

VisionWeaver: CFS-Inspired Assessment and Adaptive Training Prototype Under DR-Inspired Simulated Vision Degradation

Yuxin Yi^{1,*}, Ruoyun Long², Mike Lin³

¹ SDSZ International Department, Beijing, China

² International Curriculum Center of RDFZ, Beijing, China

³ Shanghai Pinghe Bilingual School, Shanghai, China

* Corresponding Author Email: Yuxin Yi, yimin1983@aliyun.com

All authors contributed equally

Abstract. VisionWeaver did not begin as a claim about diabetic-retinopathy diagnosis. It began with a smaller engineering problem: after a low-contrast visual-task session, a single score did not tell us enough to debug the run or explain why the software had made a difficulty change. We built a Python/PsychoPy prototype around that problem. It includes a side-by-side, CFS-inspired assessment task, a central-gate training task, keyboard fallback for development runs, and Arduino-compatible input when external buttons are attached. The first check was therefore about traceability. We used internal logs, simulated responses, ablation, parameter sweeps, and failure-log review to ask whether the task could preserve a usable trial record. In Monte Carlo tests, staircase and moving-window control reached 35.8% and 22.4% band occupancy; fixed difficulty reached 0.1%. A random forest using the first 10 synthetic trials classified simulated degradation level with 72.9% accuracy. Its warm-start setting raised band occupancy by +15.5%, reaching 98.7% of the Oracle baseline. These results support the software and analysis workflow, not clinical efficacy.

Keywords: continuous flash suppression, visual function, adaptive control, trial logging, visual training.

1. Introduction

1.1. Why a Final Score Was Too Thin

At the start of the VisionWeaver build, the fragile part was not only the stimulus drawing code. The screen could look reasonable while the saved record was still too thin to explain the run afterward. A final score might suggest that a session was easy, hard, fast, or noisy. It does not show the contrast on trial 17, whether the adaptive rule had already stepped the stimulus, which input path recorded the response, or whether an offline filter later changed the usable trials.

The medical context still matters, but it needs to stay in the right place. Fundus images show the physical side of diabetic retinopathy: microaneurysms, hemorrhages, exudates, macular edema, and neovascularization. The left panel of Figure 1 uses public-domain education

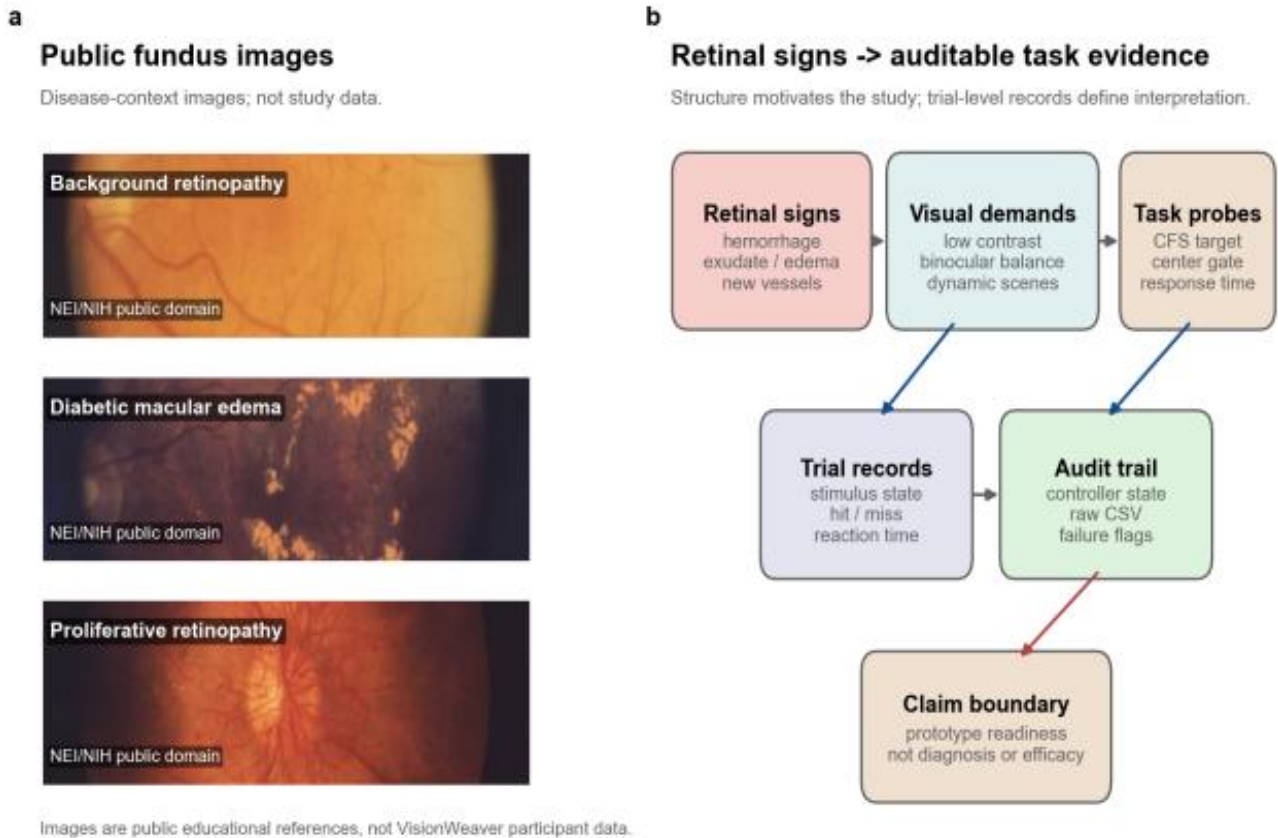


Figure 1. Diabetic eye disease and functional-vision motivation. Fundus images are public-domain NEI/NIH educational images via Wikimedia Commons; cropped and labeled for context only.

images from the U.S. National Eye Institute/National Institutes of Health so that the disease motivation is not reduced to a software feature list. VisionWeaver itself is a prototype task, not a diagnostic system; its first responsibility in this paper is to leave a record that another person can reconstruct.

