# CoGaze: Closed-Loop Eye-Face Alignment for Multi-Center Mobile Gaze-Based Cognitive Screening in Older Adults

Jiahui Yu[1], Tao Du[1], Jinfu Wang[2], You Wu[3], Xiaowen Lou[3], Ting Pang[1], and Xin Xu[1,4,5(✉)]

[1] Department of Psychiatry, School of Public Health, The Second Affiliated Hospital of school of medicine, Zhejiang University, Hangzhou 310058, China
[2] Samsung R&D Institute China-Beijing, Beijing 100022, China
[3] School of Public Health, Zhejiang University, Hangzhou 310058, China
[4] Nanhu Brain-computer Interface institute, Hangzhou 311100, China
[5] The State Key Lab of Brain-Machine Intelligence, MOE Frontier Science Center for Brain Science and Brain-Machine Integration, Zhejiang University School of Medicine, Hangzhou 310003, China
xuxinsummer@zju.edu.cn

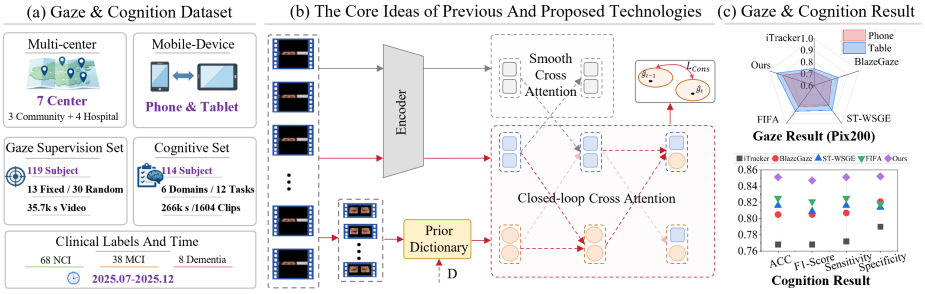


**Fig. 1.** (a) We build a real-world mobile dataset collected from seven centers, including gaze data from 119 subjects and cognitive-screening data from 114 subjects. (b) Compared with mainstream temporal modeling that relies on smoothing priors, CoGaze adopts closed-loop eye–face alignment reasoning to suppress noise-induced jitter and drift. (c) On this benchmark, CoGaze improves Pix200 by 6.3% and boosts downstream cognitive-screening accuracy by 2.6%.

**Abstract.** Mobile screening is increasingly needed for frequent, low-burden assessment of cognitive impairment in communities and households, and eye tracking offers scalable oculomotor signatures for this purpose. However, mobile gaze estimation is fragile under coupled shifts from aging-related appearance changes, uncontrolled capture noise, and fluctuating engagement, which can corrupt downstream oculomotor features.

J. Yu and T. Du contributed equally.

We propose CoGaze, a closed-loop gaze reasoning framework for mobile videos of older adults. CoGaze formulates gaze estimation as structured alignment between ocular appearance and facial geometric context. It stabilizes ocular representations with a prototype–anchor dictionary retrieved by normalized cosine attention, refines eye–face evidence via reciprocal cross-attention, and enforces temporal alignment using differential spatiotemporal attention with a segment-level consistency objective to reduce jitter and drift without blurring true saccades. We collect a deployment-oriented multi-center dataset from 7 sites (July–December 2025) with 119 gaze-task subjects and 114 clinically labeled assessments. CoGaze achieves mean error 1.29 cm/2.03 cm on phone/tablet and improves screening to 85.1% accuracy (5-fold CV). The code is available at: https://github.com/xiangxuebei/CoGaze.

**Keywords:** Cognitive screening · Mobile gaze · Alignment modeling.

## 1 Introduction

Population aging is shifting cognitive-impairment screening from occasional clinic visits to frequent, low-burden assessments in communities and households[4, 14, 16, 2]. Eye tracking provides a scalable window into cognitive function via measurable oculomotor signatures (e.g., fixation duration, scanpaths, regressions, and gaze stability), and has been increasingly supported by digital-medicine studies and reviews as a promising Alzheimer's disease digital-biomarker direction[3, 21, 19, 15].

