach of distilling expert gaze and dictations into ViT attention alignment and ontology-bounded VLM guidance is domain-agnostic. Similar gaze-supervised architectures have improved lesion detection in chest X-rays [25, 51] and pathology [23], and ontology-grounded language models have shown promise in radiology report generation [24, 43]. The key insight is that distilling expert process (where they look, what they write) rather than just outcomes (final diagnoses) improves both model faithfulness and clinical utility, generalizes beyond OCT and wAMD. We made an

intentional trade-off prioritizing ecological validity (realistic clinical sessions, detailed behavioral logging, within-subjects design depth) over cross-sectional statistical power, accepting these constraints as appropriate for a rigorous pilot establishing proof-of-concept for expert-distilled, process-oriented clinical AI.

Our study targets a binary decision (Normal vs. wAMD). While this formulation allows for precise experimental control, it simplifies the clinical reality where distinguishing "wet" from "dry" AMD or identifying "at-risk" biomarkers (e.g., assessing progression risk rather than just presence) is common. However, even within this binary scope, residents demonstrated variability in diagnosing subtle cases, confirming the task's relevance for training. Future work will expand to multi-class settings (wet/dry/normal) and longitudinal progression analysis. This scope helped isolate effects but limits external validity. A next step is a multi-class, multi-label setting with itemized uncertainty to test whether guidance helps residents resolve look-alike findings (e.g., IRF vs. SRF) rather than merely confirm "abnormal," and a multi-site replication across scanners and protocols with on-screen disclosure of domain context and model confidence.

Resident behavior and feedback point to a linked visual–text experience that defers assistance until after a first unaided pass and then reveals concise, ontology-specific biomarker tokens *each backed by visible evidence*. Concretely, we will replace diffuse heatmaps with small, non-occlusive evidence cards (tight bounding regions tied to named findings and per-item confidence) shown side-by-side with the raw image. Every item will require explicit accept/edit/reject before it can enter a note, creating an audit trail that preserves clinician agency.

A practical obstacle is the lack of ground truth for localizing biomarkers, which limits us to distributional evaluation (cross-entropy) rather than standard saliency metrics like NSS or sAUC. To bridge this gap, we will bootstrap candidate regions from gaze-derived AOIs and allow experts to "click-to-tighten" boxes during review. These micro-labels will support more rigorous evaluation using NSS/sAUC and iteratively refine AOI generation, eventually accumulating enough supervision to retire generic heatmaps in favor of precise, inspectable evidence cards.

Evaluation will shift from fluency metrics (e.g., BERTScore) to measures that reflect clinical use: ontology-exact entity scores, location-aware agreement (whether the box lands on the right structure), edit cost (distance to acceptable summary), and reliance calibration (tendency to accept or reject behavior as a function of confidence), alongside documentation effort (time and keystrokes saved). To mitigate anchoring and distraction, interaction will default to a clean image, reveal few precise regions with adaptive translucency and instant peek-through, present originals and evidence side-by-side, hide very low-confidence items under "needs review," and block off-ontology phrases by construction. Finally, deployments will label auto-generated content, never commit text without clinician approval, log proposals and edits for accountability, and minimize/de-identify gaze and dictation storage. A pre-registered, factorial field study across sites—comparing unaided-first vs. immediate guidance and text-only vs. linked visual–text evidence—will determine whether this user-first, evidence-linked design improves diagnostic work and documentation while preserving trust calibration.

Finally, our evaluation of attention alignment relied on the cross-entropy divergence used during training. While this metric ensures that the model's attention distribution mathematically approximates the expert's fixation density, it does not utilize standard saliency benchmarks such as Normalized Scanpath Saliency (NSS) or shuffled Area Under the Curve (sAUC). We acknowledge this as a limitation in comparing our results to the broader saliency literature. Future work will employ NSS and sAUC to directly quantify the similarity between resident scanpaths and model-generated AOIs, allowing us to measure the effective transfer of expert visual strategies to trainees.

9 Conclusion

We present a pilot deployment of an OCT AI co-annotator demonstrating that expert gaze and dictations can be distilled into editable, evidence-linked guidance that helps ophthalmology residents decide where to look and what to write.

We present an OCT AI co-annotator that helps clinicians decide *where to look* and *what to write* via two expert-distilled user-interface components: fixation-aligned Areas of Interest (AOIs) and an ontology-bounded VLM that drafts editable biomarker summaries. In expert elicitation, aligning a ViT to fixation density improved diagnostic performance (micro-AUC 0.95 $\rightarrow$ 0.98; single-image F1 0.8628), and dictation-derived supervision produced high-fidelity biomarker text (MedBERTScore 0.867). In deployment with residents, guidance was non-inferior with respect to accuracy and workflow-compatible: VLM support increased documented biomarkers per AMD eye (mean 2.3 $\rightarrow$ 5.8, p=8.8 $\times$ 10^{-9}) with *83.1%* retention after edits, while AOIs preserved natural viewing (comparable gaze–AOI divergence) and yielded a post-guidance speed-up. Trainees used text as a concise checklist and second reader rather than a verdict, editing freely without harming accuracy or time.

These findings, while limited by small sample size (n=11) and single-site scope, establish proof-of-concept for expert-distilled, process-oriented clinical AI. The observed effects, preserved accuracy with increased documentation breadth (2.3 to 5.8 biomarkers/eye, $p < 10^{-8}$) and high editorial engagement (83% retention), justify larger, multi-site validation studies. The methodological approach of aligning vision models to expert gaze, bounding language models to clinical ontologies, and deploying editable interfaces with explicit provenance, is domain-agnostic and addresses workflow needs shared across medical imaging specialties.

