Low Latency Gaze Tracking via Latent Optical Sensing
Supplementary Material

1 Latent Space Gaze Steering

We provide additional qualitative results for gaze steering to demonstrate the semantic richness of the learned latent representations. By applying a control network C_θ to predict a residual displacement in the $\mathcal{W}^+$ space, we can manipulate the gaze direction of a captured eye while preserving the subject's unique periocular identity and appearance. As illustrated in Fig. 1, our system successfully performs large-angle redirections (up to $\pm 20°$) across diverse subjects. The steered images maintain high anatomical plausibility, with the control network effectively re-orienting the pupil and iris while adjusting the surrounding eyelid geometry to match the new gaze target. The low angular errors reported between the target gaze and the re-estimated gaze from the steered images further confirm that the latent updates are geometrically consistent.

2 Dataset Details

Our dataset is derived from the ETH-XGaze dataset [10] that contains high-resolution RGB images under regular room illumination. Since our system is based on a sensor zoomed in on an individual eye (similar to what may be found in a near-eye display), the dataset needs to be adjusted for this setting. We specifically select images from cameras positioned frontally relative to the subjects. To ensure consistency across binocular data, we normalize both the left and right eye images [10, 11]. Left eye images are horizontally flipped, and the x-component of their corresponding gaze vectors is negated to maintain a unified coordinate system.

The data processing pipeline follows a multi-stage cropping and augmentation strategy:

- **Initial Normalization:** Images are first cropped and normalized to a resolution of 288×288 pixels, representing a physical area of 45×45 mm.
- **Augmentation and Robustness:** During training, we apply random translations and scaling to these patches to extract the final input eye patches $I \in \mathbb{R}^{256\times256}$, corresponding to a 40×40 mm region. This augmentation process is specifically designed to mimic potential optical alignment errors and mechanical tolerances in the physical prototype, ensuring the network remains robust to slight shifts in the microlens array relative to the eye.
- **Artifact Mitigation:** To eliminate corner artifacts inherent to microlens array imaging, a circular mask is applied to the final eye patches.

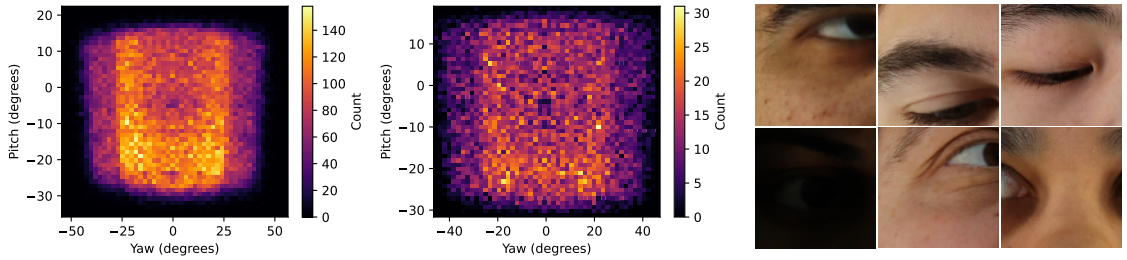
Each eye patch is annotated with a ground-truth 3D gaze vector $\mathbf{g} \in \mathbb{R}^3$ and a validity label $v \in \{0, 1\}$, where $v = 0$ indicates invalid inputs such as closed eyes, offset irises, or non-eye patches. These labels were initially manually annotated on a subset to train a classifier, which then automatically labeled the remaining corpus.

Our processed dataset consists of 68 subjects for training and 12 subjects for testing. The training set contains 142,616 images (128,379 valid), while the test set contains 25,504 images (23,828 valid). The gaze coverage spans a pitch range of $-35.87°$ to $22.45°$ and a yaw range of $-54.73°$ to $56.5°$. Figure 2 illustrates the distribution of gaze directions and examples of the automatic validity classification.

Fig. 1. Results for gaze steering in latent space.

