Low Latency Gaze Tracking via Latent Optical Sensing

YIDAN ZHENG* and MATHEUS SOUZA*, KAUST, Saudi Arabia
KAIZHANG KANG, KAUST, Saudi Arabia
QIANG FU, KAUST, Saudi Arabia
HADI AMATA, KAUST, Saudi Arabia
WOLFGANG HEIDRICH, KAUST, Saudi Arabia

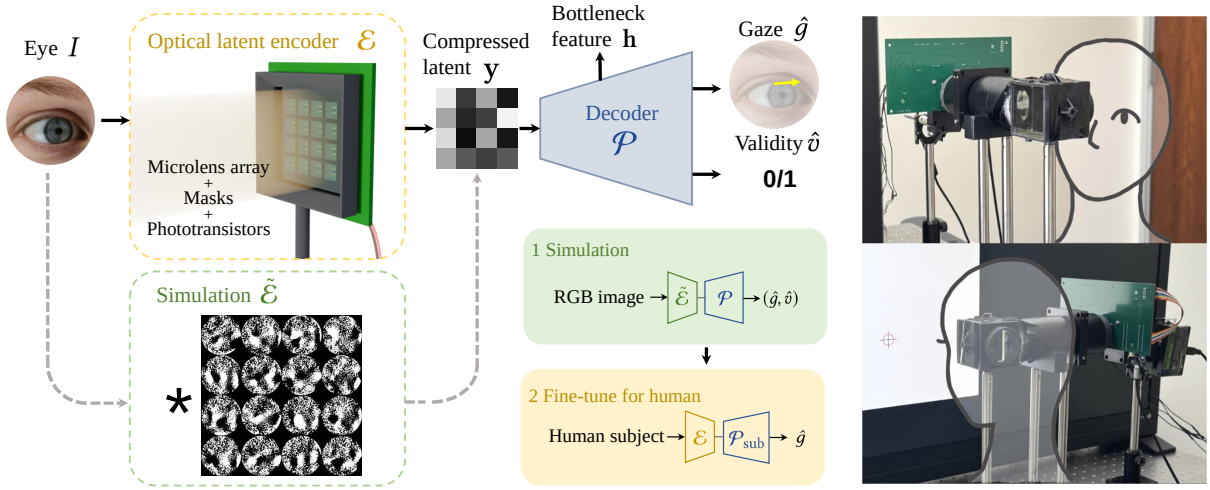


Fig. 1. **Overview of low-latency gaze tracking system.** Our system replaces conventional high-resolution cameras with a fully passive optical latent encoder $\mathcal{E}$. Light from the eye I is modulated by a microlens array and co-designed binary masks, producing a compressed latent measurement $\mathbf{y}$ captured by a 16 element phototransistor array. These features are mapped to gaze direction $\hat{g}$ and validity $\hat{v}$ via a lightweight MLP $\mathcal{F}$ and predictor $\mathcal{P}$. Our end-to-end pipeline (bottom center) uses simulation-to-real transfer with human-subject fine-tuning to achieve a total latency of less than 4 ms. (Right) Experimental setup and hardware implementation.

We present a real-time gaze tracking system that directly acquires task-relevant latent features using a fully passive optical encoder. Instead of forming and processing full-resolution images, our approach leverages a microlens array with a co-designed binary chromium mask to perform spatially multiplexed optical encoding, producing a compact set of measurements sufficient for gaze estimation. By integrating sensing and feature extraction in the optical domain, the proposed system eliminates the need for high-bandwidth image readout and substantially reduces computational overhead. The encoded measurements are captured by a 4×4 phototransistor array and mapped to gaze direction using a lightweight neural network. Our proof-of-concept prototype enables an end-to-end sensing-to-inference latency of 3.4 ms, outperforming published research systems. We demonstrate the effectiveness of our approach on both simulated and real-world data, achieving competitive gaze estimation accuracy while significantly reducing end-to-end latency compared to conventional camera-based pipelines. The proposed system establishes a new operating point in the accuracy–latency–compute trade-off for latency- and resource-constrained gaze tracking. This work highlights the potential of task-driven optical sensing for ultra-low-latency, computationally efficient human-computer interaction systems.

CCS Concepts: • **Human-centered computing** → **Displays and imagers**; • **Computing methodologies** → *Tracking*.

Additional Key Words and Phrases: Gaze tracking, Computational imaging, Optical encoding, Coded sensing, Real-time systems, Human–computer interaction

ACM Reference Format:
Yidan Zheng, Matheus Souza, Kaizhang Kang, Qiang Fu, Hadi Amata, and Wolfgang Heidrich. 2026. Low Latency Gaze Tracking via Latent Optical Sensing. In *SIGGRAPH Asia 2026 Conference Papers (SA Conference Papers '26), December 01–04, 2026, Kuala Lumpur, Malaysia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3829340.3842266

*Both authors contributed equally to this research.

Authors' Contact Information: Yidan Zheng, yidan.zheng@kaust.edu.sa; Matheus Souza, matheus.medeirosdesouza@kaust.edu.sa, KAUST, Thuwal, Saudi Arabia; Kaizhang Kang, kaizhang.kang@kaust.edu.sa, KAUST, Thuwal, Saudi Arabia; Qiang Fu, qiang.fu@kaust.edu.sa, KAUST, Thuwal, Saudi Arabia; Hadi Amata, hadi.amata@kaust.edu.sa, KAUST, Thuwal, Saudi Arabia; Wolfgang Heidrich, wolfgang.heidrich@kaust.edu.sa, KAUST, Thuwal, Saudi Arabia.

SA Conference Papers '26, Kuala Lumpur, Malaysia

ACM ISBN 979-8-4007-2842-6/2026/12
https://doi.org/10.1145/3829340.3842266

1 Introduction

Eye gaze tracking plays an important role in applications such as human–computer interaction [Kellnhofer et al. 2019; Smith et al.

2013; Zhang et al. 2015], assistive technologies [Caligari et al. 2013; Majaranta and Räihä 2002], and emerging AR/VR systems [Shadiev and Li 2023; Stengel et al. 2015; Tanriverdi and Jacob 2000; Ugwitz et al. 2022]. In particular, next-generation wearable displays require gaze estimation systems that are compact, energy-efficient, and capable of operating at ultra-low latency to support real-time interaction and foveated rendering [Guenter et al. 2012; Hong et al. 2018]. Conventional appearance-based gaze tracking systems typically rely on camera-based image acquisition followed by deep neural network inference [Davalos et al. 2025; Kellnhofer et al. 2019; Muksimova et al. 2025]. While recent advances in deep learning have significantly improved gaze estimation accuracy [Abdelrahman et al. 2023; Barkevich et al. 2024; Zhang et al. 2020, 2019], these approaches still require capturing and processing full-resolution eye images, introducing substantial sensing, bandwidth, and computational overhead.

To address these limitations, recent research has explored lightweight and event-driven gaze sensing systems. Event-based approaches leverage asynchronous vision sensors to reduce redundant image acquisition and improve temporal resolution [Angelopoulos et al. 2021; Bonazzi et al. 2024; Chen et al. 2025; Li et al. 2024; Sen et al. 2024; Zhao et al. 2023a]. Other works investigate embedded and edge-based gaze tracking pipelines for low-power operation [Bonazzi et al. 2023]. However, most existing systems still depend on reconstructing or processing image-like measurements before gaze inference, resulting in additional latency and computational cost. Furthermore, event-based methods often define latency only as the processing time after event accumulation [Li et al. 2024], without accounting for the sensing time required to collect sufficient events under varying illumination conditions.

At the same time, advances in computational imaging and task-driven optical design suggest that optical systems can be optimized directly for downstream inference objectives rather than image quality [Klotz and Nayar 2024; Shi et al. 2022; Souza et al. 2025]. Instead of treating images as mandatory intermediate representations, these approaches aim to capture compact task-relevant measurements that preserve the information necessary for a specific vision task. This perspective is particularly attractive for gaze tracking, where the final target is a low-dimensional gaze vector rather than a visually interpretable eye image. Rather than replacing high-accuracy camera-based gaze trackers, our goal is to explore a different operating point that prioritizes minimal sensing complexity and ultra-low end-to-end latency, targeting applications such as wearable devices, always-on gaze sensing, coarse gaze interaction, and attention estimation.

Motivated by these observations, we propose a passive optical gaze tracking system that directly acquires low-dimensional latent measurements using a microlens array (MLA) and optimized binary masks. Unlike conventional camera-based gaze trackers, the proposed system avoids explicit image formation and instead performs task-oriented optical encoding in a single shot. The encoded optical responses are measured by phototransistors and mapped to gaze direction using a lightweight neural network. By eliminating high-resolution image acquisition and minimizing computational overhead, the proposed system achieves ultra-low-latency gaze estimation with a compact and passive optical architecture.