A central design implication is that effective clinical explainability is *deferred, evidence-linked, and clinician-editable*. Assistance should appear after a first unaided pass; show few, non-occlusive AOIs as small evidence cards; link each biomarker token to its visible region with a per-item confidence; and require explicit accept/edit/reject before anything enters the note, all within a bounded ontology and with an audit trail. This user-first contract preserved agency, expanded coverage, and calibrated reliance—offering a practical path for clinical AI that serves as a co-annotator rather than a judge.

References

[1] Ahmet Akman. 2018. *Optical Coherence Tomography: Manufacturers and Current Systems.* Springer International Publishing, Cham, 27–37. doi:10.1007/978-3-319-94905-5_4

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[3] Mustafa Arikan, James Willoughby, Sevim Ongun, Ferenc Sallo, Andrea Montesel, Hend Ahmed, Ahmed Hagag, Marius Book, Henrik Faatz, Maria Vittoria Cicinelli, et al. 2025. OCT5k: A dataset of multi-disease and multi-graded annotations for retinal layers. *Scientific data* 12, 1 (2025), 267.

[4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.

[5] Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research* (2004).

[6] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*. IEEE, 160–169.

[7] Nora Castner, Lubaina Arsiwala-Scheppach, Sarah Mertens, Joachim Krois, Enkeleda Thaqi, Enkelejda Kasneci, Siegfried Wahl, and Falk Schwendicke. 2024. Expert gaze as a usability indicator of medical AI decision support systems: a preliminary study. *NPJ Digital Medicine* 7, 1 (2024), 199.

[8] Haomin Chen, Catalina Gomez, Chien-Ming Huang, and Mathias Unberath. 2022. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *NPJ digital medicine* 5, 1 (2022), 156.

[9] Zihui Chen, Zhi Liu, and Yingjie Song. 2026. Gaze-guided vision transformer for chest X-ray image classification. *Biomedical Signal Processing and Control* 111 (2026), 108298.

[10] Jeffrey De Fauw and et al. 2018. Clinically Applicable Deep Learning for Diagnosis and Referral in Retinal Disease. *Nature Medicine* 24, 9 (2018), 1342–1350. doi:10.1038/s41591-018-0107-6

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[12] Jojanneke Drogt, Megan Milota, Wouter Veldhuis, Shoko Vos, and Karin Jongsma. 2024. The Promise of AI for Image-Driven Medicine: Qualitative Interview Study of Radiologists' and Pathologists' Perspectives. *JMIR Human Factors* 11 (2024), e52514.

[13] Jiangxia Duan, Meiwei Zhang, Minghui Song, Xiaopan Xu, and Hongbing Lu. 2025. Eye Tracking-Enhanced Deep Learning for Medical Image Analysis: A Systematic Review on Data Efficiency, Interpretability, and Multimodal Integration. *Bioengineering* 12, 9 (2025), 954.

[14] European Society of Radiology. 2023. ESR Paper on Structured Reporting in Radiology—Update 2023. *Insights into Imaging* 14, 1 (2023), 116. doi:10.1186/s13244-023-01560-0

[15] Ning Fang, Jon Pluyter, Saskia Bakker, Igor Jacobs, Misha Luyer, Joost Nederend, Jeroen Raijmakers, Lin-Lin Chen, and Mathias Funk. 2024. From Experience to Experience: Key Insights for Improved Interaction with AI in Radiology. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[16] Ross W Filice and Raj M Ratwani. 2020. The case for user-centered artificial intelligence in radiology. *Radiology: Artificial Intelligence* 2, 3 (2020), e190095.

[17] Monika Fleckenstein, Steffen Schmitz-Valckenberg, and Usha Chakravarthy. 2024. Age-Related Macular Degeneration: A Review. *JAMA* 331, 2 (Jan. 2024), 147–157. doi:10.1001/jama.2023.26074

[18] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care. *The Lancet Digital Health* 3, 11 (2021), e745–e750. doi:10.1016/S2589-7500(21)00208-9

[19] Catalina Gomez, Sue Min Cho, Shichang Ke, Chien-Ming Huang, and Mathias Unberath. 2025. Human-AI collaboration is not very collaborative yet: a taxonomy of interaction patterns in AI-assisted decision making from a systematic review. *Frontiers in Computer Science* 6 (2025), 1521066.

[20] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[21] RLW Hanson and et al. 2023. Optical Coherence Tomography Imaging Biomarkers in Neovascular AMD: A Systematic Review. *Eye* 37 (2023), 1960–1977. doi:10.1038/s41433-022-02360-4

[22] Chihcheng Hsieh, André Luís, José Neves, Isabel Blanco Nobre, Sandra Costa Sousa, Chun Ouyang, Joaquim Jorge, and Catarina Moreira. 2024. Eyexnet: Enhancing abnormality detection and diagnosis via eye-tracking and x-ray fusion. *Machine Learning and Knowledge Extraction* 6, 2 (2024), 1055–1071.

[23] Bulat Ibragimov and Claudia Mello-Thoms. 2024. The use of machine learning in eye tracking studies in medical imaging: a review. *IEEE Journal of Biomedical and Health Informatics* 28, 6 (2024), 3597–3612.

[24] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463* (2021).