Retinal images also cannot tell us how someone behaves inside one low-contrast screen task. Prior work reports that contrast sensitivity and related functional visual metrics can change in diabetes and early retinopathy, even when standard acuity is relatively preserved [1,2]. That point shaped the data design more than the claim. We needed stimulus intensity, display time, viewing condition, response method, and analysis choices in the trial file, because otherwise hit rate and reaction time mix visual performance with attention, interface familiarity, input delay, and filtering decisions.

Continuous flash suppression (CFS) offered a useful reference because masking, target visibility, and response timing are controlled together. In a standard CFS setup, one eye receives rapidly changing high-contrast masks while the other receives a lower-contrast target [3,4,5]. The current implementation does not make that dichoptic claim. It uses a single monitor with side-by-side stimuli, borrowing the CFS logic only far enough to test contrast control, timing, input handling, and trial logging.

1.2. Why We Treated Logging as the First Test

In early testing, the interface can look stable before the data trail is stable. Task presentation, button input, difficulty updates, CSV files, and offline scripts may drift apart without producing an obvious crash. Once that happens, a later result can be hard to attribute: the task itself, input latency, a changed parameter, and a missing file all become plausible explanations.

Adaptive control increases that bookkeeping burden. Fixed difficulty is simple to replay, although it ignores whether the task is too easy or too hard. Staircase and QUEST-style methods are well established, but they require more formal assumptions than this build can justify. We

used a moving-window rule because a reviewer can inspect it directly: recent hit rate changes the next contrast, and the controller state is saved with the trial.

The input path followed the same reasoning. Keyboard fallback stayed in the code because it lets us debug the task when external hardware is unavailable. Arduino-compatible input gives a route to physical buttons without making every test depend on one hardware setup. The question for this paper is therefore narrow: after a run, can VisionWeaver save enough information to replay the controller behavior and identify where the software failed? We answer it with internal logs and simulated responders.

1.3. What Is Checked Here

The checked pieces are deliberately limited. The Python/PschoPy code runs a CFS-inspired assessment task and a central-gate training task while saving trial-level CSV logs. The moving-window difficulty rule records enough controller state for the update process to be replayed after the session. The evaluation uses simulated responders, internal logs, ablation tests, parameter sweeps, and failure-log review. These checks support engineering readiness of the workflow; they do not support an efficacy claim.

2. Related Work

2.1. Diabetic Retinopathy and Functional Visual Metrics

We used the diabetic-retinopathy literature mainly to choose the kind of visual variable worth testing. Contrast sensitivity is closer to our screen task than a single acuity number, because it asks whether a target remains identifiable as contrast changes [1]. Diabetic retinopathy gives the project a meaningful application background: it has a large disease burden [6], and recent work reports altered contrast sensitivity in diabetic retinopathy [2].

Those papers also kept the scope from drifting. Contrast sensitivity studies vary in scale, illumination, and disease staging [2]. Our data do not resolve those clinical differences. DR is the motivation for the assessment direction here; the results come from an engineering prototype and simulation.

2.2. CFS and Binocular Rivalry

The binocular-rivalry and CFS literature gave us the perceptual idea, but it also made the hardware limitation obvious. Binocular rivalry shows that incompatible images presented to the two eyes can produce alternating perceptual dominance [3]. CFS extends target suppression by presenting dynamic masks to one eye [4]. Later reviews also warn that breaking-CFS results depend on suppression depth, individual differences, and interpretation choices [5].

For that reason, VisionWeaver borrows from CFS only at the level of task design. The present hardware is a single monitor, so the implemented task is side-by-side and CFS-inspired rather than full dichoptic CFS with a stereoscope. A stereoscopic version would be needed before the project could make claims about true dichoptic presentation.

2.3. Adaptive Psychophysical Methods

Adaptive psychophysical methods shaped the controller choice, although this build stops short of threshold estimation. Transformed up-down procedures show one way to move stimulus intensity after hit/miss responses [7]. QUEST shows how a threshold can be estimated with a Bayesian model [8]. The practical lesson we needed at this stage was simpler: if difficulty never responds to performance, many trials can sit at the wrong level.

We chose a moving-window controller because it can be traced in the equation, the source code, and the CSV logs. That choice is not presented as better than staircase or QUEST. It is a deliberately inspectable rule for an early system.

Fixed and staircase modes remain useful comparisons. Fixed difficulty is replayable and exposes what happens when performance is ignored. A 1-up/2-down staircase gives a familiar adaptive reference point. QUEST and QUEST+ remain future options because they would require clearer psychometric assumptions and priors than we use here.

2.4. Visual Training, Amblyopia, and Game-Like Design

Visual training, amblyopia, and game-like intervention papers affected the interface more than the claim. Repeated practice can improve trained-task performance in some settings [9], and amblyopia studies have tested digital or game-like interventions with compliance records [9]. For VisionWeaver, the transferable design lesson is the need to record what was presented, how difficulty changed, and which sessions were actually included.