To enable robust latent-space sensing, we develop an end-to-end training framework that jointly optimizes the optical encoding masks and digital inference using both latent reconstruction and gaze supervision objectives. The learned optical encoder is trained to preserve gaze-relevant information within a compact 16-dimensional measurement space while remaining compatible with passive optical implementation. We further evaluate the system on both synthetic and real human eye data and analyze its end-to-end latency under realistic hardware conditions.

The main contributions of this work are summarized as follows:

- We propose an end-to-end latent-space gaze tracking framework that jointly optimizes passive optical encoding and neural gaze decoding, enabling direct gaze inference from compact low-dimensional measurements without reconstructing full eye images.
- We develop a compact MLA-based single-shot optical sensing system with optimized binary masks, phototransistor readout, and field-programmable gate array (FPGA) acquisition, enabling low sensing complexity and an end-to-end latency of 3.4 ms including both sensing and computation.
- We establish a two-stage simulation-to-real evaluation pipeline, achieving a cross-subject simulation error of 6.47° and a calibrated per-subject real-world error below 6°.

Although our proof-of-concept prototype is not engineered to optimize for overall system power consumption, we note that the substantial reduction of sensor measurements, the resulting reduction in data volume, and the extremely lightweight nature of the digital processing should allow for highly energy efficient implementations of the concept in the future.

2 Related Work

Low-latency and real-time gaze estimation. Recent research has underscored that for mobile XR glasses, achieving an end-to-end latency of less than 5 ms is a critical threshold to prevent the disruption of user immersion [Hong et al. 2018]. While many software-oriented frameworks such as *WebEyeTrack* [Davalos et al. 2025] and *GazeCapsNet* [Muksimova et al. 2025] report exceptionally fast inference speeds via few-shot personalization or lightweight neural architectures, they typically rely on full-image capture from conventional CMOS sensors. This reliance introduces significant power and temporal overheads due to frame-rate limitations and large-scale data readout, making them less efficient for truly real-time applications. To address these bottlenecks, hardware-oriented approaches have explored alternative sensing modalities. Li et al. [Li et al. 2020] demonstrated that spatially sparse single-pixel detectors can substitute for dense imaging sensors, while visible light sensing was proposed for energy-constrained headsets [Li et al. 2017].On the edge, systems like *TinyTracker* [Bonazzi et al. 2023] utilize in-sensor processing to enable fast, low-power operation on mobile platforms.

However, the system latency is limited not just by the computational processing, but also the exposure and readout time for the full frame. Moreover, 2D image sensor arrays are relatively power hungry due to the large number of A/D conversion needed. Event-based sensing [Bonazzi et al. 2024; Li et al. 2024; Sen et al. 2024] has been proposed to increase the frame rate and reduce the latency.

However, event-based gaze estimation still relies on temporal event accumulation and subsequent digital processing before inference, and reported latencies are often measured after sufficient events have been acquired [Angelopoulos et al. 2021; Zhao et al. 2023a]. Other work such as Pixel Processor Arrays [Bose et al. 2022] and dedicated ASICs like *JaneEye* [Han et al. 2026] have focused primarily on the sensing hardware, and achieve impressive sensing latency as low as 0.5 ms. However, for a complete eye tracking system, these sensors need to be paired heavy compute architectures requiring tens of MFLOPs per frame. In contrast, our system directly produces task-specific optical measurements rather than image frames or event streams, reducing sensing bandwidth and digital processing while achieving an end-to-end latency below 4 ms with only 0.49 MFLOPs.

At the algorithmic level, several works have pointed out that full eye images are highly redundant for gaze estimation. For example, EyeTrAES [Sen et al. 2024] introduces adaptive event slicing to extract only relevant temporal features, while dataset-driven and representation learning approaches [Cheng et al. 2022; Ghosh et al. 2022; Kim et al. 2019; Lee et al. 2022; Sun et al. 2021; Wang et al. 2022] focus on condensing eye features into compact, generalizable codes. These methods consistently emphasize that raw image data is unnecessary overhead for the task. Our work follows this trajectory by eliminating full image formation entirely, instead acquiring compact latent measurements directly through task-driven optical encoding.

Task-driven optics and latent-space sensing. Recent advances in computational imaging have demonstrated the effectiveness of jointly optimizing optical systems and downstream algorithms. By incorporating differentiable models of wave propagation, these approaches enable end-to-end learning of optical elements together with reconstruction networks, often targeting high-fidelity image formation [Dai et al. 2025; Ho et al. 2024]. This paradigm has been explored in a variety of settings, including learned diffractive optics [Wang et al. 2023], task-driven lens design [Yang et al. 2026], collaborative camera arrays [Sun et al. 2025], task-specific encoding for depth and RGBD sensing [Liu et al. 2025], and coded aperture [Asif et al. 2017; Zhao et al. 2023b] or compressive imaging systems [Duarte et al. 2008], where measurements are optimized for efficient reconstruction.

Beyond image formation, a growing body of work shifts the objective from reconstruction to task performance. In these systems, optical front-ends are optimized directly for downstream objectives such as classification or detection, as demonstrated in lensless opto-electronic neural networks [Shi et al. 2022], freeform optical designs [Klotz and Nayar 2024], and compressive and single-pixel imaging systems [Atanov et al. 2024]. This perspective recognizes that full image recovery is often unnecessary when the goal is to extract semantic information.

More recently, latent-space sensing extends this idea by learning optical encoders that map inputs directly into compact feature representations, which are then processed by lightweight neural networks [Souza et al. 2025]. By bypassing explicit image reconstruction, such approaches improve efficiency and reduce redundancy in the sensing pipeline.

Building on this perspective, we explore latent-space sensing for gaze tracking, where the optical front-end directly encodes low-dimensional gaze features instead of forming eye images. This formulation enables a compact and efficient sensing pipeline tailored for real-time applications such as AR/VR.

3 Gaze Tracking Framework and Dataset

Conventional image-based gaze estimation systems typically rely on densely sampled eye images, requiring the sensor to capture, digitize, and process a large number of pixels. However, the information needed for gaze estimation may lie on a much lower-dimensional manifold than the full image representation. Motivated by this observation, we estimate gaze from only 16 scalar measurements acquired in a single snapshot.

As shown in Figure 1, our pipeline learns optical sensing patterns that are optimized directly for gaze estimation. The optimized optical encoder $\mathcal{E}$ acquires a measurement vector $\mathbf{y} \in \mathbb{R}^{16}$. This measurement is then processed by a lightweight decoder $\mathcal{P}$, which is trained to predict the latent representation $\mathbf{h}$ of a pretrained generative model $\mathcal{G}$, together with gaze $\hat{g}$ and a validity flag $\hat{v}$. Rather than targeting pixel-level reconstruction, the system is optimized to recover a compact latent representation that preserves gaze-relevant information.

3.1 Optical Encoder and Latent Distillation

To train this system, we define a differentiable digital encoder $\tilde{\mathcal{E}}$ that emulates the measurement process of the optical encoder $\mathcal{E}$. Given an input eye image I, $\tilde{\mathcal{E}}$ is implemented as a bank of 16 learnable binary masks at the image resolution:

$$\tilde{\mathcal{E}} = \{\mathbf{m}_k\}_{k=1}^{16}, \qquad \mathbf{m}_k \in \{0,1\}^{256\times256}. \tag{1}$$

Each mask serves as an optical sensing pattern, and the corresponding scalar measurement is computed by an inner product:

$$y_k = \langle I, \mathbf{m}_k \rangle, \qquad k = 1, \ldots, 16. \tag{2}$$

The full measurement vector is

$$\mathbf{y} = [y_1, y_2, \ldots, y_{16}]^\top \in \mathbb{R}^{16}. \tag{3}$$

Since the optical encoder provides only 16 scalar measurements, direct supervision from gaze labels alone may not be sufficient to learn a robust representation. Following [Souza et al. 2025], we train a student decoder $\mathcal{P}$ to recover a compact feature aligned with the latent space of a pretrained generative inversion model. Thus, the decoder learns not only from gaze labels, but also from a structured latent representation that captures gaze-relevant eye appearance.

To obtain the generative latent space used for supervision, we train a StyleGAN2 generator $\mathcal{G}$ on eye patches [Karras et al. 2020], together with an inversion model $\mathcal{P}_\mathrm{T}$ based on the encoder architecture of [Richardson et al. 2021]. The inversion model maps an eye image to a compact latent feature and serves as the teacher for $\mathcal{P}$ during distillation training process. Given a clean eye image I, the teacher predicts a compact latent feature, gaze estimate and a validity logit:

$$\mathbf{h}_\mathrm{T}, \hat{g}_T, \hat{v}_T = \mathcal{P}_\mathrm{T}(I), \qquad \mathbf{h}_\mathrm{T} \in \mathbb{R}^{256}. \tag{4}$$

The student decoder $\mathcal{P}$ maps the compressed measurement vector $\mathbf{y}$ to a 256-dimensional bottleneck feature $\mathbf{h}$ and then predicts both gaze and validity:

$$\mathbf{h}, \hat{g}, \hat{v} = \mathcal{P}(\mathbf{y}), \qquad \mathbf{h} \in \mathbb{R}^{256}. \tag{5}$$

The decoder consists of a normalization layer, an upscaling MLP block, and a two-layer MLP residual bottleneck, followed by separate gaze and validity heads; architectural details are provided in the supplementary material. The gaze and validity heads are initialized from the corresponding pretrained heads of $\mathcal{P}_\mathrm{T}$ to accelerate convergence.