[25] Amirhossein Kiani, Bora Uyumazturk, Pranav Rajpurkar, Alex Wang, Rebecca Gao, Erik Jones, Yifan Yu, Curtis P Langlotz, Robyn L Ball, Thomas J Montine, et al. 2020. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ digital medicine* 3, 1 (2020), 23.

[26] Elmar Kotter and Erik Ranschaert. 2021. Challenges and solutions for introducing artificial intelligence (AI) in daily clinical workflow. *European radiology* 31, 1 (2021), 5–7.

[27] Harold L Kundel, Calvin F Nodine, Elizabeth A Krupinski, and Claudia Mello-Thoms. 2007. Holistic Component of Image Perception in Mammogram Interpretation: Gaze-Tracking Study. *Radiology* 242, 2 (2007), 396–402. doi:10.1148/radiol.2422051997

[28] Thomas Kurmann, Siqing Yu, Pablo Márquez-Neila, Andreas Ebneter, Martin Zinkernagel, Marion R Munk, Sebastian Wolf, and Raphael Sznitman. 2019. Expert-level automated biomarker identification in Optical Coherence Tomography scans. *Sci. Rep.* 9, 1 (Sept. 2019), 13605.

[29] Zhongwen Li, Lei Wang, Xuefang Wu, Jiewei Jiang, Wei Qiang, He Xie, Hongjian Zhou, Shanjun Wu, Yi Shao, and Wei Chen. 2023. Artificial intelligence in ophthalmology: The path to the real-world clinic. *Cell Reports. Medicine* 4, 7 (July 2023), 101095. doi:10.1016/j.xcrm.2023.101095

[30] Gilbert Lim, Kabilan Elangovan, and Liyuan Jin. 2024. Vision language models in ophthalmology. *Current Opinion in Ophthalmology* 35, 6 (Nov. 2024), 487–493. doi:10.1097/ICU.0000000000001089

[31] Damien Litchfield, Linden J Ball, Tim Donovan, David J Manning, and Trevor Crawford. 2010. Viewing another person's eye movements improves identification of pulmonary nodules in chest x-ray inspection. *Journal of Experimental Psychology: Applied* 16, 3 (2010), 251.

[32] Chong Ma, Hanqi Jiang, Wenting Chen, Yiwei Li, Zihao Wu, Xiaowei Yu, Zhengliang Liu, Lei Guo, Dajiang Zhu, Tuo Zhang, et al. 2024. Eye-gaze guided multi-modal alignment for medical representation learning. *Advances in Neural Information Processing Systems* 37 (2024), 6126–6153.

[33] Gaya Mehenni and Amal Zouaq. 2024. Ontology-Constrained Generation of Domain-Specific Clinical Summaries. In *International Conference on Knowledge Engineering and Knowledge Management*. Springer, 382–398.

[34] Cristian Metrangolo, Simone Donati, Marco Mazzola, Liviana Fontanel, Walter Messina, Giulia D'alterio, Marisa Rubino, Paolo Radice, Elias Premi, and Claudio Azzolini. 2021. OCT Biomarkers in Neovascular Age-Related Macular Degeneration: A Narrative Review. *Journal of Ophthalmology* 2021 (July 2021), 9994098. doi:10.1155/2021/9994098

[35] Manuel Paez-Escamilla, Mahima Jhingan, Denise S. Gallagher, Sumit Randhir Singh, Samantha Fraser-Bell, and Jay Chhablani. 2021. Age-related macular degeneration masqueraders: From the obvious to the obscure. *Survey of Ophthalmology* 66, 2 (2021), 153–182. doi:10.1016/j.survophthal.

2020.08.005

[36] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52.

[37] Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini. 2022. Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific Reports* 12, 1 (2022), 14952.

[38] Rikard Rosenbacke, Åsa Melhus, and David Stuckler. 2024. False conflict and false confirmation errors are crucial components of AI accuracy in medical decision making. *Nature Communications* 15, 1 (2024), 6896.

[39] Ji Seung Ryu, Hyunyoung Kang, Yuseong Chu, and Sejung Yang. 2025. Vision-language foundation models for medical imaging: a review of current practices and innovations. *Biomedical Engineering Letters* (2025), 1–22.

[40] Tetsu Sakamoto, Yukinori Harada, Taro Shimizu, et al. 2024. Facilitating Trust Calibration in Artificial Intelligence–Driven Diagnostic Decision Support Systems for Determining Physicians' Diagnostic Accuracy: Quasi-Experimental Study. *JMIR Formative Research* 8, 1 (2024), e58666.

[41] Nicole T. M. Saksens, Monika Fleckenstein, Steffen Schmitz-Valckenberg, Frank G. Holz, Anneke I. den Hollander, Jan E. E. Keunen, Camiel J. F. Boon, and Carel B. Hoyng. 2014. Macular dystrophies mimicking age-related macular degeneration. *Progress in Retinal and Eye Research* 39 (March 2014), 23–57. doi:10.1016/j.preteyeres.2013.11.001

[42] Dario D Salvucci and Joseph H Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*. 71–78.

[43] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201* (2025).

[44] Hyun Joo Shin, Kyunghwa Han, Leeha Ryu, and Eun-Kyung Kim. 2023. The impact of artificial intelligence on the reading times of radiologists for chest radiographs. *NPJ Digital Medicine* 6, 1 (2023), 82.

