GazeShift: Unsupervised Gaze Estimation and Dataset for VR Supplementary Material

A. GazeShift: Implementation Details and Qualitative Results

As VRGaze and remote-camera datasets such as Columbia and MPIIGaze differ in input resolution, we adopt architecture variations tailored to each. All experiments were conducted on a single NVIDIA RTX A5000 GPU. Our GazeShift code and VRGaze dataset are available here: https://github.com/gazeshift3/gazeshift

A.1. VR Experimental Details

The VRGaze input resolution is $1 \times 400 \times 400$. To support real-time inference on a VR headset, the gaze encoder is designed to be lightweight: it consists of 6 stride-2 MobileNetV2 blocks, each repeated twice and configured with a width multiplier of 2. A final linear layer projects the output to a gaze embedding of dimension C_d.

The appearance encoder consists of a stride-2 convolution, followed by four MobileNetV2 blocks, each repeated six times, and a final stride-1 convolutional layer, producing a feature map of size $10 \times 10 \times C_a$. The attention module employs a single-layer, single-head attention mechanism. The decoder is composed of transposed convolutional blocks, each followed by batch normalization and a `tanh` activation.

The model is trained for 20 epochs with a batch size of 30 using the AdamW optimizer, a learning rate of 0.0001, and a weight decay of 0.05.

A.1.1. VAE baseline details

The VAE baseline (Table 3 in the paper) employs an architecture analogous to GazeShift, utilizing a MobileNetV2-based image encoder and a matching decoder. It is trained strictly via a standard image reconstruction objective on the input frames.

A.1.2. Supervised Appearance-based Model

To contextualize the performance of GazeShift, we train a supervised binocular model using the same lightweight gaze encoder architecture. Training samples are selected from intervals where the gaze target remains stationary for three seconds, ensuring higher label reliability compared to periods of target motion. We adopt a simple Siamese configuration: synchronized left and right eye images serve as input, with the right-eye images horizontally flipped to reduce appearance asymmetry and improve representation learning. The resulting embeddings are concatenated and passed through a regression head to predict 3D gaze direction.

Calibration aware loss. In conventional gaze regression models, the training loss is computed directly on raw outputs, ignoring the calibration step typically performed at inference. We propose integrating calibration directly into training by applying the loss to calibrated outputs. Specifically, we implement a linear calibration module in PyTorch that fits polynomial features to the batch-level predictions. This module is treated as a differentiable layer, allowing the model to optimize for post-calibration accuracy and better align with real-world inference behavior. Calibration aware loss reduced the average error from $1.64°$ to $1.54°$. Fig. 1 depicts our supervised model architecture.

A.2. Remote Camera Experimental Details

In the remote camera setting, the input resolution of each eye crop is 64×32. The gaze encoder consists of 3 MobileNetV2 blocks, while the appearance encoder follows a ResNet-18 backbone, similar to the configuration used in Cross-Encoder. The attention module uses 2 layers with 2 heads, and the decoder is implemented using a DenseNet architecture.

Unlike the near-eye setting, both eye crops in the remote setup are extracted from the same image, resulting in similar appearance characteristics. We leverage this symmetry by generating source–target pairs not only from different time steps of the same eye, but also by using synchronized left/right eye pairs – where one image is horizontally flipped to simulate a plausible gaze shift.

B. VRGaze Dataset: Additional Characteristics and Collection Information

Target motion pattern. Figure 2 illustrates the distribution of gaze targets for a single person over multiple sessions. Linear segments are formed by a slowly moving ball to induce smooth pursuit eye movements. The nodes at the