Although the student is not trained with an image reconstruction objective, its bottleneck feature remains connected to the generator latent space. In particular, the compact feature can be repeated across the $L = 14$ StyleGAN layers to form a broadcast $\mathcal{W}^+$ code when synthesis or latent editing is required.

The training objective combines supervised gaze regression, validity classification, gaze distillation, and feature distillation. We denote the student-predicted gaze vector by $\hat{g}$, the ground-truth gaze by $\hat{g}_\mathrm{GT}$, and the teacher-predicted gaze by $\hat{g}_T$. Gaze errors are measured using cosine distance, i.e., one minus the cosine similarity between two 3D gaze vectors. The total loss is

$$\begin{aligned}\mathcal{L} =& \lambda_\mathrm{sup} v\, \mathcal{L}_\cos(\hat{g}, \hat{g}_\mathrm{GT}) + \lambda_\mathrm{cls}\mathrm{BCE}(\hat{v}, v) \\ &+ \lambda_\text{gaze-distill} v\, \mathcal{L}_\cos(\hat{g}, \hat{g}_T) + \lambda_\text{feat-distill} v\, \|\mathbf{h} - \mathbf{h}_\mathrm{T}\|_2^2.\end{aligned} \tag{6}$$

Here, the first two terms train the student using ground-truth gaze and validity labels, while the last two terms distill the teacher gaze prediction $\hat{g}_T$ and compact latent feature $\mathbf{h}_\mathrm{T}$. The gaze and feature terms are evaluated only on valid samples using v, whereas the validity classification loss is applied to all samples.

This objective trains the student to recover a compact, gaze-aware latent representation directly from the 16 optical measurements, without requiring image reconstruction during this stage. Because this representation is aligned with the teacher latent feature and remains compatible with the generator through the broadcast $\mathcal{W}^+$ construction, it also provides a natural latent coordinate system for gaze control.

3.2 Gaze Steering in Latent Space

To evaluate the ability of the latent space to effectively encode gaze directions and separate the corresponding parameters from appearance and the visual identity of the subject, we demonstrate gaze steering using the trained generator and residual displacement in latent space (see supplemental videos for examples). This experiment is intended as an analysis of the learned latent representation rather than part of the gaze estimation pipeline. Successful gaze steering indicates that the recovered latent preserves gaze-related semantic information.

Given a source gaze $\hat{g}_s$ and a target gaze $\hat{g}_t$, a control network C_θ predicts a latent update that moves the representation from the source gaze direction toward the target direction.

For the compact latent representation used above, the controller predicts

$$\Delta\mathbf{h} = C_\theta(\hat{g}_s, \hat{g}_t), \qquad \Delta\mathbf{h} \in \mathbb{R}^{256}. \tag{7}$$

The edited compact latent code is then

$$\mathbf{h}' = \mathbf{h} + \Delta\mathbf{h}, \tag{8}$$

where $\mathbf{h}$ is the compact latent representation produced by the inversion encoder. The edited latent can be decoded by the frozen generator after being repeated across the StyleGAN layers using the same broadcast construction described above.

For higher-fidelity gaze steering, we also consider editing directly in the full StyleGAN $\mathcal{W}^+$ space of the color-image generator. In this setting, the latent representation consists of $L = 14$ independent style vectors, each of dimension 512. The controller therefore predicts a layer-wise residual $\Delta\mathbf{w}^+ \in \mathbb{R}^{14\times512}$, which is added to the inverted latent code before synthesis:

$$I' = G(\mathbf{w}^+ + \Delta\mathbf{w}^+). \tag{9}$$

This full $\mathcal{W}^+$ formulation follows the same residual-control principle as the compact case, but provides additional degrees of freedom by allowing each StyleGAN layer to receive its own gaze-dependent update.

In practice, C_θ takes the concatenated gaze pair $[\hat{g}_s; \hat{g}_t]$ as input. A shared multilayer perceptron encodes this pair into a hidden representation, and the output is either a single compact displacement $\Delta\mathbf{h}$ or a set of layer-wise displacements $\Delta\mathbf{w}^+$. The generator and inversion encoder remain frozen throughout training; only the control network is optimized to produce latent edits that steer gaze while preserving identity and eye appearance.

3.3 Dataset

Our dataset is derived from ETH-XGaze [Zhang et al. 2020], consisting of high-resolution monocular eye patches. We utilize a processing pipeline involving frontal camera selection and a circular masking process to mitigate microlens-induced corner artifacts. To ensure robustness against mechanical tolerances and optical alignment errors in the physical prototype, we apply random spatial translations, scaling, and horizontal flipping as data augmentation during training. Each sample is annotated with a 3D gaze vector and a binary validity label to identify blinks or misalignments. Further technical details on the cropping geometry, coordinate normalization, and the automated classification of invalid inputs are provided in the Supplementary Material. The final processed corpus includes 68 training subjects and 12 test subjects, totaling over 168,000 images. Over valid samples, gaze spans -35.9° to 22.5° in pitch and -54.7° to 56.5° in yaw. The held-out test split covers a slightly narrower range, -31.7° to 19.1° in pitch and -46.6° to 47.2° in yaw.

3.4 Identity Augmentation via Generative Inpainting

Appearance-based gaze models are often prone to identity bias, learning shortcuts based on skin tone or facial structure rather than ocular geometry [Akgül et al. 2026]. This is exacerbated by the limited demographic diversity typical of laboratory-collected datasets [Zhang et al. 2020]. To address this, we introduce a generative augmentation scheme leveraging the Flux diffusion model [Black Forest Labs 2024] to diversify the periocular appearance associated with each ground-truth gaze label.

Our pipeline utilizes SAM3 [Carion et al. 2025] to segment the eye, creating an inpainting mask that preserves the original ocular

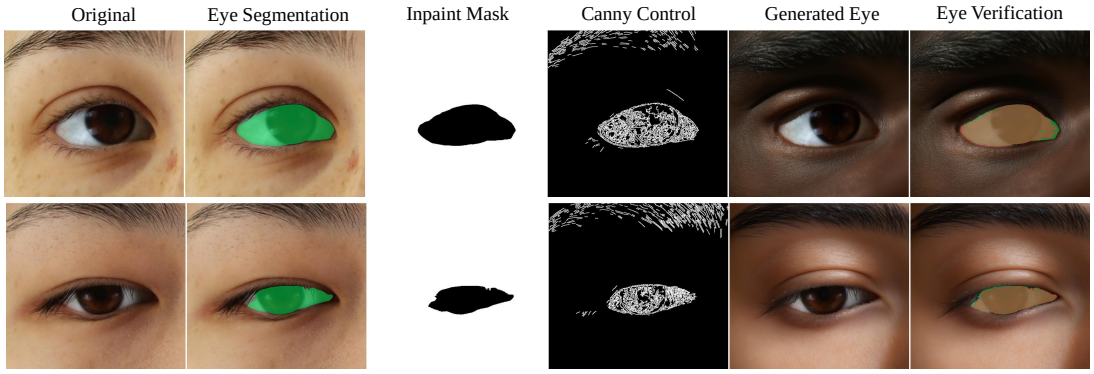


Fig. 2. **Identity-varying augmentation via generative inpainting.** From left to right: original eye crop, SAM3 eye segmentation, inpainting mask, Canny-edge ControlNet input, the Flux-generated eye (conditioned on the mask, the Canny edges, and a diversity prompt), and SAM3 re-segmentation of the generated eye overlaid on the original mask for consistency checking. Samples failing the area/IoU check are regenerated with a different seed for up to three attempts; if none passes, the original crop is retained. The gaze label therefore always remains valid. More augmentation examples can be found in the Supplementary Material.

