Efficient Nano-Optical Eye Tracking for Smart Glasses Supplementary Material

JIPENG SUN, Google Inc., Princeton University, USA
LEKANG YUAN, Princeton University, USA
RUSLAN GUSEINOV, Google Inc., Austria
BERND BICKEL and THABO BEELER, Google Inc., Switzerland
THOMAS AUZINGER, Google Inc., Austria
FELIX HEIDE, Princeton University, USA

Contents

Supplementary video. The video accompanying this document shows the nano-optical prototype tracking gaze in real time: participants seated in a head-fixed rig play a whack-a-mole game controlled by their gaze while motorized x and z translation stages move the camera to emulate glasses slippage. The ring on the display marks

the estimated gaze point, and the counter tallies targets hit. The main view is a side view of the rig; insets show close-ups of the two stages and the rig geometry (Fig. S1).

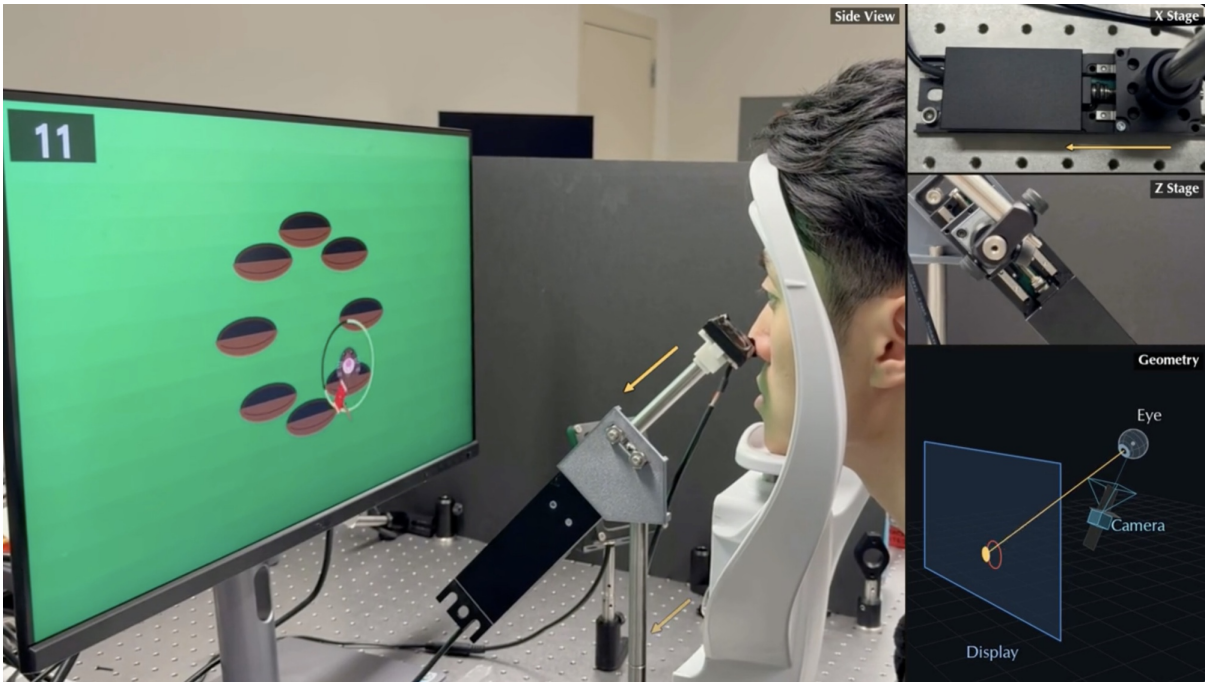

Fig. S1: Frame from the supplementary video. Left: a participant in the head-fixed rig plays the gaze-controlled whack-a-mole game; the ring on the display marks the gaze point that the nano-optical prototype below the eye estimates in real time, and the counter tallies targets hit. Right, top to bottom: the x and z translation stages that move the camera to emulate glasses slippage (arrows show the direction of motion), and the rig geometry (eye, camera, display).

A ADDITIONAL DETAILS ON FORWARD PROCESS

A.1 Effective Pixel Pitch with Noise

Resolving minute defocus gradients requires a dense PSF stack, computed via a distributed strategy [Sun et al. 2025] where partial depth blocks are synchronized to form the global volume $\mathbf{P} = \text{AllGather}(\{\widetilde{\text{PSF}}(z_k) \mid k \in \mathcal{Z}_r\})$. We accumulate the sensor intensity I_{acc} recursively using an occlusion-aware model [Ikoma et al. 2021]:

$$I_{acc}^{(k)} = (\alpha_k L_k * \mathbf{P}[k]) + (1 - \alpha_k * \mathbf{P}[k]) \cdot I_{acc}^{(k-1)}. \tag{S1}$$

To optimize the effective input resolution (spanning physical pitch, binning, or resizing), we treat the effective pitch $\hat{p}$ as a learnable parameter. We bridge the continuous/discrete gap using a spatial integration envelope $\mathcal{W}_{\hat{p}}$ (a box filter of size $\hat{p}$) as a differentiable proxy. The final measurement I is computed by applying this envelope prior to discrete downsampling $\mathcal{D}_{\downarrow}$:

$$I = \mathcal{N}_{sensor}\left(\mathcal{D}_{\downarrow}(I_{acc} * \mathcal{W}_{\hat{p}}) \cdot \kappa\right), \quad \text{with } \kappa = (\hat{p}/p_{base})^2. \tag{S2}$$

Here, κ ensures energy conservation for the larger effective pixels, and $\mathcal{N}_{sensor}$ applies photon shot noise and read noise. This formulation allows gradients to flow to $\hat{p}$ via the kernel size of $\mathcal{W}_{\hat{p}}$, enabling end-to-end optimization of the trade-off between spatial resolution and SNR.

A.2 Metasurface Profile Initialization

To prevent the end-to-end optimization from collapsing into local optima where the neural network compensates for poor optical quality, we initialize the metasurface using a dedicated physics-driven training phase. This phase relies on a custom implementation of the Fisher Information optimization loop described in the main text.

Distributed Gradient Computation. Computing the gradient of the PSF with respect to depth, $\partial\text{PSF}/\partial z$, is computationally challenging because high-resolution wave propagation simulations are memory-intensive, necessitating the distribution of depth planes across multiple GPUs. To compute the derivative across the full depth volume, we implement a custom distributed gradient layer. We utilize a differentiable `all_gather` operation to collect normalized PSF tensors from all ranks. These tensors are then sorted by their corresponding depth indices to form a global stack, allowing us to compute the numerical derivative $\Delta\text{PSF}/\Delta z$ via finite differences across the device boundary.

Initialization Loss Components. The initialization objective is a composite of three physically motivated constraints designed to shape the PSF for optimal depth encoding:

1. Precision Loss (CRLB): To maximize the sensitivity of the PSF to depth changes, we minimize the Cramér-Rao Lower Bound. This is calculated using the Fisher Information Matrix (FIM), utilizing the distributed depth gradients described above:

$$\mathcal{L}_{prec} = \log \text{Tr}\left(\mathbf{I}^{-1}\right), \quad \text{where } \mathbf{I} = \sum_{x,y} \frac{1}{\text{PSF}(x,y)} \left(\nabla\text{PSF}(x,y)\right) \left(\nabla\text{PSF}(x,y)\right)^T. \tag{S3}$$

Here, ∇PSF includes gradients with respect to spatial dimensions (x, y) and depth z.

2. Distinctness Loss: To ensure that PSFs at different depths are distinguishable and orthogonal, we minimize the cosine similarity between flattened PSF vectors at different depth indices z_i and z_j. We weigh this penalty by the physical distance $|z_i - z_j|$ to enforce sharper transitions between adjacent depth planes:

$$\mathcal{L}_{dist} = \sum_{i \neq j} |z_i - z_j| \cdot \frac{\text{PSF}(z_i) \cdot \text{PSF}(z_j)}{||\text{PSF}(z_i)||_2 \cdot ||\text{PSF}(z_j)||_2}. \tag{S4}$$

3. Compactness Loss: To conserve energy and prevent spatial aliasing (speckle noise), we encourage the energy to concentrate into tight focal spots by minimizing the Shannon Entropy of the PSF intensity distribution:

$$\mathcal{L}_{comp} = -\sum_{x,y} \widehat{\text{PSF}}(x,y) \log(\widehat{\text{PSF}}(x,y) + \epsilon), \tag{S5}$$

where $\widehat{\text{PSF}}$ is the spatially normalized probability mass function of the intensity image.

Auto-Balanced Composite Loss. To avoid manual tuning of the weighting hyperparameters between these competing objectives, we model the problem as multi-task learning with homoscedastic uncertainty. We introduce learnable weights σ_k for each loss component $k \in \{prec, dist, comp\}$ and minimize the joint objective:

$$\mathcal{L}_{init} = \sum_k \left(\frac{1}{2\sigma_k^2} \mathcal{L}_k + \log \sigma_k \right). \tag{S6}$$

This allows the optimization to autonomously balance the trade-off between maximizing depth sensitivity and maintaining a focused, energy-efficient optical spot.