Those studies do not prove that this system improves DR vision, and our data do not include patient outcomes. They support the motivation for an interactive, logged visual task; they do not supply therapeutic evidence for the current build.

2.5. Psychophysical Software Ecosystem

PsychoPy is the software decision we take most directly from the psychophysical-software literature [10]. We use it because stimulus presentation, timing control, and response collection are exactly the parts that should not be improvised in a perception experiment.

Using PsychoPy does not validate the whole VisionWeaver pipeline by itself. In our implementation, PsychoPy supports the CFS-inspired stimulus task. The moving-window controller, logging layer, and offline analysis are project-specific pieces and must be evaluated on their own.

3. System and Methods

3.1. System Architecture and Data Flow

The implementation was organized around the trial record. Stimulus presentation, response collection, parameter updates, per-trial logging, and offline analysis all write into or read from the same data chain. The system has five layers: task presentation (CFS and central-gate stimuli), input (Arduino buttons or keyboard fallback), controller (difficulty updates based on recent hit rate), logging (per-trial CSV), and analysis (session metrics and figure regeneration from logs).

This design makes the later analysis less dependent on a screenshot or a live demonstration. Task states can be traced back to raw CSV, controller behavior can be replayed offline, and figures can be regenerated by fixed scripts. Input is kept separate from logging: regardless of input method (Arduino or keyboard), the analysis pipeline receives the same trial record format, as shown in Figure 2.

Auditable prototype workflow

Stimulus presentation, adaptive control, logging, and analysis are linked by trial-level data.

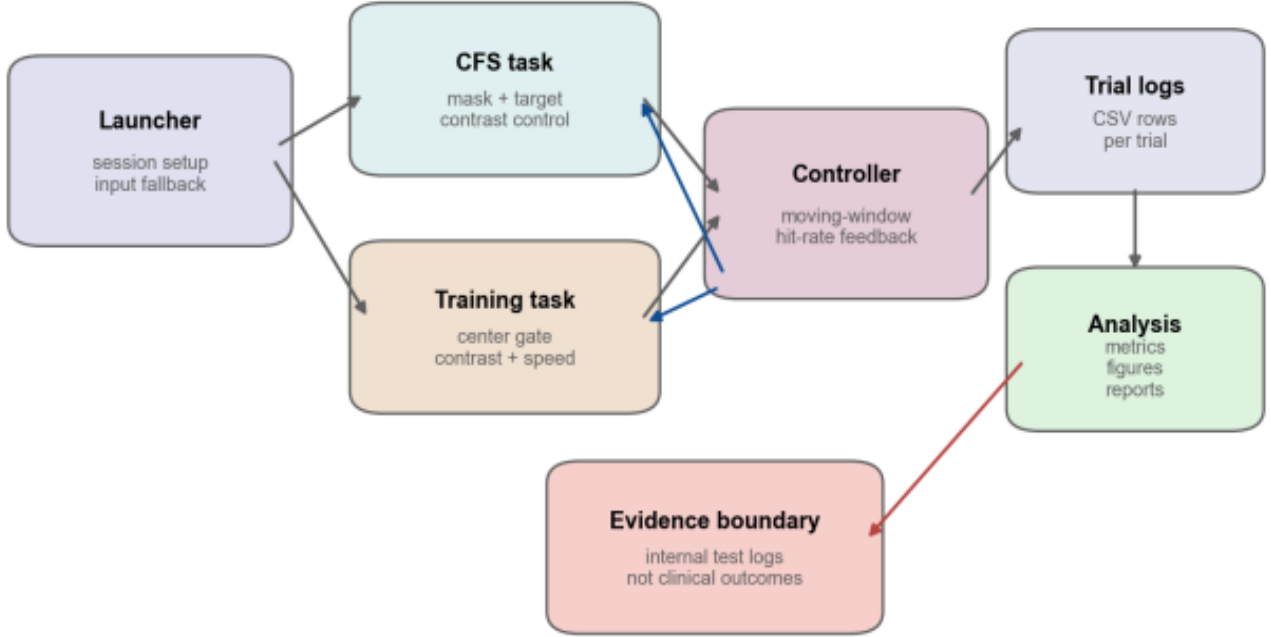


Figure 2. System architecture and data flow.

3.2. CFS-Inspired Assessment Task

The CFS-inspired task uses a side-by-side presentation format: one side displays a 12 Hz dynamic Mondrian mask while the other presents a low-contrast Gaussian target. Each trial records task condition, mask frequency, target side-swap state, target contrast, target presentation duration, hit status, reaction time, and target position. Target contrast is the primary adaptive variable: when recent hit rate is high, contrast decreases; when low, contrast increases. Contrast is bounded between 0.03 and 1.00.

3.3. Central-Gate Training Task

The central-gate training task uses a central-gate response design. A moving target traverses a dynamic background, and users respond when the target enters the vicinity of the central gate. The training module uses dual-parameter difficulty adjustment: when recent performance is high, target contrast decreases and background speed increases; when low, contrast increases and speed decreases.