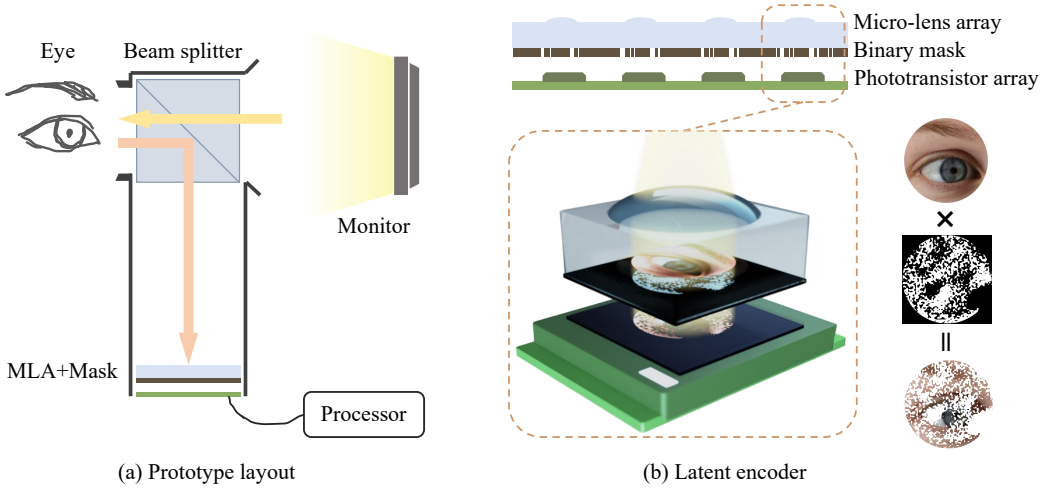
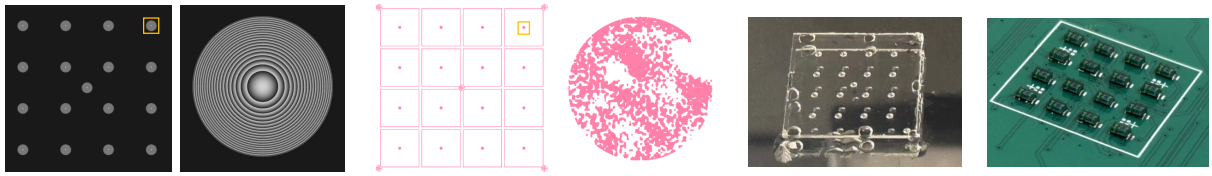


Fig. 3. **Hardware prototype.** (a) The prototype layout utilizes a beam splitter to redirect the eye's reflection toward the MLA+Mask assembly while allowing the user to view a monitor. (b) Detailed view of the latent encoder stack, consisting of a micro-lens array, a binary chromium mask, and a phototransistor array. The encoder performs spatial multiplexing, effectively convolving the eye image with a task-specific binary pattern to produce compressed measurements for direct gaze estimation.

pixels while allowing the surrounding context to be regenerated. To maintain anatomical plausibility, we employ Canny-edge [Canny 1986] ControlNet [Zhang et al. 2023] as a structural guide. Flux then inpaints the masked region using prompts spanning diverse ethnicities, ages, and lighting conditions. To ensure geometric fidelity, each sample must pass a CLIP-based [Radford et al. 2021] semantic check and an eye-consistency verification requiring 90% spatial overlap (IoU $\geq$ 0.9) with the original segmentation. This process reduces the model's reliance on identity-correlated features, as illustrated in the pipeline overview in Figure 2. Extended details on the prompt pool and verification metrics are provided in the Supplementary Material.

4 Prototype

Our prototype consists of a beam splitter and a hardware sensing module composed of a microlens array (MLA), a binary chromium (Cr) mask, and a phototransistor array, as illustrated in Fig. 3. The MLA and the Cr mask are fabricated on separate substrates and bonded together after precise alignment. In operation, the eye is illuminated and the reflected light is directed by the beam splitter into the Latent Encoder. The microlens array forms micro-images of the eye, which are selectively modulated by the binary mask. The resulting encoded intensity is captured by the phototransistor array, which is integrated with an FPGA-based acquisition circuit to yield ultra-low-latency latent optical measurements for gaze estimation.



(a) MLA (b) Cr mask (c) Optical encoder (d) Phototransistors

Fig. 4. **Fabrication of the latent optical encoder.** (a) Design layout of the Fresnel microlens array (MLA) featuring a 1 mm aperture. (b) Patterned binary Cr mask designed for spatial modulation. (c) Photograph of the assembled latent encoder, where the MLA and Cr mask wafers are precisely aligned and bonded using UV-curable glue. (d) Photograph of the 4×4 phototransistor array integrated on a custom PCB for high-speed intensity acquisition.

4.1 Microlens Array and Cr Mask

4.1.1 Design. The sensing module utilizes an MLA composed of Fresnel lenses, each featuring an aperture of 1 mm and a focal length of 1.94 mm. These lenses are designed to demagnify the circular input eye region (diameter of 40 mm) into a reduced spot of 0.256 mm, focused at the mask plane. The binary Cr mask is patterned with corresponding features of 0.256 mm to match this demagnified image size. Our architecture employs a direct mapping where each phototransistor in the 4×4 array measures the integral of light intensity over the corresponding masked image as described in Eq. (2). The phototransistors are spaced with a 4 mm horizontal pitch and a 3.88 mm vertical pitch. The layout of MLA and Cr mask design is illustrated in Fig. 4(a) and (b).

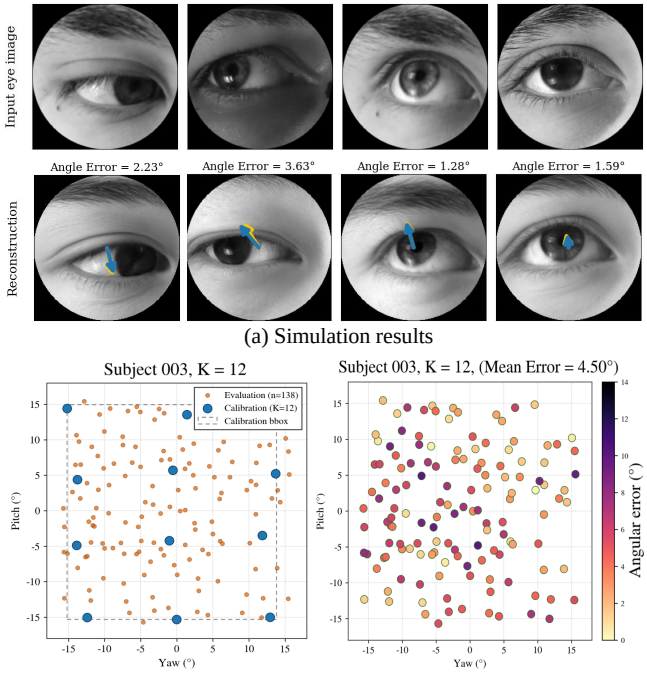
4.1.2 Fabrication. The MLA and the binary Cr mask were fabricated on separate wafers. The MLA (6.332 μm wrapping height, 1 μm fabrication pitch) was fabricated on OrmoComp using grayscale direct-write lithography followed by UV nanoimprinting, following the methodology in [Amata et al. 2024]. The binary Cr mask (1 μm fabrication pitch) was patterned on a separate wafer using the same Heidelberg DWL66+ direct-write lithography system. The MLA and mask substrates were then bonded together using careful optical alignment to place the mask near the MLA focal plane. The assembled prototype is shown in Fig. 4(c).

4.2 Phototransistors with FPGA system

We employ an array of TEMT6200FX01 phototransistors, which feature a spectral sensitivity of 450–610 nm and a radiant sensitive area of 0.75 mm^2. To minimize electronic noise and ensure signal integrity, the supporting printed circuit board (PCB) includes an amplifier and a low-pass filter, followed by an analog-to-digital converter (ADC). The converted digital signals are read out by an FPGA at a stable sampling rate of 14,400 Hz (900 Hz for 16 phototransistors). This hardware configuration enables the rapid acquisition of latent features with high signal-to-noise ratio, supporting the system's low end-to-end latency.