3 Details of Identity-Varying Augmentation via Generative Inpainting

To expand our dataset we perfomed a generative based augmentation, we know that appearance-based gaze estimation models trained on datasets such as ETH-XGaze [10] are prone to identity bias: each subject contributes thousands of images across varying gaze directions, encouraging the model to learn identity-specific shortcuts — skin tone, eye shape, eyebrow structure — rather than gaze-relevant features. This is compounded by the limited demographic



(a) Dataset gaze distribution (b) Examples of invalid input

Fig. 2. **Gaze Distribution and Validity Classification.** (a) Heatmaps showing the distribution of 3D gaze vectors for the training set (68 subjects, left) and the test set (12 subjects, right). The dataset covers a wide field-of-view with pitch ranging from $-35.87°$ to $22.45°$ and yaw from $-54.73°$ to $56.5°$. (b) Representative examples of invalid eye patches ($v = 0$) identified by our automatic classifier, including closed eyes, extreme iris offsets, and non-eye regions, which are used to improve system robustness during real-world operation.

diversity typical of laboratory-collected datasets, where certain ethnicities, age groups, and facial characteristics are underrepresented [1].

To address this, we introduce an augmentation scheme that leverages the Flux text-to-image diffusion model [2] to expand the range of subject appearances associated with each gaze label. The motivation is that appearance-based gaze estimators may rely not only on the iris, pupil, and sclera, but also on surrounding periocular cues such as skin texture, eyebrow shape, facial structure, and illumination. By varying these appearance-related factors while encouraging the synthesized images to preserve the original eye configuration, the proposed augmentation exposes the model to broader identity and contextual variation without deliberately changing the annotated gaze direction.

Our pipeline operates as follows. Given an eye crop with its associated gaze vector, we first obtain the eye region using SAM3 [4]. We then construct an inverted inpainting mask that preserves the eye pixels and marks the surrounding context for regeneration. Canny edges [3] extracted from the original image serve as a structural control signal via ControlNet [9], maintaining anatomical plausibility of the generated context. Flux then inpaints the masked region conditioned on a text prompt sampled from a curated pool of descriptions spanning diverse skin tones, age groups, makeup variations, and lighting conditions. Each generated sample undergoes two quality checks: a CLIP-based [6] prompt alignment score to ensure semantic fidelity, and an eye consistency verification that re-segments the eye in the output and requires both area preservation (within 10%) and spatial overlap (IoU ≥ 0.9) with the original. Samples failing either check are regenerated with a different seed, up to three attempts.

This approach produces augmented images that are geometrically faithful, the eye region and its gaze label are untouched, while exhibiting substantial variation in the surrounding appearance. By expanding each training sample into multiple identity-diverse variants, we reduce the model's reliance on identity-correlated features and improve generalization. Examples of the augmentation process are shown in Figure 3.

Implementation details. Before edge extraction the crop is locally contrast-equalized (CLAHE, clip limit 3, 8×8 tiles), and the Canny thresholds are lowered inside a dilated neighborhood of the eye so that lid and iris contours dominate the structural signal; edge components smaller than 30 pixels are discarded. Seeds are derived deterministically from the sample index, so the corpus can be regenerated exactly. Table 1 lists the remaining settings.

Table 1. Augmentation configuration.

Segmentation	SAM3, text prompt *"eye"*
Inpainting	FLUX.1-dev + Canny ControlNet, `bf16`
Semantic check	CLIP ViT-L/14, score ≥ 23
Generation resolution	768×768, stored at crop resolution
Steps / guidance	40 / 4.5
Strength / ControlNet scale	0.97 / 0.7
Attempts per image	up to 3, distinct seeds
Variants per source image	1

Prompt pool. All prompts share the template *"raw photo, close-up of a person's face around the eye, {variation}, shot on DSLR"*. Because every crop is right-eye oriented, each one carries the suffix *"inner eye corner and nose bridge on the right side, eyebrow sweeping from right to left"*, which keeps the synthesized periocular anatomy consistent with the crop geometry. Table 2 summarizes the pool; representative variations are:

- *"thin metal rectangular glasses frame over the brow, fair skin on cheek, indoor lighting"*
- *"natural no-makeup look on forehead and cheek, visible pores, soft natural lighting"*
- *"crow's feet wrinkles on temple skin, pale cheek, natural lighting"*
- *"pale skin, cool blue-tinted indoor fluorescent light"*
- *"slight sweat on brow, flushed cheeks, outdoor exertion"*

Table 2. Composition of the prompt pool.