Most gaze-based screening pipelines rely on gaze estimation as an intermediate step. They infer gaze direction or point-of-regard from standard RGB video and then derive oculomotor representations for downstream screening[21, 19]. Appearance-based deep models have made this feasible without dedicated eye trackers[1, 18, 6]. Recent studies focus on deployment. They handle data scarcity and distribution shift through weak supervision, self-training with pseudo-labels, and cross-domain representation learning[18, 17, 6, 22, 7, 5]. Newer work also explores geometry-aware alignment and stronger temporal modeling for egocentric and video gaze[11, 10, 13, 20]. These advances motivate evaluations that probe robustness across populations and real-world contexts[4, 14, 16, 17].

Deployable mobile screening exposes gaze estimation to coupled shifts from aging, acquisition noise, and attention fluctuations. Aging brings complex facial textures and ocular appearance changes, such as eyelid droop, wrinkles, altered eye shape, and glasses reflections, which distort pupil and iris cues and induce subject-dependent bias. Unconstrained mobile capture adds persistent noise, including handheld jitter, rapid head-pose changes, illumination shifts, occlusions, and motion blur, leading to frame-level jitter and long-term drift that corrupt fixation maps and saccadic dynamics[17, 11, 10]. In cognitively impaired populations, variable task understanding and sustained attention further increase blinks and brief lapses[9, 21], making weak supervision or pseudo-labeling more error-prone. Mainstream pipelines rely on local smoothness priors, which cannot

separate noise jitter from true saccades, causing latency, over-smoothing, and accumulated drift.

To address this, we propose CoGaze, as shown in Fig 1(b), a closed-loop gaze reasoning framework for mobile videos of older adults. We cast gaze estimation as a structured alignment problem between two evidence sources: ocular appearance and facial geometric context. CoGaze learns a canonical ocular representation and refines it through closed-loop coupling with facial context. Specifically, CoGaze introduces a prototype–anchor ocular dictionary and retrieves relevant prototypes via normalized cosine attention to stabilize the eye embedding space under appearance variation. CoGaze then performs reciprocal cross-attention: facial tokens help resolve pose and geometry ambiguity, while refined ocular evidence reweights facial responses toward gaze cues. Finally, CoGaze enforces temporal alignment with differential spatiotemporal attention and a segment-level consistency loss, reducing jitter and drift without smoothing away true saccades.

We collect a deployment-oriented, as shown in Fig 1(a), multi-center mobile dataset from 7 sites across community and hospital settings (July–December 2025), including 119 subjects for gaze tasks and 114 subjects for cognitive assessments with clinical labels. As shown in Fig 1(c), on this benchmark, CoGaze achieves strong cross-device gaze accuracy (mean error 1.29 cm on phone and 2.03 cm on tablet) and improves downstream screening (85.1% overall accuracy in 5-fold cross-validation), supporting end-to-end benefits of robust gaze recovery.

## 2 Method

### 2.1 Preliminaries

**Multi-scale full-face feature encoding** As shown in Fig 2(b), given a face video clip of length $T$, we crop a full-face region $I_t^{\text{face}}$ and an ocular region $I_t^{\text{eye}}$ for each frame $t$ as two inputs. For the full-face branch, a multi-scale encoder extracts hierarchical feature maps $\{\mathbf{F}_t^{(s)}\}_{s=1}^{S}$. To enable subsequent attention in a unified latent space, we flatten each scale into tokens and project them to a shared channel dimension $d$:

$$\mathbf{F}_t^{\text{face}} = \text{Fuse}\Big(\big\{\text{MLP}_s(\text{Flatten}(\mathbf{F}_t^{(s)}))\big\}_{s=1}^{S}\Big), \quad \mathbf{F}_t^{\text{face}} \in \mathbb{R}^{n_f \times d}, \tag{1}$$

where $\text{Flatten}(\cdot)$ converts a feature map into a token sequence, $\text{MLP}_s(\cdot)$ is the scale-specific projection layer, and $\text{Fuse}(\cdot)$ is the multi-scale fusion module. We use $S = 4$ scales.