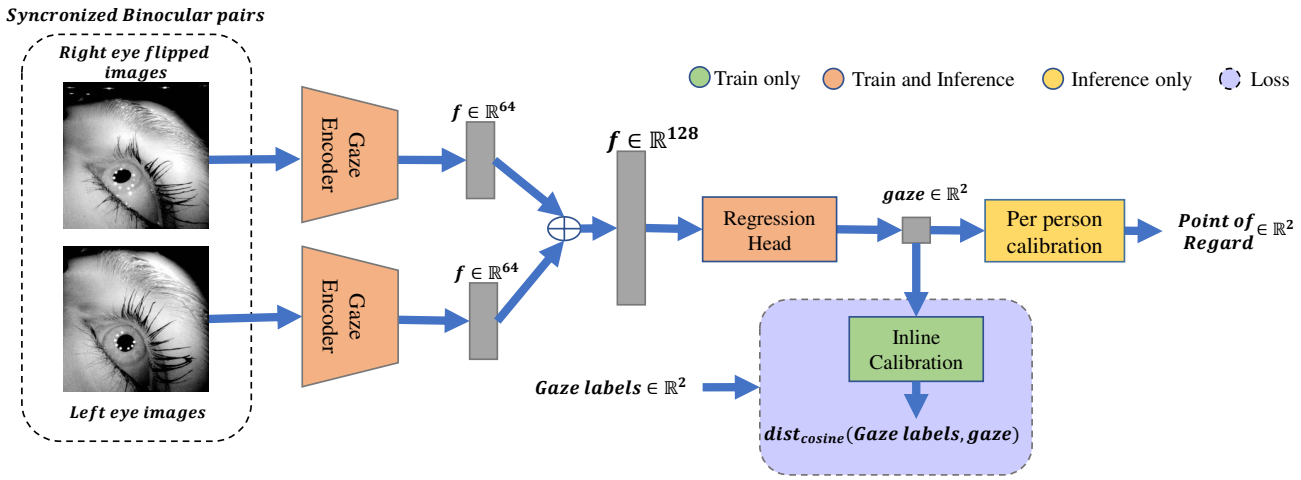


Figure 1. Supervised model architecture.

ends of the segments are stationary points intended to elicit eye fixations. Thus, each session consists of a ball moving slowly along straight trajectories, followed by an abrupt stop lasting three seconds. During each stop, the ball gradually changes in size to enhance the fixation.

Gaze direction diversity. To illustrate the range of gaze directions captured, we present in Fig. 3 nine example eye images corresponding to different target positions, including center, corners, and cardinal directions (left, right, up, down). This visualization demonstrates how eye appearance varies with gaze angle, providing qualitative insight into VRGaze dataset.

Pupil dilation diversity. To ensure the robustness of gaze tracking under varying lighting conditions and physiological responses, we collected eye images exhibiting a wide range of pupil sizes. This was achieved by varying the background brightness during data collection sessions in VR, prompting natural pupil dilation and constriction. Fig. 4 illustrates examples of eye images with different pupil diameters, highlighting the captured diversity. Such variation is essential for training models that generalize well across users and lighting environments.

Pupil Center and Glint Annotations IR LED reflections (glints) are typically visible in VR cameras and may also be used for gaze estimation. To promote future research, we also release coordinates of pupil center and glints for each image in VRGaze. To obtain the pupil and glints coordinates for the entire dataset, we manually annotated 32k images which were used to train a feature detection model to annotate the rest. We trained the model on 28k images and evaluated it on a validation set of 4k images, achieving mean accuracies of 0.54 and 0.96 pixels for pupil and glint detection, respectively. Fig. 5 illustrates annotated and predicted features. The pupil center is illustrated with a blue circle, visible glints with green circles, and undetected glints with red circles.

C. VR Device specifications

The VRGaze dataset was collected using a custom virtual reality headset prototype equipped with a dedicated off-axis eye tracking system. To validate the real-time capabilities of the GazeShift architecture, all on-device processing and inference benchmarking were executed on a Samsung Exynos 2200 System-on-Chip featuring an Xclipse 920 mobile GPU. This setup accurately reflects the thermal and computational constraints of modern standalone extended reality devices. The camera operates at 30 frames per second with a spatial resolution of 400x400 pixels, and the device provides a 90° horizontal field of view. Active illumination is supplied by an array of 10 near-infrared LEDs per eye. The camera sensor is mounted at a steep oblique angle below the eye (Figure 9) to mimic the restrictive internal layouts required by modern, slim VR form factors. Finally, strict temporal synchronization between the display rendering the visual stimulus and the camera's capture trigger ensures precise 2D Point of Regard ground truth labels across all 2.1 million recorded frames.