Skin tone and ethnicity	10
Eyeglasses	8
Everyday makeup	5
Skin features and eyebrows	10
Illumination	8
Situational appearance	3
Head coverings and hair	2
Compound (age, appearance, eyewear, lighting)	15
Total	61 (+ null)

Acceptance and fallback. If none of the three attempts satisfies both criteria, the original crop is retained unchanged rather than admitting a geometrically inconsistent sample, so a failed generation never removes a training example. The null prompt has the same effect by design. The resulting set is therefore a mixture: approximately 82% synthesized variants, 17% originals kept by fallback, and 2% pass-throughs. The pipeline was applied to the 67 training subjects; no test subject is augmented.

4 Decoder Architecture

The digital processing component of our pipeline is designed specifically for extreme computational efficiency, enabling real-time inference on resource-constrained hardware such as mobile CPUs. The decoder $\mathcal{P}$ transforms the 16-dimensional latent measurement vector $\mathbf{y}$ into a 256-dimensional bottleneck feature $\mathbf{h}$, which is subsequently processed by two dedicated heads. The architecture of decoder is shown in Fig. 4.

By utilizing only 244K parameters and 489K FLOPs, the decoder achieves a forward-pass latency of only 0.22 ms on a standard workstation CPU. This design highlights the advantage of task-driven optical sensing, as the primary feature extraction is performed passively in the optical domain, leaving only a minimal computational workload for the digital processor.

5 Simulation Results

To further validate the robustness of the latent-space sensing architecture, we present an expanded set of simulation results across the test set of 12 subjects. Fig. 5 visualizes 28 representative samples, illustrating the system's ability to generalize across varying eye shapes, skin tones, and periocular textures. An inherent advantage of our latent optical sensing paradigm is the decoupling of task-relevant information from subject identity. While the system effectively captures and retreats gaze direction with high accuracy, the resulting reconstructions in Fig. 5 exhibit a noticeable loss of fine-grained biometric features.

We also examined the effect of using fixed masks instead of jointly optimizing the mask with the model. With a fixed random mask, the angular error increased substantially to 11.87°. We further tested a Hadamard mask, commonly used in compressive sensing; however, this configuration led to unstable training, with the angular error remaining above 20°. These results indicate that joint mask design is critical for stable training and accurate gaze estimation in our setting.

6 Training Details

This section gives the training configuration of the three stages introduced in Sec. 3.1 of the main paper. All stages use grayscale 256×256 single-eye crops from the ETH-XGaze training split (Sec. 3.3 main paper), with per-sample ground-truth gaze g and validity flag v; angular errors are means over the held-out test split.

6.1 Generative model

The generator $\mathcal{G}$ [5] is trained on the 62,309 identity-augmented eye patches only (Sec. 3.4 of the main paper and Sec. 3), so that the generated images are not dominated by any specific identity. Its latent and style dimensions are both 256, the mapping network has two layers, and the synthesis network produces 256×256 single-channel images from $L = 14$ style vectors.

Patches are stored at 288×288 and augmented online in a gaze-preserving manner: a small random translation of ± 16 pixels, a 256×256 center crop, a circular aperture of radius 128 (pixels outside the aperture are set to -1, matching the physical aperture of the sensor), and per-sample brightness (± 0.2) and contrast ($\times[0.8, 1.2]$) jitter. We use Adam with learning rate 2.5×10^{-3}, $\beta = (0, 0.99)$, R_1 regularization with $\gamma = 0.8192$ and batch size 16, converging to $\mathrm{FID}_{50k} = 6.44$ at 7,000 kimg.

6.2 Inversion teacher

The backbone of $\mathcal{P}_\mathrm{T}$ is adapted to single-channel input in the style of pSp [7]: global average pooling followed by an equalized linear layer $512 \to 256$ yields the compact feature $\mathbf{h}_\mathrm{T} \in \mathbb{R}^{256}$, which is broadcast across the $L = 14$ StyleGAN layers to form the $\mathcal{W}^+$ code $\mathbf{w} \in \mathbb{R}^{14\times256}$ whenever synthesis is required. Two MLP heads read $\mathbf{h}_\mathrm{T}$: an angle head ($256 \to 128 \to 3$, SiLU) producing $\hat{g}_T$ and a validity head ($256 \to 32 \to 1$) producing $\hat{v}_T$.