A.3 Task Loss Components

The task loss $\mathcal{L}_{task}$ is a weighted sum of a segmentation term (Perception Head) and a calibration term (Calibration Head). For the perception head we use the Dice coefficient loss [Milletari et al. 2016] to handle the class imbalance between the pupil and background,

$$\mathcal{L}_{seg} = 1 - \frac{2\sum_i \mathbf{M}_i \mathbf{M}_i^*}{\sum_i \mathbf{M}_i + \sum_i \mathbf{M}_i^* + \epsilon}, \tag{S7}$$

where $\mathbf{M}^*$ is the ground-truth pupil mask. For the calibration head we use a mean-squared error on the predicted slippage vector,

$$\mathcal{L}_{calib} = \|\Delta\mathbf{t} - \Delta\mathbf{t}^*\|_2^2. \tag{S8}$$

The total task loss is $\mathcal{L}_{task} = \mathcal{L}_{seg} + \lambda_{slip}\,\mathcal{L}_{calib}$. Minimizing this objective drives the network to decouple the 2D pupil projection from 3D camera motion, providing the intermediate physical states consumed by the geometric solver in Sec. A.4.

A.4 Geometric Unprojection & Gaze Estimation

The geometric solver $\mathcal{F}_{geo}$ functions as a deterministic post-processing module that fuses the 2D segmentation mask $\mathbf{M}$ (from the Perception Head) and the 3D slippage vector $\Delta\mathbf{t}$ (from the Calibration Head) to recover the gaze vector. We model the eye as a sphere of radius $R_{eye} \approx 12$mm centered at the origin of the world coordinate system $\mathbf{O}_{world} = [0, 0, 0]^T$.

1. Centroid Extraction. First, we extract the 2D pupil center $\mathbf{c}_{pix} = (u_p, v_p)$ from the predicted binary mask $\mathbf{M}$. To maintain differentiability during the training of the perception head, we calculate the center of mass:

$$u_p = \frac{\sum_{u,v} u \cdot \mathbf{M}_{u,v}}{\sum_{u,v} \mathbf{M}_{u,v}}, \quad v_p = \frac{\sum_{u,v} v \cdot \mathbf{M}_{u,v}}{\sum_{u,v} \mathbf{M}_{u,v}}. \tag{S9}$$

2. Ray Casting in Camera Frame. We back-project the centroid $\mathbf{c}_{pix}$ into a normalized direction vector $\mathbf{d}_{cam}$ in the local camera coordinate frame using the optimized intrinsic matrix $\mathbf{K}$:

$$\mathbf{d}_{cam} = \frac{\mathbf{K}^{-1}[u_p, v_p, 1]^T}{\left\|\mathbf{K}^{-1}[u_p, v_p, 1]^T\right\|_2}. \tag{S10}$$

Here, $\mathbf{K}$ is defined by the effective pixel pitch $\hat{p}$ and the focal length f determined by the meta-optics design.

3. Slippage Correction. The key innovation of our method is the dynamic correction of the camera pose. We define the camera's optical center $\mathbf{O}_{cam}$ by applying the predicted slippage $\Delta\mathbf{t}$ to the nominal camera position $\mathbf{T}_{nom}$:

$$\mathbf{O}_{cam} = \mathbf{T}_{nom} + \Delta\mathbf{t}. \tag{S11}$$

Assuming the slippage is primarily translational (as rotational slippage is coupled with gaze rotation in the spherical model), the ray direction in the world frame remains $\mathbf{d}_{world} \approx \mathbf{d}_{cam}$.

4. Ray-Sphere Intersection. The 3D pupil position $\mathbf{P}_{3D}$ lies at the intersection of the camera ray and the eye sphere. We define the ray as $\mathbf{r}(s) = \mathbf{O}_{cam} + s \cdot \mathbf{d}_{world}$. Solving for the intersection with the sphere $\|\mathbf{r}(s)\|^2 = R_{eye}^2$ leads to the quadratic equation:

$$\|\mathbf{d}_{world}\|^2 s^2 + 2(\mathbf{O}_{cam} \cdot \mathbf{d}_{world})s + (\|\mathbf{O}_{cam}\|^2 - R_{eye}^2) = 0. \tag{S12}$$

We solve for s and select the smaller positive root (representing the front surface of the eye) to obtain the intersection distance s^*. The 3D pupil centroid is then:

$$\mathbf{P}_{3D} = \mathbf{O}_{cam} + s^* \cdot \mathbf{d}_{world}. \tag{S13}$$

5. Gaze Vector Computation. Finally, the gaze vector $\mathbf{g}$ is defined as the normalized vector extending from the center of the eye to the 3D pupil centroid:

$$\mathbf{g} = \frac{\mathbf{P}_{3D} - \mathbf{O}_{world}}{\|\mathbf{P}_{3D} - \mathbf{O}_{world}\|_2} = \frac{\mathbf{P}_{3D}}{\|\mathbf{P}_{3D}\|_2}. \tag{S14}$$

This vector $\mathbf{g}$ represents the optical axis, which can be further corrected to the visual axis by a fixed kappa angle offset if personalization data is available.

A.5 Geometric Interpretation of the Fisher-Information Objective

This subsection expands the one-sentence geometric reading of the Fisher Information Matrix (FIM) given in main-paper Sec. 3.2. Geometrically, we interpret the sensitivity fields $\mathbf{J}_1$ and $\mathbf{J}_2$ as vectors residing in the high-dimensional pixel space $\mathbb{R}^{|\Omega|}$. The diagonal elements of the FIM correspond to the squared magnitudes of these vectors, such that $\mathbf{F}_{ii} = \|\mathbf{J}_i\|^2$. This quantity represents the sensitivity of the optical system to the i-th motion parameter. The off-diagonal term $\mathbf{F}_{12}$ captures the geometric alignment between these vectors. The orthogonality of the encoding is determined by the span angle ϕ between $\mathbf{J}_1$ and $\mathbf{J}_2$, defined as

$$\cos\phi = \frac{\mathbf{F}_{12}}{\|\mathbf{J}_1\|\|\mathbf{J}_2\|} = \frac{\mathbf{F}_{12}}{\sqrt{\mathbf{F}_{11}\mathbf{F}_{22}}}. \tag{S15}$$

As illustrated in main-paper Fig. 3, to facilitate passive optical computing, our optimization goal is two-fold: maximize the vector magnitudes ($\|\mathbf{J}_i\| \to \infty$) to ensure high sensitivity, and drive the angle $\phi \to 90^\circ$ (where $\cos\phi \to 0$) to ensure the encodings are orthogonal and statistically independent. The optical embedding loss of main-paper Eq. 6 implements these two goals through its sensitivity and orthogonality terms.

B ADDITIONAL DETAILS ON BI-LEVEL OPTIMIZATION

B.1 Bi-Level Objectives

We formalize the bi-level optimization used in main-paper Sec. 3.5. Given a discrete architecture $\alpha \in \mathcal{A}$ and continuous parameters $\theta = \{\phi_{opt}, \hat{p}, \mathbf{w}_{net}\}$ (metasurface phase, effective pixel pitch, and network weights), the inner loop solves a budget-penalized task loss and the outer loop maximizes validation accuracy $\mathcal{V}$:

$$\theta^*(\alpha) = \underset{\theta}{\operatorname{argmin}} \Big[\mathcal{L}_{task}(\alpha, \theta) + \lambda \, \max\big(0, \, C_{total}(\alpha, \hat{p}) - C_{budget}\big) \Big], \tag{S16}$$

$$\alpha^* = \underset{\alpha \in \mathcal{A}}{\operatorname{argmax}} \, \mathcal{V}\big(\alpha, \theta^*(\alpha)\big), \tag{S17}$$

where λ is the FLOPs-budget penalty. The inner objective thus binds optics, sensor pitch, and subnet weights to a single continuous optimization, and the outer objective searches the architecture space with REA (Sec. B.3).

B.2 Inner Loop: Weight Inheritance and Short-Horizon Fine-Tuning

We solve the inner minimization in Eq. S16 with end-to-end stochastic first-order optimization via autograd differentiation [Sitzmann et al. 2018]. Because we model both the optical wavefront propagation and the sensor integration as fully differentiable modules, gradients can be computed with respect to the heterogeneous parameters $\theta = \{\phi_{opt}, \hat{p}, \mathbf{w}_{net}\}$ and the loss minimized by standard backpropagation.