[45] Suganya Subramaniam, Sara Rizvi, Ramya Ramesh, Vibhor Sehgal, Brinda Gurusamy, Hikmatullah Arif, Jeffrey Tran, Ritu Thamman, Emeka C Anyanwu, Ronald Mastouri, et al. 2025. Ontology-guided machine learning outperforms zero-shot foundation models for cardiac ultrasound text reports. *Scientific Reports* 15, 1 (2025), 5456.

[46] Malliga Subramanian, Kogilavani Shanmugavadivel, Obuli Sai Naren, K Premkumar, and K Rankish. 2022. Classification of Retinal OCT Images Using Deep Learning. In *2022 International Conference on Computer Communication and Informatics (ICCCI)*. 1–7. doi:10.1109/ICCCI54379.2022.9740985 ISSN: 2329-7190.

[47] Ali S Tejani, Tessa S Cook, Mohannad Hussain, Teri Sippel Schmidt, and Kevin P O'Donnell. 2024. Integrating and adopting AI in the radiology workflow: a primer for standards and integrating the healthcare enterprise (IHE) profiles. *Radiology* 311, 3 (2024), e232653.

[48] Andreas Theissler, Anna-Lena Kraft, Max Rudeck, and Fabian Erlenbusch. 2020. VIAL-AD: Visual Interactive Labelling for Anomaly Detection – An approach and open research questions. *Proceedings of the Workshop on Interactive Adaptive Learning co-located with European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2020)* (2020), 84–89.

[49] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. 2020. Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns* 1, 4 (2020).

[50] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. 2020. Human–computer collaboration for skin cancer recognition. *Nature Medicine* 26, 8 (2020), 1229–1234.

[51] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* 8, 12 (2024), 2293–2303.

[52] Annelien van der Gijp and et al. 2017. A Narrative Systematic Review of Eye-Tracking Research in Radiology. *Advances in Health Sciences Education* 22, 3 (2017), 765–789. doi:10.1007/s10459-016-9698-1

[53] Bin Wang, Hongyi Pan, Armstrong Aboah, Zheyuan Zhang, Elif Keles, Drew Torigian, Baris Turkbey, Elizabeth Krupinski, Jayaram Udupa, and Ulas Bagci. 2024. GazeGNN: A gaze-guided graph neural network for chest x-ray classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2194–2203.

[54] Sheng Wang, Xi Ouyang, Tianming Liu, Qian Wang, and Dinggang Shen. 2022. Follow my eye: Using gaze to supervise computer-aided diagnosis. *IEEE Transactions on Medical Imaging* 41, 7 (2022), 1688–1698.

[55] Katharina Wenderott, Jim Krups, Julian A Luetkens, and Matthias Weigl. 2024. Radiologists' perspectives on the workflow integration of an artificial intelligence-based computer-aided detection system: A qualitative study. *Applied Ergonomics* 117 (2024), 104243.

[56] Catherine C Wu, Jeremy M Wolfe, and Karla K Evans. 2019. Eye Movements in Medical Image Perception. *Cognitive Research: Principles and Implications* 4, 1 (2019), 7. doi:10.1186/s41235-019-0154-4

[57] Rui Yang, Edison Marrese-Taylor, Yuhe Ke, Lechao Cheng, Qingyu Chen, and Irene Li. 2023. Integrating UMLS knowledge into large language models for medical question answering. *arXiv preprint arXiv:2310.02778* (2023).

[58] Nilufer Yildirim and et al. 2024. Multimodal Healthcare AI: Identifying and Designing Clinically Useful Systems. In *CHI Conference on Human Factors in Computing Systems*. 1–18. doi:10.1145/3613904.3642013

[59] Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M Padilla, Jeffrey Caterino, Ping Zhang, et al. 2024. Rethinking human-AI collaboration in complex medical decision making: a case study in sepsis diagnosis. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

A Dataset Curation and Preprocessing

This section provides a detailed description of the datasets curated and preprocessed for the multi-task finetuning of the MedGemma Vision Language Model (VLM). Our goal was to build a specialized ophthalmic AI assistant capable of interpreting retinal Optical Coherence Tomography (OCT) scans across three distinct tasks: binary diagnosis, biomarker generation, and biomarker discrimination. To achieve this, we sourced data from publicly available medical datasets, including OCT-C8 and OCT-5K, and combined them with fine-grained biomarker information extracted from physician transcripts.

A.1 Diagnosis Dataset

The foundation for the model's visual understanding and diagnostic capability was established using a large-scale dataset for a binary classification task.

- **Source:** A curated subset of the publicly available Retinal OCT-C8 dataset.
- **Task:** Binary classification to determine the primary pathology of an OCT scan. The model is constrained to respond with either 'Normal' or 'wet-AMD'.
- **Size and Composition:** The dataset comprises a total of 4,600 OCT images, balanced perfectly between the two classes (2,300 images for 'Normal' and 2,300 for 'wet-AMD').
- **Purpose:** This dataset's primary role is to provide a robust, domain-specific visual foundation, training the model to recognize the core features differentiating healthy retina from those affected by wet Age-related Macular Degeneration.

A.2 Biomarker Generation Dataset

A key challenge in developing specialized medical AI is the availability of high-quality, annotated data for identifying specific pathological features (biomarkers). Our biomarker dataset was designed to address this challenge for the generative identification task.