3.4. Moving-Window Controller

The controller operates on binary trial outcomes $h_t \in \{0, 1\}$, where $h_t = 1$ indicates a hit. For trial t , the hit rate over the most recent u trials is:

$$\hat{p}_t = \frac{1}{m_t} \sum_{i=\max(1, t-w+1)}^t h_i, \quad m_t = \min(w, t) \quad (1)$$

Figure 3. VisionWeaver Prototype — Actual Running Screenshots

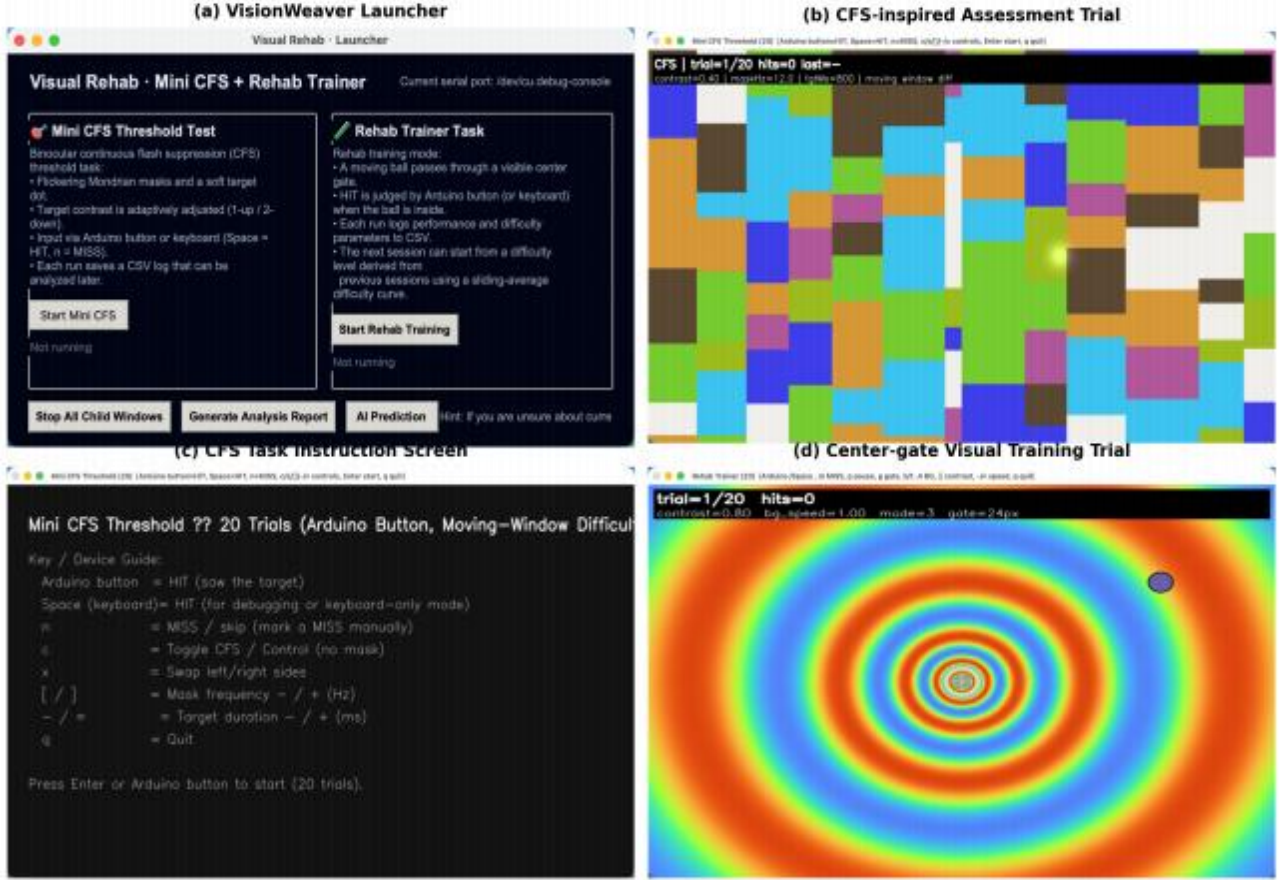


Figure 3. CFS task, central-gate training task, and moving-window controller (actual project screenshots).

In the CFS-inspired task, as shown in Figure 3, the target hit rate is $p^* = 0.75$ with tolerance band $\varepsilon = 0.05$:

$$c_{t+1} = \begin{cases} c_t - \Delta_c & \text{if } \hat{p}_t > p^* + \varepsilon \\ c_t + \Delta_c & \text{if } \hat{p}_t < p^* - \varepsilon \\ c_t & \text{otherwise} \end{cases} \quad (2)$$

In the central-gate training task, the controller simultaneously updates target contrast c_t and background speed s_t with target band $[0.60, 0.80]$:

$$(c_{t+1}, s_{t+1}) = \begin{cases} (c_t - \Delta_c, s_t + \Delta_s) & \text{if } \hat{p}_t > 0.80 \\ (c_t + \Delta_c, s_t - \Delta_s) & \text{if } \hat{p}_t < 0.60 \\ (c_t, s_t) & \text{otherwise} \end{cases} \quad (3)$$

3.5. Data Inclusion Rules and Offline Verification

We define log inclusion rules before reporting results to avoid selective presentation of favorable sessions. Complete logs are used for primary engineering verification; legacy logs for compatibility auditing; partial/special logs for failure mode analysis. All March 2026 clean sessions are labeled as internal test evidence, not clinical participant data.

3.6. DR-Inspired Simulated Vision Degradation Levels

We built a parameterized image degradation filter with five DR-inspired simulated vision degradation levels (R0–R4) to test system adaptation without real patients. These levels are simula-

Table 1. DR-Inspired Simulated Vision Degradation Filter Parameters

Simulated Level	Blur σ	Contrast	Noise	DR Inspiration
R0 (baseline)	0	1.00	0	Normal-retina reference
R1	0.8	0.85	0.05	Microaneurysm-inspired
R2	1.5	0.70	0.10	Hemorrhage-inspired
R3	2.2	0.55	0.15	Cotton-wool-spot-inspired
R4	3.0	0.40	0.20	Neovascularization-inspired

tion conditions, not clinical grading labels. The filter maps three visual degradation dimensions to image operations: resolution loss, contrast reduction, and increased visual noise, as shown in Table 1.