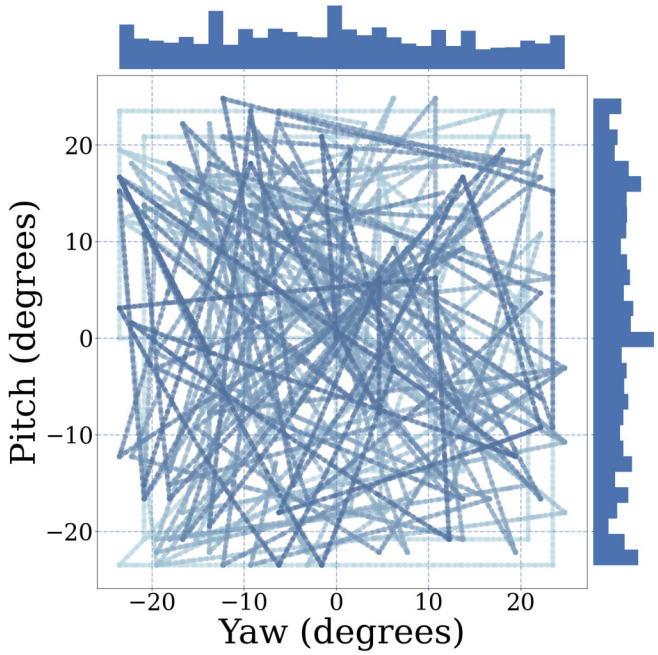


Figure 2. Single person data visualization.

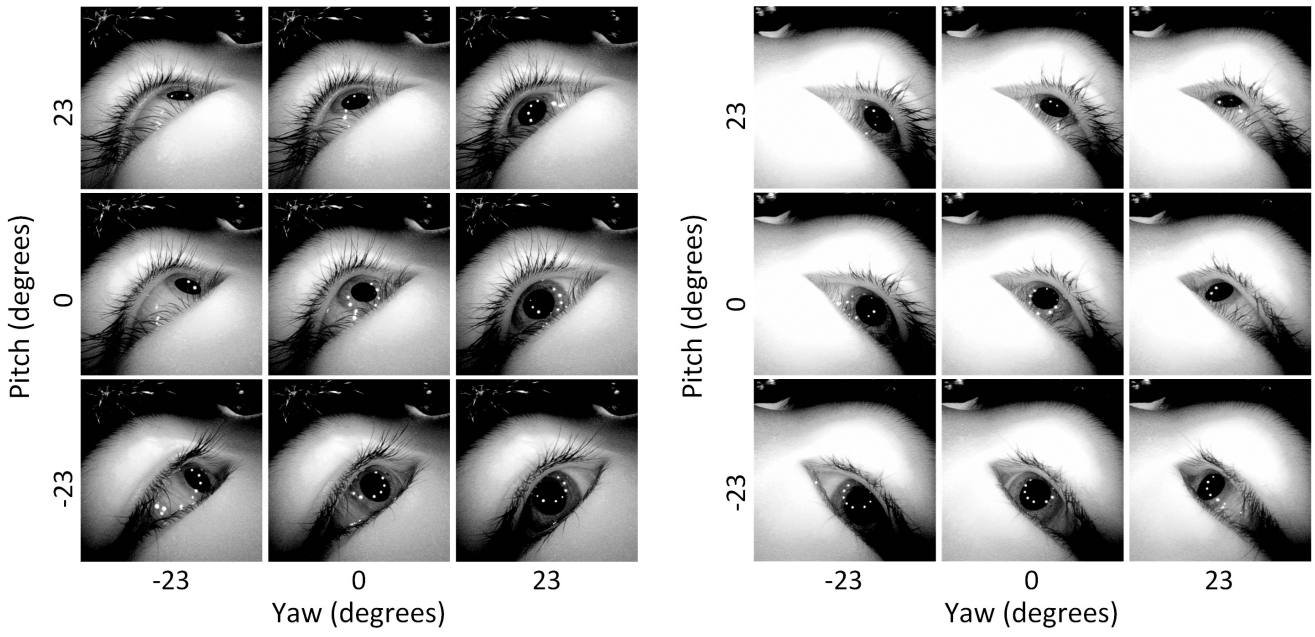


Figure 3. Gaze direction visualization.