- **Source:** The dataset corpus was created by combining annotations from the OCT-5K dataset with a pre-processed biomarker list extracted from physician transcripts from our internal dataset.
- **Task:** A generative task where the model must identify all relevant pathological features from an OCT scan and output a comma-separated list of biomarkers.
- **Size and Composition:** The combined dataset contains a total of 573 unique samples.
- **Preprocessing:** To manage complexity and focus the model on the most relevant features, biomarkers with fewer than 10 occurrences in the combined dataset were excluded from the final training set. This resulted in a final vocabulary of 12 key biomarkers: "Drusen", "Photoreceptor Degeneration", "Pigment Epithelial Detachment", "Geographic Atrophy", "Choroidal Fold", "Epiretinal Membrane", "Hyperfluorescent Spots", "Intraretinal Fluid", "Posterior Vitreous Detachment", "Fluid", "Subretinal Fluid", and "Choroidal Neovascularization".

Table S2 provides a detailed frequency count of all identified biomarkers in the raw combined dataset before the final filtering step was applied.

A.3 Biomarker Discrimination Dataset

To facilitate targeted verification of specific features and improve the model's visual grounding, a VQA dataset was created.

	Healthy	SRF	IRF	HF	Drusen	RPD	ERM	GA	ORA	IRC	FPED
Training Set (23,030)	6480	1142	2947	5668	5077	1995	6139	1093	2280	4321	4766
Test Set (1029)	165	65	48	178	376	153	140	62	200	54	359

Table S1. Distribution of biomarkers in Kurmann et al.

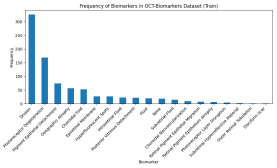
Table S2. Frequency of Biomarkers in the Combined Dataset (Pre-filtering)

Biomarker	**Frequency**
Drusen	410
Photoreceptor Degeneration	209
Pigment Epithelial Detachment	99
Geographic Atrophy	73
Choroidal Fold	60
Posterior Vitreous Detachment	33
Hyperfluorescent Spots	32
Epiretinal Membrane	31
Intraretinal Fluid	29
Fluid	25
None	25
Subretinal Fluid	20
Choroidal Neovascularization	12
Retinal Pigment Epithelial Migration	*9*
Retinal Pigment Epithelium Atrophy	*6*
Subretinal Hyperreflective Material	*4*
Photoreceptor Layer Disruption	*4*
Disciform Scar	*2*
Hemorrhage	*1*
Outer Retinal Tubulation	*1*

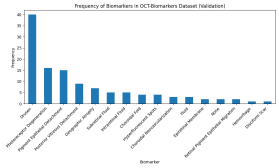
- **Source:** The biomarker discrimination dataset was dynamically generated from the curated biomarker generation Dataset.
- **Size and Composition:** The dataset comprises a total of 1664 unique samples.
- **Task:** A discriminative task where the model is presented with a direct question of the form, "Is [BIOMARKER_NAME] present in this OCT scan?" and must respond with a definitive 'Yes' or 'No'.
- **Generation Strategy:** An initial, exhaustive generation process revealed that an uncurated discriminative dataset would be overwhelmingly populated with "No" answers, potentially teaching the model a trivial strategy of always predicting "No". To circumvent this, a balancing strategy was employed. For each image, an equal number of "Yes" instances (for biomarkers present) and "No" instances (randomly selected from biomarkers that were absent) were generated. This approach forced the model to learn the actual visual features corresponding to each biomarker rather than relying on statistical priors.

B VLM Model Training Configurations and Supplementary

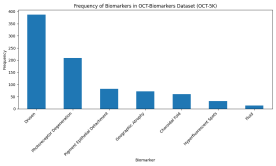
The Vision Language Model (VLM) finetuning was conducted in two stages. The detailed configurations for each stage are provided in Table S3.



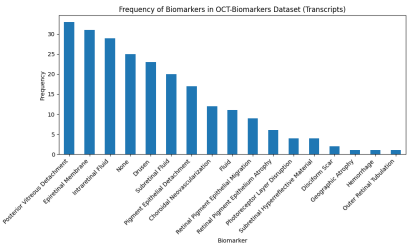
(a) Training data distribution.



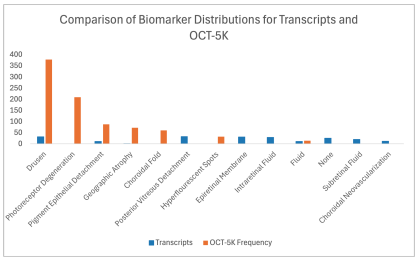
(b) Validation data distribution.



(c) OCT-5K dataset distribution.



(d) OCT transcript dataset distribution.



(e) Distribution of transcript and OCT-5K against each other.

Fig. S1. Biomarker Dataset Distribution

Table S3. Training configurations for the VLM finetuning.

Parameter	Stage One Configs	Stage Two Configs
Model Configuration		
Model Name	unsloth/medgemma-4b-it-unsloth-bnb-4bit	unsloth/medgemma-4b-it-unsloth-bnb-4bit
LoRA Configuration		
r	32	32
alpha	64	64
dropout	0.05	0.05
Trainer Configuration		
Num Epochs	3	8
Train Batch Size	56	56
Eval Batch Size	224	224
Grad Accum Steps	1	1
Warmup Ratio	0.1	0.1
Learning Rate	5e-6	1e-6

B.1 Per-Biomarker Evaluation Performance

To address the need for clinical specificity beyond semantic similarity metrics (e.g., BERTScore), we conducted a granular, ontology-exact evaluation of the VLM. We calculated F1, Precision, and Recall scores for individual biomarkers across two distinct tasks: *Biomarker Generation* (where the model spontaneously lists findings) and *Biomarker Discrimination* (where the model answers "Yes/No" to specific queries, e.g., "Is Drusen present?").