5 Evaluation

To thoroughly evaluate the performance of our all-optical latent feature acquisition system, we design a two-stage assessment pipeline followed by a dedicated analysis of system latency. The first stage consists of a fully simulation-based evaluation, where the optical encoder $\tilde{\mathcal{E}}$ and its associated binary masks are optimized and trained



(a) Simulation results

(b) Real measurement results (Subject 003)

Fig. 5. **Evaluation results in simulation and real-world hardware.** (a) Representative gaze estimation results from the cross-subject simulation. The top row shows the input eye patches, while the bottom row displays the corresponding reconstructions overlaid with ground-truth (yellow) and predicted (blue) gaze vectors. (b) Real hardware measurement results for Subject 003. The left plot illustrates the spatial distribution of $n = 138$ evaluation points relative to $K = 12$ calibration points. The right heatmap visualizes the per-point angular error across the field of view, spanning approximately ±15° in pitch and yaw, achieving a mean error of 4.50° after subject-specific fine-tuning.

alongside the decoder $\mathcal{P}$ using our curated dataset of eye images, as discussed in Section 3. The second stage transitions to real-world deployment through human-subject fine-tuning. We fabricated the optimized masks from the first stage and assemble them with MLA and phototransistors. The system processes direct optical input from human eyes; for each participant, we apply a rapid subject-specific calibration to fine-tune the weights of $\mathcal{P}$, yielding the final inference

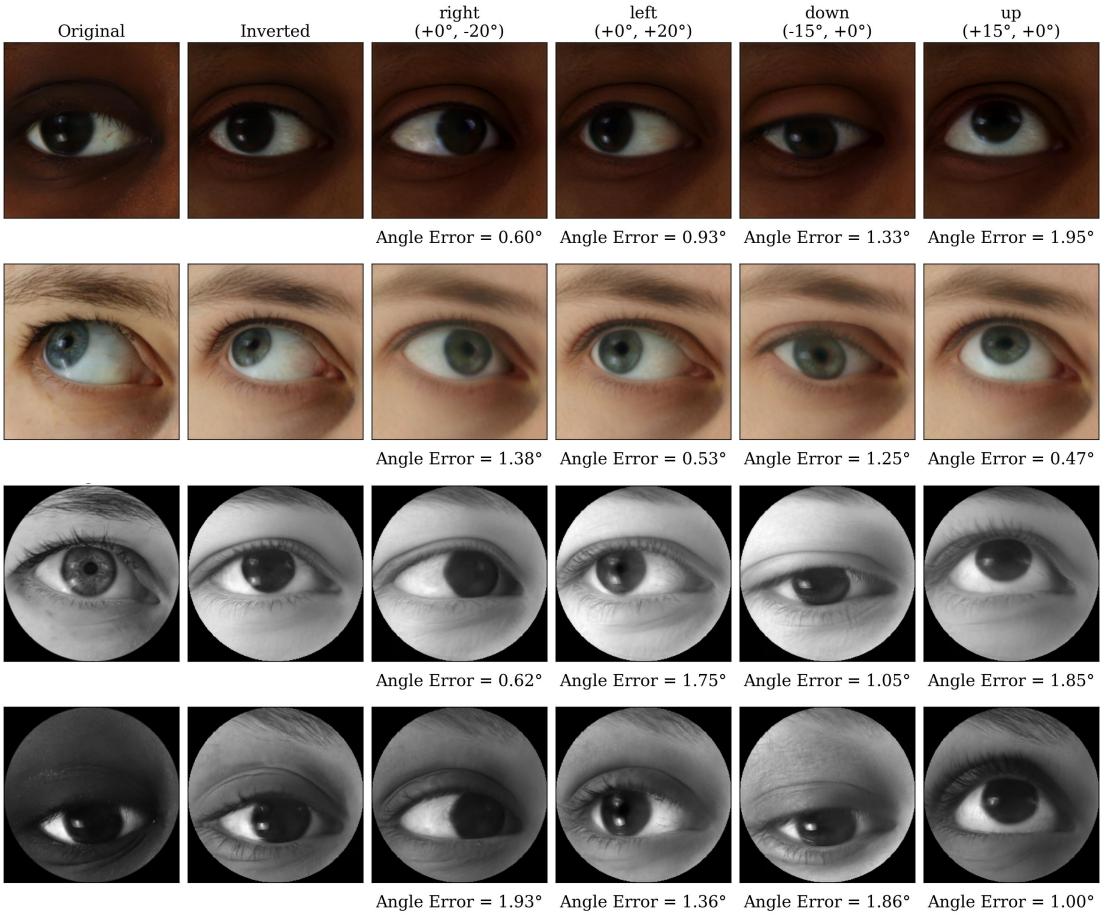


Fig. 6. **Gaze steering example.** Original / Inverted shows the source eye and its latent inversion from the simulated 16-dimensional sensor measurements. We steer gaze to right $(0^\circ, -20^\circ)$, left $(0^\circ, +20^\circ)$, down $(-15^\circ, 0^\circ)$, and up $(+15^\circ, 0^\circ)$. Grayscale results show that the compact latent space preserves steerable gaze information while varying the appearance, thus maintaining privacy, whereas color results use full image-space inversion for higher fidelity in the identity of the subject.

$\mathcal{P}_{\text{sub}}$. This two-step approach allows us to bridge the gap between theoretical optical design and practical low-latency gaze tracking.

5.1 Fully Simulated Evaluation

We evaluate whether gaze can be estimated from a compact set of learned optical measurements by comparing our method with image-based baselines under controlled sensing and computation budgets. The goal is to test whether task-optimized optical sensing can preserve gaze-relevant information while reducing both the number of acquired measurements and the complexity of the downstream predictor.

All models are trained and evaluated on monocular ETH-XGaze eye patches [Zhang et al. 2020], using the subject split described in Section 3.3, unless otherwise specified, we train using the original training set together with the generative augmentation strategy described in Section 3.4. For the full-resolution ResNet18 baseline, this augmentation improves the angular error from 3.33° to 3.08°, suggesting that diversifying periocular appearance while preserving gaze labels improves cross-subject generalization. We compare against representative convolutional backbones used in modern gaze-estimation pipelines, including ResNet-family models [Athavale et al. 2022; Cheng et al. 2022; Park et al. 2020], as well as MobileNetV3 [Howard et al. 2017] and EfficientNet-B0 [Tan and Le 2019]. These architectures are relevant baselines because they are commonly used as compact encoders or bottlenecks for edge-oriented vision systems. Image-based baselines operate either on 16×16 or 256×256 eye patches. In contrast, our method does not acquire a dense image; it estimates gaze from 16 learned optical measurements captured by a 4×4 phototransistor array and decoded by a lightweight MLP trained with latent-space distillation, as described in Section 3.1.

Fig 5(a) qualitatively illustrates this sensing regime. The learned measurements preserve enough gaze information to recover a representative eye sample with a similar gaze direction, but the reconstruction is not intended to preserve the subject's identity or detailed appearance. Instead, it corresponds to a plausible eye within the

Table 1. Compactness–accuracy comparison on ETH-XGaze eye patches. Lower angular error is better. MobileNetV3 and EfficientNet-B0 are included as representative edge-oriented image backbones. Params and FLOPs refer to the digital predictor after sensing.

Method	Output	Params (M)	FLOPs (M)	Angular error (°) ↓
ResNet18	16×16 img.	11.697	280.10	5.54
ResNet18	256×256 img.	11.697	4546	3.08
MobileNetV3	16×16 img.	1.503	4.27	6.87
MobileNetV3	256×256 img.	1.503	143.04	3.11
EfficientNet-B0	16×16 img.	5.124	19.51	5.87
EfficientNet-B0	256×256 img.	5.124	1029.00	**2.96**
Ours	16 meas.	**0.244**	**0.49**	6.47

generative model distribution that matches the inferred gaze. This is desirable for privacy-preserving gaze sensing, since eye-tracking data can reveal sensitive personal information [Kröger et al. 2020; Liebling and Preibusch 2014], whereas our sensing pipeline avoids capturing or reconstructing a faithful biometric eye image (Figure 6).

Table 1 summarizes the compactness–accuracy tradeoff. Our method achieves 6.47° angular error using only 16 measurements, 0.244M parameters, and 0.49 MFLOPs. Compared with the 16×16 ResNet18 baseline, it reduces sensing dimensionality by 16×, parameters by approximately 48×, and predictor computation by approximately 573×, with a moderate error increase from 5.54° to 6.47°. Compared with the full-resolution 256×256 ResNet18 baseline, it reduces sensing dimensionality by 4096× and predictor computation by approximately 9300×.

The compact-backbone comparisons show that the proposed approach is not simply a smaller neural network. At 16×16, MobileNetV3 obtains 6.87° error with 1.503M parameters and 4.27M FLOPs, while our method achieves lower error with approximately 6.2× fewer parameters and 8.7× fewer FLOPs. EfficientNet-B0 improves the error to 5.87°, but requires approximately 21× more parameters and 40× more FLOPs. Thus, even aggressively downsampled image-based baselines still require forming and processing an eye image, whereas our system directly acquires a small set of task-optimized optical measurements.

We further examined whether accuracy depends on the gaze range by evaluating the same model on test subsets restricted to increasingly wide ranges. The error varies by less than 0.3° across these subsets, from 6.17° within ±15°—the range covered by the hardware evaluation in Sec. 5.3—to 6.47° over the full test span, indicating that accuracy is limited by the sensing budget rather than by how far the eye is rotated.

Overall, these results show that the proposed system targets an extreme low-measurement regime rather than the unconstrained accuracy regime of dense image-based gaze estimators. By combining learned optical sensing with latent-space supervision, the system provides a compact, identity-obfuscating pipeline for low-power and low-latency gaze estimation.

Table 2. Real-measurement calibration results across 12 subjects. K denotes the number of subject-specific calibration points used to fine-tune the decoder; MAE is the mean angular error.