Writing $\hat{I} = \mathcal{G}(\mathbf{w} + \bar{\mathbf{w}})$ for the reconstruction obtained by decoding the predicted latent (offset by the mean latent $\bar{\mathbf{w}}$) with the frozen generator, the teacher is trained with

Table 3. Training configuration of the three stages.

	$\mathcal{G}$	$\mathcal{P}_{\mathrm{T}}$	$\tilde{\mathcal{E}} + \mathcal{P}$
Architecture	StyleGAN2	pSp encoder $\rightarrow \mathcal{W}^+$	16 masks + MLP
Optimizer	Adam	AdamW	AdamW (×2)
Learning rate	$2.5{\cdot}10^{-3}$	10^{-4}	10^{-4} / 10^{-3}
Weight decay	–	$3{\cdot}10^{-3}$	10^{-2} / 0
Schedule	constant	one-cycle	plateau + decay
Batch size	16	32	128
Iterations	8,000 kimg	100k	100k
GPUs	1×A100	2×A100	1×A100
Metric	FID_{50k} 6.44	2.97°	6.47°

$$\begin{aligned}\mathcal{L}_{\mathcal{P}_{\mathrm{T}}} =& \lambda_{\mathrm{ang}} v \, \|\hat{g}_T - g\|_1 + \lambda_{\mathrm{cls}} \mathrm{BCE}(\hat{v}_T, v) \\ &+ \lambda_{\mathrm{rec}} v \, \|\hat{I} - I\|_2^2 + \lambda_{\mathrm{gaze}} v \, \|E(\hat{I}) - E(I)\|_1,\end{aligned} \tag{1}$$

with $\lambda_{\mathrm{ang}} = 1$, $\lambda_{\mathrm{cls}} = 10^{-3}$, $\lambda_{\mathrm{rec}} = 1$ and $\lambda_{\mathrm{gaze}} = 0.5$, where E is a frozen single-eye gaze estimator (EfficientNet-B0 [8], trained separately on the same data). The last term is a perceptual *gaze* consistency: it uses no labels, but forces the reconstruction to be read by E as the same direction as the real image, which keeps the inverted latent gaze-faithful rather than merely photometrically close.

We optimize with AdamW (learning rate 10^{-4}, weight decay 3×10^{-3}) under a one-cycle schedule (peak 10^{-4}, warm-up fraction 0.01) for 100k iterations at batch size 32 on two A100 GPUs. The encoder input is augmented with mild blur and additive noise, while the reconstruction target and the gaze-consistency branch always use the clean image I. The retained teacher attains 2.97° on the test split.

6.3 Learned masks and gaze decoder

During training, each mask $\mathbf{m}_k$ of $\tilde{\mathcal{E}}$ is parameterized as a 64×64 map of values in $[0, 1]$ (initialized uniformly at random), bilinearly upsampled to 256×256, binarized to one bit with a straight-through estimator, and multiplied by the circular aperture. Gradients therefore reach the continuous parameterization while the forward pass uses the binary pattern that the optical encoder $\mathcal{E}$ can realize; the mask bank holds 65,536 trainable values.

Completing the decoder description of Sec. 3.1, the normalization layer of $\mathcal{P}$ is a LayerNorm over $\mathbf{y}$, the upscaling block is $16 \rightarrow 256 \rightarrow 256$ with SiLU and dropout 0.2, and the residual bottleneck consists of two pre-norm blocks of width 256, yielding $\mathbf{h} \in \mathbb{R}^{256}$. The decoder has 244,324 parameters.

In the training objective of Sec. 3.1 we set $\lambda_{\mathrm{sup}} = 1$, $\lambda_{\mathrm{cls}} = 10^{-3}$, $\lambda_{\text{gaze-distill}} = 0.5$ and $\lambda_{\text{feat-distill}} = 0.1$. The teacher always sees the *clean* image while $\tilde{\mathcal{E}}$ receives the augmented one, so the two distillation terms additionally act as an augmentation-invariance constraint.