Figure 4. Pupils dilation illustration.

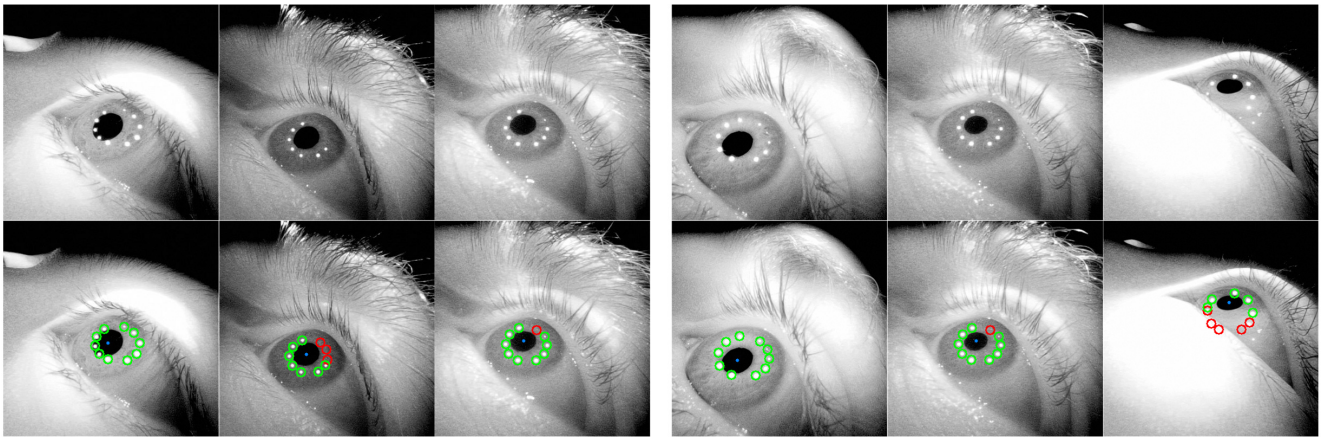


Figure 5. Pupil center and glints from manual annotations (left) and model predictions (right).