The filter output is fed into an ideal observer model that converts image-level degradation into synthetic psychophysical parameters. For each simulated degradation level, we compute d' as a function of contrast, then fit a psychometric function to extract contrast threshold $c_{\text{threshold}}$ (75% correct) and slope:

Level	$c_{\text{threshold}}$	β
R0	0.184	0.100
R1	0.259	0.138
R2	0.480	0.244
R3	0.491	0.249
R4	0.712	0.327

3.7. Monte Carlo Synthetic Responder Controller Evaluation

The Monte Carlo synthetic responder is a sigmoid psychometric function: given current trial contrast c , the synthetic responder generates a hit with probability:

$$P(\text{hit} | c) = \frac{1}{1 + \exp\left(-\frac{c - c_{\text{threshold}}}{\beta}\right)} \quad (4)$$

Each controller (Fixed, Staircase, Moving-window) $\times$ each DR-inspired simulated vision degradation level (R0–R4) runs 200 independent simulations of 30 trials each (total 3,000 simulations). We record band occupancy, MAE, and convergence trials.

3.8. Behavior-Sequence-Based Simulated Degradation Level Infer- ence

We train a random forest classifier to infer a synthetic session’s DR-inspired simulated vision degradation level (R0–R4) from the first 10 trials of behavioral data. Training data: 7,500 synthetic ML sessions (1,500/level). From 10 trials we extract 53 features (per-trial values, cumulative statistics, sequence features), as shown in Table 2 and Table 3. The classifier outputs simulated degradation level probabilities, and recommended starting parameters are computed as:

$$c_{\text{recommended}} = \sum_i P(R_i) \times c_{\text{threshold}}(R_i) \quad (5)$$

Table 2. Data Quality Audit Summary

Task	Group	Sessions	Trials	Hit Rate
CFS	Clean	6	120	0.783
CFS	Legacy	14	298	0.346
CFS	Partial	15	17	0.000
Training	Clean	9	180	0.411
Training	Legacy	1	20	0.000
Training	Partial	10	30	0.033

Table 3. CFS Controller Comparison (R0 Simulated Degradation Level, $n = 200$ simulations)

Controller	Band Occupancy	Interpretation
Fixed	0.1%	Almost never in target band
Staircase	35.8%	$\sim 1/3$ trials in band
Moving-window	22.4%	$\sim 1/5$ trials in band

4. Internal Verification Results

4.1. Session Scale and Data Quality

From real internal project logs, the metrics script parsed 55 session logs with 665 per-trial records. CFS: 6 clean complete sessions (120 trials); central-gate training task: 9 clean complete sessions (180 trials). In these internal training logs, the initial difficulty ($c_0 = 0.6$, $S_0 = 0.5$) was set empirically without bench calibration; the measured mean hit rate of 41% is below the target band (60–80%), indicating initial difficulty was too high. This is a known engineering constraint to be resolved through calibration and longer sessions.

4.2. Monte Carlo Controller Comparison

The Monte Carlo runs gave the controller comparison a clearer shape. In this synthetic setting, both adaptive controllers spent far more time in the target band than fixed difficulty. Staircase reached the highest band occupancy, while moving-window showed larger interpretability benefits and a less favorable occupancy score. When the simulated degradation level increased, every controller spent less time in band, although the adaptive modes still remained ahead of fixed difficulty.

4.3. Ablation Analysis

From the R0 synthetic responder ablation, Speed-only achieves 0% band occupancy (identical to Fixed) because the synthetic responder’s sigmoid takes only contrast as input. Contrast-only and Dual-parameter are identical (10.4%) because the speed channel has no effect on the synthetic responder. This confirms that, under the current simulation only, contrast is the only effective adaptive dimension. Real users may be affected by background motion, but this requires human experiments.

4.4. Results Boundaries

The current evidence is strongest for the internal engineering loop. The system generates structured per-trial logs, the analysis scripts compute metrics from those logs, the controller