**Prototype–anchor ocular prior dictionary** For the ocular branch, an eye encoder produces eye tokens $\mathbf{F}_t^{\text{eye}} \in \mathbb{R}^{n_e \times d}$ (left/right eyes share weights and are fused). Aging-related appearance changes enlarge intra-class variation even

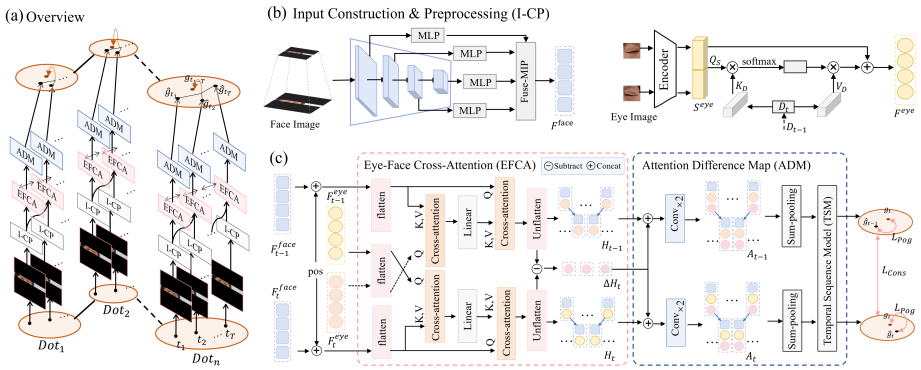


**Fig. 2.** Architecture of CoGaze. (a) Overall pipeline: given a face video sequence under a multi-point paradigm, CoGaze predicts the gaze-point trajectory from the sequence. (b) Input construction and preprocessing (I-CP): extracting multi-scale facial tokens and fusing them into a unified facial representation, while enhancing ocular tokens by querying a learnable prior dictionary. (c) Closed-loop reasoning module: Eye–Face Cross-Attention (EFCA) performs reciprocal cross-attention; the Attention Difference Map (ADM) explicitly models inter-frame consistency via differential spatiotemporal attention.

under the same gaze state. To provide stable anchors, we introduce a learnable prior dictionary:

$$\mathbf{D} = \{(\mathbf{p}_m, \mathbf{a}_m)\}_{m=1}^{M}, \tag{2}$$

where $\mathbf{p}_m$ is a learnable *content prototype* and $\mathbf{a}_m$ is an *anchor* in a normalized ocular coordinate system. We enhance ocular tokens by cosine cross-attention over the prior dictionary:

$$\tilde{\mathbf{F}}_t^{\text{eye}} = \text{Softmax}\left(\frac{\overline{\mathbf{F}_t^{\text{eye}}\mathbf{W}_Q}\;\overline{\mathbf{D}\mathbf{W}_K}^{\top}}{\tau}\right)\mathbf{D}\mathbf{W}_V, \tag{3}$$

where $\bar{\cdot}$ denotes $\ell_2$-normalization and $\tau$ is a temperature parameter.

### 2.2 Closed-loop cross spatiotemporal attention

**Eye-to-face cross-attention (EFCA)** As shown in Fig 2(c), we align modalities by using the enhanced ocular tokens from an adjacent frame as queries and the current fused face tokens as keys/values:

$$\hat{\mathbf{F}}_t^{\text{face}} = \text{Attn}\Big(\mathbf{Q} = \tilde{\mathbf{F}}_{t-1}^{\text{eye}},\ \mathbf{K} = \mathbf{F}_t^{\text{face}},\ \mathbf{V} = \mathbf{F}_t^{\text{face}}\Big). \tag{4}$$

This makes gaze-relevant geometry and compact facial features interact directly, promoting robust gaze-related feature extraction under occlusion or blur.