Solving the inner problem from scratch for each candidate architecture is expensive, since random initialization requires many iterations to converge. We therefore use a weight-inheritance strategy inspired by one-shot architecture search [Cai et al. 2019; Pham et al. 2018]: the inner-loop parameters are initialized using a projection mapping $\mathcal{M}$ from the pre-trained Supernet $\mathcal{S}$ with an initial sensor pixel pitch p_{init},

$$\theta_{init}(\alpha) \;=\; \{\, \phi_{opt}^{\mathcal{S}}, \; p_{init}, \; \mathcal{M}(\mathcal{S}, \alpha) \,\}. \tag{S18}$$

The parameters θ are then tuned via SGD for a short horizon (3 epochs). This fast adaptation lets the continuous variables — specifically the optical phase ϕ_{opt} and subnet weights $\mathbf{w}_{net}$ — co-adapt to the capacity of the sampled architecture α without incurring the cost of full retraining.

B.3 Regularized Evolutionary Algorithm Details

In the outer loop of our bi-level optimization framework, we employ a Regularized Evolutionary Algorithm (REA) to efficiently navigate the discrete, non-differentiable search space of neural architectures $\mathcal{A}$. Unlike standard genetic algorithms that often suffer from premature convergence due to elitism, REA maintains population diversity by strictly removing the oldest individuals rather than the lowest-performing ones.

The evolutionary process begins by initializing a population $\mathcal{P}_0$ of size P with architectures sampled uniformly from the search space. The search then proceeds through a cyclical process for a fixed number of iterations N_{iter}. In each step t, we employ a tournament selection strategy where a random subset $\mathcal{T} \subset \mathcal{P}_t$ of size k is sampled. The parent α_{parent} is selected by maximizing a penalized fitness metric that accounts for the hardware constraints, consistent with the outer objective defined in the main text:

$$\alpha_{parent} = \underset{\alpha \in \mathcal{T}}{\operatorname{argmax}} \left[\mathcal{V}(\alpha, \theta^*) - \gamma \cdot \mathbb{I}(C_{total}(\theta^*) > C_{budget}) \right], \tag{S19}$$

where $\mathcal{V}$ is the validation accuracy, θ^* represents the optimized parameters from the inner loop, γ is the penalty coefficient, and $\mathbb{I}(\cdot)$ is the indicator function that activates if the estimated cost C_{total} exceeds the budget.

Once a parent is selected, we apply a stochastic mutation operator $\mathbb{M}(\cdot)$ to generate a child architecture $\alpha_{child} = \mathbb{M}(\alpha_{parent})$. The child is then evaluated via the inner loop mechanism. To avoid the computational cost

of training from scratch, we initialize the child's parameters θ_{init} by projecting weights from the Supernet $\mathcal{S}$ using the mapping function $\mathcal{M}$:

$$\theta_{init}(\alpha_{child}) = \{\phi^{\mathcal{S}}_{opt}, p_{init}, \mathcal{M}(\mathcal{S}, \alpha_{child})\}. \tag{S20}$$

These parameters undergo a short-horizon optimization (3 epochs) to co-adapt the optical phase ϕ_{opt} and neural weights $\mathbf{w}_{net}$ to the new topology, yielding the final fitness score.

Upon completion of the evaluation, the population is updated. We formally define the update step at iteration t as replacing the oldest individual α_{oldest} in the history with the new candidate:

$$\mathcal{P}_{t+1} = (\mathcal{P}_t \setminus \{\alpha_{oldest}\}) \cup \{\alpha_{child}\}. \tag{S21}$$

This aging mechanism ensures that the algorithm continuously explores new regions of the search space rather than stagnating around early local optima.

To ensure robust exploration, the mutation operator $\mathbb{M}$ stochastically modifies specific attributes of the network. We define three primary mutation types: *kernel mutation*, which alters the kernel size (e.g., $3 \times 3 \rightarrow 5 \times 5$) or operation type; *width mutation*, which modifies the expansion ratio to scale the network width; and *depth mutation*, which adds or removes residual blocks within a stage, constrained to the range $[D_{min}, D_{max}]$. Table S1 summarizes the hyperparameters used.

Table S1: REA Hyperparameters. Settings used for the discrete architecture search.

Parameter	Value
Evolution Cycles (N_{iter})	150
Population Size (P)	100
Tournament Sample Size (k)	10
Inner Loop Fine-tuning Epochs	3
Budget Penalty (γ)	10.0
Mutation Rate (per layer)	0.1

B.4 Sandwich-Rule Supernet Training

Main-paper Sec. 3.3 summarizes the weight-sharing supernet training in two sentences; we give the full rationale here. Training this large space of sub-networks independently is infeasible. Instead, we train the supernet as a stochastic weight-sharing graph. To ensure that every possible sub-network is well-trained, we employ the "Sandwich Rule" strategy [Cai et al. 2019]. At each training step, we simultaneously optimize the largest possible sub-network ($\mathcal{N}_{max}$), the smallest ($\mathcal{N}_{min}$), and sampled intermediate sub-networks ($\mathcal{N}_{sample}$). The gradients are accumulated to update both the shared supernet weights θ and the differentiable optical parameters ϕ as

$$\mathcal{L}_{train} = \mathcal{L}_{task}(\mathcal{N}_{max}) + \mathcal{L}_{task}(\mathcal{N}_{min}) + \sum_{i=1}^{n} \mathcal{L}_{task}(\mathcal{N}^{(i)}_{sample}). \tag{S22}$$

This strategy anchors the optimization on the upper and lower performance bounds. $\mathcal{N}_{max}$ pushes the limit of task accuracy, while $\mathcal{N}_{min}$ ensures the system remains functional even in ultra-low-cost states. Crucially, applying this summation to the optical optimization enforces a robust synergy. Because the metasurface sees gradients from all sub-networks simultaneously, the optics are forced to become a "generalist" encoder: the lens must preserve enough high-frequency detail for $\mathcal{N}_{max}$ to exploit, while maintaining sufficient contrast for $\mathcal{N}_{min}$ to decode. This makes the lens a robust warm start point for the subsequent specializations of the bi-level search (Sec. B.1).

B.5 Computational Cost Model

We expand the cost model that enters the budget penalty of Eq. S16 (main-paper Sec. 3.4). We formulate the system efficiency objective C_{total} in terms of the total number of floating-point operations required for a single inference pass [Ma et al. 2018]. Our cost is a function of two coupled variables: the effective pixel pitch $\hat{p}$, which dictates the dimensionality of the input data [Sze et al. 2017] (via physical readout or digital resizing), and the active neural topology $\alpha = \{w, d, k\}$, which defines the capacity of the reconstruction network [Tan and Le 2019].

For a chosen sub-network architecture α and input tensor pitch $\hat{p}$, the total inference cost is calculated by summing the operations required by the specific layers of the MobileNetV3 backbone [Howard et al. 2019]. Since our architecture relies on inverted residual blocks with depthwise separable convolutions, the total cost is

$$C_{\text{total}}(\hat{p}, \alpha) = \sum_{\ell=1}^{d} \left[\text{Ops}_{\ell}^{\text{pw}}(\cdot) + \text{Ops}_{\ell}^{\text{dw}}(\cdot) + \text{Ops}_{\ell}^{\text{se}}(\cdot) \right], \tag{S23}$$

where each term represents the floating-point operations for pointwise expansion/projection, depthwise convolution, and squeeze-and-excitation modules, respectively. The input resolution scales as $H_s W_s \propto \hat{p}^{-2}$, so the cost couples pixel density to network depth and width, which is what lets the co-design trade one against the other.

B.6 Architecture Sampling for the Accuracy–Compute Benchmark

The following protocol underlies the Pareto frontiers of main-paper Sec. 4.3 and Fig. 8. To move beyond theoretical PSF analysis and quantify downstream performance, we benchmark the trade-off between optical encoding accuracy and computational cost, benchmarking our design against fixed hyperbolic and double-helix baselines on the challenging ambiguity dataset (Sec. C.3). To ensure a fair comparison of architectural capacity, we employ a shared supernet "sandwich" training scheme (Sec. B.4). We evaluate 100 REA selected subnetworks—spanning 5 stages with kernel sizes $\{3, 5, 7\}$, expansion ratios $\{3, 4, 6\}$, and depths $\{2, 3, 4\}$—coupled with continuous resolution scaling drawn uniformly from $[0.3\times, 1.0\times]$. Each configuration undergoes batch normalization recalibration prior to inference. We quantify computational cost in units of multiply-accumulate operations (MACs), following standard protocols [Howard et al. 2017].

C ADDITIONAL DETAILS ON SYNTHETIC DATASET GENERATION

C.1 Synthetic Data Generation

Evaluating our design requires precise supervision of gaze vectors and sensor displacements across a diverse user population, which we obtain from high-quality synthetic data. We employ a data-driven pipeline in which high-fidelity 3D scans of human heads (geometry and appearance) are acquired from a large and diverse population using a One-Light-At-A-Time (OLAT) rig of 15 cameras and 15 lights [Sarkar et al. 2023]. The detailed geometry and appearance are registered against a parametric head/eye model, preserving biologically accurate skin features and a matching periocular region; fine details (corneal roughness, iris texture, eyelashes) are augmented from manually created templates. Fig. S2 shows 60 representative samples.