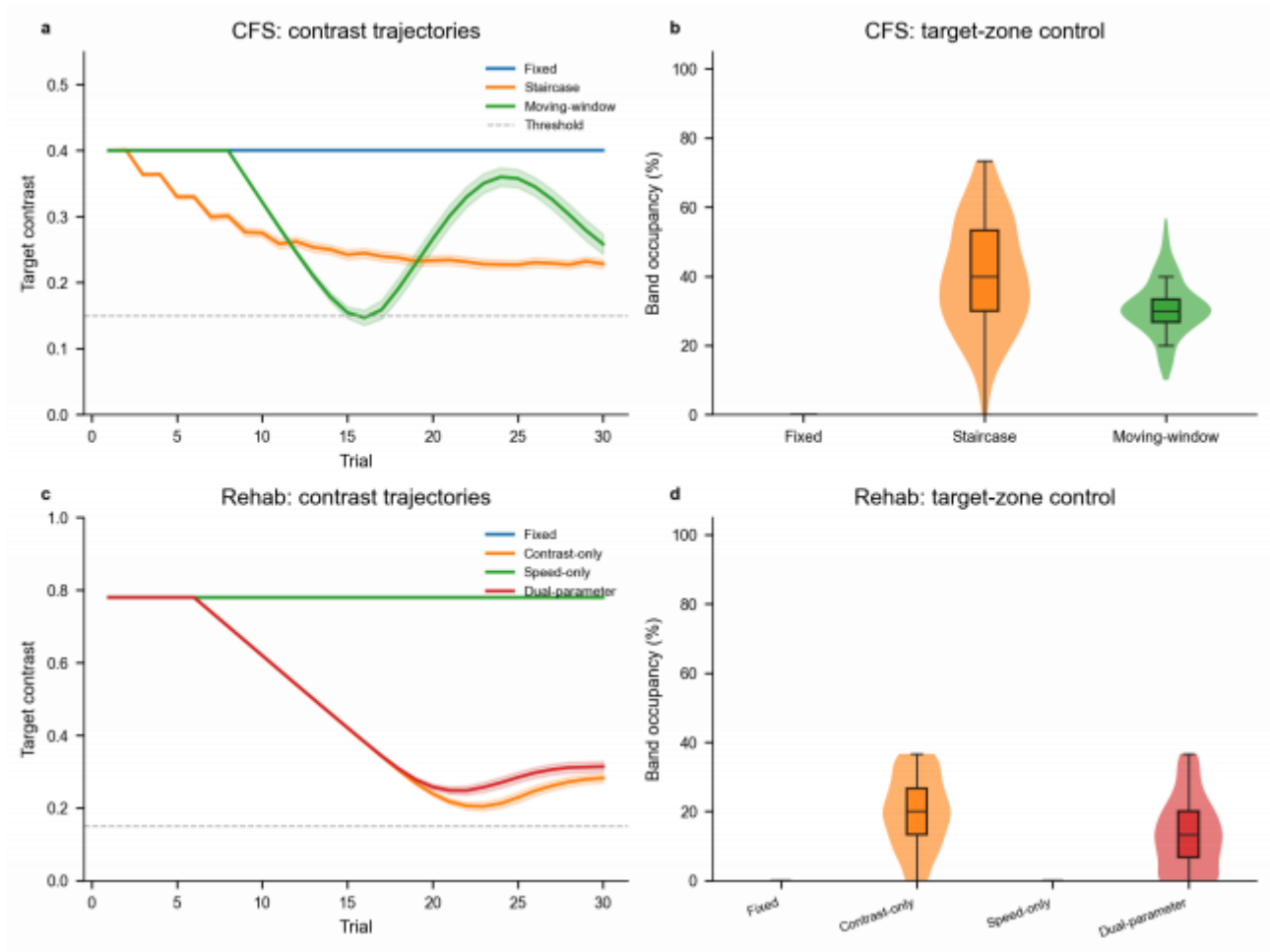


Figure 4. Monte Carlo controller comparison (R0 simulated degradation level): Staircase (35.8%) > Moving-window (22.4%) >> Fixed (0.1%).

can be replayed offline, and failure logs expose weaknesses in the build. This evidence does *not* support clinical efficacy, patient outcomes, or behavioral improvement claims, as shown in Figure 4.

4.5. Controller Behavior Under DR-Inspired Simulated Vision Degradation

From Monte Carlo synthetic responder runs across DR-inspired simulated vision degradation levels, the moving-window controller adjusted toward different contrast ranges: R0 converged to contrast ≈ 0.22 , R1 to 0.30, R2 to 0.55, R3 to 0.58, and R4 to 0.97. This shows simulated controller adaptation across the parameter range, not per-user calibration in real participants. as shown in Figure 5 and Table 4.

Table 4. Central-Gate Training Task Ablation (R0 Synthetic Responder)

Controller	Band Occupancy	Note
Fixed	0.0%	Both fixed
Contrast-only	10.4%	Only contrast adjusted
Speed-only	0.0%	Only speed adjusted
Dual-parameter	10.4%	Both adjusted

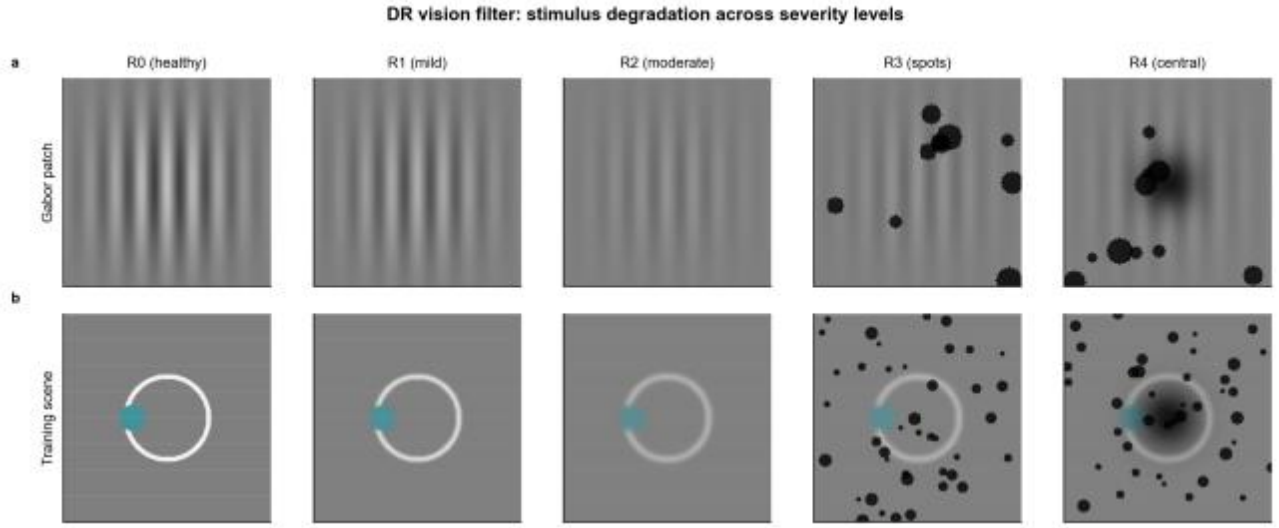


Figure 5. DR-inspired simulated vision degradation: same Gabor stimulus filtered through R0–R4 simulated degradation levels.