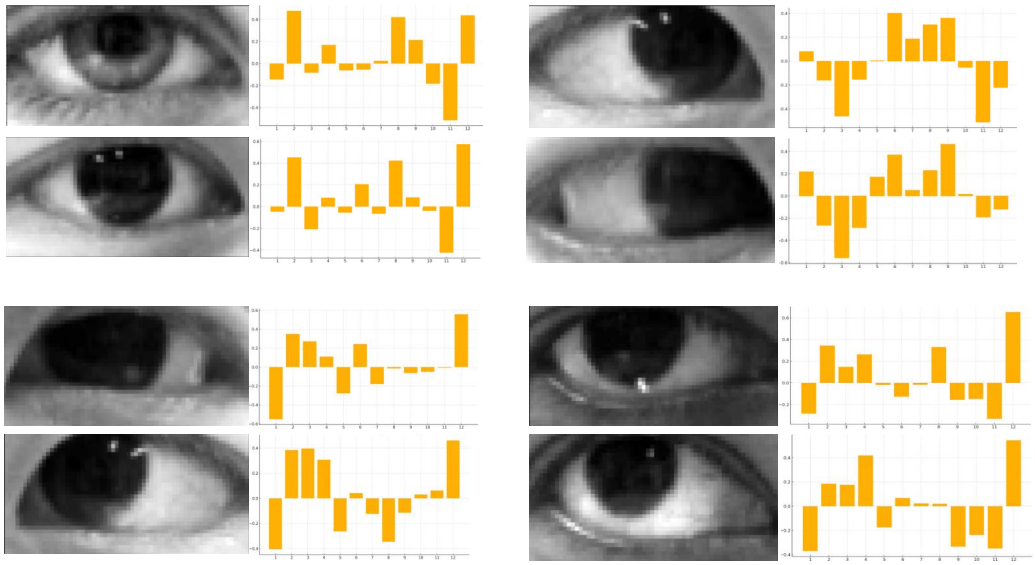
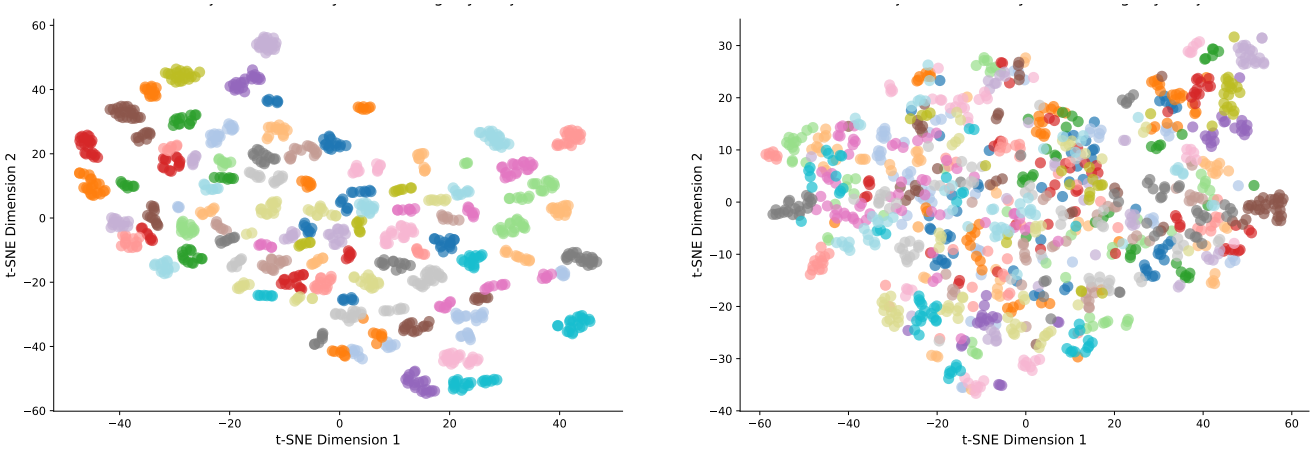


Figure 6. Gaze embeddings across directions. Each quadrant shows two eye images with similar gaze directions, along with their corresponding 12D gaze embeddings. Within a quadrant, the embeddings align closely, while embeddings across quadrants differ, reflecting distinct gaze directions. Notably, in the top-left quadrant, the gaze embeddings remain similar despite clear differences in eye appearance.



(a) Appearance embeddings

(b) Gaze embeddings

Figure 7. **Latent space visualization.** A t-SNE projection of GazeShift's gaze and appearance embeddings. Each point corresponds to an eye image from the Columbia Gaze dataset and is colored according to the subject's identity, illustrating the clustering of appearance features by the subject's identity.

Re: **Letter of Waiver and Consent**

I, the undersigned, hereby grant my consent to the publication of all my VR sessions (eye recording, no surroundings) and related datasets, including any derivative works thereof (the "**Materials**"), collected by the Company during the course of the VR gaze estimation project and the use of the XR PoC device, such Materials will be made publicly available and accessible by the public, including by way of publications in academic papers, journals, business conventions, etc., without any other identifiable details.

The Company is and will always remain the sole owner all rights in the Materials. I hereby waive any rights that I may have for privacy or any moral rights or rights for publication that I may have in the Materials or in connection with their publication as aforesaid, and confirm that the Company and/or its employees and/or office holders and/or shareholders shall not be responsible in any way as a result of the publication or any use of the Materials by any third party.

I shall not be entitled to any remuneration and/or compensation and/or royalties in connection with the publication of the Materials as aforesaid.

APPROVED AND ACKNOWLEDGED BY:

Name: _____
Signature: _____
Date: _____

Figure 8. Every person participating in the VRGaze dataset signed the consent form

Figure 9. **VRGaze hardware setup.** The internal view of the custom virtual reality headset used for data collection. The ring of 850 nm near-infrared LEDs (visible as purple lights) provides active illumination for robust corneal glint formation, while the restrictive form factor necessitates the off-axis camera placement.