We control the virtual subject with a state vector

$$\mathbf{s} = (\phi, \theta, \omega, t_x, t_y, t_z) \in \mathbb{R}^6,$$

in which (ϕ, θ) are eye azimuth and elevation constrained by Listing's law [Tweed and Vilis 1990], ω is a small torsional perturbation modeling biological variation (Listing's thickness [Straumann et al. 1996]), and $\mathbf{t} = (t_x, t_y, t_z)$ is a translation along the subject's nasal bridge coupled with a systematic pitch rotation around the interaural axis to model glasses slippage; a small rigid perturbation models random mounting deviations. The camera is mounted along the lower frame bezel of a eyewear prototype rig – below the wearer's nominal line of sight – thereby preventing occlusions from the supraorbital ridge. Relative to the pupil center of a nominal

Fig. S2: Synthetic Eye-Tracking Dataset Samples. 60 representative samples drawn from our synthetic dataset. Each is rendered by path tracing from a randomly sampled state vector $\mathbf{s}$, yielding large variation in subject identity, gaze rotation, and device slippage.

user model at primary gaze, the camera's optical center is positioned at a spatial offset of $(-10, 10, -14)mm$, corresponding to a distance of $20mm$.

Scenes are rendered with physically-based path tracing (Blender Cycles [Blender Online Community 2025]) to produce the radiance and depth maps that feed the discretized radiance layers I_k used by the diffractive optical simulation (main-paper Sec. 3.1). A perspective pinhole camera yields distortion-free input for the wave-optics forward model. Additional ground-truth signals are emitted per frame: landmark positions in world and image space (eye center, pupil center, cornea tip), gaze directions, and image-space segmentation masks (visible pupil, unoccluded pupil).

C.2 Variational Training via Triplet Sampling

Optimizing the optical frontend with Fisher Information (main-paper Eq. 6) requires the partial derivatives of the sensor image I with respect to the gaze and slippage motion parameters, $\nabla_{\text{gaze}} I$ and $\nabla_{\text{slip}} I$. We approximate these gradients with a finite-difference *triplet* strategy, which sidesteps differentiable path tracing [Nimier-David et al. 2019]. For each virtual subject we render the triplet $(I_{\text{anchor}}, I_{\text{slip}+}, I_{\text{gaze}+})$ corresponding to the state vectors $\mathbf{s}_{\text{anchor}}$, $\mathbf{s}_{\text{slip}+}$, $\mathbf{s}_{\text{gaze}+}$, where $\mathbf{s}_{\text{anchor}}$ is randomly sampled and $\mathbf{s}_{\text{slip}+}$, $\mathbf{s}_{\text{gaze}+}$ deviate from it only on the slippage and gaze subspace:

$$\mathbf{s}_{\text{slip}+} = \mathbf{s}_{\text{anchor}} + (\mathbf{0}, \boldsymbol{\delta}_{\text{slip}}) = \mathbf{s}_{\text{anchor}} + (0, 0, 0, \delta_{t_x}, \delta_{t_y}, \delta_{t_z}),$$
$$\mathbf{s}_{\text{gaze}+} = \mathbf{s}_{\text{anchor}} + (\boldsymbol{\delta}_{\text{gaze}}, \mathbf{0}) = \mathbf{s}_{\text{anchor}} + (\delta_\phi, \delta_\theta, \delta_\omega, 0, 0, 0).$$

The perturbation magnitudes $\boldsymbol{\delta}_\cdot$ are chosen small enough to remain in the local linear regime of the intensity function yet large enough to exceed the sensor noise floor. This construction lets us numerically approximate the sensitivity fields and directly enforce the orthogonality constraint $\langle \nabla_{\text{gaze}} I,\ \nabla_{\text{slip}} I \rangle \to 0$ during optimization. Fig. S3 illustrates the sampling.

C.3 Ambiguity Evaluation Dataset Construction

To rigorously evaluate the method's ability to resolve the ill-posed inverse problem of passive tracking, we construct a targeted evaluation dataset focused on geometric ambiguity. In the absence of active glints, the pupil centroid is the primary high-frequency geometric anchor for gaze estimation, so the hardest inference conditions arise when distinct 3D physical states project to the *same* 2D pupil centroid on the sensor. Under these conditions a standard refractive camera captures nearly identical intensity distributions, forcing the solver onto subtle

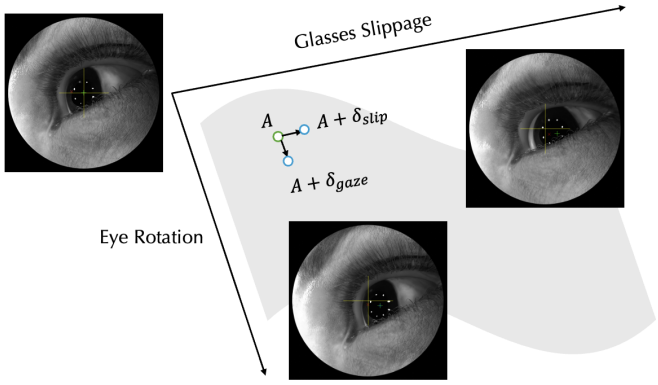

Fig. S3: Variational Training Data Generation. Training data is generated as differential triplets (*Anchor*, *Gaze-perturbed*, *Slip-perturbed*) used to numerically approximate the Fisher-information gradients required by the optical objective.

secondary cues (e.g., iris perspective distortion) that are typically indistinguishable at the sensor noise floor — a phenomenon we call *projection ambiguity*.

We generate ambiguous image pairs $(I_{\text{slip}}, I_{\text{gaze}})$ with strict pupil-centroid projection equivalence. For a given base state with pupil center p_c, we first apply a random slippage translation $\mathbf{t}_{\text{slip}}$ to render I_{slip}. We compute the resulting 2D pupil location $\hat{p}_c = \Pi_{\text{slip}}(p_c)$ under the translated camera projection. To build the ambiguous counterpart I_{gaze}, we keep the camera in its neutral (non-slipped) position and analytically solve for the eye-rotation angles $(\delta_\phi, \delta_\theta)$ that drive the pupil centroid to the same sensor coordinate $\hat{p}_c$. Because the dominant feature — the pupil — occupies the same pixel in both I_{slip} and I_{gaze}, 2D positional cues are removed and disentanglement performance depends entirely on the depth-dependent PSF signature encoded by the optics.

C.4 Eye Tracking Dataset Generation

This section lists the distributions according to which the state vector $\mathbf{s} = (\phi, \theta, \omega, t_x, t_y, t_z) \in \mathbb{R}^6$ was sampled.

We utilize truncated normal distributions TruncNorm to avoid unnatural eye poses and excessive slippage caused by outliers. A normal distribution with mean μ and standard deviation σ is bounded from below and above by l and u and renormalized.

Table S2: Ambiguity Dataset. An pose with minimal Listing's thickness and a slippage along the nasal bridge is randomly sampled. The additional slippage of $\mathbf{s}_{slip+}$ is comparatively smaller to avoid unnatural eye rotation when matching the pupil center location in image space. The yaw distribution is not symmetric to discard eye rotations away from the camera, which would be handled by eye tracking system of the other eye.

Parameter	Description	Distribution	Parameters
ϕ_{anchor}	Eye Yaw	TruncNorm	$\mu = 0°, \sigma = 15°, l = -35°, u = 45°$
θ_{anchor}	Eye Pitch	TruncNorm	$\mu = 0°, \sigma = 5°, l = -15°, u = 15°$
ω_{anchor}	Eye Roll	TruncNorm	$\mu = 0°, \sigma = 0.3°, l = -0.8°, u = 0.8°$
d_{anchor}	Slip Distance	TruncNorm	$\mu = 2\ mm, \sigma = 5\ mm, l = 0\ mm, u = 20\ mm$
d_{slip+}	Slip Distance	TruncNorm	$\mu = 1\ mm, \sigma = 2\ mm, l = 0.2\ mm, u = 5\ mm$

For the variational training dataset, we chose much smaller deviations to ensure that I_{slip+} and I_{gaze+} stay in the local linear regime.

Table S3: Variational Training Dataset. The small slippage distance of $\mathbf{s}_{slip+}$ ensures local linearity above the sensor noise level.

Parameter	Description	Distribution	Parameters
d_{slip+}	Slip Distance	Uniform	$l = 0.1\ mm, u = 0.3\ mm$

D ADDITIONAL EXPERIMENTAL PROTOTYPE DETAILS

Here, we describe additional details on the implemented hardware prototype.

D.1 Meta-Optic Structure Parameter Search

While the end-to-end optimization designs a phase map ϕ_{meta} based on the diameter map $\mathbf{D}$, the physical properties of the individual nanopillars (i.e., height, pitch, materials, and structure-to-phase mapping) were determined in a preprocessing stage.

We selected amorphous silicon (a-Si) as the pillar material due to its high refractive index and transparency in the 940 nm range. Fused silica was used as the substrate, providing a standard, compatible non-crystalline foundation. We employed RCWA [Liu and Fan 2012] to establish the relationship between a periodic arrangement of pillars of uniform height h and diameter d on a 2D regular grid with pitch s, and the resulting phase shift $\phi(\lambda, h, s, d)$ and transmittance $\tau(\lambda, h, s, d)$ under normal incidence at wavelength λ.