4.6. ML Simulated Degradation Level Inference and Warm-Start

From 7,500 synthetic ML sessions, the 5-class classifier for DR-inspired simulated vision degradation levels achieves 72.9% overall accuracy (5-fold CV: $71.9\% \pm 1.0\%$; chance: 20%). Errors concentrate between adjacent simulated degradation levels. The most informative features are cumulative hit rate, total miss count, and mean contrast, as shown in Figure 6 and Table 5.

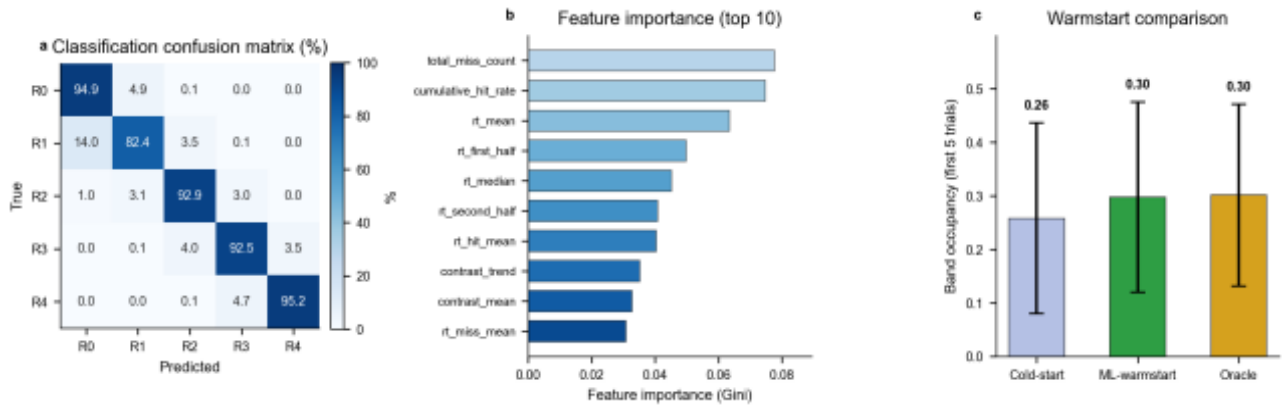


Figure 6. Synthetic ML pipeline: 5×5 simulated degradation level confusion matrix (accuracy 72.9%) and top-10 feature importance.

Table 5. Synthetic ML Warm-Start Comparison

Strategy	Band Occupancy	Convergence Trials
Cold-start	0.258	16.3
ML-warmstart	0.298 (+15.5%)	15.2
Oracle	0.302	15.0

From synthetic ML warm-start simulations, ML-warmstart improves band occupancy from 0.258 to 0.298 (+15.5%), reaching 98.7% of the simulated Oracle upper bound. The system automatically selects appropriate difficulty starting points after 10 synthetic trials, as shown in Figure 7.

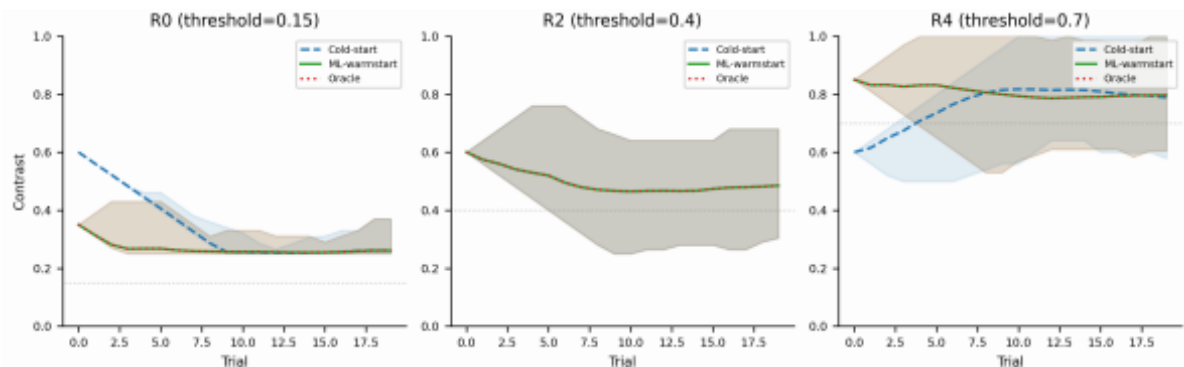


Figure 7. Synthetic ML warm-start comparison: ML-recommended (0.298, +15.5%) vs Cold-start (0.258) vs Oracle (0.302).

5. Discussion

5.1. What the Current Build Actually Proves

The strongest outcome of this round is practical: a VisionWeaver run can be replayed as an engineering record. The internal logs preserve task parameters, trial outcomes, controller decisions, and failure cases, so the session does not have to be judged from interface behavior alone. That record made it possible to compare controllers, inspect bad runs, and connect failure-log review back to concrete implementation issues such as overly difficult initial settings, zero-hit sessions, and parameter-boundary behavior. The result is still an engineering result, not a clinical claim.