Optimization uses two AdamW groups — the decoder at learning rate 10^{-4} with weight decay 10^{-2}, the masks at 10^{-3} without weight decay — both on a schedule that holds the peak rate for the first 80% of training and then decays linearly to 0.1×, with gradient clipping at 1.0 over the union of both parameter sets. We train for 100k iterations at batch size 128 on a single A100. Training images are augmented with random Gaussian blur (kernel 11, $\sigma \in [0.1, 10]$) and additive Gaussian noise ($\sigma \in [0.01, 0.10]$), each applied with probability 0.5, modeling the defocus and photon noise of the optical encoder.

Table 4. Per-subject calibration of $\mathcal{P}_{\text{sub}}$ on the physical prototype: mean angular error as a function of the calibration budget K, evaluated on the $150 - K$ samples held out from each subject's calibration set. Values are the mean ± standard deviation across the 12 subjects.

K	MAE (°) ↓
9	6.55 ± 1.28
12	**5.96 ± 1.26**
15	5.92 ± 1.41
18	5.69 ± 1.29
21	5.64 ± 1.22

7 Per-Subject Calibration

For a budget of K points we select, among the 150 recorded samples of a subject, the K whose reference gaze best tiles the gaze range: a regular grid is placed over the bounding box of the subject's pitch–yaw distribution and each grid position is snapped to the nearest unused sample. The selection is deterministic, and the remaining $150 - K$ samples are held out for evaluation, so all reported errors are measured on gaze directions never seen during calibration.

The decoder consumes the normalized phototransistor readings directly, its input LayerNorm absorbing the per-sample gain and offset, so no radiometric calibration of the sensor is required. Starting from the distilled weights, all parameters of $\mathcal{P}$ are fine-tuned on the K pairs by minimizing the mean angular error,

$$\mathcal{L}_{\text{cal}} = \frac{1}{K} \sum_{j=1}^{K} \arccos\big(\cos(\hat{g}_j, g_j) \big), \tag{2}$$

with AdamW at learning rate 10^{-4}, batch size 16 and gradient clipping at 0.05. Fitting 0.24M parameters to as few as nine samples requires strong regularization, so we use a weight decay of 0.5, which together with the small clipping threshold keeps the solution close to the distilled initialization. Each (subject, K) pair is calibrated independently with a fixed seed. We run 3,500 epochs, evaluate on the held-out samples every 50 epochs, and retain the best checkpoint.

Table 4 extends the budgets of the main paper to $K = 18$ and $K = 21$ and reports the spread across subjects. Beyond $K = 12$ the error stays within $0.33°$ of the operating point, well inside the between-subject standard deviation of $\approx 1.3°$. The residual error is therefore dominated by between-subject variation rather than by the budget.

References

[1] 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

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

[3] 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

[4] 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

[5] 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.

[6] 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.

[7] 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)*.

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

[9] 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.

[10] 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)*.

[11] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2018. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications* (Warsaw, Poland) *(ETRA '18)*. Association for Computing Machinery, New York, NY, USA, Article 12, 9 pages. doi:10.1145/3204493.3204548