To identify the *optimal* uniform height H and pitch S, we performed an exhaustive search of the design space. We simulated heights h from 100 nm to 1.5 μm (in 10 nm steps), pitches s from 100 nm to 1 μm (in 10 nm steps), and diameters ranging from 10% to 90% of the pitch (in 0.5% steps) at $\lambda = 940$ nm. This totaled approximately 2 million RCWA simulations.

In the subsequent analysis, the phase shifts $\phi_{\lambda,\hat{h},\hat{s}}(d)$ and transmittances $\tau_{\lambda,\hat{h},\hat{s}}(d)$ of candidate pairs $(\hat{h}, \hat{s})$ were evaluated. We enforced the following criteria:

- **High Transmittance:** Only diameters with transmittance $\geq 90\%$ were considered to ensure high optical efficiency and prevent the metasurface from acting as an amplitude grating.
- **Injectivity:** The phase response must be strictly monotonic with respect to diameter. This ensures the inverse mapping from phase to diameter is well-defined.
- **Phase Sensitivity:** The local phase sensitivity $\frac{\partial \phi}{\partial d}$ must remain below $9^\circ\ \text{nm}^{-1}$. This ensures minor fabrication errors do not yield catastrophic phase shifts.
- **Safety Margins:** We excluded a 5 nm buffer zone around any diameters rejected by the previous criteria to prevent fabrication tolerances from pushing pillars into prohibited regimes.
- **Phase Coverage:** The viable diameters must cover the full 2π phase range (plus a 5% margin) to support arbitrary phase map designs.

Visual inspection of the valid candidates identified parameters $H = 700$ nm and $S = 530$ nm, located centrally within the region of viable designs (see Fig. S4). Consequently, systematic fabrication shifts (e.g., deviations in a-Si coating thickness) do not cause immediate performance degradation. This robustness is further confirmed by the corresponding diameter-to-phase mappings (Fig. S5). The final chosen parameters were 700 nm height, 530 nm pitch, and a usable diameter range of approximately 100 nm to 275 nm.

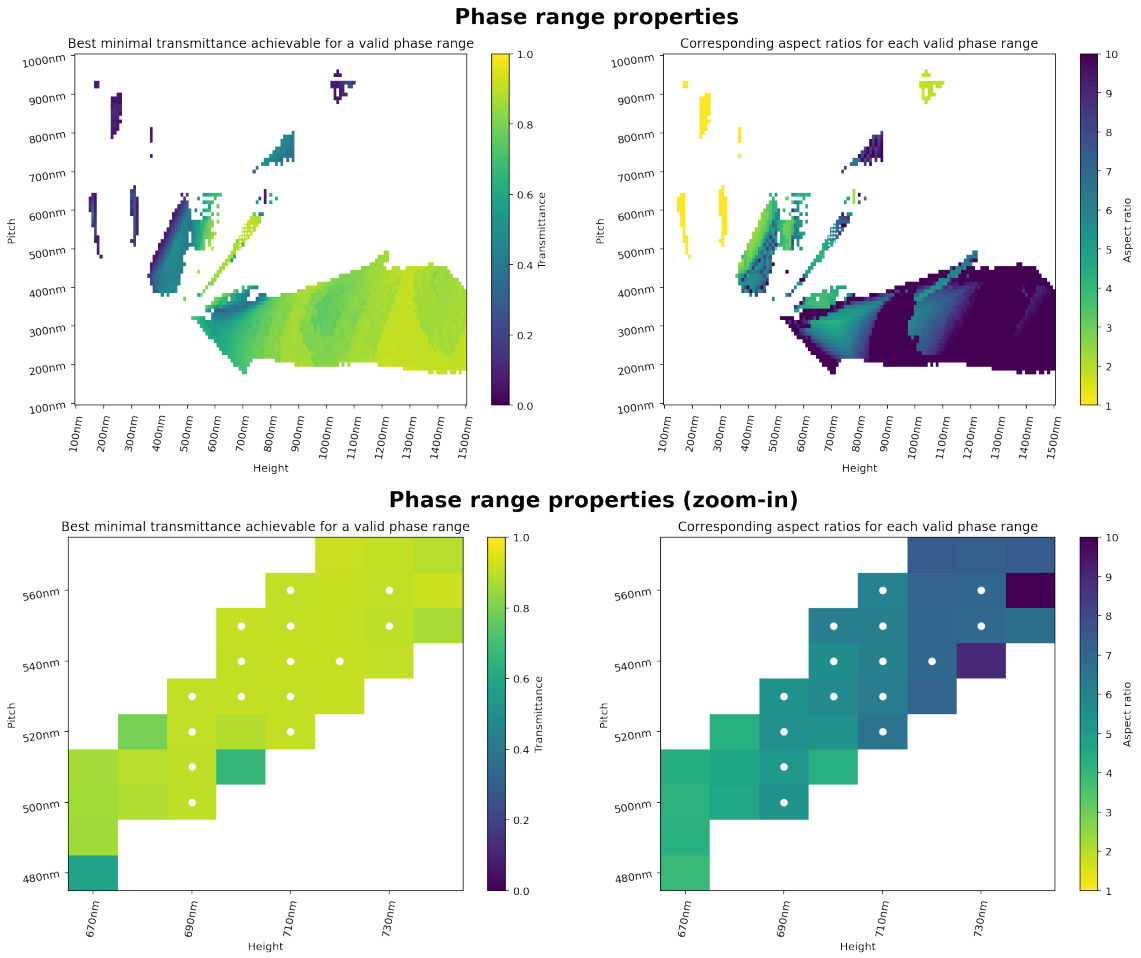

Fig. S4: Visual analysis of (pillar height and pillar pitch) candidate pairs. **Upper left:** Minimal transmittance over the 2π phase range. White regions indicate pairs failing to produce a viable diameter-to-phase mapping. **Upper right:** Maximal aspect ratio. Low aspect ratios are preferred for fabrication stability. **Bottom:** Zoomed views of the viable regions (white points mark $\geq 90\%$ transmittance). The selected parameters ($H = 700$ nm, $S = 530$ nm) are centered in this region to maximize fault tolerance.

D.2 Meta-Optic Fabrication

To fabricate the device, 0.5 mm thick 4 inch Fused Silica wafers with 700 nm amorphous Silicon (aSi) thin film were commercially acquired.

D.2.1 Aperture fabrication. A VM652 adhesion promoter (primer) was spin-coated onto the substrate. A layer of LOR 5B lift-off resist was then spin-coated at 2000 rpm for 60 s. The sample was baked at 170 °C for 5 min and allowed to cool for 3 min at room temperature. Subsequently, S1818 photoresist was spin-coated at 3000 rpm for 1 min. The resist was then soft-baked at 115 °C for 1 min and allowed to cool for 3 min. Photolithographic exposure was performed using a EVG610 mask aligner, to pattern the 1 mm apertures (see Fig. S6 (Left)). After