K	9	12	15
MAE (°) ↓	6.55	5.96	5.92

5.2 Gaze Steering Evaluation

To assess identity and appearance preservation in the gaze-steering operator, we report a cyclic round-trip evaluation. Starting from the source reconstruction, we first steer the gaze to a target direction and then steer it back to the original direction, comparing the cycled image with the source reconstruction. This provides a meaningful reference for image-quality metrics such as PSNR, SSIM, and LPIPS, since the forward edit alone has no ground-truth target image. The forward and backward angular errors, 1.18° and 1.23°, confirm accurate steering in both directions, while the photometric scores, PSNR 29.02 dB, SSIM 0.879, and LPIPS 0.125, indicate that appearance is well preserved through the full round trip. Qualitative examples are shown in Fig. 6.

We use gaze steering as a lightweight latent-space editing module rather than a dedicated steering network. Given the source and target gaze directions, a small MLP predicts a gated offset in $\Delta\mathbf{w}^+$, enabling controlled gaze edits with minimal architectural complexity. Prior work used gaze steering to synthesize adaptation data and reported small improvement in gaze-estimation accuracy [Yu et al. 2019]; however, we did not observe a comparable adaptation gain in our setting. Instead, our results highlight gaze steering as a practical editing tool for graphics applications, requiring only a compact MLP rather than a specialized steering architecture.

5.3 Evaluation with Human Subjects

To evaluate the real-world performance of our system and bridge the simulation-to-reality gap, we conducted a series of experiments with human subjects using our hardware prototype as shown on the right of Fig. 1.

During acquisition, subjects stabilized their heads and observed a monitor through the beam splitter. Calibration targets were randomly presented across the screen to approximately uniformly cover the gaze range, and the known screen geometry was used to compute the ground-truth gaze direction. No eye images were captured during evaluation. Illumination was provided by standard broadband room lighting. For each target, the FPGA collected continuous raw intensity signals for less than one second at 900 Hz. The phototransistor outputs were normalized using white and black reference measurements, filtered for electronic noise and to form the final 16-dimensional latent measurements.

We conducted experiments with 12 human subjects using the hardware prototype. Data were collected under ordinary indoor lighting using subjects of different genders and ethnicities. The evaluation naturally included variations in eye color, contact lenses, eye makeup, and small involuntary head movements. Subjects self-initiated each measurement once they believed they had stably fixated on the displayed target; trials affected by obvious blinks or

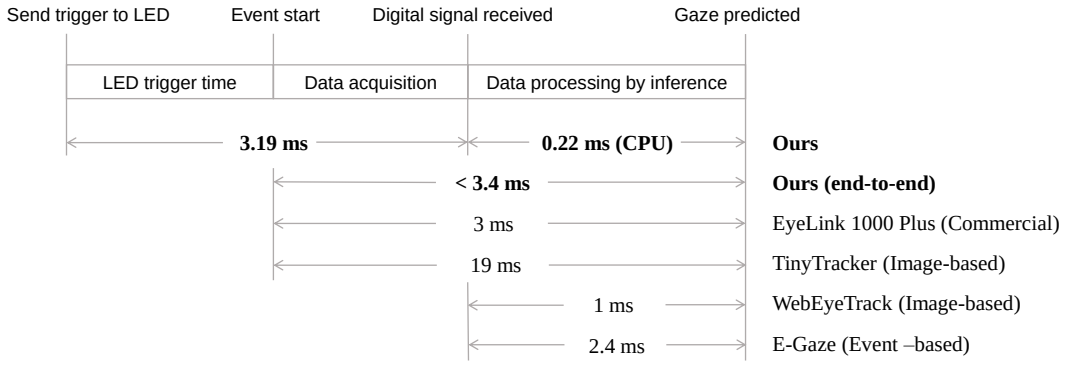


Fig. 7. **Timing diagram and latency comparison.** Our end-to-end pipeline is measured from the initial LED trigger through physical acquisition (3.19 ms) to final CPU inference (0.22 ms), totaling less than 3.4 ms. This performance is compared against state-of-the-art commercial [SR Research Ltd. 2013], image-based [Bonazzi et al. 2023; Davalos et al. 2025], and event-based systems [Li et al. 2024], demonstrating that our all-optical latent sensing approach matches specialized high-speed trackers while significantly outperforming conventional mobile edge vision systems.

large unintended movements were simply repeated. For each subject, we captured 150 high-quality data points, with a small number of remaining outliers automatically rejected based on deviations from adjacent samples. We selected K data points from this set for rapid subject-specific calibration of $\mathcal{P}_{sub}$. These points were distributed across the entire gaze area to ensure comprehensive coverage, reporting results for $K = 9, 12,$ and 15 in Table 2. Notably, with $K = 12$ calibration points, the system already achieves a mean angular error of 5.96°, with some subjects reaching a minimum error of 4.5°. Fig. 5(b) illustrates this process for a representative participant (Subject 003), showing the spatial distribution of the $K = 12$ calibration points alongside the evaluation points (left), as well as the resulting gaze estimation errors across the field of view (right).

Two geometric factors are compensated by the per-subject calibration. Without a corneal glint as reference, residual sensor–eye misalignment is geometrically ambiguous from eye rotation [Niehorster et al. 2020], which glint-based methods mitigates by construction [Guestrin and Eizenman 2006]; and the optical axis observed by appearance-based sensing differs from the visual axis by the kappa angle, ∼5°, individual-specific, and learnable only on average across subjects. Both are constant within a session, so twelve calibration points recover most of the error, with little gain beyond. The ±15° range is determined by the geometry of our proof-of-concept setup, not the encoding principle: the same encoder covers the full test span with nearly uniform error (Sec. 5.1), making the calibrated 5.96° comparable to the matched-range simulation error of 6.17°. Since foveal vision spans only ∼2° [Patney et al. 2016], this accuracy supports region-level gaze applications—attention, dwell, and coarse selection—rather than the sub-degree regime of clinical work or precise pointing; at 6.3°, LiGaze distinguishes 4, 9, and 16 screen regions with 100%, 99.5%, and 91.7% accuracy [Li et al. 2017].

5.4 System Speed and Latency

We evaluate the end-to-end latency of the proposed system, including optical sensing, signal acquisition, and computational inference. All latency is measured over 2000 runs after 50 warm-up iterations.

Measurement Setup. To measure the physical sensing latency, we use an LED as a controlled optical trigger. A digital signal is sent to turn on the LED, and the resulting change in measured intensity is recorded by the phototransistor array and read by the PC. The time difference between the trigger signal and the recorded intensity change reflects the sensing and acquisition latency.

Optical Sensing Latency. Across repeated measurements, the average latency from LED trigger to PC readout is 3.19 ms. This includes phototransistor response, analog-to-digital conversion, FPGA readout, and data transfer. The sensing pipeline operates at 900 Hz, corresponding to a frame interval of approximately 1.11 ms.

Inference Latency. Inference is evaluated on a workstation equipped with an AMD EPYC 7763 CPU and an NVIDIA RTX A4500 GPU. On a CPU, the forward pass takes approximately 0.22 ms. On a GPU, the network inference itself requires 0.17 ms. When accounting for data transfer between CPU and GPU, an additional overhead is observed. The total GPU-based inference latency, including data transfer, is approximately 0.68 ms.

End-to-End Latency. Combining sensing and inference, the total system latency is approximately 3.4 ms on CPU and 3.87 ms on GPU. This demonstrates that the proposed system achieves sub-5 ms latency, significantly outperforming conventional camera-based gaze tracking systems that typically operate at tens of milliseconds due to image acquisition and processing overhead. A comparison with representative systems is provided in Figure 7. The temporal breakdown of our trigger-to-output pipeline and its comparison with state-of-the-art image-based and event-based trackers are further illustrated in Fig. 7.

It is important to note that latency definitions vary across prior work. For example, in recent event-based gaze tracking systems

such as [Li et al. 2024], latency is defined as *"the time required to estimate gaze from an event set"*, which corresponds to processing time after events have been accumulated. As a result, this definition does not explicitly account for the event acquisition process itself. In practice, event generation depends on brightness changes and can become significantly slower under low-light or low-contrast conditions, introducing additional sensing delay that is not reflected in the reported latency.

While high-end commercial systems such as the EyeLink 1000 Plus [SR Research Ltd. 2013] report an end-to-end latency of approximately 3 ms, our measured sensing latency of 3.19 ms includes the inherent activation delay of the LED trigger used in our setup. Since this trigger introduces additional delay in the measurement loop, the reported latency should be considered an upper bound. Accounting for this measurement overhead, the actual latency of the optical sensing-to-inference pipeline is expected to be lower, suggesting that the proposed system can match or potentially exceed the temporal performance of specialized hardware while maintaining a fully passive and compact architecture.