**Face-to-eye cross-attention** We then reverse the direction to let facial geometry refine local ocular representation:

$$\hat{\mathbf{F}}_t^{\text{eye}} = \text{Attn}\Big(\mathbf{Q} = \hat{\mathbf{F}}_t^{\text{face}},\ \mathbf{K} = \tilde{\mathbf{F}}_t^{\text{eye}},\ \mathbf{V} = \tilde{\mathbf{F}}_t^{\text{eye}}\Big)\,, \tag{5}$$

and form the joint representation

$$\mathbf{H}_t = \text{Concat}\Big(\hat{\mathbf{F}}_t^{\text{face}}, \hat{\mathbf{F}}_t^{\text{eye}}\Big)\,. \tag{6}$$

When eye regions are partially occluded or affected by blinking, facial structure provides relative spatial cues that reduce reliance on invalid ocular tokens.

**Attention Difference Map (ADM): differential spatiotemporal attention** To emphasize regions that change most and are gaze-relevant, we compute feature differences between adjacent frames and predict a spatial attention map:

$$\Delta\mathbf{H}_t = \mathbf{H}_t - \mathbf{H}_{t-1}, \quad \mathbf{A}_t = \sigma(\text{Conv}(\Delta\mathbf{H}_t))\,, \tag{7}$$

where $\sigma(\cdot)$ is the element-wise sigmoid. We reweight features using element-wise product $\odot$:

$$\bar{\mathbf{H}}_t = \mathbf{A}_t \odot \mathbf{H}_t. \tag{8}$$

Let $\mathbf{w}_t = \text{Flatten}(\mathbf{A}_t)$ be the flattened attention weights. We construct a spatiotemporal attention representation that aggregates the previous state, the change term, and the current state:

$$\mathbf{S}_t = \text{Concat}\big(\bar{\mathbf{H}}_{t-1},\ \Delta\mathbf{H}_t,\ \bar{\mathbf{H}}_t\big)\,. \tag{9}$$

This design suppresses noise-induced jitter and long-term drift while preserving true saccades.

### 2.3 Gaze prediction and objective

Given per-frame spatiotemporal representations $\{\mathbf{S}_t\}_{t=1}^{T}$, we apply a temporal sequence module (TSM) to model oculomotor dynamics:

$$\mathbf{z}_t = \text{TSM}\big(\{\mathbf{S}_t\}_{t=1}^{T}\big)\,, \quad \hat{\mathbf{g}}_t = f_{\text{MLP}}(\mathbf{z}_t) \in \mathbb{R}^2, \tag{10}$$

where $\hat{\mathbf{g}}_t$ is the predicted 2D PoR and $\mathbf{g}_t$ is the ground truth.

The training objective consists of: (i) a regression term and (ii) an adjacent-difference consistency term within the same target-point segment to suppress frequent jitter:

$$\mathcal{L}_{\text{reg}} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\hat{\mathbf{g}}_{i,t} - \mathbf{g}_{i,t}\|_1\,, \tag{11}$$

$$\mathcal{L}_{\text{cons}} = \frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=2}^{T}\|\hat{\mathbf{g}}_{i,t} - \hat{\mathbf{g}}_{i,t-1}\|_1\,. \tag{12}$$

where $N$ is the number of target points per trial, $T$ is the number of frames collected per target point. The final objective is $\mathcal{L} = \mathcal{L}_{\text{reg}} + \lambda\,\mathcal{L}_{\text{cons}}$, with $\lambda$ balancing the two losses.

**Table 1.** Gaze estimation performance on older adults collected with different mobile devices. Errors are reported in cm; Pix100/Pix200 denote accuracies under 100/200-pixel thresholds.