The controller comparison is best read as a stress test of the closed-loop logic under simulated responders. In the R0 Monte Carlo setting, Staircase reached 35.8% band occupancy, Moving-window reached 22.4%, and Fixed difficulty reached only 0.1%. That result supports the engineering choice to keep adaptive control in the system. It also shows that controller “smoothness” and target-band performance are different objectives. Staircase was more effective by the current metric; Moving-window was easier to replay and interpret, while still capable of drifting toward boundaries.

5.2. Simulated Data Flow Review

The verification path was staged and simulated. Internal logs first showed whether the tasks generated analyzable per-trial records. Simulated responders then made the same tasks repeatable enough for Monte Carlo comparison. Parameter sweeps checked whether the controller moved through plausible contrast ranges instead of staying fixed. The ablation isolated which control channel mattered under the current responder model. Finally, the ML warm-start experiment tested whether early synthetic trial summaries could choose a better starting point.

That trail matters during review because every stage leaves an inspectable artifact: logs, controller trajectories, sweep outputs, ablation tables, confusion matrices, and warm-start comparisons. The 72.9% simulated degradation-level classifier accuracy is meaningful in that context: it was high enough to improve synthetic warm-start band occupancy by +15.5% and reach 98.7% of the simulated Oracle. It should not be read as diagnostic accuracy for diabetic retinopathy; the labels are simulated degradation levels and the input behavior comes from synthetic sessions.

5.3. What Still Does Not Transfer

The main non-transfer point is the responder. The current simulated responders assume stable thresholds and do not model fatigue, learning, attention shifts, hesitation, misunderstanding, or motor/touchscreen variability. The ablation makes this boundary explicit: Speed-only had no effect because the R0 synthetic responder used contrast as the active input. That is a useful internal check, not evidence that speed is irrelevant for real users. A real participant may react to motion, timing, background clutter, or task anxiety in ways this responder cannot express.

The internal project logs are also uneven. Clean sessions support replay and metric extraction, but legacy, partial, zero-hit, and low-hit runs show that the prototype still needs calibration before human-facing interpretation. These logs are valuable because they expose engineering failure modes. They are not a substitute for supervised usability or clinical data collection, and they do not establish DR treatment efficacy, diagnostic performance, disease staging, or improvement in real patient outcomes.

5.4. Next Engineering Step

The next round should keep the same evidence discipline. Before any clinical framing, the system needs bench calibration and a small healthy-volunteer pilot focused on usability, task comprehension, session length, failure handling, and whether the logged metrics remain stable with real human responses. If that pilot is acceptable, the following stage can be a supervised DR participant study that evaluates feasibility and behavior under clinical oversight. Clinical outcome work should come only after those steps.

6. Conclusion

VisionWeaver was tested as a simulation-based engineering prototype. The current evidence supports three narrow findings:

1. In synthetic Monte Carlo experiments, the virtual responder separated controller behavior clearly enough to compare band occupancy (Staircase 35.8%, Moving-window 22.4%, Fixed 0.1%).
2. The DR-inspired simulated vision degradation levels converted DR motivation into a five-level simulation range ($c_{\text{threshold}}$ from 0.184 to 0.712), allowing controller behavior to be checked across levels.
3. In the simulated setting, the ML classifier reached 72.9% accuracy and enabled warm-start, improving band occupancy by 15.5% (98.7% of Oracle).

All results are simulation/internal engineering evidence, not clinical efficacy evidence. The next step should be usability testing with healthy volunteers before any supervised clinical scenario validation.

References

- [1] D. G. Pelli and P. Bex, “Measuring contrast sensitivity,” *Vision Research*, vol. 90, pp. 10–14, 2013.
- [2] M.-C. Silva-Viguera, et al., “Contrast sensitivity function in diabetic retinopathy,” *J. Clinical Medicine*, vol. 12, no. 12, 4136, 2023.
- [3] R. Blake and N. K. Logothetis, “Visual competition,” *Nature Reviews Neuroscience*, vol. 3, pp. 13–21, 2002.
- [4] N. Tsuchiya and C. Koch, “Continuous flash suppression reduces negative afterimages,” *Nature Neuroscience*, vol. 8, pp. 1096–1101, 2005.
- [5] S. Gayet, S. Van der Stigchel, and C. L. E. Paffen, “Breaking continuous flash suppression: Competing for consciousness,” *Frontiers in Psychology*, vol. 5, 460, 2014.
- [6] Z. L. Teo, Y.-C. Tham, M. Yu, et al., “Global prevalence of diabetic retinopathy and projection of burden through 2045,” *Ophthalmology*, vol. 128, no. 11, pp. 1580–1591, 2021.
- [7] H. Levitt, “Transformed up-down methods in psychoacoustics,” *J. Acoustical Society*, vol. 49, pp. 467–477, 1971.
- [8] A. B. Watson and D. G. Pelli, “QUEST: A Bayesian adaptive psychometric method,” *Perception & Psychophysics*, vol. 33, pp. 113–120, 1983.
- [9] T. Y. Gao, C. J. A. Guo, et al., “Effectiveness of a binocular video game vs placebo video game for improving visual functions in older children, teenagers, and adults with amblyopia,” *JAMA Ophthalmology*, vol. 136, pp. 172–181, 2018.
- [10] J. W. Peirce, et al., “PsychoPy2: Experiments in behavior made easy,” *Behavior Research Methods*, vol. 51, pp. 195–203, 2019.