The evaluation was conducted on a validation set of 58 samples. As shown in Table S4, the model demonstrates strong performance on common pathologies but exhibits performance degradation on rare classes, highlighting the impact of data imbalance in the training corpus.

Table S4. Per-Biomarker Performance Metrics. Comparison of the Generative (List) task vs. the Discriminative (VQA) task. *N* represents the number of support examples in the validation set. Evaluation is ontology-exact.

	Biomarker Generation Task				**Biomarker Discrimination Task**			
Biomarker	**F1**	**Precision**	**Recall**	**N**	**F1**	**Precision**	**Recall**	**N**
Drusen	0.87	0.83	0.91	44	0.98	0.96	1.00	46
Photoreceptor Disruption	0.67	0.58	0.79	24	0.88	0.82	0.96	29
Pigment Epithelial Detachment	0.40	0.33	0.50	10	0.38	0.50	0.30	18
Intraretinal Fluid	0.50	0.33	1.00	1	1.00	1.00	1.00	16
Epiretinal Membrane	0.00	0.00	0.00	2	0.67	1.00	0.50	12
Choroidal Fold	0.50	0.40	0.67	3	0.00	0.00	0.00	15
Posterior Vitreous Detachment	0.29	0.20	0.50	2	0.00	0.00	0.00	17
Geographic Atrophy	0.00	0.00	0.00	9	0.20	1.00	0.11	14
Fluid (Generic)	0.29	0.33	0.25	4	0.00	0.00	0.00	11
Subretinal Fluid	0.00	0.00	0.00	1	0.00	0.00	0.00	13

B.2 Analysis of Task Performance

We observed two distinct trends in the model's behavior:

Task Complexity (Discrimination vs. Generation). The model generally achieved higher performance on the *Biomarker Discrimination* task compared to *Biomarker Generation*. For example, detection of Intraretinal Fluid improved from an F1 of 0.50 in the generative setting to 1.00 in the discriminative setting. This performance gap is expected, as the binary classification nature of the discrimination task is inherently simpler than the open-ended generation required to list all present biomarkers. Additionally, the Discrimination dataset is larger and balanced with adversarial "No" examples, providing stronger supervision for specific features.

Frequency Bias. Biomarkers with high prevalence in the training set, such as Drusen ($N = 46$ in discrimination test) and Photoreceptor Disruption, achieved high F1 scores across both tasks (Drusen Discrimination F1: 0.98). Conversely, low-frequency biomarkers such as Subretinal Fluid and Epiretinal Membrane suffered from low detection rates. Note that biomarkers with zero support samples in the pilot test set (e.g., CNV, Hemorrhage) were excluded from this analysis. Future work will necessitate a larger, stratified test set to ensure robust evaluation of these rare pathologies.

C Vision Transformer (ViT) Gaze-Alignment Training Details - Complete Section Added

This section provides the technical specifications for reproducing the gaze-aligned ViT training described in Section 3.2 of the main paper.

C.1 Attention Rollout Algorithm

Attention rollout computes the effective attention from input patches to the final layer by recursively multiplying attention matrices across all transformer layers.

Given:

- $A^{(\ell)} \in \mathbb{R}^{N \times N}$: average attention matrix at layer ℓ (averaged over all 12 heads),
- $N = 1025$ tokens (1 [CLS] token + $32 \times 32 = 1024$ patches),
- $L = 12$ transformer layers.

Algorithm:

(1) **Initialize:**

$$\tilde{A}^{(0)} = I$$

(2) **Recursive aggregation** for each layer $\ell = 1, 2, \ldots, 12$:

$$\tilde{A}^{(\ell)} = \tilde{A}^{(\ell-1)} \cdot A^{(\ell)}$$

(3) **Extract attention** from the [CLS] token to image patches:

$$\hat{A} = \tilde{A}^{(L)}[0,\ 1:]$$

i.e., the first row of $\tilde{A}^{(L)}$ excluding the [CLS] position.

(4) **Reshape and normalize:**

$$\hat{A}_{\text{grid}} = \text{reshape}(\hat{A},\ (32, 32))$$

$$\hat{A}^{(L)}(x) = \frac{\hat{A}_{\text{grid}}}{\sum_{i,j} \hat{A}_{\text{grid}}[i, j]}$$

This yields a normalized 32×32 probability distribution $\hat{A}^{(L)}(x)$ over image patches, directly comparable to the expert fixation-density target A^*.

C.2 Training Configuration

Table S5. ViT gaze-alignment training hyperparameters and configuration.