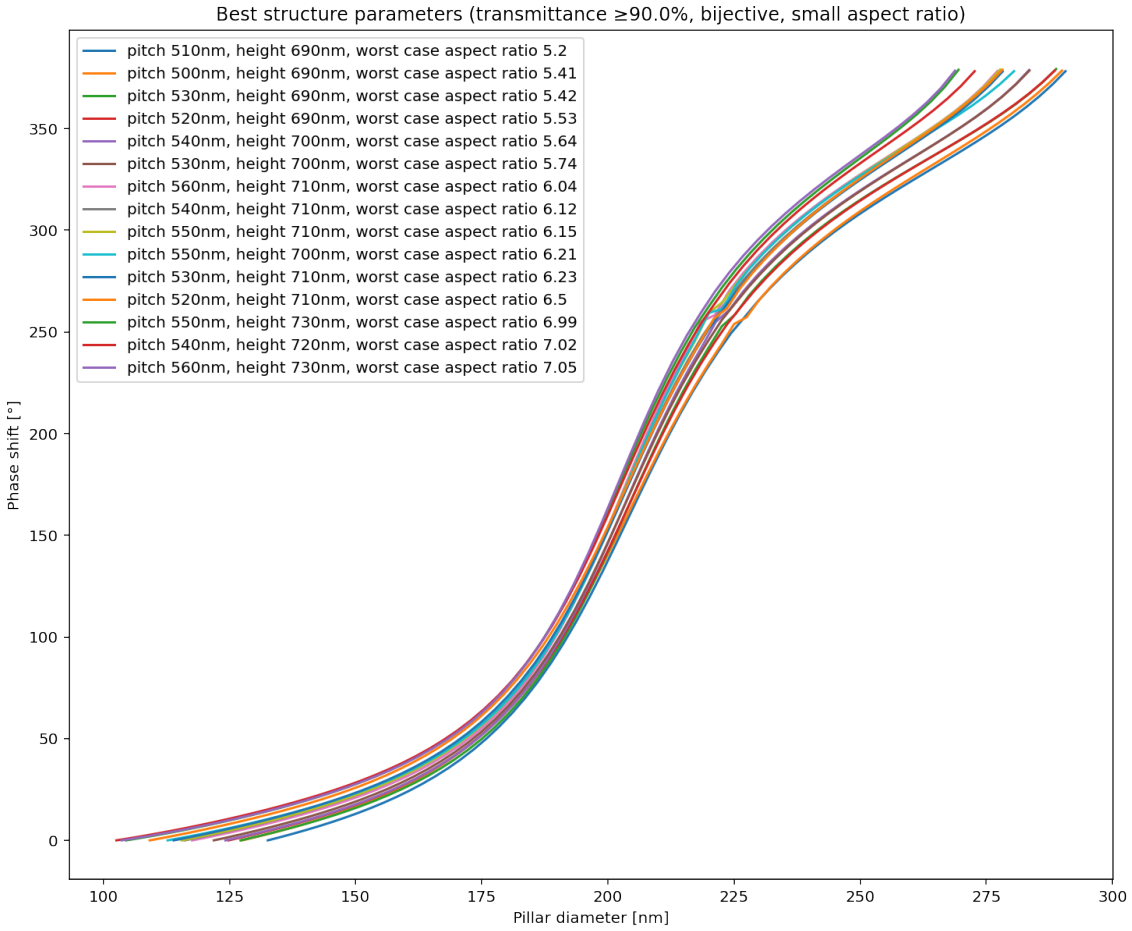

Fig. S5: Diameter-to-phase mappings for various height and pitch candidates identified by our search. The strict monotonicity allows for a simple inversion to obtain the phase-to-diameter mapping required to convert optimized designs into nanofabrication blueprints.

UV exposure, the sample underwent a post-exposure bake at 115 °C for 1 min, followed by development in MF-319 developer for 40 s. The sample was rinsed twice in deionized water and dried. A 80 nm Chromium layer was deposited using a Plassys MEB550S electron-beam evaporator at a deposition rate of 0.1 nm/s. Lift-off was subsequently performed to strip the resist and excess chromium, resulting in a patterned circular aperture with 1 mm diameter. A sacrificial resist layer was then spin-coated at 4000 rpm and baked at 115 °C for 3 min. Finally, the wafer was diced into 1×1 cm² chips using a semi-automatic dicer Accretech SS10, using GM blade at a feed rate of 0.5 mm/s.

D.2.2 Metalens fabrication. The aperture sample was thoroughly cleaned, first in acetone, then isopropyl alcohol, both with mild sonication. Oxygen plasma treatment was performed, at 340 W power for 5 min. This was done to ensure no residual polymer would be present on the surface. A 20 nm Cr layer was deposited using a Plassys MEB550S electron-beam evaporator at a deposition rate of 0.1 nm/s, to be used as a hard mask for the metalens

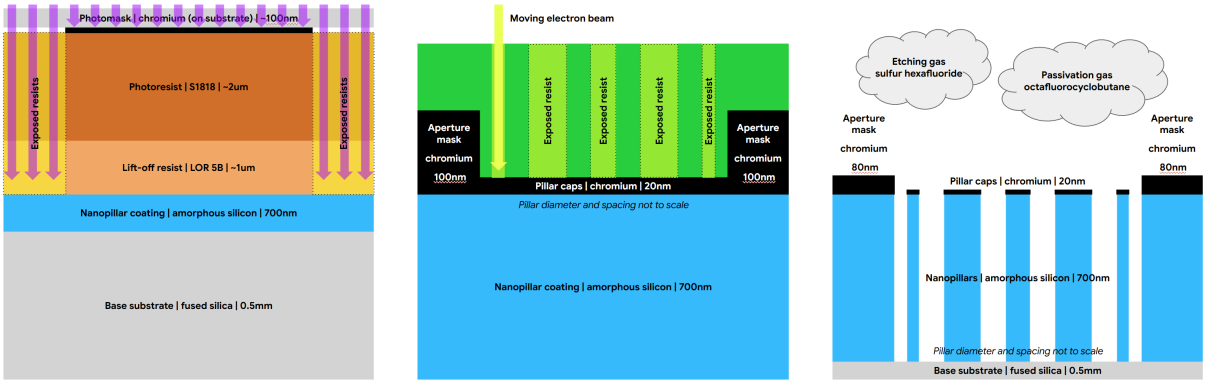

Fig. S6: Nanofabrication Process Steps. (Left) Photolithography removes the resist from the surrounding field, preparing the surface for the opaque Cr deposition while protecting the 1 mm aperture. (Center) Electron-beam lithography exposes and cross-links the negative-tone resist, hardening the caps that define the nanopillars. (Right) A Pseudo-Bosch etch removes the amorphous silicon from the unmasked regions surrounding the pillars.

pattern. A bilayer of ma-N 2401 negative-tone resist was spincoated and soft-baked at 90 °C for 90 s. Electron-beam lithography of the metalens pattern was carried out using a Raith 5150 EBPG 100kV system, using a 1 nm resolution (see Fig. S6 (Center)). Development was performed in ma-D 525 developer and DI H2O as a stopper. The Cr layer was then etched with OxIn PlasmaPro 100 Cobra ICP-RIE with Cl_2/O_2 gas mixture to transfer the pattern from a 'soft mask' polymer layer to a 'hard mask' Cr layer. Residual resist was removed by soaking in mr-Rem 700 for 30 min. O2 Plasma cleaning was performed for 5 min at 340 W. ICP-RIE was then performed to etch the metalens pattern into the underlying 700 nm a-Si layer, using a SF_6/C_4F_8 gas mixture ((see Fig. S6 (Right))), followed by a final chromium etch step of 20 nm, to strip the Cr cap on top of the pillars.

The final structures are 700 nm a-Si pillar arrays on-top of a transparent, fused Silica 0.5 mm substrate, within a 1 mm circular aperture in the 60 nm Cr coating to avoid light leaks during subsequent use.

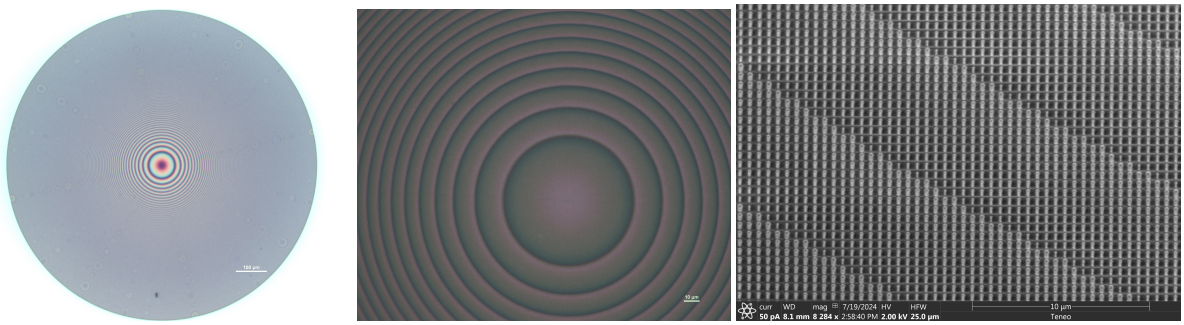

Fig. S7: Microscopy of Nanophotonic Structures. **Left:** Optical micrograph of the fabricated amorphous silicon metalens (1 mm aperture) with hyperbolic phase profile on a fused silica substrate. The structural coloration arises from the interaction with visible light, though the design is optimized for 940 nm NIR operation. **Middle:** High-magnification optical view (scale bar: 10;μm) revealing the sub-NIR-wavelength nanopillar array. **Right:** Scanning Electron Microscope (SEM) oblique view showing the high-aspect-ratio nanopillars with vertical sidewalls achieved via pseudo-Bosch etching.

D.3 Nano-Optical Camera Prototype

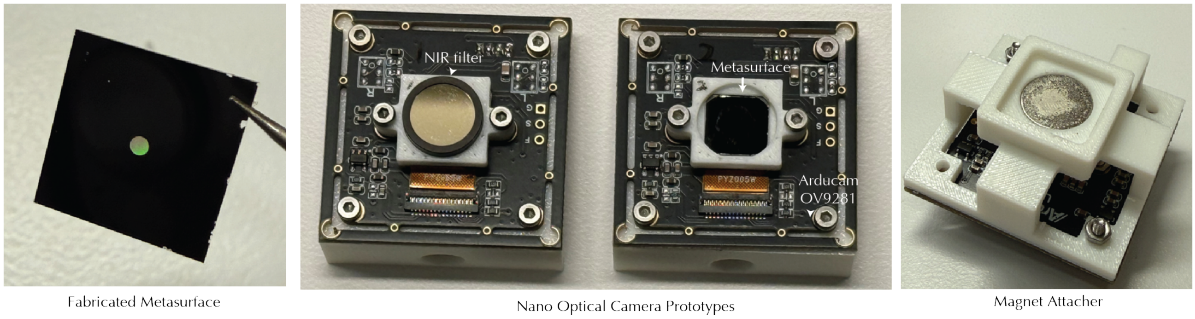

Fig. S8: Nano-optical camera prototype. Left: Fabricated metasurface. **Middle:** Two nano-optical camera prototypes built on monochrome CMOS sensors, corresponding to the baseline hyperbolic and the co-optimized designs. The near-infrared bandpass filter was removed in one prototype for visualization. **Right:** Magnetically attached holder integrated on the rear of the assembly, enabling rapid switching between the two prototypes during data capture.