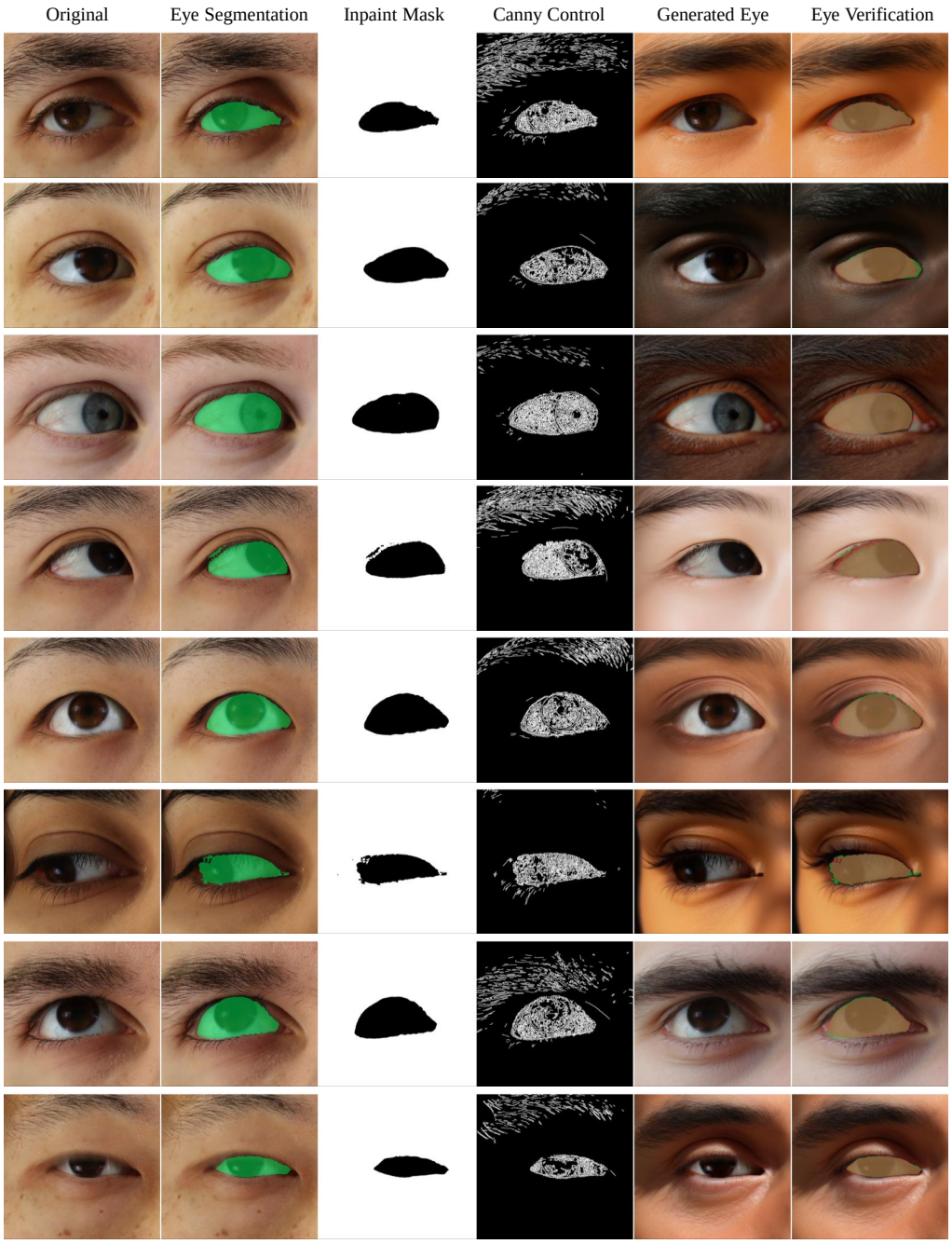


Fig. 3. **Identity-varying augmentation via generative inpainting.** From left to right: original eye crop, SAM3 eye segmentation, inpainting mask, Canny-edge ControlNet input, Flux-generated eye conditioned on the mask, Canny edges, and a diversity prompt, and SAM3 re-segmentation of the generated eye overlaid on the original mask for consistency checking. Samples failing the area/IoU check are replaced by the original crop, so the gaze label always remains valid.

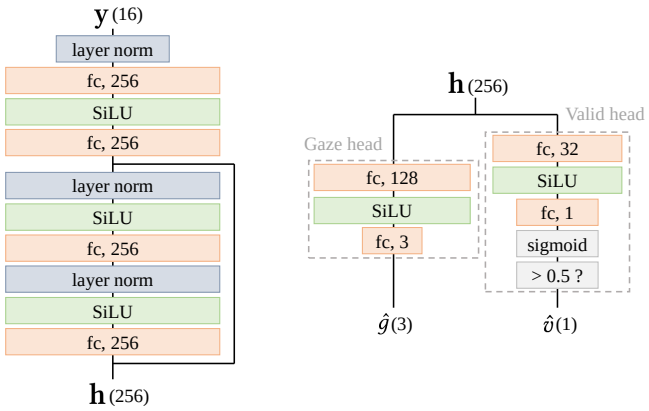


Fig. 4. **Detailed schematic of the lightweight decoder $\mathcal{P}$.** (Left) The input measurement vector $\mathbf{y} \in \mathbb{R}^{16}$ is processed through a series of LayerNorm and FC-256 blocks to generate the bottleneck representation $\mathbf{h}$. (Right) The bottleneck feature is split into a Gaze Head for 3D vector prediction and a Valid Head that utilizes a sigmoid threshold (> 0.5) to filter artifacts such as blinks or misalignments