| Device | Method | Fixed point | | | | Random point | | | | Mean | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Fix err. | Dot err. | Pix100 Acc. | Pix200 Acc. | Fix err. | Dot err. | Pix100 Acc. | Pix200 Acc. | Fix err. | Pix200 Acc. |
| **Phone** | iTracker[12] | 1.93 | 1.87 | 0.551 | 0.737 | 2.21 | 2.11 | 0.531 | 0.693 | 2.07 | 0.715 |
| | BlazeGaze[8] | 1.62 | 1.58 | 0.618 | 0.775 | 1.77 | 1.69 | 0.591 | 0.741 | 1.70 | 0.758 |
| | ST-WSGE[17] | 1.47 | 1.40 | 0.662 | 0.822 | 1.58 | 1.51 | 0.649 | 0.804 | 1.53 | 0.813 |
| | FIFA[10] | 1.42 | 1.36 | 0.668 | 0.833 | 1.54 | 1.43 | 0.654 | 0.812 | 1.48 | 0.823 |
| | **Ours** | **1.26** | **1.15** | **0.703** | **0.891** | **1.32** | **1.24** | **0.681** | **0.874** | **1.29** | **0.883** |
| **Tablet** | iTracker[12] | 2.78 | 2.63 | 0.712 | 0.755 | 2.95 | 2.77 | 0.693 | 0.732 | 2.87 | 0.744 |
| | BlazeGaze[8] | 2.42 | 2.26 | 0.751 | 0.828 | 2.58 | 2.40 | 0.731 | 0.805 | 2.41 | 0.817 |
| | ST-WSGE[17] | 2.23 | 2.07 | 0.782 | 0.867 | 2.39 | 2.21 | 0.776 | 0.854 | 2.31 | 0.861 |
| | FIFA[10] | 2.19 | 2.02 | 0.797 | 0.872 | 2.34 | 2.17 | 0.781 | 0.869 | 2.23 | 0.871 |
| | **Ours** | **1.94** | **1.79** | **0.841** | **0.947** | **2.11** | **1.94** | **0.824** | **0.926** | **2.03** | **0.937** |

## 3 Experiments

### 3.1 Dataset

We collected real-world mobile multi-modal cognitive-assessment data from seven centers over **six months** (July–December 2025) under a standardized protocol. Participants were older adults aged 50–90 years, with a mean age of **72.7** years (SD, 6.7 years), and had basic comprehension and communication ability; we excluded severe visual/hearing impairment, communication disorders, major depression, and other psychiatric conditions that could affect task completion. For gaze estimation, **119** subjects completed 13 fixed-point and 30 random-point trials, yielding ∼**35.7k** seconds of face video for training. For cognitive screening, **114** subjects completed demographics and **12 tasks** spanning six cognitive domains (seven tasks are gaze-related), producing ∼**266k** seconds of video (1604 segments) with psychiatrist-confirmed diagnoses or CDR labels (NCI/MCI/Dementia: 68/38/8).

### 3.2 Implementation Details

We construct two mobile gaze-estimation tasks (fixed-point and random-point) to obtain supervision signals. During acquisition, we synchronously record the front-camera face video and the target-point coordinates/timestamps in screen coordinates, followed by frame-level quality control.

The gaze estimator takes clips of 30 frames as input. For each frame, we crop a $128 \times 128$ full-face image and a $128 \times 128$ eye image. We use ResNet-18 as the backbone, aggregate temporal features, and regress 2D gaze coordinates. The model is trained end-to-end with SGD (lr=0.001, momentum=0.9, batch size=16) for 10,000 iterations using cosine-annealing scheduling. For downstream cognitive-status classification, the trained gaze estimator was fixed and applied to cognitive-screening videos to infer gaze trajectories, which were then aligned with task logs and converted into subject-level oculomotor features (latencies, error/correction, saccade dynamics, and ROI dwell statistics). We used a strict

subject-level stratified 5-fold split based on the NCI/MCI/dementia distribution, ensuring that all data from the same participant remained in the same fold. All compared gaze estimators shared the same folds, feature categories, and downstream classifier.

We compare CoGaze with iTracker[12] and three representative SOTA gaze trackers (BlazeGaze[8], ST-WSGE[17], and FIFA[10]) under the same protocol on both phone and tablet. For gaze estimation, we report mean localization errors (Fix err., Dot err.) and threshold accuracies (Pix100/Pix200) for both fixed-point and random-point paradigms.