6 Discussion and Outlook

In this work, we have demonstrated an all-optical latent feature acquisition system that achieves low-latency gaze tracking by shifting the computational burden of feature extraction into the optical domain. By utilizing a task-driven optical encoding, we bypass the conventional image-formation bottleneck, directly capturing a compact 16-dimensional representation of the eye.

The primary advantage of the proposed pipeline is its extreme efficiency on both the sensing and computational sides. As shown in our latency analysis, the physical acquisition takes only 3.19 ms, while the lightweight digital predictor requires a mere 0.22 ms on a CPU. This combined performance results in a total system latency of less than 3.4 ms, successfully meeting the critical 5 ms threshold required for immersive XR applications. Unlike image-based systems that are limited by CMOS frame rates and massive data readout, our sensing-as-inference paradigm enables a high sampling rate of 900 Hz with minimal data volume.

Our evaluation results highlight the effectiveness of the two-stage training strategy. In a fully simulated cross-subject setting, the system achieves a mean angular error of 6.47° without individual fine-tuning. Our real-world experiments demonstrate that rapid subject-specific fine-tuning successfully bridges the simulation-to-reality gap. With only 12 to 15 calibration points, the per-subject mean angular error is reduced to below 6°. This confirms that the learned latent features are not only robust to physical fabrication tolerances but are also expressive enough to capture gaze-relevant information across different human subjects.

While our current hardware serves as a proof-of-concept prototype, it establishes a blueprint for more efficient commercial implementations. The massive reduction in measurement dimensionality—from a 256×256 image to just 16 scalars—represents a 4096× decrease in sensing data. This reduction directly translates to lower power consumption in data transmission and processing, which is vital for battery-operated mobile XR glasses. State-of-the-art camera-based gaze trackers achieve substantially higher accuracy, but rely on richer sensing: high-resolution eye images, active infrared illumination, larger input regions (both eyes or the full face), and more extensive calibration, together with considerably more computational processing [Cheng et al. 2022; Kellnhofer et al. 2019; Zhang et al. 2020]. Our work instead explores a different point on the accuracy-latency-compute trade-off. Consequently, direct comparison of angular accuracy alone does not fully reflect the different design objectives of these systems. Future implementations could further improve accuracy and integration through better optical fabrication, alignment, and miniaturization while preserving the core sensing paradigm. Within the regime identified above, such systems are particularly attractive for latency- and power-constrained applications—smart glasses, near-eye displays, attention monitoring, coarse gaze interaction, and foveated rendering with conservatively padded foveal regions—where extremely low end-to-end latency and compact passive hardware matter more than the highest attainable gaze accuracy.

Acknowledgments

This work was supported by King Abdullah University of Science and Technology (Individual Baseline Funding) and partly done in the KAUST Nanofabrication Corelabs.

The authors would like to thank all the participants for providing the human gaze datasets that made this research possible.

Author contributions. Y.Zheng prepared the simulation dataset, designed and assembled the optical prototype, and conducted the real experiments. M.Souza designed and trained the model, and evaluated the simulation and real data. K.Kang designed and implemented the PCB and FPGA system. Q.Fu assisted in design, analysis and manuscript writing. H.Amata fabricated the Cr mask and microlenses. W.Heidrich supervised the project.

References

Ahmed A Abdelrahman, Thorsten Hempel, Aly Khalifa, Ayoub Al-Hamadi, and Laslo Dinges. 2023. L2cs-net: Fine-grained gaze estimation in unconstrained environments. In *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*. IEEE, 98–102.

Burak Akgül, Erol Şahin, and Sinan Kalkan. 2026. Investigating Bias and Fairness in Appearance-based Gaze Estimation. arXiv:2604.10707 [cs.CV] doi:10.48550/arXiv.2604.10707

Hadi Amata, Qiang Fu, and Wolfgang Heidrich. 2024. Comparative Performance Analysis of Multi-level Diffractive Lens and Lens Fabricated by Grayscale Lithography and Soft-imprinting, In Optica Imaging Congress 2024 (3D, AOMS, COSI, ISA, pcAOP). *Optica Imaging Congress 2024 (3D, AOMS, COSI, ISA, pcAOP)*, ITh4D.1. doi:10.1364/ISA.2024.ITh4D.1

Anastasios N. Angelopoulos, Julien N.P. Martel, Amit P. Kohli, Jörg Conradt, and Gordon Wetzstein. 2021. Event-Based Near-Eye Gaze Tracking Beyond 10,000 Hz. *IEEE Transactions on Visualization and Computer Graphics* 27, 5 (2021), 2577–2586. doi:10.1109/TVCG.2021.3067784

M. Salman Asif, Ali Ayremlou, Aswin Sankaranarayanan, Ashok Veeraraghavan, and Richard Baraniuk. 2017. Flatcam: Thin bare-sensor cameras using coded aperture and computation. *IEEE Transactions on Computational Imaging* 3, 3 (2017), 384–397.

Andrei Atanov, Jiawei Fu, Rishubh Singh, Isabella Yu, Andrew Spielberg, and Amir Zamir. 2024. How Far Can a 1-Pixel Camera Go? Solving Vision Tasks Using Photoreceptors and Computationally Designed Visual Morphology. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29 – October 4, 2024, Proceedings, Part LXXIV* (Milan, Italy). Springer-Verlag, Berlin, Heidelberg, 458–476. doi:10.1007/978-3-031-72904-1_27

Rishi Athavale, Lakshmi Sritan Motati, and Rohan Kalahasty. 2022. One eye is all you need: Lightweight ensembles for gaze estimation with single encoders. *arXiv preprint arXiv:2211.11936* (2022).

Kevin Barkevich, Reynold Bailey, and Gabriel J Diaz. 2024. Using deep learning to increase eye-tracking robustness, accuracy, and precision in virtual reality. *Proceedings*

of the ACM on computer graphics and interactive techniques 7, 2 (2024), 1–16.

Black Forest Labs. 2024. FLUX. https://github.com/black-forest-labs/flux.

Pietro Bonazzi, Sizhen Bian, Giovanni Lippolis, Yawei Li, Sadique Sheik, and Michele Magno. 2024. Retina : Low-Power Eye Tracking with Event Camera and Spiking Hardware. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 5684–5692.

Pietro Bonazzi, Thomas Rüegg, Sizhen Bian, Yawei Li, and Michele Magno. 2023. TinyTracker: Ultra-Fast and Ultra-Low-Power Edge Vision In-Sensor for Gaze Estimation. In *2023 IEEE SENSORS*. 1–4. doi:10.1109/SENSORS56945.2023.10325167

Laurie Bose, Jianing Chen, Stephen J. Carey, and Piotr Dudek. 2022. Pixel Processor Arrays For Low Latency Gaze Estimation. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 970–971. doi:10.1109/VRW55335.2022.00336

Marco Caligari, Marco Godi, Simone Guglielmetti, Franco Franchignoni, and Antonio Nardone. 2013. Eye tracking communication devices in amyotrophic lateral sclerosis: impact on disability and quality of life. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* 14, 7-8 (2013), 546–552.

John Canny. 1986. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8, 6 (1986), 679–698. doi:10.1109/TPAMI.1986.4767851

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. 2025. SAM 3: Segment Anything with Concepts. arXiv:2511.16719 [cs.CV] https://arxiv.org/abs/2511.16719

Ning Chen, Yiran Shen, Tongyu Zhang, Yanni Yang, and Hongkai Wen. 2025. EXGaze: High-Frequency and Low-Latency Gaze Tracking with Hybrid Event-Frame Cameras for On-Device Extended Reality. *IEEE Transactions on Visualization and Computer Graphics* 31, 5 (2025), 2299–2309. doi:10.1109/TVCG.2025.3549565

Yihua Cheng, Yiwei Bao, and Feng Lu. 2022. Puregaze: Purifying gaze feature for generalizable gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 436–443.

Jun Dai, Liqun Chen, Xinge Yang, Yuyao Hu, Jinwei Gu, and Tianfan Xue. 2025. Tolerance-Aware Deep Optics. *arXiv preprint arXiv:2502.04719* (2025).

Eduardo Davalos, Yike Zhang, Namrata Srivastava, Yashvitha Thatigotla, Jorge A Salas, Sara McFadden, Sun-Joo Cho, Amanda Goodwin, Ashwin TS, and Gautam Biswas. 2025. WEBEYETRACK: Scalable Eye-Tracking for the Browser via On-Device Few-Shot Personalization. *arXiv preprint arXiv:2508.19544* (2025).

Marco F Duarte, Mark A Davenport, Dharmpal Takhar, Jason N Laska, Ting Sun, Kevin F Kelly, and Richard G Baraniuk. 2008. Single-pixel imaging via compressive sampling. *IEEE signal processing magazine* 25, 2 (2008), 83–91.

Shreya Ghosh, Munawar Hayat, Abhinav Dhall, and Jarrod Knibbe. 2022. Mtgls: Multitask gaze estimation with limited supervision. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 3223–3234.

Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. 2012. Foveated 3D graphics. *ACM Trans. Graph.* 31, 6, Article 164 (Nov. 2012), 10 pages. doi:10.1145/2366145.2366183

E.D. Guestrin and M. Eizenman. 2006. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering* 53, 6 (2006), 1124–1133. doi:10.1109/TBME.2005.863952

Tao Han, Ang Li, Qinyu Chen, and Chang Gao. 2026. JaneEye: A 12-nm 2K-FPS 18.9-μJ/Frame Event-based Eye Tracking Accelerator. In *2026 31st Asia and South Pacific Design Automation Conference (ASP-DAC)*. 170–176. doi:10.1109/ASP-DAC66049.2026.11420815

Chi-Jui Ho, Yash Belhe, Steve Rotenberg, Ravi Ramamoorthi, Tzu-Mao Li, and Nicholas Antipa. 2024. A Differentiable Wave Optics Model for End-to-End Computational Imaging System Optimization. *arXiv preprint arXiv:2412.09774* (2024). Also presented at SPIE OPTO 2024 (PC12903).

Injoon Hong, Kyeongryeol Bong, and Hoi-Jun Yoo. 2018. Challenges of eye tracking systems for mobile XR glasses. In *Applications of Digital Image Processing XLI*, Andrew G. Tescher (Ed.), Vol. 10752. International Society for Optics and Photonics, SPIE, 1075217. doi:10.1117/12.2322657

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).

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.

Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. 2019. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6912–6921.

Joohwan Kim, Michael Stengel, Alexander Majercik, Shalini De Mello, David Dunn, Samuli Laine, Morgan McGuire, and David Luebke. 2019. NVGaze: An Anatomically-Informed Dataset for Low-Latency, Near-Eye Gaze Estimation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300780

Jeremy Klotz and Shree K. Nayar. 2024. Minimalist Vision with Freeform Pixels. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXIV* (Milan, Italy). Springer-Verlag, Berlin, Heidelberg, 329–346. doi:10.1007/978-3-031-73039-9_19

Jacob Leon Kröger, Otto Hans-Martin Lutz, and Florian Müller. 2020. What Does Your Gaze Reveal About You? On the Privacy Implications of Eye Tracking. In *Privacy and Identity Management. Data for Better Living: AI and Privacy*. Springer International Publishing, 226–241. doi:10.1007/978-3-030-42504-3_15

Isack Lee, Jun-Seok Yun, Hee Hyeon Kim, Youngju Na, and Seok Bong Yoo. 2022. Latentgaze: Cross-domain gaze estimation through gaze-aware analytic latent code manipulation. In *Proceedings of the asian conference on computer vision*. 3379–3395.

Nealson Li et al. 2024. E-Gaze: Gaze Estimation with Event Camera. *IEEE TPAMI* (2024).

Richard Li, Eric Whitmire, Michael Stengel, Ben Boudaoud, Jan Kautz, David Luebke, Shwetak Patel, and Kaan Akşit. 2020. Optical Gaze Tracking with Spatially-Sparse Single-Pixel Detectors. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 117–126. doi:10.1109/ISMAR50242.2020.00033

Tianxing Li, Qiang Liu, and Xia Zhou. 2017. Ultra-Low Power Gaze Tracking for Virtual Reality. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems* (Delft, Netherlands) *(SenSys '17)*. Association for Computing Machinery, New York, NY, USA, Article 25, 14 pages. doi:10.1145/3131672.3131682

Daniel J. Liebling and Sören Preibusch. 2014. Privacy Considerations for a Pervasive Eye Tracking World. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 1169–1177. doi:10.1145/2638728.2641688

Yuhui Liu, Liangxun Ou, Qiang Fu, Hadi Amata, Wolfgang Heidrich, and Yifan Peng. 2025. Learned Binocular-Encoding Optics for RGBD Imaging Using Joint Stereo and Focus Cues. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 15833–15842.

Päivi Majaranta and Kari-Jouko Räihä. 2002. Twenty years of eye typing: systems and design issues. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications* (New Orleans, Louisiana) *(ETRA '02)*. Association for Computing Machinery, New York, NY, USA, 15–22. doi:10.1145/507072.507076

Shakhnoza Muksimova, Yakhyokhuja Valikhujaev, Sabina Umirzakova, Jushkin Baltayev, and Young Im Cho. 2025. GazeCapsNet: A lightweight gaze Estimation framework. *Sensors* 25, 4 (2025), 1224.

Diederick C Niehorster, Thiago Santini, Roy S Hessels, Ignace TC Hooge, Enkelejda Kasneci, and Marcus Nyström. 2020. The impact of slippage on the data quality of head-worn eye trackers. *Behavior research methods* 52, 3 (2020), 1140–1160.

Seonwook Park, Emre Aksan, Xucong Zhang, and Otmar Hilliges. 2020. Towards end-to-end video-based eye-tracking. In *European conference on computer vision*. Springer, 747–763.

Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. 2016. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions On Graphics (TOG)* 35, 6 (2016), 1–12.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2021. Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Argha Sen, Nuwan Sriyantha Bandara, Ila Gokarn, Thivya Kandappu, and Archan Misra. 2024. EyeTrAES: Fine-grained, Low-Latency Eye Tracking via Adaptive Event Slicing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 150 (Nov. 2024), 32 pages. doi:10.1145/3699745

Rustam Shadiev and Dandan Li. 2023. A review study on eye-tracking technology usage in immersive virtual reality learning environments. *Computers & education* 196 (2023), 104681.

Wanxin Shi, Zheng Huang, Honghao Huang, Chengyang Hu, Minghua Chen, Sigang Yang, and Hongwei Chen. 2022. LOEN: Lensless opto-electronic neural network empowered machine vision. *Light: Science & Applications* 11, 1 (2022), 121.

Brian A Smith, Qi Yin, Steven K Feiner, and Shree K Nayar. 2013. Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 271–280.

Matheus Souza, Yidan Zheng, Kaizhang Kang, Yogeshwar Nath Mishra, Qiang Fu, and Wolfgang Heidrich. 2025. Latent space imaging. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 28295–28305.

SR Research Ltd. 2013. *EyeLink 1000 Plus [Apparatus and software].* Mississauga, Ontario, Canada. https://www.sr-research.com/eyelink-1000-plus/ Running Host Software version 5.50.

Michael Stengel, Steve Grogorick, Martin Eisemann, Elmar Eisemann, and Marcus A Magnor. 2015. An affordable solution for binocular eye tracking and calibration in head-mounted displays. In *Proceedings of the 23rd ACM international conference on Multimedia*. 15–24.

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.

Yunjia Sun, Jiabei Zeng, Shiguang Shan, and Xilin Chen. 2021. Cross-encoder for unsupervised gaze representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3702–3711.

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

Vildan Tanriverdi and Robert JK Jacob. 2000. Interacting with eye movements in virtual environments. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 265–272.

Pavel Ugwitz, Ondřej Kvarda, Zuzana Juříková, Čeněk Šašinka, and Sascha Tamm. 2022. Eye-tracking in interactive virtual environments: implementation and evaluation. *Applied Sciences* 12, 3 (2022), 1027.

Tianyu Wang, Mandar M Sohoni, Logan G Wright, Martin M Stein, Shi-Yuan Ma, Tatsuhiro Onodera, Maxwell G Anderson, and Peter L McMahon. 2023. Image sensing with multilayer nonlinear optical neural networks. *Nature Photonics* 17, 5 (2023), 408–415.

Yaoming Wang, Yangzhou Jiang, Jin Li, Bingbing Ni, Wenrui Dai, Chenglin Li, Hongkai Xiong, and Teng Li. 2022. Contrastive Regression for Domain Adaptation on Gaze Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19376–19385.

Xinge Yang, Qiang Fu, Yunfeng Nie, and Wolfgang Heidrich. 2026. Task-driven lens design. *Opt. Express* 34, 5 (Mar 2026), 8961–8975. doi:10.1364/OE.588912

Yu Yu, Gang Liu, and Jean-Marc Odobez. 2019. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11937–11946.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3836–3847.

Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. 2020. ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation. In *European Conference on Computer Vision (ECCV)*.

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based Gaze Estimation in the Wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4511–4520. doi:10.1109/CVPR.2015.7299081

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2019. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 41, 1 (2019), 162–175. doi:10.1109/TPAMI.2017.2778103

Guangrong Zhao, Yurun Yang, Jingwei Liu, Ning Chen, Yiran Shen, Hongkai Wen, and Guohao Lan. 2023a. EV-Eye: rethinking high-frequency eye tracking through the lenses of event cameras. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 2716, 14 pages.

Ruixuan Zhao, Chengshuai Yang, R Theodore Smith, and Liang Gao. 2023b. Coded aperture snapshot spectral imaging fundus camera. *Scientific Reports* 13, 1 (2023), 12007.