Model Architecture	
ViT variant	Base/16
Number of layers	12
Hidden dimension	768
Number of heads	12
MLP dimension	3072
Patch size	16×16
Input resolution	512×512
Patch grid size	32×32
Total tokens	1025 (1 [CLS] + 1024 patches)
Initialization	
Pretrained weights	ImageNet-21k
Training Configuration	
Optimizer	AdamW
Base learning rate	3×10^{-4}
Minimum learning rate	1×10^{-6}
β_1, β_2	0.9, 0.999
Weight decay	0.05
Gradient clipping	max_norm = 1.0
Batch size	32
Total epochs	100
Warmup epochs	10
LR schedule	Cosine annealing
Loss Function	
Classification loss	Binary cross-entropy (BCE)
Alignment loss	Cross-entropy between distributions
α values tested	{0, 0.01, 0.05, 0.1, 0.2, 0.3}
Selected α	0.05
Data Augmentation	
Horizontal flip	$p = 0.5$
Color jitter	brightness=0.2, contrast=0.2
Random rotation	$\pm 10^{\circ}$
Random crop	scale=[0.9, 1.0]

C.3 Cross-Entropy Loss for Probability Distributions

The alignment loss compares two discrete probability maps over the 32×32 patch grid:

$$\mathcal{L}_{\text{align}} = -\sum_{i=1}^{32}\sum_{j=1}^{32} A^*[i,j]\ \log\left(\hat{A}^{(L)}[i,j] + \epsilon\right) \tag{4}$$

where:

- $A^*[i,j]$ is the expert fixation density, normalized such that $\sum_{i,j} A^*[i,j] = 1$
- $\hat{A}^{(L)}[i,j]$ is the model attention from attention rollout
- $\epsilon = 10^{-8}$ ensures numerical stability

C.4 Training Procedure

For each training image x with label y and expert target A^*:

(1) **Forward pass:**

$$\hat{y} = \text{ViT}(x), \qquad \hat{A}^{(L)} = \text{AttentionRollout}(\text{ViT}, x)$$

(2) **Compute losses:**

$$\mathcal{L}_{\text{cls}} = \text{BCE}(\hat{y}, y), \qquad \mathcal{L}_{\text{align}} = \text{CrossEntropy}\left(\hat{A}^{(L)}, A^*\right)$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \alpha\, \mathcal{L}_{\text{align}}$$

(3) **Backward pass and optimization:**

- Compute gradients via backpropagation
- Clip gradients with max_norm = 1.0
- Update weights using AdamW

Learning rate schedule:

- Linear warmup (epochs 1–10) from 0 to 3×10^{-4}
- Cosine annealing (epochs 11–100) down to 1×10^{-6}

Model selection: The checkpoint with highest validation accuracy across the 5 folds is retained. Final results are averaged across folds.

C.5 Implementation Notes

- **Attention extraction overhead:** Attention rollout adds ~15% computational overhead.
- **Patient-level CV:** Data split at the patient level prevents leakage, since each patient contributes 5 OCT images.
- **Baseline:** $\alpha = 0$ corresponds to a standard ViT trained solely with classification loss.
- **Mixed precision:** AMP reduces memory usage and improves training speed without affecting final accuracy.
- **Attention normalization:** Rollout-generated attention maps are always normalized to sum to 1 before computing loss.

D Inter-participant agreement

The results of calculation of Cohen's κ across the image sets rated by two different participants in each user study set are presented in Table S6.

	κ	97.5% CI	N
post-guidance control	0.48	(0.15, 0.81)	18
AOI-guidance	0.66	(0.36,0.96)	24

Table S6. κ values and confidence intervals for inter-reader agreement across user study conditions.

E Hypothesis tests for User Study 2

	AOI			VLM
	pre vs. guided	guided vs. post	pre vs. post	pre vs. guided
accuracy	0.321 (0.936)	0.702 (0.936)	0.527 (0.936)	0.269 (0.936)
FPR	1 (1)	1 (1)	0.474 (0.936)	1 (1)
FNR	0.42 (0.936)	0.692 (0.936)	0.668 (0.936)	0.322 (0.936)
Time per eye	0.608 (0.751)	1.42×10^{-4} (7.44×10^{-4})	6.38×10^{-6} (1.34×10^{-4})	0.198 (0.519)
Correct dx per min	0.912 (0.957)	0.113 (0.395)	0.065 (0.274)	0.486 (0.751)
Time to final Dx	0.296 (0.622)	1.48×10^{-5} (1.55×10^{-4})	1.31×10^{-4} (7.44×10^{-4})	0.477 (0.751)
Confidence	0.71 (0.785)	0.968 (0.968)	0.678 (0.785)	0.25 (0.583)
comment edit time	0.359 (0.686)	0.523 (0.751)	0.141 (0.424)	0.542 (0.751)
Effort				0.582 (0.751)

Table S7. *P*-values and Benjamini-Hochberg adjusted *P*-values (in parentheses) for measured quantities by experiment blocks in AOI and VLM conditions in **US2**. Values for accuracy, FPR, and FNR were generated by Fisher's Exact test. The Wilcoxon signed-rank test was used for all other quantities.

F Mixed Effects Modeling in User Study 2

The results of mixed effects modeling for the AOI and VLM conditions on diagnostic correctness (whether participant diagnosis agrees with ground truth diagnosis) and time-to-diagnosis are presented in Tables S8–S9. We allowed for random by-participant effects on the diagnostic correctness while controlling for ground truth classication (normal vs wet AMD) and image position in the experimental block (to control for possible fatigue effects). The latter two were considered fixed effects to ensure model convergence. Modeling was performed using the 'lmer' package in R.

G VLM comment analysis

Vocabulary overlaps between VLM-provided biomarkers, user-provided biomarkers (from control block), and user-modified VLM-provided biomarkers are shown in Figure S2 below.

H User Survey Questions

The following items were queried during the post-experiment surveys in **User Study 2**.