### 3.3 Gaze Estimation Performance

Table 1 summarize gaze-point prediction on real-world mobile data. CoGaze achieves the best overall performance across devices and paradigms, with a mean error of 1.29 cm on phone and 2.03 cm on tablet. Moreover, CoGaze attains 91.0% Pix200 accuracy on average, improving by ∼6.3% .These gains indicate fewer high-deviation frames and more stable gaze trajectories, providing a more reliable basis for downstream oculomotor feature extraction and cognitive screening.

### 3.4 Downstream Cognitive Screening Performance

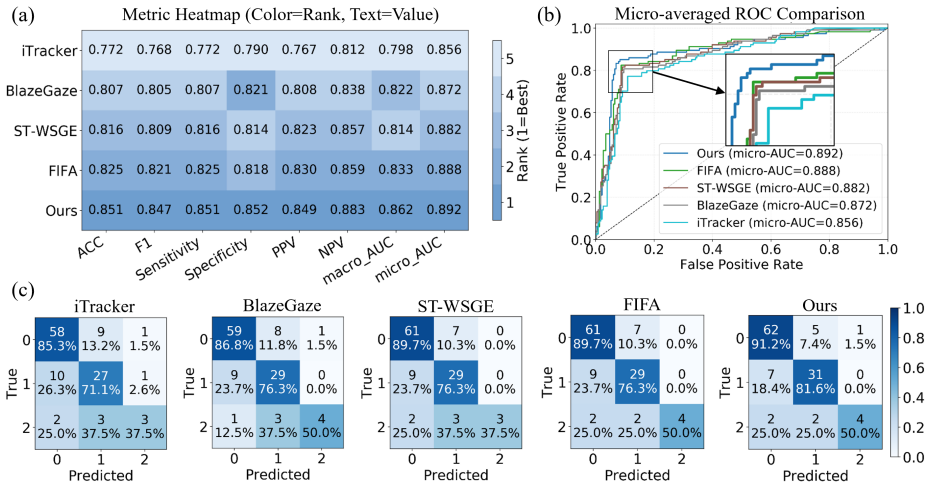


**Fig. 3.** Cognitive screening results. (a) Comparison of classification accuracy across different models under 5-fold cross-validation. (b) Visualization of micro-AUC. (c) Confusion matrix aggregated over the five folds.

We use each gaze estimator to derive subject-level oculomotor features and train a classifier for cognitive-status prediction. Under the same experimental

setting, we report the performance of competing methods and ours using 5-fold cross-validation. As shown in Fig 3, features extracted by CoGaze achieve higher accuracy and AUC, reaching an overall accuracy of 85.1% and improving performance for both NCI and MCI (91.2% and 81.6%, respectively). As illustrated in

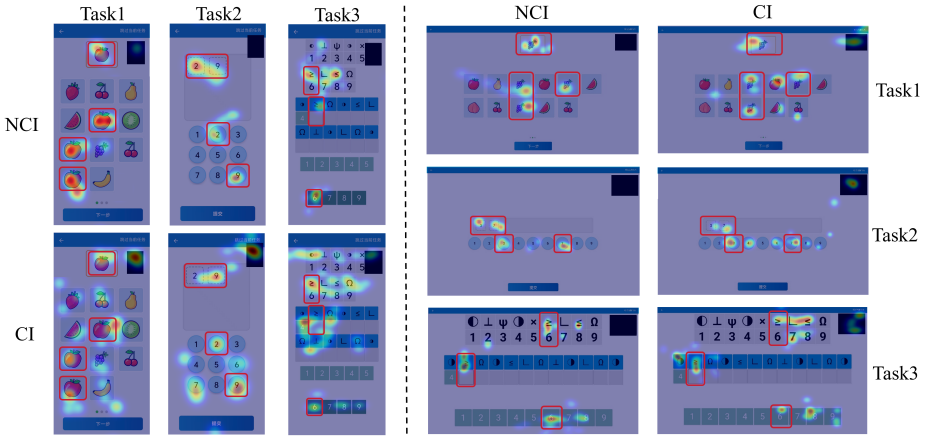


**Fig. 4.** Group-level gaze heatmap comparison between NCI and CI participants for selected tasks.

Fig 4, we further visualize group-level average gaze heatmaps for selected task segments. Compared with CI, NCI typically exhibits more concentrated and consistent fixations on task-relevant regions, whereas CI shows more dispersed attention, shifting regions of interest, and insufficient dwell time on critical areas.