In the prototyping stage, the proposed metalens was designed with a 1 mm aperture and a lens-to-sensor distance of 2 mm. According to the thin lens equation, an $f/1.85$ lens is required to focus an object at a distance of 25 mm onto the sensor plane. Consequently, we fabricated a custom lens holder using 3D printing and mounted it to an Arducam OV9281 UVC camera module. A Thorlabs FBH05940-10 spectral filter was integrated into the assembly using a printed housing. To facilitate rapid switching between sensors, a magnetic mount was also fabricated. The fully assembled prototype is shown in Fig. S8.

D.4 Data Acquisition Protocol

We captured the controlled-slippage dataset using the experimental setup shown in main-paper Fig. 9. During data collection, each subject's head was fixed with a head stage while viewing a 7×7 stimulus grid displayed on a monitor. The monitor was a 27-inch, 1080p LCD positioned 40 cm in front of the pupil, with its center horizontally and vertically aligned with the pupil center. The grid spanned a $25° \times 35°$ field of view. To prevent occlusion caused by the close camera placement, the stimulus grid was shifted upward by 5° in the vertical direction. Each stimulus was shown as a 5×5-pixel white cross at one of the 49 grid locations. The prototype camera was mounted approximately 2 cm below the eye to mimic a near-eye camera placement in an AR-glasses configuration, and translated using motorized translation stages (MTS25-Z8, Thorlabs) to emulate seven camera slippage states sharing the same origin: the origin state, two lateral translations along d_x, and four axial translations along d_z. A near-infrared illuminator provided the illumination. The translation step size for both d_x and d_z was 2.5 mm. For each stimulus position, the stimulus was presented, the system waited for 1.5 s to allow fixation, and one eye image was recorded. This produced 7×49 captures for each prototype. The full capture procedure was repeated for both the co-optimized metasurface prototype and the hyperbolic baseline prototype, with the prototype order randomized across subjects to reduce ordering bias. In total, we collected data from 20 subjects.

D.5 Glint Corneal Reflection Comparison

While glint-based tracking is ubiquitous and highly successful in enclosed VR/AR headsets, deploying it in everyday smart glasses introduces severe practical challenges. These include eyelash occlusion, strict power budgets, spatial constraints, and most notably, extreme ambient sunlight interference. To illustrate the limitations

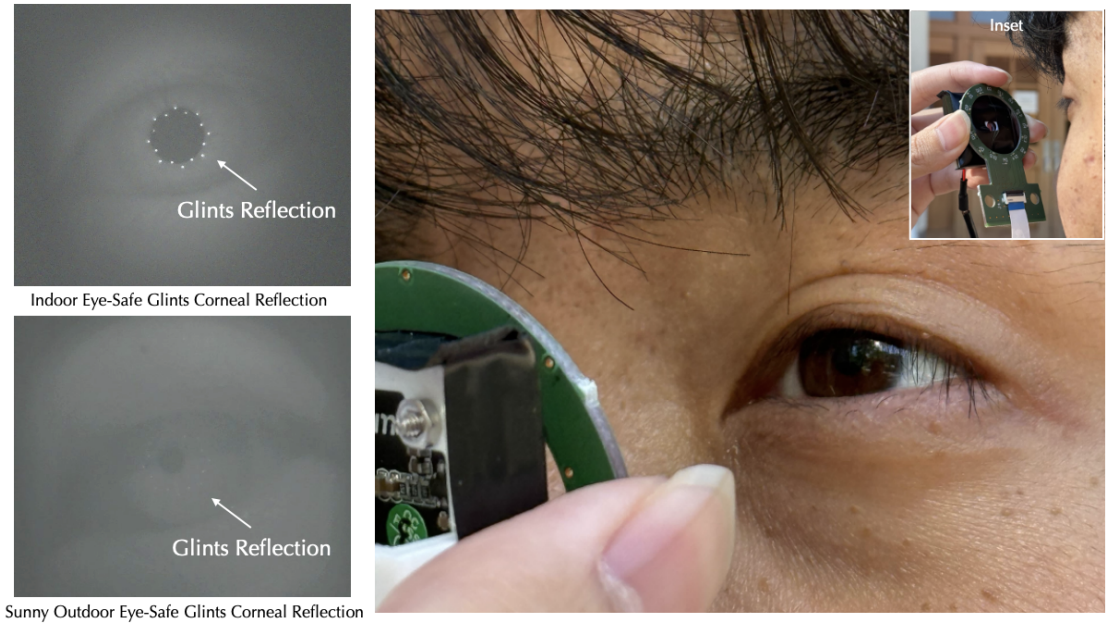

Fig. S9: Comparison of indoor and outdoor eye-safe glint measurements. The assembled prototype camera captures high-contrast glint corneal reflections indoors (left, top), where ambient infrared (IR) noise is minimal. Conversely, in a sunny outdoor environment (left, bottom), severe ambient solar IR floods the sensor, drastically degrading the Signal-to-Noise Ratio (SNR) and rendering the glints nearly indistinguishable when the LED power is throttled to maintain strict eye safety.

of outdoor glint reflection, we evaluated our assembled prototype camera under identical eye-safe illumination levels in both indoor and outdoor environments.

Indoors, the ambient near-infrared (NIR) illumination is negligible, meaning the primary source of corneal irradiance comes directly from the eye-tracking LEDs. Under the IEC 62471 standard for continuous photobiological safety (Exempt Group), the maximum allowable corneal irradiance is $E_{\text{limit}} = 10\ \text{mW/cm}^2$. At an eye relief distance of $d = 1.6$ cm, the maximum safe radiant intensity I_{max} for the LED is given by the inverse-square law:

$$I_{\text{max}} = E_{\text{limit}} \times d^2 = 10\ \text{mW/cm}^2 \times (1.6\ \text{cm})^2 = 25.6\ \text{mW/sr}$$

Operating near this limit indoors yields bright, high-contrast glints with an excellent Signal-to-Noise Ratio (SNR), as shown in Figure S9.

However, in outdoor scenarios—particularly on a sunny afternoon—the sun acts as a massive broadband IR radiator. The IEC 62471 safety limits are additive; the total irradiance from both the environment ($E_{\text{ambient\_IR}}$) and the device ($E_{\text{LED\_IR}}$) must not exceed the 10 mW/cm^2 threshold:

$$E_{\text{ambient\_IR}} + E_{\text{LED\_IR}} \leq 10\ \text{mW/cm}^2$$

On a typical bright day, the ambient solar IR reaching the cornea (even when shaded by the brow or frames) can easily reach $E_{\text{ambient\_IR}} \approx 4\ \text{mW/cm}^2$. This severe ambient penalty restricts the remaining permissible LED irradiance to merely 6 mW/cm^2. Consequently, the maximum safe radiant intensity for the tracking LED drops to 15.36 mW/sr.

At this heavily throttled power level, the camera struggles to isolate the weak LED reflection from the overwhelming solar background noise, washing out the glints entirely. Because maintaining a reliable glint SNR under strict eye-safety regulations outdoors is highly impractical, traditional glint-based eye tracking is insufficient for all-day wearable smart glasses. This fundamental limitation motivates our shift toward the robust, glint-free appearance-based method utilizing our fabricated metasurfaces, which we consider the practical solution for real-world egocentric vision applications.

E ADDITIONAL PSF ANALYSIS

To evaluate the robustness of the learned optics against sensor quantization, we analyze the Point Spread Functions (PSFs) in the XY plane across a depth range of 17mm to 26mm. Figure S10 presents a comparison between our proposed lens and two standard engineered designs: the Hyperbolic lens and the Double Helix (DH) lens. This comparison is performed at two spatial resolutions: the native sensor resolution (3 μm pitch) and a downsampled resolution (9 μm pitch) simulating 3×3 pixel binning.

The performance degradation observed in standard engineered PSFs during downsampling can be formalized by considering the image formation model. Let the continuous PSF at depth z be denoted by $h(x, y; z)$. The discrete intensity measured at pixel (i, j) is given by the integration of the PSF over the pixel area $\mathcal{A}_{i,j}$:

$$I_{i,j}(z) = \iint_{(x,y)\in\mathcal{A}_{i,j}} h(x, y; z)\, dx\, dy. \tag{S24}$$

Depth estimation relies on the sensitivity of this intensity profile to changes in z, often quantified by the Fisher Information $\mathcal{I}(z)$:

$$\mathcal{I}(z) = \sum_{i,j} \frac{1}{I_{i,j}(z) + \sigma^2} \left(\frac{\partial I_{i,j}(z)}{\partial z}\right)^2, \tag{S25}$$

where σ^2 represents noise variance. For engineered PSFs like the Double Helix, depth is encoded in high-frequency spatial features, specifically the rotation of two distinct lobes. These features correspond to rapid spatial variations in $h(x, y; z)$.