Participant demographics

AOI experiment	Variable	Log-Odds	Std. Error	p
Fixed effects	Intercept	4.098	0.997	3.98×10^{-5}
	AOI guidance	0.647	0.887	0.466
	post-guidance control	1.260	0.928	0.174
	GT = wAMD	-2.979	0.727	4.21×10^{-5}
	within-block position	0.175	0.087	0.044
	AOI guidance:within-block position	-0.162	0.109	0.137
	post-guidance control:within-block position	-0.191	0.114	0.093
	Group	Name	Variance	Std. Dev.
Random effects	Participant	Intercept	1.337	1.156
		AIC	BIC	log(likelihood)
Model fit		232.7	265.3	-108.3

VLM experiment	Variable	Log-Odds	Std. Error	p
Fixed effects	Intercept	2.565	0.899	4.35×10^{-3}
	VLM guidance	0.441	1.349	0.744
	GT = wAMD	-0.337	0.616	0.584
	within-block position	-0.087	0.083	0.296
	VLM guidance control:within-block position	0.043	0.137	0.754
	Group	Name	Variance	Std. Dev.
Random effects	Participant	Intercept	0.0343	0.185
		AIC	BIC	log(likelihood)
Model fit		98.7	115.4	-43.4

Table S8. Mixed effects modeling of diagnostic correctness in each experimental condition of **US2**. "GT" = Ground Truth; ":" = interaction between, AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion.

(1) Participant level of experience
(2) Participant confidence in reading OCT scans [1,4]

General UI & Control Comments

(1) What were your impressions of the system's interface and usability?
(2) Any comments about your experience interpreting the images in the first block today?

AOI-specific questions

(1) How helpful was the guidance provided by the heatmaps? [1,5]
(2) Did the heatmap guidance assist you in identifying potential cases of wet AMD more efficiently? Please provide specific examples or scenarios if possible.
(3) Do you have any suggestions on how the guidance could be improved?
(4) Which of the following changes to the guidance method do you feel could improve its ability to help you interpret OCT scans? (A graphical change in how the heatmaps are presented, Heatmaps with fewer highlighted areas, Heatmaps with more highlighted areas, An auto-generated text summary of the OCT images, Ability to change the heatmap in real time)
(5) Any additional comments or feedback you would like to provide about the system and experiment?

VLM-specific questions

(1) How would you rate the accuracy of the text summaries? [1,5]

AOI experiment	Variable	Coefficient	Std. Error	p
Fixed effects	Intercept	22.553	2.295	8.38×10^{-16}
	AOI guidance	-6.237	2.720	0.022
	post-guidance control	-11.351	2.806	6.22×10^{-5}
	GT = wAMD	8.847	1.091	5.75×10^{-15}
	within-block position	-0.915	0.212	1.92×10^{-5}
	AOI guidance:within-block position	0.585	0.299	0.051
	post-guidance control:within-block position	0.595	0.308	0.054
	Group	Name	Variance	Std. Dev.
Random effects	Participant	Intercept	13.01	3.607
	Residual		125.25	11.91
		R^2_{model}	R^2_{fixed}	R^2_{random}
Model fit		0.263	0.188	0.075
VLM experiment	Variable	Coefficient	Std. Error	p
Fixed effects	Intercept	26.478	6.333	8.44×10^{-4}
	VLM guidance	-8.320	6.691	0.216
	GT = wet AMD	13.036	3.296	1.34×10^{-4}
	within-block position	-0.172	0.521	0.742
	VLM guidance:within-block position	0.636	0.735	0.389
	Group	Name	Variance	Std. Dev.
Random effects	Participant	Intercept	60.98	7.81
	Residual		302.80	14.40
		R^2_{model}	R^2_{fixed}	R^2_{random}
Model fit		0.246	0.107	0.139

Table S9. Mixed effects modeling of time spent per eye in each experimental condition of **US2**. "GT" = Ground Truth; ":" = interaction between.

(2) What comments do you have about the accuracy of the text summaries?
(3) How did your workflow change when text summaries were provided?
(4) Could you see yourself using a system that auto-generates text summaries in clinical practice? Why or why not?
(5) How would the following changes to the text prompt method change its ability to help you interpret OCT scans? [1,5] × (More succinct descriptions, More verbose descriptions, A confidence score on the description, Highlighted areas of relevance from which the descriptions are drawn, A conversational AI agent to interactively clarify and update the description)
(6) Any additional comments or feedback you would like to provide about the text system and experiment?

Post-guidance Questions

(1) In the unguided block (after the guided block), what difference did you feel in reading the OCT images?
(2) To what extent did you want the guidance back? [1,5]
(3) How do you feel about your efficiency in reading the OCT images in the unguided condition again?

Received 11 September 2025

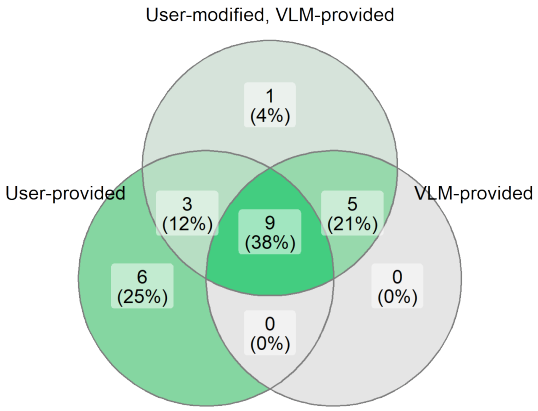


Fig. S2. Overlap of unique identified biomarkers across all VLM user study conditions