### 3.5 Ablation Study

**Table 2.** Ablation study. Msc: multi-scale face representation; Pd: ocular prior dictionary;

| Method | EFCA | Msc | Pd | ADM | Fix err. | Pix200 Acc | Acc | mac-AUC |
|---|---|---|---|---|---|---|---|---|
| baseline | ✓ | × | × | × | 1.52 | 0.815 | 0.825 | 0.833 |
| w/o. Pd | ✓ | × | × | ✓ | 1.48 | 0.827 | 0.825 | 0.841 |
| w/o. Msc | ✓ | × | ✓ | ✓ | 1.39 | 0.855 | 0.843 | 0.857 |
| CoGaze | ✓ | ✓ | ✓ | ✓ | **1.32** | **0.874** | **0.851** | **0.862** |

Table 2 quantifies the interaction between each module and the eye–face cross-attention. Compared with the baseline, the full model improves Pix200 by 5.9%. Introducing the ocular prior dictionary (Pd) yields the largest gain,

boosting Pix200 by 2.9%; adding the multi-scale face representation (Msc) further improves Pix200 by 1.9%, indicating complementary benefits from ocular priors and multi-scale facial context.

## 4 Discussion and Conclusion

Our findings suggest that the main challenge of mobile gaze-based cognitive screening in older adults is not only reducing gaze error, but preserving meaningful gaze dynamics under aging-related appearance changes and unconstrained capture noise. CoGaze addresses this challenge by explicitly coupling ocular and facial cues to produce more stable and reliable gaze estimation. Overall, CoGaze provides an effective front end for scalable, low-burden mobile cognitive screening. However, the current cognitive-screening cohort remains relatively small. Future work will explore larger-scale external multi-center validation and cross-device stability analysis, as well as prospective studies to assess its utility in real-world clinical and community screening scenarios.

**Acknowledgments.** This work was partly supported by the Vision-Based Assessment of Cognitive Impairment project under Grant SRC-Beijing-DVL-2025-00230, the Zhejiang Provincial Key Research and Development Program of China under Grant 2025C02108, and the National Natural Science Foundation of China under Grant 62406280. This work was also partly supported by the Nanhu Brain-computer Interface Institute, Hangzhou, China.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Bao, Y., Lu, F.: Unsupervised gaze representation learning from multi-view face images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1419–1428 (2024)
2. Berron, D., Glanz, W., Clark, L., et al.: A remote digital memory composite to detect cognitive impairment in memory clinic samples in unsupervised settings using mobile devices. npj Digital Medicine **7**, 79 (2024)
3. Boujelbane, M.A., Trabelsi, K., Salem, A., Ammar, A., Glenn, J.M., Boukhris, O., AlRashid, M.M., Jahrami, H., Chtourou, H.: Eye tracking during visual paired-comparison tasks: A systematic review and meta-analysis of the diagnostic test accuracy for detecting cognitive decline. Journal of Alzheimer's Disease **99**(1), 207–221 (2024)
4. Butler, P.M., Yang, J., Brown, R., et al.: Smartwatch- and smartphone-based remote assessment of brain health and detection of mild cognitive impairment. Nature Medicine **31**, 829–839 (2025)
5. Catruna, A., Cosma, A., Radoi, E.: Crossgaze: A strong method for 3d gaze estimation in the wild. In: The 18th IEEE International Conference on Automatic Face and Gesture Recognition (2024)