When the sensor undergoes 3×3 downsampling, the integration area $\mathcal{A}_{i,j}$ increases nine-fold. This operation acts as a spatial low-pass filter. If the spatial frequency of the depth-encoding features (e.g., the gap between DH lobes) exceeds the Nyquist limit of the downsampled grid (9 μm), the high-frequency gradients $\nabla_{xy} h$ are averaged out. Consequently, the derivative term $\frac{\partial I_{i,j}}{\partial z}$ diminishes significantly because the subtle shifts in the PSF structure are no longer resolved by the coarse pixel grid. As seen in Rows 4–6 of Figure S10, the distinct lobes of the DH lens merge, and the sharp edges of the Hyperbolic lens blur, leading to a collapse in Fisher Information and a loss of depth discriminability.

In contrast, our proposed lens is optimized end-to-end with the sensor model included. This allows the optical design to learn a depth encoding strategy that distributes information into lower spatial frequencies—manifesting as macro-scale aberrations rather than fine-edge diffractive features. As a result, the term $\frac{\partial I_{i,j}}{\partial z}$ remains significant even after integration over larger pixel areas, ensuring the PSF retains its depth-dependent signature despite quantization.

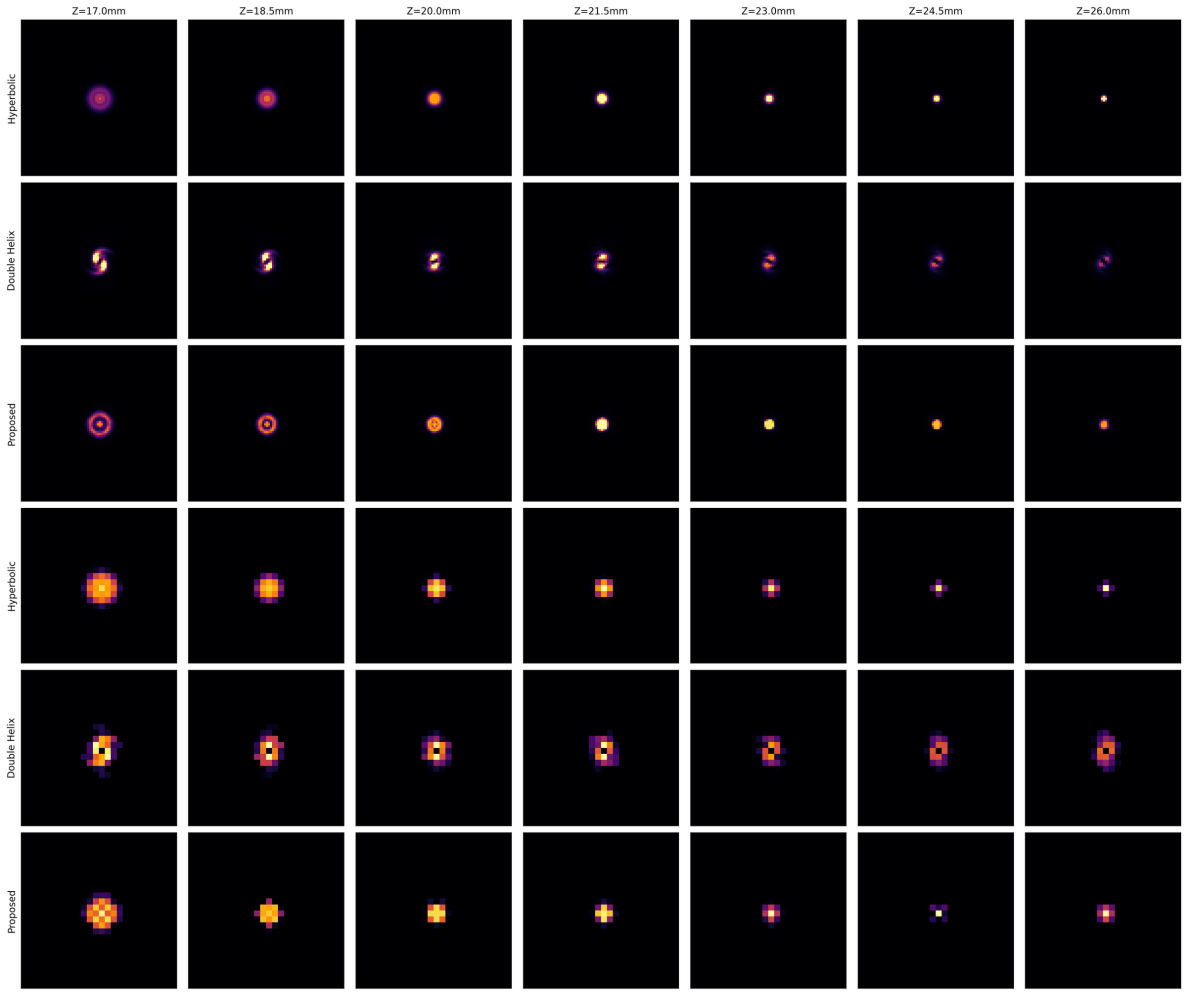

Fig. S10: Depth PSF Robustness to Downsampling. We visualize the PSFs of the hyperbolic lens, Double Helix (DH) lens, and our proposed lens at 1.5mm depth intervals from 17mm to 26mm. The top three rows display PSFs at the original spatial sampling resolution ($3\,\mu\text{m} \times 3\,\mu\text{m}$), where all lenses effectively encode depth via defocus, rotation, or aberration. The bottom three rows show the PSFs after 3×3 downsampling ($9\,\mu\text{m} \times 9\,\mu\text{m}$). While the Hyperbolic and Double Helix lenses lose critical high-frequency detail required for depth decoding due to spatial aliasing, the proposed lens remains robust, effectively encoding depth through varying aberrations that survive quantization.

REFERENCES

Blender Online Community. 2025. *Blender - a 3D modelling and rendering package.* Blender Foundation, Stichting Blender Foundation, Amsterdam. https://www.blender.org

Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. 2019. Once-for-all: Train one network and specialize it for efficient deployment. *arXiv preprint arXiv:1908.09791* (2019).

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1314–1324.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

Hayato Ikoma, Cindy M Nguyen, Christopher A Metzler, Yifan Peng, and Gordon Wetzstein. 2021. Depth from defocus with learned optics for imaging and occlusion-aware depth estimation. In *2021 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–12.

Victor Liu and Shanhui Fan. 2012. S4: A free electromagnetic solver for layered periodic structures. *Computer Physics Communications* 183, 10 (2012), 2233–2244.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. 2018. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*. 116–131.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*. Ieee, 565–571.

Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. 2019. Mitsuba 2: a retargetable forward and inverse renderer. *ACM Trans. Graph.* 38, 6, Article 203 (Nov. 2019), 17 pages. https://doi.org/10.1145/3355089.3356498

Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. 2018. Efficient neural architecture search via parameters sharing. In *International conference on machine learning*. PMLR, 4095–4104.

Kripasindhu Sarkar, Marcel C. Bühler, Gengyan Li, Daoye Wang, Delio Vicini, Jérémy Riviere, Yinda Zhang, Sergio Orts-Escolano, Paulo Gotardo, Thabo Beeler, and Abhimitra Meka. 2023. LitNeRF: Intrinsic Radiance Decomposition for High-Quality View Synthesis and Relighting of Faces. In *SIGGRAPH Asia 2023 Conference Papers* (Sydney, NSW, Australia) *(SA '23)*. Association for Computing Machinery, New York, NY, USA, Article 42, 11 pages. https://doi.org/10.1145/3610548.3618210

Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen P. Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. 2018. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics (SIGGRAPH)* 37, 4 (2018), 1–13.

D Straumann, DS Zee, D Solomon, and PD Kramer. 1996. Validity of Listing's law during fixations, saccades, smooth pursuit eye movements, and blinks. *Experimental brain research* 112, 1 (1996), 135–146.

Jipeng Sun, Kaixuan Wei, Thomas Eboli, Congli Wang, Cheng Zheng, Zhihao Zhou, Arka Majumdar, Wolfgang Heidrich, and Felix Heide. 2025. Collaborative On-Sensor Array Cameras. *ACM Transactions on Graphics (TOG)* 44, 4 (2025), 1–18.

Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer. 2017. Efficient processing of deep neural networks: A tutorial and survey. *Proc. IEEE* 105, 12 (2017), 2295–2329.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

Douglas Tweed and Tutis Vilis. 1990. Geometric relations of eye position and velocity vectors during saccades. *Vision research* 30, 1 (1990), 111–127.