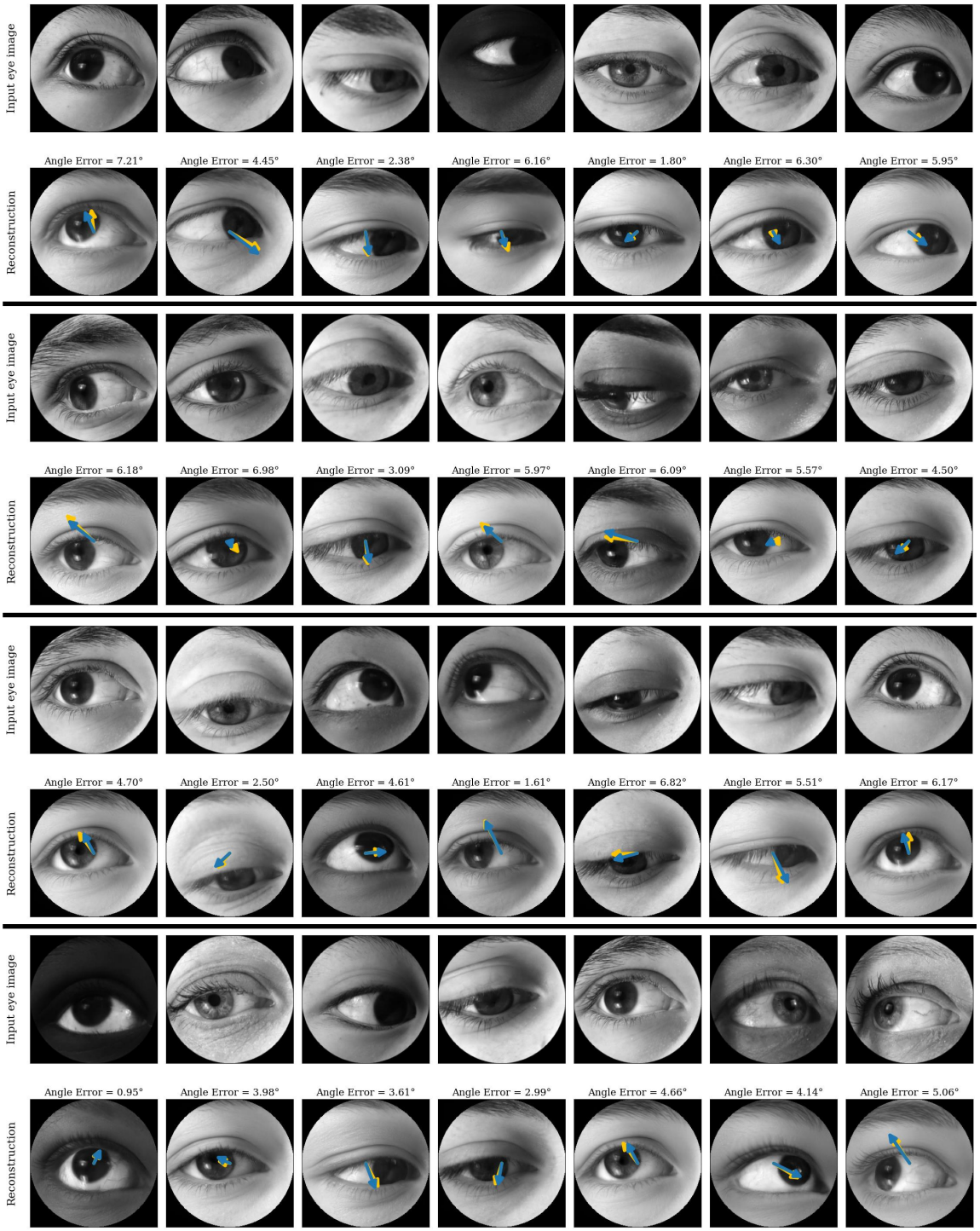


Fig. 5. Gallery of latent sensing results in simulation.