6. Cheng, Y., Wang, H., Zhang, Z., Yue, Y., Kim, B., Lu, F., Chang, H.J.: 3d prior is all you need: Cross-task few-shot 2d gaze estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23891–23900 (2025)
7. Cho, H., Ahn, H., Wu, G., Kim, W.H.: Adaptive adversarial data augmentation with trajectory constraint for alzheimer's disease conversion prediction. In: Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15966, pp. 13–23. Springer Nature Switzerland (2025)
8. Davalos, E., Zhang, Y., Srivastava, N., Thatigotla, Y., Salas, J.A., McFadden, S., Cho, S.J., Goodwin, A., Ashwin, T.S., Biswas, G.: Webeyetrack: Scalable eye-tracking for the browser via on-device few-shot personalization (2025), arXiv preprint
9. Erickson, C.M., Wexler, A., Largent, E.A.: Digital biomarkers for neurodegenerative disease. JAMA Neurology **82**(1), 5–6 (01 2025)
10. Hu, D., Cui, M., Huang, K.: Fifa: Fine-grained inter-frame attention for driver's video gaze estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18760–18769 (2025)
11. Kawana, Y., Shiba, S., Kong, Q., Kobori, N.: Ga3ce: Unconstrained 3d gaze estimation with gaze-aware 3d context encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3081–3090 (2025)
12. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., B.S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2176–2184 (2016)
13. Liu, S., Chen, W., Liu, J., Luo, X., Shen, L.: Gem: Context-aware gaze estimation with visual search behavior matching for chest radiograph. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15001, pp. 525–535. Springer Nature Switzerland (2024)
14. Polk, S.E., Öhman, F., Hassenstab, J., König, A., Papp, K.V., Schöll, M., Berron, D.: A scoping review of remote and unsupervised digital cognitive assessments in preclinical alzheimer's disease. npj Digital Medicine **8**, 266 (2025)
15. Qi, W., Zhu, X., Wang, B., Shi, Y., Dong, C., Shen, S., Li, J., Zhang, K., He, Y., Zhao, M., Yao, S., Dong, Y., Shen, H., Kang, J., Lu, X., Jiang, G., Boots, L.M.M., Fu, H., Pan, L., Chen, H., Yan, Z., Xing, G., Cao, S.: Alzheimer's disease digital biomarkers: multidimensional landscape and ai model scoping review. npj Digital Medicine **8**(1), 366 (2025)
16. van den Berg, R.L., van der Landen, S.M., Keijzer, M.J., van Gils, A.M., van Dam, M., Ziesemer, K.A., Jutten, R.J., Harrison, J.E., de Boer, C., van der Flier, W.M., Sikkes, S.A.M.: Smartphone- and tablet-based tools to assess cognition in individuals with preclinical alzheimer disease and mild cognitive impairment: Scoping review. Journal of Medical Internet Research **27**, e65297 (2025)
17. Vuillecard, P., Odobez, J.M.: Enhancing 3d gaze estimation in the wild using weak supervision with gaze following labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13508–13518 (2025)
18. Wang, Y., Li, J., Dai, W., Shi, B., Zhang, X., Li, C., Xiong, H.: Bootstrap autoencoders with contrastive paradigm for self-supervised gaze estimation. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 50794–50806. PMLR (2024)

19. Wei, M., Li, J., You, T., Yu, Y., Lu, J., Liang, S., Jin, Z., Han, Q., Zuo, C., Ye, J., Yu, J., Chen, X., Dong, Q., Wu, W., Wang, Y., Jiang, Y., Cui, M.: An accessible and efficient mobile eye-tracking application for community-based cognitive impairment screening in china. Alzheimer's Research & Therapy **17**(1), 250 (2025)
20. Wu, S., Zhang, X., Wang, B., Jin, Z., Li, H., Feng, J.: Gaze-directed vision gnn for mitigating shortcut learning in medical image. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15001, pp. 514–524. Springer Nature Switzerland (2024)
21. Xu, Y., Zhang, C., Pan, B., Yuan, Q., Zhang, X.: A portable and efficient dementia screening tool using eye tracking machine learning and virtual reality. npj Digital Medicine **7**, 219 (2024)
22. Zhong, Y., Tang, C., Yang, Y., Qi, R., Zhou, K., Gong, Y., Heng, P.A., Hsiao, J.H., Dou, Q.: Weakly-supervised medical image segmentation with gaze annotations. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15003, pp. 530–540. Springer Nature Switzerland (2024)